Publ. RIMS, Kyoto Univ.
18 (1982), 101-117

A Posteriori Componentwise Error Estimate for a Computed Solution of a System of Linear Equations

By

Tetsuro YAMAMOTO*

Introduction

Let $x^{(0)}=(x_1^{(0)}, \cdots, x_n^{(0)})^t$ be a computed solution of a system of n linear equations

$$(0.1) \qquad Ax=b$$

where $A=(a_{ij})$, $x=(x_1, \cdots, x_n)^t$ and $b=(b_1, \cdots, b_n)^t$. Then a question naturally arises as to whether the approximate solution $x^{(0)}$ is a satisfactory one. Let A be nonsingular and L be an approximation for the inverse of A. In practical computation, L may be chosen as a computer result for A^{-1}. Let $R=I_n-LA$ and $r=Ax^{(0)}-b$ where I_n denotes the $n\times n$ identity. If R has the spectral radius which is smaller than one, then L is nonsingular and

$$A^{-1}=(I_n-R)^{-1}L\,.$$

Hence, if we denote by x^* the exact solution of (0.1), then we have

$$(0.2) \qquad x^*-x^{(0)}=-A^{-1}r=-(I_n-R)^{-1}Lr\,,$$

or

$$(0.3) \qquad \|x^*-x^{(0)}\|\leqq\|(I_n-R)^{-1}\|\cdot\|Lr\|\leqq(1-\|R\|)^{-1}\|Lr\|$$

with some vector norm $\|\cdot\|$, provided that $\|R\|<1$. Therefore, if $\|Lr\|$ and $\|R\|$ are small enough, then we can conclude from (0.3) that $x^{(0)}$ is accurate. However, if there are large and small values among $|x_1^{(0)}|, \cdots,$

Communicated by S. Hitotumatu, December 22, 1980.
* Department of Mathematics, Faculty of Science, Ehime University, Matsuyama 790, Japan.

$|x_n^{(0)}|$, then (0.3) does not give a sharp estimate for a specified component of $x^{(0)}$. Therefore, in such a case, the use of (0.2) is desirable. However, (0.2) requires the computation of the inverse of $I_n - R$, which is troublesome.

In this paper, we shall first prove a result for finding the componentwise error bounds of $x^{(0)}$ without using $(I_n - R)^{-1}$. Next, we shall perform its error analysis for a machine having a floating-point arithmetic device with the base β in which the results are chopped to t β-digits. The results of the analysis show that our method works well if $\|R\|_\infty < 1$ and

$$\|L\|_\infty (\|A\|_\infty + \|A\|_\infty \cdot \|x^{(0)}\|_\infty + \|r\|_\infty) n\beta^{1-t}$$

is not large, where $\|\cdot\|_\infty$ denotes the maximum norm. Further, based on this result, we shall propose a practical algorithm for estimating rigorously the error of $x^{(0)}$. Finally, numerical examples are given, which illustrate our results.

§ 1. Notation

Throughout this paper, we shall use the following notation (cf. Urabe [5] and Yamamoto [8]): Let $x = (x_1, \cdots, x_n)^t$ and $y = (y_1, \cdots, y_n)^t$ be two vectors. Then we write $x \geqslant y$ or $y \leqslant x$ if $x_i \geqq y_i$ for all i. We put $\nu[x] = (|x_1|, \cdots, |x_n|)^t$. The same notation is used for matrices $A = (a_{ij})$ and $B = (b_{ij})$: $A \geqslant B$ or $B \leqslant A$ if $a_{ij} \geqq b_{ij}$ for all i, j and we put $\nu[A] = (|a_{ij}|)$.

§ 2. A Result

Let $x^{(0)}$ be an approximate solution of the system (0.1) which has the unique solution x^* and L be an approximation for the inverse of A. We put $K = \nu[I_n - LA]$. Then we have the following theorem.

Theorem 1. *Let* $\|\cdot\|$ *be a monotonic vector norm and* κ *be a vector such that*

$$Kx \leqslant \|x\| \kappa \tag{2.1}$$

for all $x \geqslant 0$. *We assume that* $\|\kappa\| < 1$ *and put*

$$\varepsilon = \nu[L(Ax^{(0)} - b)],\ a = (1 - \|\kappa\|)^{-1}\|\varepsilon\| \quad \text{and} \quad \alpha = \varepsilon + a\kappa .$$

Then we have

$$\nu[x^* - x^{(0)}] \leqslant \alpha .$$

Further, if we define a sequence of vectors $\{\alpha^{(k)}\}$ *by*

$$\alpha^{(0)} = \alpha, \quad \alpha^{(k+1)} = \varepsilon + K\alpha^{(k)}, \quad k = 0, 1, 2, \cdots, \tag{2.2}$$

then

$$\alpha^{(0)} \geqslant \alpha^{(1)} \geqslant \cdots \to \alpha^* = (I_n - K)^{-1}\varepsilon = (\alpha_1^*, \cdots, \alpha_n^*)^t$$

and

$$\nu[x^* - x^{(0)}] \leqslant \alpha^* \leqslant \alpha^{(k)}, \quad k \geqq 0 .$$

That is, we have

$$|x_i^* - x_i^{(0)}| \leqq \alpha_i^* \leqq \alpha_i^{(k)}, \quad 1 \leqq i \leqq n ,$$

for every $k \geqq 0$, *where* x_i^*, $x_i^{(0)}$ *and* $\alpha_i^{(k)}$ *denote the i-th component of* x^*, $x^{(0)}$ *and* $\alpha^{(k)}$, *respectively.*

Proof. We first remark that $\|K\| < 1$, because the norm is monotonic and (2.1) implies $\|Kx\| \leqq \|x\| \cdot \|\kappa\| < \|x\|$ for $x \neq 0$. Therefore, it follows from (0.2) that

$$\begin{aligned} \nu[x^* - x^{(0)}] &\leqslant \nu[(I_n - R)^{-1}]\nu[Lr] \\ &= \nu[I_n + R + R^2 + \cdots]\varepsilon \\ &\leqslant (I_n + \nu[R] + \nu[R^2] + \cdots)\varepsilon \\ &\leqslant (I_n + K + K^2 + \cdots)\varepsilon \\ &\leqslant \varepsilon + \|\varepsilon\|\kappa + \|K\varepsilon\|\kappa + \cdots \\ &\leqslant \varepsilon + (\|\varepsilon\| + \|\varepsilon\| \cdot \|\kappa\| + \|\varepsilon\| \cdot \|\kappa\|^2 + \cdots)\kappa \\ &= \varepsilon + a\kappa = \alpha . \end{aligned} \tag{2.3}$$

Next, the monotone decreasing property of the sequence $\{\alpha^{(k)}\}$ is proved by induction on k: In fact, by noting that

$$\|\alpha\| \leqq \|\varepsilon\| + a\|\kappa\| = a ,$$

we have

$$\alpha^{(1)}=\varepsilon+K\alpha\leqslant\varepsilon+\|\alpha\|\kappa\leqslant\varepsilon+a\kappa=\alpha=\alpha^{(0)}$$

and $\alpha^{(k)}\leqslant\alpha^{(k-1)}$ implies that

$$\alpha^{(k+1)}=\varepsilon+K\alpha^{(k)}\leqslant\varepsilon+K\alpha^{(k-1)}=\alpha^{(k)}.$$

Therefore, $\{\alpha^{(k)}\}$ converges to a vector $\alpha^*\geqslant 0$, which satisfies

$$\alpha^*=\varepsilon+K\alpha^*.$$

It follows from this that

$$\alpha^*=(I_n-K)^{-1}\varepsilon=(I_n+K+K^2+\cdots)\varepsilon\,.$$

Consequently we obtain from (2.3)

$$\nu[x^*-x^{(0)}]\leqslant\alpha^*\leqslant\cdots\leqslant\alpha^{(k)}\leqslant\cdots\leqslant\alpha^{(1)}\leqslant\alpha^{(0)}=\alpha\,.$$

Q.E.D.

Remark. For the maximum norm $\|\cdot\|_\infty$, the i-th component κ_i of the vector κ is given by

$$\kappa_i=\sum_{j=1}^{n}\kappa_{ij}$$

where κ_{ij} denote the (i,j) elements of the matrix K. Hence, in this case, we have $\|\kappa\|_\infty=\|K\|_\infty$.

§ 3. Floating-Point Error Analysis

In practice, we cannot obtain the exact values of the vectors $\alpha^{(k)}$ ($k\geqq 0$), because of the rounding errors made in the computation. So the floating-point error analysis would be necessary. We shall call it out for the result of Theorem 1 by choosing the maximum norm $\|\cdot\|_\infty$. We assume that we work with a computer in which numbers are represented in the form $\pm d\beta^m$ where β is the base of the number system and d is the mantissa consisting of t digits and $0\leqq d<1$. We use the techniques due to Wilkinson [6], [7], Forsythe and Moler [2] and Paige [3]. Thus, if $\circ$ denotes any of the four arithmetic operations $+,-,\times,/$, then $a=\mathrm{fl}(b\circ c)$ means that a, b and c are floating-point numbers and a is obtained from b and c using the appropriate floating-point operation.

We assume that $n \geqq 2$,

$$\text{fl}(a \circ b) = a \circ b(1+\xi), \quad |\xi| < \beta^{1-t}, \tag{3.1}$$

and

$$1.006(n+1)\beta^{1-t} < 0.01 . \tag{3.2}$$

Note that (3.1) reflects a machine in which the results are chopped to t β-digits. If we consider a machine in which the results are rounded to t β-digits, then we should replace β^{1-t} in (3.1) and (3.2) by $2^{-1}\beta^{1-t}$, and the inequality $<$ in (3.1) by $\leqq$. Observe also that (3.2) means that $\beta^{1-t} < 0.01/3.018 < 0.0034$.

In the following, for the sake of convenience, we shall write $\theta_n = n\beta^{1-t}$ and use the following inequalities:

$$\text{If } 0 \leqq na < 0.01 , \text{ then } (1+a)^n \leqq e^{na} < 1 + 1.006na . \tag{3.3}$$

The following two lemmas are essentially proved in Wilkinson [6].

Lemma 1. *If a_i, $i=1, 2, \cdots, n$ are the floating-point numbers, then*

$$\text{fl}(a_1 + \cdots + a_n) = \sum_{i=1}^{n} a_i(1+\xi_i) \tag{3.4}$$

where

$$1+\xi_i = \begin{cases} (1+\eta_2)\cdots(1+\eta_n) & (i=1) \\ (1+\eta_i)\cdots(1+\eta_n) & (2 \leqq i \leqq n), \quad |\eta_j| < \beta^{1-t} \quad (1 \leqq j \leqq n). \end{cases} \tag{3.5}$$

Furthermore

$$\text{fl}|(a_1 + \cdots + a_n) - \sum_{i=1}^{n} a_i| < 1.006\theta_{n-1} \sum_{i=1}^{n} a_i \tag{3.6}$$

If $a_i \geqq 0$ $(1 \leqq i \leqq n)$, then

$$\sum_{i=1}^{n} a_i < (1 - 1.006\theta_{n-1})^{-1} \text{fl}(a_1 + \cdots + a_n).$$

Proof. The equality (3.4) is proved by induction on n. (3.6) follows from (3.4) since (3.3) and (3.5) imply that

$$1 + |\xi_i| < (1+\theta_1)^{n-1} \quad (1 \leqq i \leqq n),$$

Q.E.D.

Lemma 2. *If a_i and b_i are the floating-point numbers, then*

$$\mathrm{fl}(a_1b_1+\cdots+a_nb_n)=\sum_{i=1}^{n}a_ib_i(1+\xi_i)$$

where

$$1+\xi_i=\begin{cases}(1+\eta_1)(1+\zeta_2)\cdots(1+\zeta_n) & (i=1)\\ (1+\eta_i)(1+\zeta_i)\cdots(1+\zeta_n) & (2\leqq i\leqq n)\end{cases}$$

with $|\eta_i|, |\zeta_j|<\beta^{1-t}$, $i=1,2,\cdots,n$, $j=2,\cdots,n$. *Hence*

$$|\mathrm{fl}(a_1b_1+\cdots+a_nb_n)-\sum_{i=1}^{n}a_ib_i|<1.006\theta_n\sum_{i=1}^{n}a_ib_i,$$

so that, if $a_ib_i\geqq 0$, *then*

$$\sum_{i=1}^{n}a_ib_i<(1-1.006\theta_n)^{-1}\,\mathrm{fl}(a_1b_1+\cdots+a_nb_n).$$

In the following, we denote by $\tilde{a}$ the computer result for an expression a. Thus, if a is a number, then $\tilde{a}$ means the floating-point representation of a in the machine. For simplicity, we assume that $\tilde{x}^{(0)}=x^{(0)}$, $\tilde{A}=A$, $\tilde{b}=b$ and $\tilde{L}=L$.

Lemma 3. *Let* $r=Ax^{(0)}-b$ *and*

$$c=1.006\nu[A]\nu[x^{(0)}]+1.004n^{-1}\tilde{r}\,. \tag{3.7}$$

Then

$$\tilde{r}=r+\delta r\,,\quad \nu[\delta r]\leqslant\theta_n c\,.$$

Proof. Let $r=(r_1,\cdots,r_n)^t$ and $s_i=\sum_{j=1}^{n}a_{ij}x_j^{(0)}$. Then we have

$$\begin{aligned}\tilde{r}_i&=\mathrm{fl}(\tilde{s}_i-b_i)=\tilde{s}_i-b_i+(\tilde{s}_i-b_i)\xi_0\\ &=\sum_{j=1}^{n}a_{ij}x_j^{(0)}(1+\xi_j)-b_i+\frac{\tilde{r}_i}{1+\xi_0}\xi_0\\ &=r_i+\delta r_i\quad(1\leqq i\leqq n)\end{aligned}$$

where $|\xi_0|<\beta^{1-t}$, ξ_j are defined in Lemma 2 and

$$\delta r_i=\sum_{j=1}^{n}a_{ij}x_j^{(0)}\xi_j+(1+\xi_0)^{-1}\tilde{r}_i\xi_0\,.$$

Hence

$$|\delta r_i| < 1.006\theta_n \sum_{j=1}^{n} |a_{ij}x_j^{(0)}| + (1-\beta^{1-t})^{-1}\tilde{r}_i\beta^{1-t}$$

$$< 1.006\theta_n \sum_{j=1}^{n} |a_{ij}| \cdot |x_j^{(0)}| + (1-0.0034)^{-1}\tilde{r}_i n^{-1}\theta_n$$

$$< c_i\theta_n$$

where c_i denotes the i-th component of the vector c defined in (3.7). This implies $\nu[\delta r] \leqslant \theta_n c$ where $\delta r = (\delta r_1, \cdots, \delta r_n)'$.

Q.E.D.

Lemma 4. *Let* $\varepsilon = \nu[Lr]$ *and*

$$d = (1.006\|\tilde{r}\|_\infty + \|c\|_\infty)\nu[L](1, \cdots, 1)^t.$$

Then

$$\tilde{\varepsilon} = \varepsilon + \delta\varepsilon, \quad \nu[\delta\varepsilon] \leqslant \theta_n d.$$

Proof. Let ε_i and d_i be the i-th components of the vectors ε and d respectively and set $L = (l_{ij})$. Then we have from Lemma 2

$$\mathrm{fl}(\sum_{j=1}^{n} l_{ij}\tilde{r}_j) = \sum_{j=1}^{n} l_{ij}\tilde{r}_j + \sum_{j=1}^{n} l_{ij}\tilde{r}_j\xi_j \, (|\xi_j| < 1.006\theta_n)$$

$$= \sum_{j=1}^{n} l_{ij}(r_j + \delta r_j) + \sum_{j=1}^{n} l_{ij}\tilde{r}_j\xi_j$$

so that

$$|\tilde{\varepsilon}_i - \varepsilon_i| \leqq \left| |\mathrm{fl}(\sum_{j=1}^{n} l_{ij}\tilde{r}_j)| - |\sum_{j=1}^{n} l_{ij}r_j| \right|$$

$$\leqq \left| \sum_{j=1}^{n} l_{ij}(\delta r_j + \tilde{r}_j\xi_j) \right|$$

$$< \sum_{j=1}^{n} |l_{ij}| \{\theta_n c_j + |\tilde{r}_j| (1.006\theta_n)\} \leqq d_i\theta_n. \qquad \text{Q.E.D.}$$

Lemma 5. *Let*

$$E = 1.006\nu[L]\nu[A] + 1.004n^{-1}\begin{pmatrix} \tilde{\kappa}_{11} & & \\ & \ddots & \\ & & \tilde{\kappa}_{nn} \end{pmatrix}.$$

Then we have

$$\tilde{K}=K+\delta K\,,\quad \nu[\delta K]\leqslant\theta_n E\,.$$

Proof. We denote the (i,j) element of a matrix M by M_{ij}, etc. Then we have

$$(I_n-\widetilde{LA})_{ij}=\delta_{ij}-\sum_{k=1}^{n} l_{ik}a_{kj}(1+\xi_{ikj})\ (|\xi_{ikj}|<1.006\theta_n)$$

$$=\delta_{ij}-\sum_{j=1}^{n} l_{ik}a_{kj}-\sum_{k=1}^{n} l_{ik}a_{kj}\xi_{ikj}$$

$$=(I_n-LA)_{ij}-\sum_{k=1}^{n} l_{ik}a_{kj}\xi_{ikj}$$

where δ_{ij} denotes the Kronecker symbol. Therefore

$$\mathrm{fl}(I_n-\widetilde{LA})_{ij}=(I_n-\widetilde{LA})_{ij}+\delta_{ij}(I_n-\widetilde{LA})_{ii}\eta_i \quad (|\eta_i|<\beta^{1-t})$$

$$=(I_n-LA)_{ij}-\sum_{k=1}^{n} l_{ik}a_{kj}\xi_{ikj}+\delta_{ij}(I_n-\widetilde{LA})_{ii}\eta_i$$

so that we can write

$$\tilde{K}=K+\delta K$$

where

$$|(\delta K)_{ij}|\leqq\sum_{k=1}^{n}|l_{ik}a_{kj}\xi_{ikj}|+|\delta_{ij}(I_n-\widetilde{LA})_{ii}\eta_i|$$

$$<1.006\theta_n\sum_{k=1}^{n}|l_{ik}|\,|a_{kj}|+\delta_{ij}(1-\beta^{1-t})^{-1}\tilde{K}_{ii}\beta^{1-t}$$

$$<\theta_n E_{ij}\,. \qquad \text{Q.E.D.}$$

Lemma 6. *Let κ_i be defined as in the remark at the end of Section 2 and $e=(e_1,\cdots,e_n)^t$ where*

$$e_i=\sum_{j=1}^{n}E_{ij}+1.006\sum_{j=1}^{n}\tilde{\kappa}_{ij}\,.$$

Then we have

$$\tilde{\kappa}=\kappa+\delta\kappa\,,\quad \nu[\delta\kappa]\leqslant\theta_n e\,.$$

Proof. We have

$$\tilde{\kappa}_i=\mathrm{fl}(\tilde{\kappa}_{i1}+\cdots+\tilde{\kappa}_{in})$$

$$=\sum_{j=1}^{n}\tilde{\kappa}_{ij}+\sum_{j=1}^{n}\tilde{\kappa}_{ij}\xi_j\ (|\xi_j|<1.006\theta_{n-1})$$

$$=\kappa_i+\delta\kappa_i$$

where

$$\delta\kappa_i=\sum_{j=1}^{n}(\delta K)_{ij}+\sum_{j=1}^{n}\tilde{\kappa}_{ij}\xi_j\ .$$

Hence

$$|\delta\kappa_i|\leqq\sum_{j=1}^{n}E_{ij}\theta_n+1.006\theta_{n-1}\sum_{j=1}^{n}\tilde{\kappa}_{ij}<\theta_n e_i\ . \qquad \text{Q.E.D.}$$

Lemma 7. *Let*

$$a=\frac{\|\varepsilon\|_\infty}{1-\|K\|_\infty}$$

and

$$f=1.004\left\{\frac{\|d\|_\infty+n^{-1}\|\tilde{\varepsilon}\|_\infty}{1-\|\tilde{\kappa}\|_\infty}+\frac{n^{-1}\|\varepsilon\|_\infty}{1-\|\kappa\|_\infty}+\frac{\|\varepsilon\|_\infty\|e\|_\infty}{(1-\|\tilde{\kappa}\|_\infty)(1-\|\kappa\|_\infty)}\right\}. \tag{3.8}$$

Then we have

$$\tilde{a}=a+\delta a\ . \quad |\delta a|<f\theta_n\ .$$

Proof. Let $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)^t$

$$\max_i\tilde{\varepsilon}_i=\tilde{\varepsilon}_p \quad\text{and}\quad \max_i\varepsilon_i=\varepsilon_q\ .$$

Then

$$\delta\varepsilon_p\geqq\delta\varepsilon_p+(\varepsilon_p-\varepsilon_q)=\tilde{\varepsilon}_p-\varepsilon_q\geqq\delta\varepsilon_q$$

so that

$$|\tilde{\varepsilon}_p-\varepsilon_q|\leqq\max(|\delta\varepsilon_p|,|\delta\varepsilon_q|)\leqq\|d\|_\infty\theta_n\ .$$

Hence we can write $\tilde{\varepsilon}_p=\varepsilon_q+\delta\varepsilon_\infty$ or

$$\|\tilde{\varepsilon}\|_\infty=\|\varepsilon\|_\infty+\delta\varepsilon_\infty$$

where $|\delta\varepsilon_\infty|\leqq\|d\|_\infty\theta_n$. Similarly we have

$$\|\tilde{\kappa}\|_\infty=\|\kappa\|_\infty+\delta\kappa_\infty$$

where $|\delta\kappa_\infty| \leqq \|e\|_\infty \theta_n$. Therefore we have

$$\tilde{a} = \frac{\|\tilde{\varepsilon}\|_\infty}{\mathrm{fl}(1-\|\tilde{\kappa}\|_\infty)}(1+\xi_1) \quad (|\xi_1| < \beta^{1-t})$$

$$= \frac{(\|\varepsilon\|_\infty + \delta\varepsilon_\infty)(1+\xi_1)}{(1-\|\kappa\|_\infty - \delta\kappa_\infty)(1+\xi_2)} \quad (|\xi_2| < \beta^{1-t})$$

so that

$$\tilde{a} - a = \frac{(1-\|\kappa\|_\infty)(\delta\varepsilon_\infty + \|\tilde{\varepsilon}\|_\infty \xi_1) + \|\varepsilon\|_\infty \delta\kappa_\infty - \|\varepsilon\|_\infty (1-\|\tilde{\kappa}\|_\infty)\xi_2}{(1-\|\tilde{\kappa}\|_\infty)(1-\|\kappa\|_\infty)(1+\xi_2)}.$$

It follows from this that

$$|\tilde{a} - a| < f\theta_n$$

where f is defined in (3.8). Q.E.D.

We are now in a position to prove the following theorem.

Theorem 2. *Let α be the vector defined in Theorem* 1 *with the maximum norm. Then, under the assumption of Theorem* 1, *we have*

$$\alpha \leqslant \tilde{\alpha} + \delta\tilde{\alpha}$$

where

$$\delta\tilde{\alpha} = \{1.004 n^{-1}\tilde{\alpha} + d + ae + (f + \tilde{a} n^{-1})\tilde{\kappa}\}\theta_n : \tag{3.9}$$

Proof. We have

$$\tilde{\alpha}_i = \mathrm{fl}\{\tilde{\varepsilon}_i + \mathrm{fl}(\tilde{a}\tilde{\kappa}_i)\}$$

$$= \{\tilde{\varepsilon}_i + \mathrm{fl}(\tilde{a}\tilde{\kappa}_i)\}(1+\xi)$$

$$= \varepsilon_i + \delta\varepsilon_i + (a+\delta a)(\kappa_i + \delta\kappa_i)(1+\eta) + \{\tilde{\varepsilon}_i + \mathrm{fl}(\tilde{a}\tilde{\kappa}_i)\}\xi$$

$$= \alpha_i + \delta\varepsilon_i + a\delta\kappa_i + \delta a\tilde{\kappa}_i + \tilde{a}\tilde{\kappa}_i\eta + \{\tilde{\varepsilon}_i + \mathrm{fl}(\tilde{a}\tilde{\kappa}_i)\}\xi$$

where $|\xi|, |\eta| < \beta^{1-t}$. Hence

$$|\tilde{\alpha}_i - \alpha_i| < d_i\theta_n + ae_i\theta_n + (f\theta_n)\tilde{\kappa}_i + \tilde{a}\tilde{\kappa}_i\beta^{1-t} + (1-\beta^{1-t})^{-1}\tilde{\alpha}_i\beta^{1-t}$$

$$< \{d_i + ae_i + f\tilde{\kappa}_i + \tilde{a}\tilde{\kappa}_i n^{-1} + 1.004\tilde{\alpha}_i n^{-1}\}\theta_n$$

which means

$$\alpha \leqslant \tilde{\alpha} + \delta\alpha$$

where $\delta\alpha$ is defined by (3.9). Q.E.D.

We have from Lemmas 3–7 that

$$d_i \leqq (1.006\|\tilde{r}\|_\infty + \|c\|_\infty)\|L\|_\infty$$
$$< [(1.006 + 1.004n^{-1})\|\tilde{r}\|_\infty + 1.006\|A\|_\infty\|x^{(0)}\|_\infty]\|L\|_\infty$$

and

$$e_i < \|E\|_\infty + 1.006(1 - 1.006\theta_{n-1})^{-1}\mathrm{fl}(\tilde{\kappa}_{i1} + \cdots + \tilde{\kappa}_{in})$$
$$< 1.006\|L\|_\infty\|A\|_\infty + 1.004n^{-1}\max_i \tilde{\kappa}_{ii} + 1.02\|\tilde{\kappa}\|_\infty .$$

Therefore, we can say that, if $\|\tilde{\kappa}\|_\infty \ll 1$ and

$$\|L\|_\infty(\|A\|_\infty + \|A\|_\infty\|x^{(0)}\|_\infty + \|\tilde{r}\|_\infty)\theta_n$$

is small enough, then each component of the vector $\delta\alpha$ is small as compared with that of α and our method works well. Observe also that, for our purpose, we need not know the exact $\delta\alpha$. It suffices to know the order of each component. Hence, in practice, the following result may be useful:

Theorem 3. *Let*

$$\tilde{A}_\infty = \max_i \mathrm{fl}(|a_{i1}| + \cdots + |a_{in}|),$$

$$\tilde{L}_\infty = \max_i \mathrm{fl}(|l_{i1}| + \cdots + |l_{in}|),$$

$$\tilde{\kappa}_\infty = \max_i \mathrm{fl}(\tilde{\kappa}_{i1} + \cdots + \tilde{\kappa}_{in}) \quad (= \|\tilde{\kappa}\|_\infty),$$

$$\tilde{c}_\infty = 1.02\,\tilde{A}_\infty\|x^{(0)}\|_\infty + 0.502\|\tilde{r}\|_\infty ,$$

$$\tilde{d}_\infty = 1.02\,\tilde{L}_\infty(\|\tilde{r}\|_\infty + \tilde{c}_\infty),$$

and

$$\tilde{e}_\infty = 1.03(\tilde{L}_\infty\tilde{A}_\infty + \tilde{\kappa}_\infty) + 0.502\max \tilde{\kappa}_{ii}$$

Further, assume that $\tilde{e}_\infty\theta_n < 1$ *and there exists a positive number* m *such that* $\tilde{\kappa}_\infty < 1 - m^{-1} - \tilde{e}_\infty\theta_n$. *Set*

$$\tilde{f}_\infty = 1.004(m-1)\{(1 + n^{-1} + m\tilde{e}_\infty\theta_n)\tilde{d}_\infty + (2n^{-1} + m\tilde{e}_\infty)\|\tilde{\varepsilon}\|_\infty\}$$

and

$$\Delta\tilde{\alpha}_i = 1.004\beta^{1-t}\tilde{\alpha}_i + \{\tilde{d}_\infty(1+m\tilde{e}_\infty\theta_n) + m\|\tilde{\varepsilon}\|_\infty\tilde{e}_\infty + \tilde{f}_\infty + n^{-1}\tilde{\alpha}\tilde{\kappa}_\infty\}\theta_n .$$

Then

$$|x_i^* - x_i^{(0)}| \leqq \tilde{\alpha}_i + \Delta\tilde{\alpha}_i , \quad i=1, 2, \cdots, n.$$

Proof. We obtain from Theorem 2,

$$\|c\|_\infty \leqq 1.006\|A\|_\infty\|x^{(0)}\|_\infty + 1.004n^{-1}\|\tilde{r}\|_\infty$$
$$< 1.006(1-1.006\theta_{n-1})^{-1}\tilde{A}_\infty\|x^{(0)}\|_\infty < \tilde{c}_\infty ,$$
$$\|d\|_\infty \leqq \|L\|_\infty(1.006\|r\|_\infty + \|c\|_\infty) < 1.02\tilde{L}_\infty\|\tilde{r}\|_\infty + 1.011\tilde{c}_\infty$$
$$< \tilde{d}_\infty$$

and

$$\|e\|_\infty < \|E\|_\infty + 1.006(1-1.006\theta_{n-1})^{-1}\tilde{\kappa}_i$$
$$< 1.006\|L\|_\infty\|A\|_\infty + 1.004n^{-1}\max_i \tilde{\kappa}_{ii} + 1.02\tilde{\kappa}_i$$
$$< 1.006(1-0.01)^{-1}\tilde{L}_\infty(1-0.01)^{-1}\tilde{A}_\infty + 0.502\max_i \tilde{\kappa}_{ii} + 1.02\tilde{\kappa}_i$$
$$< \tilde{e}_\infty .$$

Moreover, if $\tilde{\kappa}_\infty < 1 - m^{-1} - \tilde{e}_\infty\theta_n$ for some $m>0$, then we have

$$(1-\|\tilde{\kappa}\|_\infty)^{-1}\|\tilde{\kappa}\|_\infty < m-1, \quad (1-\|\kappa\|_\infty)^{-1} < (1-\|\tilde{\kappa}\|_\infty - \tilde{e}_\infty\theta_n)^{-1} < m$$

and

$$(1-\|\kappa\|_\infty)^{-1}\tilde{\kappa}_\infty < m-1 .$$

Hence

$$ae_i \leqq m\|\varepsilon\|_\infty\tilde{e}_\infty \leqq m(\|\tilde{\varepsilon}\|_\infty + \tilde{d}_\infty\theta_n)\tilde{e}_\infty$$

and

$$f\tilde{\kappa}_i \leqq f\|\tilde{\kappa}\|_\infty < 1.004(m-1)\{\tilde{d}_\infty + n^{-1}\|\tilde{\varepsilon}\|_\infty + n^{-1}\|\varepsilon\|_\infty + m\|\varepsilon\|_\infty\tilde{e}_\infty\}$$
$$< 1.004(m-1)\{\tilde{d}_\infty + n^{-1}\|\tilde{\varepsilon}\|_\infty + (n^{-1} + m\tilde{e}_\infty)(\|\tilde{\varepsilon}\|_\infty + \tilde{d}_\infty\theta_n)\} = \tilde{f}_\infty .$$

The result now follows from Theorem 2. Q.E.D.

§ 4. Numerical Examples

We shall illustrate our results by simple examples.

Example 1. Consider the linear system $Ax=b$ given by

$$0.51273x_1+0.62137x_2=0.14012$$

$$0.41835x_1+0.50701x_2=0.34827$$

which is due to Peters and Wilkinson [4]. As is remarked there, this is extremely ill-conditioned and has the exact solution vector $x^*=(-15977.7406\cdots, 13184.4264\cdots)^t$. We solve this system by Gaussian ellimination. A single precision computation (chopping the results to 6 hexadecimal digits in the mantissa) on FACOM 230-28 computer of Ehime University yields

$$x^{(0)}=(-15594.90, 12868.53)^t. \tag{4.1}$$

The matrix L, a numerical result for A^{-1}, is also given by

$$L=\begin{pmatrix} 0.5439359E+5 & -0.6666244E+5 \\ -0.4488187E+5 & 0.5500726E+5 \end{pmatrix}.$$

We then compute $K=\nu[I_2-LA]$, $\varepsilon=\nu[L(Ax^{(0)}-b)]$ and α, etc., with double precision arithmetic (chopping the results to 14 hexadecimal digits in the mantissa). Then $\|\tilde{\kappa}\|_\infty(=\tilde{\kappa}_\infty)=0.028\cdots<1$ and Theorem 1 is applicable. The results are shown in Table 1.

Table 1. Error bounds for $x^{(0)}$ given by (4.1).

i	$\tilde{\alpha}$	$\tilde{\alpha}^{(1)}$	$\tilde{\varepsilon}$	$\tilde{r}$
1	384.5585…	382.8805…	373.669…	0.686…E-2
2	317.2004…	315.9270…	308.326…	−0.226…E-5

Next, we apply Theorem 3 to estimate the effect of the errors made in the computation. Observe that, in our computer, $\beta=16$ and $t=14$. Then we have

$$\tilde{d}_\infty=0.558\cdots E+5,\ \tilde{e}_\infty=0.141\cdots E+6,\ \tilde{\kappa}_\infty<0.5-\tilde{e}_\infty\theta_n,\ \text{etc.},$$

so that we take $m=2$ for simplicity to compute f_∞ and obtain

$$\tilde{f}_\infty=0.74\cdots E+5 \quad \text{and} \quad \|\Delta\tilde{\alpha}\|_\infty=0.252\cdots E-5<0.253E-5\,.$$

This implies that

$$\nu[x^*-x^{(0)}]\leqslant\tilde{\alpha}+0.253E-5\begin{pmatrix}1\\1\end{pmatrix}\leqslant\begin{pmatrix}384.5586\\317.2005\end{pmatrix},$$

or

$$\begin{pmatrix}-15979.46\\12551.32\end{pmatrix}\leqslant x^*\leqslant\begin{pmatrix}-15210.35\\13185.74\end{pmatrix}.$$

On the other hand, if we use the double precision arithmetic to compute $x^{(0)}$ and L, then we have

$$x^{(0)}=(-15977.74063\cdots, 13184.42647\cdots)^t, \tag{4.2}$$

$$L=\begin{pmatrix}0.557288\cdots E+5 & -0.682989\cdots E+5\\ -0.459836\cdots E+5 & 0.563575\cdots E+5\end{pmatrix},$$

and

$$\tilde{\kappa}_\infty=0.109\cdots E-10.$$

The large change of $x^{(0)}$ from (4.1) to (4.2) (as well as L) reflects the ill-conditionality of the system. The results of Theorem 1 applied to $x^{(0)}$ given by (4.2) and the above L are shown in Table 2.

Table 2. Error bounds for $x^{(0)}$ given by (4.2).

i	$\tilde{\alpha}$	$\tilde{\alpha}^{(1)}$	$\tilde{\varepsilon}$	$\tilde{r}$
1	0.392526…E-7	0.392526…E-7	0.392526…E-7	0.181…E-11
2	0.323868…E-7	0.323868…E-7	0.323868…E-7	0.909…E-12

In this case, we have $\tilde{e}_\infty=0.144\cdots E+6$ and again take $m=2$ to compute f_∞. Then we obtain $\|\Delta\tilde{\alpha}\|_\infty=0.260\cdots E-5$ (which is larger than that of the single precision arithmetic). Thus we can assert that

$$\begin{pmatrix}-15977.74064\\13184.42646\end{pmatrix}\leqslant x^*\leqslant\begin{pmatrix}-15977.74062\\13184.42648\end{pmatrix}.$$

Example 2. Consider the linear system given by

$$0.876543x_1+0.617341x_2+0.589973x_3=0.863257$$

$$0.612314x_1+0.784461x_2+0.827742x_3=0.820647$$

$$0.317321x_1+0.446779x_2+0.476349x_3=0.450098$$

which is found in Wilkinson [7] and is discussed also in Yamamoto [9]. This system is ill-conditioned, too. We again solve this by Gaussian ellimination with single precision arithmetic. Then we obtain a numerical solution

$$(4.3)\qquad x^{(0)}=(0.6363233,\ -0.2946413E-1,\ 0.5486381)^t.$$

At the same time, we have a matrix L, approximation for A^{-1}, such that $\tilde{L}_\infty=0.657\cdots E+5$ (see Yamamoto [9]). In this case, by the double precision computation, we have

$$\tilde{\kappa}_\infty=0.967\cdots E-2 \quad\text{and}\quad \tilde{e}_\infty=0.150\cdots E+6.$$

The vectors $\tilde{\alpha}$, $\tilde{\alpha}^{(1)}$, $\tilde{\varepsilon}$ and $\tilde{r}$ are shown in Table 3.

Table 3. Error bounds for $x^{(0)}$ given by (4.3).

i	$\tilde{\alpha}$	$\tilde{\alpha}^{(1)}$	$\tilde{\varepsilon}$	$\tilde{r}$
1	0.573591…E-5	0.570495…E-5	0.560957…E-5	0.192…E-7
2	0.427810…E-4	0.426081…E-4	0.423670…E-4	0.303…E-7
3	0.362315…E-4	0.361321…E-4	0.359655…E-4	0.165…E-7

Further, if we apply Theorem 3 by taking $m=2$, then we have

$$\Delta\tilde{\alpha}_i=0.150\cdots E-1<0.151E-9$$

which implies that

$$\nu[x^*-x^{(0)}]\leqslant\tilde{\alpha}^{(0)}+0.151E-9\begin{pmatrix}1\\1\\1\end{pmatrix}\leqslant\begin{pmatrix}0.573607E-5\\0.427813E-4\\0.362317E-4\end{pmatrix},$$

or

$$\begin{pmatrix}0.6363177\\-0.295069E-1\\0.5486019\end{pmatrix}\leqslant x^*\leqslant\begin{pmatrix}0.6363291\\-0.2942135E-1\\0.5486743\end{pmatrix}.$$

If we compute $x^{(0)}$ and L by the double precision arithmetic, then we have

$$(4.4)\qquad x^{(0)}=(0.63632896\cdots,\ -0.29506656\cdots E-1,\ 0.54867420\cdots)^t,$$

and

$$\tilde{L}_\infty=0.66\cdots E+5.$$

The double precision computation yields $\tilde{\kappa}_\infty = 0.272\cdots E-11$ and $\tilde{e}_\infty = 0.151\cdots E+6$. The vectors $\tilde{\alpha}$ and $\tilde{\alpha}^{(1)}$, etc., are shown in Table 4.

Table 4. Error bounds for $x^{(0)}$ given by (4.4).

i	$\tilde{\alpha}$	$\tilde{\alpha}^{(1)}$	$\tilde{\varepsilon}$	$\tilde{r}$
1	0.157211…E-14	0.157211…E-14	0.157…E-14	−0.138…E-16
2	0.126332…E-13	0.126332…E-13	0.126…E-13	0.416…E-16
3	0.108600…E-13	0.108600…E-13	0.108…E-13	0.277…E-16

Further we have

$$\Delta\tilde{\alpha}_i = 0.151\cdots E-9$$

where we have taken $m=2$ to compute f_∞. Thus we obtain

$$\nu[x^* - x^{(0)}] \leqslant \tilde{\alpha} + 0.152E-9\begin{pmatrix}1\\1\end{pmatrix} \leqslant 0.153E-9\begin{pmatrix}1\\1\end{pmatrix}.$$

Example 3. Consider the linear system

$$33x_1 + 16x_2 + 72x_3 = 152.833$$
$$-24x_1 - 10x_2 - 57x_3 = -94.324$$
$$-8x_1 - 4x_2 - 17x_3 = -38.308$$

which has the exact solution $x^* = (-0.001, 10, -0.1)^t$. Then, by the single precision computation, we obtain

$$x^{(0)} = (0.1018889E-2,\ 0.9999983E+1,\ -0.1000051)^t \tag{4.5}$$

and

$$L = \begin{pmatrix} -9.667\cdots & -2.666\cdots & -32.001\cdots \\ 8.003\cdots & 2.500\cdots & 25.501\cdots \\ 2.666\cdots & 0.666\cdots & 9.000\cdots \end{pmatrix},$$

so that the system is well-conditioned. The double precision computation yields

$$\tilde{\kappa}_\infty = 0.213\cdots E-3$$

and the vector $\tilde{\alpha}$, $\tilde{\alpha}^{(1)}$ and $\tilde{\varepsilon}$, etc., are shown in Table 5. Further we have from Theorem 3

$$\|\Delta\tilde{\alpha}\|_\infty = 0.869\cdots E-10 < 0.87E-10.$$

Table 5. Error bounds for $x^{(0)}$ given by (4.5).

i	$\tilde{\alpha}$	$\tilde{\alpha}^{(1)}$	$\tilde{\varepsilon}$	$\tilde{r}$
1	0.188865…E-4	0.188848…E-4	0.188825…E-4	−0.223…E-4
2	0.171678…E-4	0.171674…E-4	0.171667…E-4	0.120…E-4
3	0.515085…E-5	0.515052…E-5	0.515004…E-5	0.515…E-5

Therefore, our method works well in this case, too.

References

[1] Collatz, L., *Functional analysis and numerical mathematics*, Academic Press, New York, 1966.

[2] Forsythe, G. and Moler, C., *Computer solution of linear algebraic systems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[3] Paige, C. C., Error analysis of the symmetric Lanczos process for the eigenproblem, *London Univ. Inst. of Computer Science, Tech. Note ICSI*, **209**, 1969.

[4] Peters, G. and Wilkinson, J. H., Inverse iteration, ill-conditioned equations and Newton method, *SIAM Rev.*, **21** (1979), 339-360.

[5] Urabe, M., A posteriori component-wise error estimation of approximate solutions to nonlinear equations, *Lecture Notes in Computer Sci.*, **29**, Springer, New York, 1975.

[6] Wilkinson, J. H., *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[7] ———, *The algebraic eigenvalue problem*, Oxford Univ. Press, London, 1965.

[8] Yamamoto, T., Error bounds for computed eigenvalues and eigenvectors, *Numer. Math.*, **34** (1980), 189-199.

[9] ———, Error bounds for approximate solutions of systems of equations, to appear.

[10] ———, Componentwise error estimates for approximate solutions of systems of equations, *Lecture Notes in Num. Appl. Anal.*, **3** (1981), 1-22.