RICE UNIVERSITY

ROUNDING ERRORS IN THE SOLUTION OF MATRIX EQUATIONS
WITH DIAGONALLY DOMINANT MATRICES HAVING
POSITIVE ELEMENTS ON THE PRINCIPAL DIAGONAL
AND NON-POSITIVE ELEMENTS ELSEWHERE

by

Randa Suzanne Randolph

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

Thesis Director's signature:

Houston, Texas
May, 1968
Abstract

ROUNDING ERRORS IN THE SOLUTION OF MATRIX EQUATIONS WITH DIAGONALLY DOMINANT MATRICES HAVING POSITIVE ELEMENTS ON THE PRINCIPAL DIAGONAL AND NON-POSITIVE ELEMENTS ELSEWHERE

Randa Suzanne Randolph

In solving the matrix equation \( A\bm{w} = \bm{d} \), where \( A = \{ a_{ij} \} \) is an irreducible \( J \times J \) matrix such that

(i) \( \bar{c} + \sum_{j \neq i} a_{ij} \leq a_{ii}, \ i = 1, 2, \ldots, J \)

(ii) \( -1 \leq \sum_{j \neq i} a_{ij} < 0, \ i = 1, 2, \ldots, J \)

(iii) \( a_{ij} \leq 0, \ i \neq j \)

(iv) \( \bar{c} > 0 \),

we wish to find a bound which is independent of \( J \) for \( \frac{|| \hat{\bm{w}} - \bm{w} ||_{\infty}}{|| \hat{\bm{w}} ||_{\infty}} \), where \( \hat{\bm{w}} \) is the computed solution and \( \bm{w} \) is the exact solution. The solution is to be computed using floating point arithmetic with a \( t \)-digit mantissa. All inner products are computed in double precision.

Factoring \( A \) into the product of a lower triangular matrix, \( \hat{L} \), and a unit upper triangular matrix, \( \hat{U} \), results in \( \hat{L}\hat{U} = A+E \), where \( E \) is the matrix whose elements are the errors introduced
in the computation of \( \hat{L} \) and \( \hat{U} \). By computing \( \hat{y} \), which solves exactly \((\hat{L}+\delta\hat{L})\hat{y} = d\), and then \( \hat{w} \), which solves exactly \((\hat{U}+\delta\hat{U})\hat{w} = \hat{y} \), we see that \( \hat{w} \) is the exact solution of \((A+K)\hat{w} = d\), where \( K = E + \hat{L} \delta\hat{U} + \delta\hat{L} \hat{U} + \delta\hat{L} \delta\hat{U} \). When \( \bar{\gamma} > 5(N^{-1-t_1}) \), where \( N \) is the base of the number system and \( t_1 = t - \log_N 1.053 \), we are able to exhibit bounds for \( \|A^{-1}\|_\infty, \|E\|_\infty, \|\hat{L}\|_\infty, \|\hat{U}\|_\infty, \|\delta\hat{L}\|_\infty \) and \( \|\delta\hat{U}\|_\infty \) so that

\[
\frac{\|\hat{w} - w\|_\infty}{\|\hat{w}\|_\infty} < (8.95 \max_{i \leq 1} a_{ii} + 3.03) \frac{\nu}{\delta} + 6.17 \frac{\nu}{\delta^2},
\]

where \( \delta = \bar{\delta} - 5\nu \). This bound is independent of \( J \).
Acknowledgments

Of the many people who deserve my gratitude, I would especially like to thank Professor Henry H. Rachford, Jr., my thesis director, whose understanding guidance encouraged me in this undertaking. I am also indebted to Professors B. Frank Jones, Jr., and Marc Q. Jacobs, who served on my thesis committee, and to Miss Janet Gordon, who patiently typed the very difficult manuscript. I would also like to thank the National Science Foundation whose generous support in the form of an N.S.F. Traineeship enabled me to continue my education on the graduate level at Rice University.
We wish to solve $A\omega = d$ for $\omega$, where $A = \{a_{ij}\}$, irreducible such that

(i) $\chi + |\sum_{j \neq i} a_{ij}| \leq a_{ii}$, $i=1,2,\ldots,J$

(ii) $1 \leq \sum_{j \neq i} a_{ij} < 0$, $i=1,2,\ldots,J$

(iii) $a_{ij} \leq 0$, $i \neq j$

(iv) $\chi > 0$.

We shall let $A = LU$, where $U$ is a unit upper triangular matrix, and $L$ is a lower triangular matrix. We shall first determine $\gamma$ such that $L\gamma = d$ and then $\omega$ such that $U\omega = \gamma$. The vector $\omega$ thus determined should satisfy $A\omega = d$.

The following lemma is a generalization of Lemma 3.1 of (1).

Lemma 1: For $A$ satisfying 1.1, $\|A^{-1}\|_\omega \leq \frac{1}{\delta}$.

Proof: Recall that for $S = \{s_{ij}\}$, $\|S\|_\omega = \max_i \sum_j |s_{ij}|$ and for a vector $v = \{v_k\}$, $\|v\|_\omega = \max_k |v_k|$. Let $H = A - \delta I$, then $H$ and $A$ are both diagonally dominant and irreducible. An elementary argument (2, page 85) shows $A^{-1} > 0$, where the matrix inequality is to be interpreted element-by-element. Let $y = A^{-1}\xi$, where $\xi = 1, i = 1,2,\ldots,J$. Then $\|y\|_\omega = \|A^{-1}\|_\omega$, and since $y = (H + \delta I)^{-1}\xi$, $\|y\|_\omega = (H + \delta I)^{-1}\xi$.
we have \(0 < \delta y = \xi - (H + \delta I)^{-1}H\xi\). As \((H + \delta I)^{-1} > 0\) and \(H\xi \geq 0\), \(\delta y \leq \xi\). Therefore \(\|y\|_\infty \leq \frac{1}{\xi}\), and the lemma is proved.

In order to study in detail the algorithm induced by the factorization \(A = LU\) and by \(LY = d\), \(U_w = \gamma\), we set

\[
L = \begin{pmatrix}
\ell_{11} & & \\
\ell_{21} & \ell_{22} & \\
\vdots & & \ddots \\
\ell_{J1} & \ell_{J2} & \cdots & \ell_{JJ}
\end{pmatrix}, \quad U = \begin{pmatrix}
1 & u_{12} & u_{13} & \cdots & u_{1J} \\
1 & u_{23} & \cdots & u_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
1 & \cdots & \cdots & 1 & u_{J-1,J}
\end{pmatrix},
\]

and the recursions follow as \(r = 1, 2, \ldots, J\).

1.2

(i) \(\ell_{ir} = a_{ir} - \sum_{k=1}^{r-1} \ell_{ik}u_{kr}\), \(i=r, r+1, \ldots, J\)

(ii) \(u_{rj} = \frac{a_{rj} - \sum_{k=1}^{r-1} \ell_{rk}u_{kj}}{\ell_{rr}}\), \(j=r+1, r+2, \ldots, J\)

(iii) \(\gamma_r = \frac{d_r - \sum_{k=1}^{r-1} \ell_{rk}\gamma_k}{\ell_{rr}}\)

where \(\sum_{k=1}^{r-1} \ell_{ik}u_{kj} = \sum_{k=1}^{r-1} \ell_{rk}\gamma_k = 0\).

The solution \(w\) is then determined by

1.3

(i) \(w_J = \gamma_J\)

(ii) \(w_r = \gamma_r - \sum_{k=r+1}^{J} u_{rk}w_k\), \(r=J-1, J-2, \ldots, 1\).
We shall now analyze the error introduction and propagation in the factorization of $A$ into $LU$. We shall assume that all computation, with the exception of inner products, is carried out in floating point base $N$ arithmetic with $t$ digits in the mantissa. All inner products are computed with a $f_l^2$ operator, i.e., the sum is produced as a double-precision number which is then used in subsequent calculations with the final answer being given in single precision. It is shown in (3) that the following holds

(i) $f_l(a \cdot b) = (a \cdot b)(1 + \eta)$

$$1.4$$

(ii) $f_l^2(\Sigma ab) = \Sigma ab(1 + u_k)$,

where $\square$ is $+,-,\times$, or $\div$ and $|\eta| \leq \nu' < \nu$, $|u_k| \leq (n-k+2) \nu_2$, where $\nu' = N^{1-t}$, $\nu = N^{1-t_1}$, $t_1 = t - \log N 1.053$, and $\nu n < 1$. The value of $\nu_2$ is $N^{1-t_2}$, where $t_2 = 2t - \log N 1.053$. We shall also assume that $J \nu_2 < \nu$, so that $|f_l^2(\Sigma ab)| \leq (1 + \nu) \Sigma |ab|$ for $n < J$.

Following Wilkinson (3) in computing $l_{ir}$ and $u_{rj}$, and using a symbol with a caret, e.g., $\hat{u}$, to denote the computed value of the exact element $u$, we have

(i) $\hat{l}_{ir} = f_l(a_{ir} - f_l^2(\Sigma \hat{l}_{ik} \hat{u}_{kr}))$

$$1.5$$

$$= (1 + \eta)(a_{ir} - \Sigma \hat{l}_{ik} \hat{u}_{kr}(1 + u_k)),$$ $i=1,2,...,r-1$
(ii) \( \hat{u}_{rj} = \frac{a_{rj} - f_2(\sum_{k=1}^{r-1} \hat{u}_{rk} u_{kj})}{\hat{\ell}_{rr}} \)

\[
= (1+\varepsilon)(1+\rho) \left( \frac{a_{rj} - \sum_{k=1}^{r-1} \hat{u}_{rk} u_{kj}(1+\beta_k)}{\hat{\ell}_{rr}} \right), \quad j=r+1, r+2, \ldots, J,
\]

where \( |\eta|, |\xi|, |\rho| \leq \nu', \quad |u_k|, |\beta_k| \leq (r-k+1)\nu_2 \), and it is assumed that the elements of \( A \) are stored and enter the computations as exact quantities.

To proceed, we shall postulate something trivially stronger than 1.1. Let 1.1' designate 1.1 with the following substitution

\[ 1.1 \text{ iv}' \quad \gamma > 5\nu. \]

We shall also assume that \( \nu' < 0.005 \), so that \((1+\nu')(1+\nu)(1+3\nu) < 1+5\nu. \) From 1.5 we have

(i) \( |\hat{\ell}_{ir}| \leq (1+\nu')(|a_{ir}|+(1+\nu) \sum_{k=1}^{r-1} |\hat{u}_{ik}| |\hat{u}_{kr}|), \quad i=2, \ldots, J, \quad r=1, 2, \ldots, i-1 \)

(ii) \( |\hat{\ell}_{rr}| \geq (1-\nu')(a_{rr}-(1+\nu) \sum_{k=1}^{r-1} |\hat{u}_{rk}| |\hat{u}_{kr}|), \quad r=1, 2, \ldots, J \)

(iii) \( a_{rj} \leq (1+\nu')(1+\nu') \left( \frac{|a_{rj}|+(1+\nu) \sum_{k=1}^{r-1} |\hat{u}_{rk}| |\hat{u}_{kj}|}{|\hat{\ell}_{rr}|} \right) \leq \left( \frac{(1+\nu')(1+\nu')}{(1-\nu')} \right) \left( \frac{a_{rr}-(1+\nu) \sum_{k=1}^{r-1} |\hat{u}_{rk}| |\hat{u}_{kr}|}{a_{rr}-(1+\nu) \sum_{k=1}^{r-1} |\hat{u}_{rk}| |\hat{u}_{kr}|} \right), \)
\( r = 2, \ldots, J - 2, j = r + 1, \ldots, J \)

(iv) \( |u_{1j}| = (1 + \nu') \frac{|a_{1j}|}{a_{11}}, \quad |\hat{\ell}_{11}| = a_{11}, \quad j = 2, \ldots, J, i = 1, 2, \ldots, J. \)

**Lemma 2:** Let \( A \) satisfy 1.1', and let \( \hat{\ell}_{ir}, \hat{u}_{rj} \) be computed by by 1.5, then

\[
|\hat{\ell}_{rr}| > 0 \quad \text{and} \quad \sum_{j=r+1}^{J} |\hat{u}_{rj}| < \frac{1 + 3 \nu}{1 + \delta}, \quad r = 1, 2, \ldots, J - 1.
\]

**Proof:** The proof will be by induction from 1.1' and 1.6.

Clearly both assertions are true for \( r = 1 \), as \( |\hat{\ell}_{11}| = a_{11} > 0 \), and

\[
\sum_{j=2}^{J} |\hat{u}_{1j}| \leq (1 + \nu') \sum_{j=2}^{J} \frac{|a_{1j}|}{a_{11}} < \frac{(1 + \nu') \sum_{i=2}^{J} |a_{1j}|}{1 + \delta} < \frac{1 + 3 \nu}{1 + \delta}.
\]

Assume the assertions are true for \( 1 \leq r < n \). Then

\[
\sum_{j=n+1}^{J} |\hat{u}_{nj}| \leq \frac{(1 + \nu')(1 + \nu')}{(1 - \nu')(1 - \nu)} \left( \sum_{j=n+1}^{J} \left( \frac{|a_{nj}| + (1 + \nu') \sum_{k=1}^{n-1} |\hat{\ell}_{nk}|}{a_{nn} - (1 + \nu') \sum_{k=1}^{n-1} |\hat{\ell}_{nk}|} \right) \right)
\]

\[
\leq \frac{(1 + 3 \nu)(\sum_{j=n+1}^{J} |a_{nj}| + (1 + \nu') \sum_{k=1}^{n-1} |\hat{\ell}_{nk}|)}{1 + \delta} \sum_{j=n+1}^{J} |\hat{u}_{kj}| + \frac{n - 1}{\sum_{j=n+1}^{J} |\hat{u}_{kj}|} \left( \frac{|a_{nj}| + (1 + \nu') \sum_{k=1}^{n-1} |\hat{\ell}_{nk}|}{1 + \delta} \sum_{j=n+1}^{J} |\hat{u}_{kj}| \right)
\]

so long as the denominator remains positive, since \( (-|\hat{u}_{kn}|) \)

\[
> - \frac{1 + 3 \nu}{1 + \delta} \sum_{j=n+1}^{J} |\hat{u}_{kj}| + \sum_{j=k+1}^{n-1} |\hat{u}_{kj}| \quad \text{and} \quad a_{nn} \geq \delta + \sum_{j \neq n} |a_{nj}|.
\]
Let us restate the above inequality as

\[(i) \quad \sum_{j=n+1}^{J} |\mathbf{u}_{nj}| < \frac{1+3\nu}{\delta + \alpha_n} \]

\[+ \frac{n-1}{1+\frac{\delta}{\beta_n}} \]

1.7 \quad (ii) \quad \alpha_n = \sum_{j=1}^{n-1} |a_{nj}| + (1+\nu) \sum_{k=1}^{n-1} \hat{\mathbf{u}}_{nk} \left( \frac{1+3\nu}{1+\delta} \right) + \sum_{j=k+1}^{n} |\mathbf{u}_{kj}| \]

(iii) \quad \beta_n = \sum_{j=n+1}^{J} |a_{nj}| + (1+\nu) \sum_{k=1}^{n-1} \hat{\mathbf{u}}_{nk} \sum_{j=n+1}^{J} |\mathbf{u}_{kj}| ,

for \( \beta_n \neq 0 \). We note that if \( \beta_n = 0 \), \( \sum_{j=n+1}^{J} |\mathbf{u}_{nj}| = 0 < \frac{1+3\nu}{1+\delta} \)

Let us now consider

\[\alpha_n = \sum_{j=1}^{n-1} |a_{nj}| + (1+\nu) \sum_{k=1}^{n-1} \hat{\mathbf{u}}_{nk} \left( \frac{1+3\nu}{1+\delta} \right) + (1+\nu) \sum_{k=1}^{n-1} \hat{\mathbf{u}}_{nk} \sum_{j=k+1}^{n} |\mathbf{u}_{kj}| \]

\[= \sum_{j=1}^{n-1} \left( |a_{nj}| + (1+\nu) \sum_{k=1}^{j-1} \hat{\mathbf{u}}_{nk} \hat{\mathbf{u}}_{kj} \right) + \sum_{j=1}^{n-1} \hat{\mathbf{u}}_{nj} \left( \frac{1+\nu}{1+\delta} \right) \]

We note that since \((1-\nu')(\beta_n + \delta + \alpha_n) \leq \hat{\mathbf{u}}_{nn} \), \( \alpha_n \geq 0 \), \( \beta_n \geq 0 \), we have shown \( |\mathbf{u}_{nn}| > 0 \). Therefore

\[1.7i' \quad \sum_{j=n+1}^{J} |\mathbf{u}_{nj}| < \frac{1+3\nu}{\frac{\delta}{\beta_n}} \]

and the lemma will follow if we can establish \( \beta_n \leq 1 \).
We shall now prove by induction that \( |\hat{\nu}_{nk}| = \sum_{s=1}^{k} |a_{ns}|g_{sk} \), where we may determine appropriate bounds on \( g_{sk} \), \( k=1, \ldots, n-1 \), \( s=1, \ldots, k \). The value of \( g_{11} \) is 1, since \( |\hat{\nu}_{n1}| = |a_{n1}| \). Since

\[ |\hat{\nu}_{n2}| \leq (1+\nu')(|a_{n2}|+(1+\nu)|\hat{\nu}_{n1}||u_{12}|), \]

we see that \( |\hat{\nu}_{n2}| = \sum_{s=1}^{k} |a_{ns}|g_{s2} \), where \( g_{22} \leq (1+\nu') \), \( g_{21} \leq (1+\nu')(1+\nu)|u_{12}| \). Let us assume that \( |\hat{\nu}_{nk}| = \sum_{s=1}^{k} |a_{ns}|g_{sk} \), \( 1 \leq k < j \). Consider

\[ |\hat{\nu}_{nj}| \leq (1+\nu')(a_{nj}|+(1+\nu)\sum_{k=1}^{j-1} |\hat{\nu}_{nk}|u_{kj}|) \]

\[ = (1+\nu')|a_{nj}|+(1+\nu')(1+\nu)\sum_{k=1}^{j-1} \sum_{s=1}^{k} |a_{ns}|g_{sk}|u_{kj}| \]

\[ = (1+\nu')|a_{nj}|+(1+\nu')(1+\nu)\sum_{s=1}^{j-1} |a_{ns}|(\sum_{k=s}^{j-1} g_{sk}|u_{kj}|) \].

Therefore \( |\hat{\nu}_{nj}| = \sum_{s=1}^{j} |a_{ns}|g_{sj} \), where \( g_{jj} \leq (1+\nu') \) and \( g_{sj} \)

\[ \leq (1+\nu')(1+\nu)\sum_{k=s}^{j-1} g_{sk}|u_{kj}| \].

The next step is to show that

\[ (1+\nu)\sum_{k=1}^{n-1} |\hat{\nu}_{nk}| \sum_{j=n+1}^{J} |u_{kj}| \leq \sum_{s=1}^{n-1} |a_{ns}|. \]

We note first that \( (1+\nu)\sum_{k=1}^{n-1} |\hat{\nu}_{nk}| \sum_{j=n+1}^{J} |u_{kj}| = \sum_{k=1}^{n-1} k \sum_{s=1}^{n-1} |a_{ns}|g_{sk}(1+\nu)\sum_{j=n+1}^{J} |u_{kj}| \)

\[ = \sum_{s=1}^{n-1} |a_{ns}|(1+\nu)\sum_{k=s}^{J} |u_{kj}| = \sum_{s=1}^{n-1} |a_{ns}|f_{sn}, \] where
\[ f_{sn} = (1+\nu)^{n-1} \sum_{k=s}^{J} g_{sk} \sum_{j=n+1}^{\wedge} |u_{kj}|. \]  
We shall find an upper bound for the value of \( f_{sn} \) by replacing \( g_{sk} \) at each step with the appropriate bound as follows

\[ f_{sn} = (1+\nu)^{n-1} \sum_{k=s}^{J} g_{sk} \sum_{j=n+1}^{\wedge} |u_{kj}| \]

\[ \leq (1+\nu)^{n-1} \sum_{k=s}^{J} g_{sk} \sum_{j=n+1}^{\wedge} |u_{tk}| \]

\[ \leq (1+\nu)^{n-2} \sum_{k=t+1}^{J} g_{sk} \sum_{j=n+1}^{\wedge} |u_{tk}| \]

\[ = (1+\nu)^{s+2} \sum_{i=s}^{i+1} g_{si} \sum_{j=\wedge}^{s+2} |u_{ij}| \]

\[ \leq (1+\nu)^{s+1} \sum_{i=\wedge}^{s+2} g_{sm} \sum_{j=i+1}^{s+2} |u_{ij}| \]

\[ \leq (1+\nu)^{s+1} \sum_{i=\wedge}^{s+1} g_{sm} |u_{mi}| \]

\[ = (1+\nu)^{s+1} \sum_{i=\wedge}^{s+1} g_{sm} |u_{mi}| \]

\[ \leq (1+\nu)(1+\nu')(\frac{1+3\nu}{1+\delta}) < 1, \text{ proving 1.8.} \]
Therefore, from 1.7iii, 1.8, 1.1i, and 1.7i', we have $\beta_n < 1$

and $\sum_{j=n+1}^{J} |u_{nj}| < \frac{1+3\nu}{1+\delta}$, proving the second lemma.

From Lemma 2, we see that $\|\hat{U}\|_\infty < 1 + \frac{1+3\nu}{1+\delta}$. In order to find a bound for $\|L\|_\infty$, we consider

$$\sum_{r=1}^{i-1} |\lambda_{ir}| \leq (1+\nu') \sum_{r=1}^{i-1} |a_{ir}| + (1+\nu')(1+\nu) \sum_{r=1}^{i-1} \sum_{k=1}^{r-1} |\lambda_{ik}| |u_{kr}|$$

$$\leq (1+\nu') + (1+\nu')(1+\nu) \sum_{r=1}^{i-2} |\lambda_{ir}| + \sum_{r=1}^{i-1} |\lambda_{ir}| \left( \frac{1+5\nu}{1+\delta} \right) \leq (1+\nu') + \sum_{r=1}^{i-1} |\lambda_{ir}| \left( \frac{1+5\nu}{1+\delta} \right) .$$

From this we have $\sum_{r=1}^{i-1} |\lambda_{ir}| < (1+\nu') \left( \frac{1+5\nu}{1+\delta} \right)$, and $\sum_{r=1}^{i-1} |\lambda_{ir}| < \frac{(1+\delta')(1+\nu')}{\delta'-5\nu}$.

Therefore $\|L\|_\infty < (1+\nu') \max_i a_{ii} + \frac{(1+\delta')(1+\nu')}{\delta'-5\nu}$, since $|\lambda_{rr}| < (1+\nu') a_{rr}, r=1,2,\ldots,J$. We have thus proved

Lemma 3: For $A$ satisfying 1.1', and the elements of $\hat{L}$ and $\hat{U}$ computed by 1.5,

(i) $\|\hat{U}\|_\infty < 1 + \frac{1+3\nu}{1+\delta}$

(ii) $\|L\|_\infty < (1+\nu') \max_i a_{ii} + \frac{(1+\delta')(1+\nu')}{\delta'-5\nu}$.

Let us now rewrite 1.5 as
(i) $\hat{u}_{ir} = a_{ir} + \sum_{k=1}^{r-1} u_{ik}^r + e_{ir}, \quad i=r$

1.10

(ii) $\hat{u}_{ir} = \frac{a_{ir} + \sum_{k=1}^{i-1} u_{ik}^r}{\hat{u}_{ii}} + e_{ir}, \quad i<r$

where $e_{ir} = fl(a_{ir} - fl_2(\sum_{k=1}^{r-1} u_{ik}^r)) - (a_{ir} - \sum_{k=1}^{r-1} u_{ik}^r)$,

$\epsilon_{ir} = fl\left(\frac{a_{ir} - fl_2(\sum_{k=1}^{i-1} u_{ik}^r)}{\hat{u}_{ii}}\right) - \left(\frac{a_{ir} - \sum_{k=1}^{i-1} u_{ik}^r}{\hat{u}_{ii}}\right)$.

From 1.10, we note

(i) $\sum_{k=1}^{r} u_{ik}^r = a_{ir} + e_{ir}, \quad i=r$

1.11

(ii) $\sum_{k=1}^{i} u_{ik}^r = a_{ir} + e_{ir}, \quad i<r$

where $e'_{ir} = \hat{u}_{ii} e_{ir}$. Hence we see that in factoring $A$, the errors involved in the computation result in $\hat{\hat{a}} = A + E$, where $E$ is the matrix composed of the error terms in 1.11. In order to determine a bound for $\|E\|_\infty$, we must find bounds for $e_{ir}$ and $e'_{ir}$. From 1.10i, we have

$e_{ir} = (1+\eta)(a_{ir} - \sum_{k=1}^{r-1} u_{ik}^r (1+\omega_k)) - (a_{ir} - \sum_{k=1}^{r-1} u_{ik}^r)$

$= \eta(a_{ir} - \sum_{k=1}^{r-1} u_{ik}^r) - (1+\eta) \sum_{k=1}^{r-1} u_{ik}^r \omega_k$
Then \((l+\eta)e_{ir} = \eta \hat{e}_{ir} - (l+\eta) \sum_{k=1}^{r-1} \hat{u}_{ik} \hat{u}_{kr}^k\) and \(|e_{ir}| \leq \sqrt{\sum_{k=1}^{r} \hat{u}_{ik} \hat{u}_{kr}^k}\).

For \(i > r\), \(|e_{ii}| \leq \frac{\nu}{1-\nu}(1+\nu')a_{ii} + \nu \sum_{k=1}^{i-1} |\hat{u}_{ki}| \leq \nu a_{ii} + \nu \sum_{k=1}^{i-1} |\hat{u}_{ki}| u_{ki}^k|.

Therefore, \(\sum_{r=1}^{i} \frac{1}{e_{ir}} \leq \nu a_{ii} + \nu \sum_{k=1}^{i-1} |\hat{u}_{ki}| + \nu \sum_{k=1}^{i-1} \frac{1}{\hat{u}_{kr}^k} = \nu a_{ii} + \nu \sum_{k=1}^{i-1} |\hat{u}_{ki}| + \nu \sum_{k=1}^{i-1} \frac{1}{\hat{u}_{kr}^k},\)

From 1.10ii, we have

\[
e'_{ir} = \hat{\theta}_{ii} e_{ir} = (1+\xi)(1+\rho) (a_{ir} - \sum_{k=1}^{i-1} \hat{u}_{ik} \hat{u}_{kr}^k(1+\beta_k)) - (a_{ir} - \sum_{k=1}^{i-1} \hat{u}_{ik} \hat{u}_{kr}^k)
\]

\[
= (a_{ir} - \sum_{k=1}^{i-1} \hat{u}_{ik} \hat{u}_{kr}^k)(1+\theta) \sum_{k=1}^{i-1} \hat{u}_{kr}^k,
\]

where \((1+\xi)(1+\rho) = 1+\theta, \frac{|\theta|}{1+\theta} \leq 2\nu' + 5.1\nu'^2\) and \((1+\theta)e_{ir} = \theta(1+\nu')a_{ii} \hat{u}_{ir} - (1+\nu')a_{ii} \hat{u}_{ir}\).

Then \(|e'_{ir}| \leq (2\nu' + 5.1\nu'^2)(1+\nu')a_{ii} \hat{u}_{ir} | + \nu \sum_{k=1}^{i-1} |\hat{u}_{kr}^k| \) and \(\sum_{r=i+1}^{J} |e'_{ir}| \leq (2\nu' + 5.1\nu'^2)(1+\nu')a_{ii} \hat{u}_{ir} | + \nu \sum_{r=i+1}^{J} |u_{ir}| \).
\[ + \nu \sum_{k=1}^{i-1} \hat{t}_{ik} \sum_{r=i+1}^{J} \hat{u}_{kr}. \] 

Therefore,

\[ \sum_{r=1}^{i} |e_{ir}| + \sum_{r=i+1}^{J} |e'_{ir}| \leq \nu \alpha_{ii} + \nu \sum_{k=1}^{i-1} \hat{t}_{ik} \sum_{r=k}^{J} \hat{u}_{kr} + (2\nu' + 5.1\nu^2) a_{ii} \]

\[ + \sum_{k=1}^{i-1} \hat{t}_{ik} \sum_{r=i+1}^{J} \hat{u}_{kr} \]

\[ = (\nu + 2\nu' + 5.1\nu'^2) a_{ii} + \nu \sum_{k=1}^{i-1} \hat{t}_{ik} \sum_{r=k}^{J} \hat{u}_{kr} \]

\[ < (\nu + 2\nu' + 5.1\nu'^2) a_{ii} + \nu(1+\nu')(1+\hat{\delta}) \frac{1}{\hat{\delta} - 5\nu} (1 + \frac{1 + 3\nu}{1 + \hat{\delta}}) \]

\[ < (\nu + 2\nu' + 5.1\nu'^2) a_{ii} + \nu + \nu \nu' + \frac{2\nu + 9.9\nu^2}{\hat{\delta} - 5\nu}. \]

The preceeding remarks are collected into

**Lemma 4:** Given A satisfying 1.1', there exists a matrix E satisfying \[ \|E\|_{\infty} < (\nu + 2\nu' + 5.1\nu'^2) a_{ii} + \nu + \nu \nu' + \frac{2\nu + 9.9\nu^2}{\hat{\delta} - 5\nu}, \]

such that \[ \hat{L} \hat{U} = A + E, \] where \( \hat{L} \) and \( \hat{U} \) are computed from 1.5.

We shall now be concerned with the errors involved in the computation of \( \hat{\gamma} \) and \( \hat{\omega} \) such that \( \hat{L}\hat{\gamma} = d, \hat{U}\hat{\omega} = \gamma \). Using a symbol with a caret to denote the computed value of the exact element, we have

\[ (i) \quad \hat{\gamma}_r = f_1 \left( \frac{d_r - f_1 \left( \sum_{k=1}^{r-1} \hat{t}_{rk} \hat{\gamma}_k \right)}{\hat{\hat{r}}_{rr}} \right) \]
\[ r-1 \sum_{k=1}^{r-1} d_k \gamma_k (l+\rho_k) = (1+\eta)(1+\mu) \sum_{k=1}^{r-1} \gamma_k (l+\rho_k), \; r=1,\ldots,J \]

\[ (1+\eta) (1+\mu) \sum_{k=1}^{r-1} d_k \gamma_k (l+\rho_k) = (1+\eta)(1+\mu) \sum_{k=1}^{r-1} \gamma_k (l+\rho_k), \; r=1,\ldots,J \]

\[ (1+\varepsilon) (\gamma_r - \sum_{k=r+1}^{J} \gamma_k (1+\varepsilon_k)), \; r=J,J-1,\ldots,1, \]

where \(|\eta|, |\mu|, |\varepsilon| \leq \nu', |\rho_k| \leq (r-k+1)\nu_2, |\varepsilon_k| \leq (J-k+2)\nu_2.\]

Let \(1+\theta = \frac{1}{(1+\eta)(1+\mu)}, \; |\theta(1+\nu')| < 2\nu\) then

\[ (1+\theta) \sum_{rr} d_{rr} \gamma_r + \sum_{k=1}^{r-1} d_k \gamma_k (l+\rho_k) = d_r. \]

Rewriting 1.13 in matrix form, we have \((\hat{L}+\delta \hat{L})\gamma = \hat{d},\) where

\((\delta \hat{L})_{rr} = \delta \hat{L}_{rr}, \; (\delta \hat{L})_{rk} = \rho_k \hat{L}_{rk}.\) Therefore,

\[ \|\delta \hat{L}\|_{\infty} < |\theta| \max(1+\nu')a_{ii} + \nu \sum_{k=1}^{r-1} |\hat{d}_{rk}| \]

\[ < 2\nu \max_{i} a_{ii} + \frac{\nu(1+\delta)(1+\nu')}{\delta-5\nu}. \]

Let \(1+\theta' = \frac{1}{1+\varepsilon}, \; |\theta'| < \varepsilon,\) then

\[ (1+\theta') \sum_{rr} d_{rr} \gamma_r + \sum_{k=r+1}^{J} d_k \gamma_k (1+\varepsilon_k) = \gamma_r. \]

Then 1.14 in matrix form is \((\hat{U}+\delta \hat{U})\omega = \hat{\omega},\) where \((\delta \hat{U})_{rr} = \theta', \; (\delta \hat{U})_{rk} = \varepsilon_k \hat{U}_{rk}.\) Therefore, \(\|\delta \hat{U}\|_{\infty} \leq \nu + \nu \sum_{j=r+1}^{J} |\hat{d}_{rk}| < \nu + \frac{\nu(1+3\nu)}{1+\delta}.\)
We have thus shown

**Lemma 5:** For \( \hat{L}, \hat{U} \) computed by 1.5 and \( \hat{\gamma}, \hat{\omega} \) computed by 1.12, \( \hat{\gamma} \) and \( \hat{\omega} \) solve exactly the matrix equations \((\hat{L}+\delta\hat{L})\hat{\gamma} = \delta\hat{\omega} = \gamma\), where \( \|\delta\hat{L}\|_\infty < 2\nu \max a_{1i} + \frac{\nu(1+\delta)(1+\nu')}{\delta-5\nu} \),

\[ \|\delta\hat{U}\|_\infty < \nu + \frac{\nu(1+3\nu)}{1+\delta} \]

Before stating the following theorem, let us restate 1.1' with \( \delta = \gamma-5\nu \),

(i) \( \delta + 5\nu + \left| \sum_{j\neq i} a_{ij} \right| < a_{ii}, \ i=1,2,\ldots,J \)

1.15 (ii) \(-1 \leq \sum_{j\neq i} a_{ij} < 0, \ i=1,2,\ldots,J \)

(iii) \( \delta > 0, \ a_{ij} < 0, \ i\neq j \).

Since \( \delta < \gamma \), Lemma 1 holds for \( \delta \) and \( \|A^{-1}\|_\infty < \frac{1}{\delta} \).

**Theorem 1:** Let \( \nu = N^{1-t_1} \), where \( N \) is the base of the number system, \( t_1 = t - \log_1.053 \). For the matrix equation \( A\omega = d \), where \( A \) satisfies 1.15, if the solution is computed by 1.5 and 1.12, using floating point arithmetic with a \( t \)-digit mantissa and having all inner products computed with a \( f_{l_2} \) operator, then the computed solution, \( \hat{\omega} \), satisfies

\[
\frac{||\hat{\omega}-\omega||_\infty}{||\hat{\omega}||_\infty} < (8.95 \max a_{1i} + 3.03)\frac{\nu}{\delta} + 6.17\frac{\nu}{\delta^2} .
\]

Proof: From Lemma 5, we have \((\hat{L}+\delta\hat{L})(\hat{U}+\delta\hat{U})\hat{\omega} = \delta\hat{\omega} = \gamma\). Therefore, \( \hat{\omega} \) satisfies exactly the equation \((A+K)\hat{\omega} = d\), where \( K = E + \delta\hat{L} \hat{U} + \delta\hat{L} \delta\hat{U} + \delta\hat{L} \delta\hat{U} \). Since \( A\omega = d \), we have \((A+K)\hat{\omega} = A\omega\), and we see that \( \frac{\hat{\omega}-\omega}{\hat{\omega}} = -A^{-1}K \). Therefore, to find a bound for
\[ \frac{\|\hat{\omega} - \omega\|}{\|\hat{\omega}\|}, \text{ we must find a bound for} \]

\[ \|A^{-1}K\|_\infty < \|A^{-1}\|_\infty (\|E\|_\infty + \|\delta \hat{L}\|_\infty \|\hat{U}\|_\infty + \|\hat{L}\|_\infty \|\delta \hat{U}\|_\infty + \|\delta \hat{L}\|_\infty \|\delta \hat{U}\|_\infty). \]

From Lemmas 3 and 5, we can determine the following bounds

\[ (i) \quad \|\delta \hat{L}\|_\infty \|\hat{U}\|_\infty < (2\nu \max_i a_{ii} + \frac{\nu (1+\delta)(1+\nu')}{\delta-5\nu}) (1 + \frac{1+3\nu}{1+\delta}) \]

\[ < 4\nu \max_i a_{ii} + \nu + \nu'\nu + \frac{2\nu+9.9\nu^2}{\delta} \]

\[ 1.16 \quad (ii) \quad \|\hat{L}\|_\infty \|\delta \hat{U}\|_\infty < ((1+\nu')\max_i a_{ii} + \frac{(1+\delta)(1+\nu')}{\delta-5\nu}) \nu + \frac{\nu(1+3\nu)}{1+\delta} \]

\[ < (2\nu+\nu\nu')\max_i a_{ii} + \nu + \nu'\nu + \frac{2\nu+9.9\nu^2}{\delta} \]

\[ (iii) \quad \|\delta \hat{L}\|_\infty \|\delta \hat{U}\|_\infty < (2\nu \max_i a_{ii} + \frac{\nu (1+\delta)(1+\nu')}{\delta-5\nu}) (\nu + \frac{\nu(1+3\nu)}{1+\delta}) \]

\[ < 4\nu^2 \max_i a_{ii} + \nu^2 + \nu'\nu^2 + \frac{2.1\nu^2}{\delta}. \]

From 1.16 and Lemmas 1 and 4, we have

\[ \|A^{-1}K\|_\infty < (8.95 \max_i a_{ii} + 3.03)\frac{\nu}{\delta} + 6.17\frac{\nu}{\delta^2} \]

proving the theorem.

Therefore, we have shown that for A satisfying 1.15 and the solution of \( Aw = d \) computed by 1.5 and 1.12, the bound on the rounding error for the computed solution \( \hat{\omega} \) depends only on \( \delta \) and \( \nu \), and is independent of \( J \).
References

