Improved error bounds for linear systems with $H$-matrices

Atsushi Minamihata$^1$, Kouta Sekine$^2$, Takeshi Ogita$^3,5$, Siegfried M. Rump$^{2,4}$, and Shin’ichi Oishi$^{2,5}$

$^1$Graduate School of Fundamental Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

$^2$Faculty of Science and Engineering, Waseda University

$^3$School of Arts and Sciences, Tokyo Woman’s Christian University
2-6-1 Zempukuji, Suginami-ku, Tokyo 167-8585, Japan

$^4$Institute for Reliable Computing, Hamburg University of Technology
Schwarzenbergstraße 95, Hamburg 21071, Germany

$^5$CREST, JST

Received October 31, 2014; Published July 1, 2015

Abstract: Improved componentwise error bounds for approximate solutions of linear systems are derived in the case where the coefficient of a given linear system is an $H$-matrix. One of the error bounds presented in this paper proves to be tighter than the existing error bound, which is effective especially for ill-conditioned cases. Numerical experiments are performed to illustrate the effect of the improvements.

Key Words: linear system, error bound, $H$-matrix, verified numerical computations

1. Introduction

Let $\mathbb{R}$ be the set of real numbers. For a linear system

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n,$$

we can efficiently obtain an approximate solution $\tilde{x}$ of (1) by a standard numerical algorithm such as the Gaussian elimination with partial pivoting. However, when floating-point arithmetic is used for numerical computations, the computed solution involves rounding errors. In general, without using verified numerical computations, we do not know how accurate the computed solution is.

In order to compute the accuracy of $\tilde{x}$, it is essential to evaluate an upper bound of $|A^{-1}|$, since

$$|A^{-1}b - \tilde{x}| \leq |A^{-1}||b - A\tilde{x}|,$$

where the residual $b - A\tilde{x}$ is easy to compute. For this purpose, several methods of calculating upper bounds of $|A^{-1}|$ have been proposed (cf. e.g., [4, 7]) using the property of an $H$-matrix. It is important
to consider both computational cost for calculating the bounds and their quality. Therefore, the bound in [7, Theorem 2.1] is very useful in that sense. In this paper, we try to refine it and provide improved error bounds. One of the methods derived in this paper has already been implemented in the routine verifylss in INTLAB Version 7 [6]. In a similar way to the method in [7], computational costs of the proposed methods are \(O(n^2)\) if a given matrix \(A \in \mathbb{R}^{n \times n}\) is an \(H\)-matrix.

The paper is organized as follows. In Section 2, we explain the notation and state definitions used in this paper. In Section 3, we propose several verification methods based on the Rump’s theorem [7, Theorem 2.1]. Finally, in Section 4, we present some numerical results showing the performance of the proposed methods.

2. Notation and definitions

Let \(I\) denote the \(n \times n\) identity matrix, \(O\) the \(n \times n\) matrix of all zeros and \(0\) the \(n\)-vector of all zeros. Inequalities for matrices are understood componentwise, e.g. for real \(n \times n\) matrices \(A = (a_{ij})\) and \(B = (b_{ij})\) the notation \(A \leq B\) means \(a_{ij} \leq b_{ij}\) for all \((i,j)\). In particular, the notation \(A \geq O\) (or \(A > O\)) means that all the elements of \(A\) are nonnegative (or positive). Moreover, the notation \(|A|\) means \(|A| = (|a_{ij}|) \in \mathbb{R}^{n \times n}\), the nonnegative matrix consisting of componentwise absolute values of \(A\). Similar notation is applied to real vectors.

Let \(A \in \mathbb{R}^{n \times n}\). The spectral radius of \(A\), which is the largest magnitude of the eigenvalues of \(A\), is denoted by \(\rho(A)\). The comparison matrix \((\hat{A}_{ij})\) of \(A\) is defined as

\[
\hat{a}_{ij} = \begin{cases} |a_{ij}| & \text{if } i = j \\ -|a_{ij}| & \text{if } i \neq j \end{cases}.
\]

**Definition 2.1** (monotone). A real \(n \times n\) matrix \(A\) is called monotone if \(Ax \geq 0\) implies \(x \geq 0\) for \(x \in \mathbb{R}^n\).

**Definition 2.2** (Z-matrix). Let \(A = (a_{ij})\) be a real \(n \times n\) matrix with \(a_{ij} \leq 0\) for \(i \neq j\). Then \(A\) is called a Z-matrix.

**Definition 2.3** (M-matrix). If a Z-matrix \(A\) is monotone, then \(A\) is called an M-matrix.

**Definition 2.4** (H-matrix). If \((A)\) is an M-matrix, then \(A\) is called an H-matrix.

**Lemma 2.1** (e.g. [3, p.113]). If \(A\) is an H-matrix, then \(|A^{-1}| \leq (A)^{-1}\).

3. Modified Rump’s theorem

In this section, we derive componentwise error bounds for an approximate solution of a linear system \(Ax = b\) where \(A\) is an \(H\)-matrix. For an arbitrary linear this can usually be achieved by using the preconditioned linear system \(RAx = Rb\). The bounds are improved versions of those in the following theorem:

**Theorem 3.1** (Neumaier [4], Rump [7]). Let \(A \in \mathbb{R}^{n \times n}\) be given. Assume \(v \in \mathbb{R}^n\) with \(v > 0\) satisfies \(u := (A)v > 0\). Let \(D \in \mathbb{R}^{n \times n}\) denote the diagonal part of \((A)\), and define \(w \in \mathbb{R}^n\) by

\[
w_k := \max_{1 \leq i \leq n} \frac{G_{ik}}{u_i}, \quad \text{for } 1 \leq k \leq n,
\]

where \(G := I - (A)D^{-1} \geq O\). Then \(A\) is nonsingular and

\[
|A^{-1}| \leq D^{-1} + vw^T. \tag{3}
\]

From the definitions of \(u\) and \(w\) in Theorem 3.1, it follows

\[
I - (A)D^{-1} \leq uw^T, \tag{4}
\]

which is used for deriving (3). We will improve this inequality (4) focussing on the diagonal elements, because the diagonal elements of \(I - (A)D^{-1}\) are zeros, while those of \(uw^T\) are nonnegative.

The following theorem is the first modified version of Theorem 3.1.
Theorem 3.2. Let \( A \in \mathbb{R}^{n \times n} \) be given. Let further \( D, u, v, w \) be defined as in Theorem 3.1. Define \( D_s := \text{diag}(s_1, \ldots, s_n) \in \mathbb{R}^{n \times n} \) with
\[
s_k := u_k w_k \geq 0 \quad \text{for } 1 \leq k \leq n.
\] (5)

Then \( A \) is nonsingular and
\[
|A^{-1}| \leq (D^{-1} + vw^T)(I + D_s)^{-1}.
\] (6)

Proof. From the definition of \( D_s \), it holds that
\[
D_s \leq uw^T - (I - \langle A \rangle D^{-1})
\]
and
\[
I + D_s \leq \langle A \rangle D^{-1} + uw^T.
\] (7)

From the assumption, \( A \) is an \( H \)-matrix, so that \( \langle A \rangle^{-1} \geq O \). Multiplying (7) from the left by \( \langle A \rangle^{-1} \) yields
\[
\langle A \rangle^{-1}(I + D_s) \leq D^{-1} + \langle A \rangle^{-1}uw^T = D^{-1} + vw^T
\]
and
\[
\langle A \rangle^{-1} \leq (D^{-1} + vw^T)(I + D_s)^{-1}
\]
since \( I + D_s \) is nonsingular and \( (I + D_s)^{-1} \geq O \). Combining this and Lemma 2.1 yields (6). \( \square \)

Remark 1. The bound (6) is always sharper than (3), because \( (I + D_s)^{-1} \leq I \).

Next, we consider a further modification of estimating \( \langle A \rangle^{-1} \). Let \( \Delta \) be defined as
\[
\Delta := uw^T - (I - \langle A \rangle D^{-1})
\]
Then it holds
\[
I + \Delta = \langle A \rangle D^{-1} + uw^T.
\]

From the matrix determinant for a rank-one update [2, p.475], we have
\[
det(I + \Delta) = det((A)D^{-1} + uw^T) = (1 + w^TD\langle A \rangle^{-1}u) det((A)D^{-1})
\]
\[
= (1 + w^TDv) det((A)) det(D^{-1}).
\]

Since \( 1 + w^TDv > 0 \), \( det(I + \Delta) \neq 0 \) and therefore \( I + \Delta \) is nonsingular. Then
\[
I = ((A)D^{-1} + uw^T)(I + \Delta)^{-1}.
\] (8)

Multiplying (8) from the left by \( \langle A \rangle^{-1} \) yields
\[
\langle A \rangle^{-1} = (D^{-1} + vw^T)(I + \Delta)^{-1},
\] (9)
since \( \langle A \rangle^{-1}u = v \). By using the Neumann series expansion, we have
\[
(I + \Delta)^{-1} = I - \Delta + \Delta^2 - \Delta^3 + \cdots.
\] (10)

Note that (9) is an equality. Therefore, if \( \rho(\Delta) < 1 \), it can be expected that an approximation of \( (I + \Delta)^{-1} \) by few terms of the right-hand side in (10) provides a sufficiently accurate approximation of \( \langle A \rangle^{-1} \).

Based on the above-mentioned discussions, we present the second modified version of Theorem 3.1.

Theorem 3.3. Let \( A \in \mathbb{R}^{n \times n} \) be given. Let further \( D, u, v, w \) be defined as in Theorem 3.1. Define \( \Delta := uw^T - (I - \langle A \rangle D^{-1}) \). Then it holds for any \( m \in \mathbb{N} \) that
\[
|A^{-1}| \leq (D^{-1} + vw^T)(I - \sum_{k=1}^{m} (\Delta^{2k-1} - \Delta^{2k})).
\] (11)

379
Proof. Multiplying (9) from the right by $I + \Delta^{2m+1}$ yields

$$\langle A \rangle^{-1}(I + \Delta^{2m+1}) = (D^{-1} + vw^T)(I - \sum_{k=1}^{m}(\Delta^{2k-1} - \Delta^{2k})).$$

(12)

Since $\Delta \geq O$, we have

$$\langle A \rangle^{-1} \leq \langle A \rangle^{-1}(I + \Delta^{2m+1}).$$

(13)

Inserting (12) into (13) yields (11).

Remark 2. In practice, it is not necessary to compute $\Delta$ explicitly. For any $z \in \mathbb{R}^n$, the computation of $\Delta z$ only requires a dot product and a matrix-vector product.

Remark 3. It is not certain whether Theorem 3.3 gives a more tight bound of $|A^{-1}|$ than Theorem 3.1 for arbitrary $\Delta$. However, in the case of $A \approx I$, it is expected that $\rho(\Delta) < 1$ and Theorem 3.3 gives more tight bounds than Theorems 3.1 and 3.2.

Next, we present the following lemma to improve further the error bounds for linear systems.

Lemma 3.1. Let $A \in \mathbb{R}^{n \times n}$ and $b, \tilde{x} \in \mathbb{R}^n$ be given. Let further $u, v$ be defined as in Theorem 3.1. Suppose $C \in \mathbb{R}^{n \times n}$ satisfies $\langle A \rangle^{-1} \leq C$. For any $y \in \mathbb{R}^n$ such that $|b - A\tilde{x}| \leq y$, assume $\beta \in \mathbb{R}$ satisfies $0 \leq \beta \leq \min_{1 \leq i \leq n} y_i/u_i$. Then

$$|A^{-1}b - \tilde{x}| \leq \beta v + C(y - \beta u).$$

(14)

Proof. Recalling $u = \langle A \rangle v$ and Lemma 2.1 implies

$$|A^{-1}b - \tilde{x}| \leq |A^{-1}| |b - A\tilde{x}| \leq \langle A \rangle^{-1} y$$

(15)

$$= \beta v + \langle A \rangle^{-1} y - \beta v$$

$$= \beta v + \langle A \rangle^{-1}(y - \beta(A)v) = \beta v + \langle A \rangle^{-1}(y - \beta u)$$

$$\leq \beta v + C(y - \beta u),$$

which proves the lemma.

Remark 4. The bound (14) is always sharper than the standard estimate $Cy$ using (15), since it holds that

$$\beta v + C(y - \beta u) = \beta \langle A \rangle^{-1} u + C(y - \beta u) \leq \beta Cu + C(y - \beta u) = Cy.$$

4. Computational results

In this section, we present some computational results. The numerical experiments are carried out using MATLAB R2012b and INTLAB Version 7 [6] on a PC with 2.8 GHz Intel Core i7 CPU and 16 GB of main memory. We compare the results using the error bounds listed in Table I for several matrices.

|   |   |   |   |
|---|---|---|---|
| I | $(D^{-1} + vw^T)y$ by Theorem 3.1 [7] and (2) |   |   |
| II | $(D^{-1} + vw^T)(I + D_2)^{-1}y$ by Theorem 3.2 and (2) |   |   |
| III | $\beta v + (D^{-1} + vw^T)(I + D_2)^{-1}(y - \beta u)$ by Theorem 3.2 and Lemma 3.1 |   |   |
| IV | $(D^{-1} + vw^T)(y - \Delta(y - \Delta y))$ by Theorem 3.3 with $m = 1$ and (2) |   |   |
| V | $\beta v + (D^{-1} + vw^T)((y - \beta u) - \Delta((y - \beta u) - \Delta (y - \beta u)))$ by Theorem 3.3 with $m = 1$ and Lemma 3.1 |   |   |

380
4.1 Results for random matrices
First, we discuss the numerical behavior of the error bounds in Table I for random matrices. We generate random matrices of specified condition number by Higham’s `randsvd` using the MATLAB function `gallery` as follows:

\[ B = \text{gallery('randsvd', n, cnd)} \]

Then \( B \) becomes an \( n \times n \) pseudo-random matrix with \( \text{cond}(B) \approx cnd \) where \( \text{cond}(B) = \| B \|_2 \| B^{-1} \|_2 \). We fix \( n = 100 \) and vary \( cnd \) up to \( 10^{15} \). Moreover, the right-hand side vector \( c \) is generated by \( c = \text{randn}(n, 1) \). An approximate solution \( \tilde{x} \) is computed by the MATLAB command \( \tilde{x} = B \backslash c \) with iterative refinement to attain maximum accuracy. The residual \( c - B\tilde{x} \) is calculated by using the algorithm Dot2 [5] that can compute an improved approximation of the residual in twice the working precision.

After generating \( B \), we compute an approximate inverse \( R \) of \( B \) by the MATLAB’s function `inv`. Then we set \( A := RB \) and \( b := Rc \), respectively. Here \( A \) is expected to be almost the identity matrix and therefore an \( H \)-matrix. To ensure whether \( A \) is an \( H \)-matrix, we adopt the method in [7], which computes an approximate Perron vector of \( D^{-1}E \) for the splitting \( \langle A \rangle = D - E \) for finding \( v > 0 \) such that \( \langle A \rangle v > 0 \). See [7] for detail.

In Fig. 1, we display the medians of the results (the entries of the result vectors) obtained by the error bounds II, III, IV and V, with that by the error bound I normed to 1. From Fig. 1, it can be seen that the error bounds III, IV and V give better results than the error bound I, especially for larger condition numbers. Among them, the error bounds V provides the best results in these cases.

For these examples, the matrix \( A \) becomes almost the identity matrix \( I \), which means \( \| G \| = 1 \). As a result, \( \rho(\Delta) < 1 \) holds, so that the estimation by the error bounds IV and V using (11) for \( \langle A \rangle^{-1} \) works very well. On the other hand, the magnitude of the diagonal entries \( s_k \) of \( D_s \) in (5) also becomes much less than 1. Therefore, the effect of \( (I + D_s)^{-1} \) in the error bound II becomes almost negligible.

![Fig. 1. Medians of the results by the error bounds II, III, IV and V, the result using the error bound I normed to 1 (randsvd, n = 100).](image)

4.2 Results for sparse \( H \)-matrices
Next, we deal with sparse linear systems whose coefficients are \( H \)-matrices. We take test problems from the University of Florida Sparse Matrix Collection [1] as the matrix \( A \). In Table II, we display the list of test matrices with their properties. The right-hand side vector \( b \) is generated as before by \( b = \text{randn}(n, 1) \). To compute approximate solutions of linear systems, we use the MATLAB’s function `bicg` as the preconditioned Bi-CG method as follows:
Table II. Test sparse matrices from the University of Florida Sparse Matrix Collection.

| Problem               | n      | nnz    | cond       | Class   |
|-----------------------|--------|--------|------------|---------|
| Bourchtein/atmosmodd  | 1,270,432 | 8,814,880 | 9.02 · 10^3 | SDD/H   |
| Bourchtein/atmosmodl  | 1,489,752 | 10,319,760 | 1.47 · 10^3 | SDD/H   |
| HB/sherman3          | 5,005 | 20,033 | 5.01 · 10^17 | SDD/M   |
| Sandia/ASIC,100ks    | 99,190 | 578,890 | 9.30 · 10^9  | SDD/H   |
| Wang/wang3           | 26,064 | 177,168 | 6.18 · 10^3  | M       |
| Wang/wang4           | 26,068 | 177,196 | 4.02 · 10^5  | M       |

nnz: the number of nonzero elements, cond: condition number
SDD: strictly diagonally dominant, M: M-matrix, H: H-matrix

Table III. Medians of the results, the result by the error bound I normed to 1.

| Problem               | II        | III       | IV         | V         |
|-----------------------|-----------|-----------|------------|-----------|
| Bourchtein/atmosmodd  | 7.98 · 10^{-1} | 7.98 · 10^{-1} | 3.01 · 10^{18} | 3.01 · 10^{18} |
| Bourchtein/atmosmodl  | 7.58 · 10^{-1} | 7.58 · 10^{-1} | 9.91 · 10^{20} | 9.91 · 10^{20} |
| HB/sherman3          | 6.12 · 10^{-1} | 6.12 · 10^{-1} | 2.45 · 10^{6}  | 2.45 · 10^{6}  |
| Sandia/ASIC,100ks    | 2.11 · 10^{-6} | 2.11 · 10^{-6} | 2.65 · 10^{22} | 2.65 · 10^{22} |
| Wang/wang3           | 5.96 · 10^{-1} | 5.96 · 10^{-1} | 4.41 · 10^{7}  | 4.41 · 10^{7}  |
| Wang/wang4           | 6.32 · 10^{-1} | 6.32 · 10^{-1} | 1.77 · 10^{8}  | 1.77 · 10^{8}  |

[L,U] = ilu(A);
x = bicg(A,b,1e-10,1000,L,U);

We choose v = e if the test matrix A is strictly diagonally dominant. Otherwise, we set v as an approximate Perron vector of \( D^{-1}E \) for the splitting \( \langle A \rangle = D - E \) in the same way as the previous example in Section 4.1.

In Table III, we display the computational results obtained by the error bounds II, III, IV and V, with that by the error bound I normed to 1. As seen from Table III, both the error bounds IV and V provide much worse results than the error bound I. This is because \( \rho(\Delta) > 1 \) in these cases. On the other hand, the error bounds II and III always give better results than the error bound I, as ensured in theory. There is almost no difference in the results between the error bounds II and III, because \( \beta \) is very small in these cases. In particular, they work very effectively for the problem Sandia/ASIC,100ks, since the diagonal part of \( I + D_s \) becomes much greater than 1 in that case.

References

[1] T.A. Davis and Y. Hu, “The University of Florida Sparse Matrix Collection,” ACM Trans. Math. Software, vol. 38, no. 1:1, pp. 1–25, 2011. http://www.cise.ufl.edu/research/sparse/matrices.
[2] C.D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, Philadelphia, 2000.
[3] A. Neumaier, Interval Methods for Systems of Equations, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 1990.
[4] A. Neumaier, “A simple derivation of the Hansen-Blick-Rohn-Ning-Kearfott enclosure for linear interval equations,” Reliable Computing, vol. 5, no. 2, pp. 131–136, 1999.
[5] T. Ogita, S.M. Rump, and S. Oishi, “Accurate sum and dot product,” SIAM Journal on Scientific Computing, vol. 26, no. 6, pp. 1955–1988, 2005.
[6] S.M. Rump, INTLAB – INTerval LABoratory. In Tibor Csendes, editor, Developments in Reliable Computing, pp.77–104. Kluwer Academic Publishers, Dordrecht, 1999. http://www.ti3.tuhh.de/rump/.
[7] S.M. Rump, “Accurate solution of dense linear systems,” Part II: Algorithms using directed rounding, J. Comp. Appl. Math., 242, pp. 185–212, 2013.