Journal Pre-proof

Modified error bounds for approximate solutions of dense linear systems

Atsushi Minamihata, Takeshi Ogita, Siegfried M. Rump,
Shin'ichi Oishi

PII: S0377-0427(19)30551-5
DOI: https://doi.org/10.1016/j.cam.2019.112546
Reference: CAM 112546

To appear in: *Journal of Computational and Applied Mathematics*

Received date : 22 April 2019
Revised date : 15 October 2019

Please cite this article as: A. Minamihata, T. Ogita, S.M. Rump et al., Modified error bounds for approximate solutions of dense linear systems, *Journal of Computational and Applied Mathematics* (2019), doi: https://doi.org/10.1016/j.cam.2019.112546.

Modified error bounds for approximate solutions of dense linear systems*

Atsushi Minamihata[a,*], Takeshi Ogita[b], Siegfried M. Rump[c,d], Shin'ichi Oishi[d]

[a] *Department of Information and System Engineering, Chuo University, Tokyo, Japan*
[b] *Division of Mathematical Sciences, Tokyo Woman's Christian University, Tokyo, Japan*
[c] *Institute for Reliable Computing, Hamburg University of Technology, Hamburg, Germany*
[d] *Faculty of Science and Engineering, Waseda University, Tokyo, Japan*

Abstract

We derive verified error bounds for approximate solutions of dense linear systems. There are verification methods using an approximate inverse of a coefficient matrix as a preconditioner, where the preconditioned coefficient matrix is likely to be an H-matrix (also known as a generalized diagonally dominant matrix). We focus on two inclusion methods of matrix multiplication for the preconditioning and propose verified error bounds adapted to the inclusion methods. These proposed error bounds are tighter than conventional ones, especially in critically ill-conditioned cases. Numerical results are presented showing the effectiveness of the proposed error bounds.

Keywords: Linear system, Verified solution, Error bound, H-matrix
2010 MSC: 65G20

1. Introduction

We consider a linear system

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n. \tag{1}$$

We aim to compute an error bound of an approximate solution $\tilde{x}$ of (1). There exist some verification methods (e.g. [1, 2, 3]) using an approximate inverse R of A as a preconditioner, i.e., the linear system (1) is transformed into $RAx = Rb$. If A is not extremely ill-conditioned, then $RA \approx I$. In other words, it is likely that RA is an H-matrix, because a subclass of an H-matrices is the class of strictly diagonally dominant matrices.

Several verification methods [3, 4, 5] have been proposed using the property of an H-matrix. Such verification methods consist of the following steps:

(i) Computing an approximate inverse R of A

(ii) Calculating an inclusion of RA

(iii) Verifying that A is nonsingular

(iv) Computing an error bound of an approximate solution $\tilde{x}$

*This research was partially supported by MEXT as "Exploratory Issue on Post-K computer" (Development of verified numerical computations and super high-performance computing environment for extreme researches) and JSPS KAKENHI Grant Numbers 17K12692.

*Corresponding author
Email address: minamihata.71e@g.chuo-u.ac.jp (Atsushi Minamihata)

Although the verification method in [4] often provides a tight error bound of $\tilde{x}$, it requires $O(n^3)$ operations in (iv). On the other hand, the methods in [3, 5] require only $O(n^2)$ operations in (iv). In [3, 5], the following inequality is used:

$$|A^{-1}b - \tilde{x}| \leq |(RA)^{-1}||R(b - A\tilde{x})| \tag{2}$$

Here, the absolute values and the inequality for vectors and matrices are understood componentwise. If $RA \approx I$, the methods in [3, 5] provide tight error bounds. However, if A is large and ill-conditioned, $RA \approx I$ does not hold, which causes overestimation in (2). To overcome this, we aim to improve the error bound and reduce the overestimation.

In floating-point arithmetic, the computation of RA suffers from rounding errors. Instead of calculating RA exactly, as in (ii), we can obtain an inclusion $\mathbf{C}$ of RA such that $RA \in \mathbf{C} = [\underline{C}, \overline{C}]$ with $\underline{C}, \overline{C} \in \mathbb{R}^{n\times n}$, which means $\underline{C}_{ij} \leq (RA)_{ij} \leq \overline{C}_{ij}$ for all (i, j). Let $\mathbb{F}$ be a set of floating-point numbers with some fixed precision with u being the unit roundoff. For $A, R \in \mathbb{F}^{n\times n}$, we focus on the following two inclusion methods:

(a) $\mathbf{C} = [\bigtriangledown(RA), \bigtriangleup(RA)]$ (See [2, 6].)
(b) $\mathbf{C} = [\bigtriangledown(RA), \bigtriangledown(RA) + 2nu|R||A|]$ (See [7, 8].)

Here, $\bigtriangledown(RA)$ and $\bigtriangleup(RA)$ denote the computed results of the matrix product RA in floating-point arithmetic with rounding downwards and upwards, respectively. Thus, $(\bigtriangledown(RA))_{ij} \leq (RA)_{ij} \leq (\bigtriangleup(RA))_{ij}$ hold for all (i, j). These inclusion methods allow to use BLAS routines for matrix multiplication with standard algorithms. Note that we do not need to compute $|R||A|$ explicitly when using Method (b), because we compute $|R|(|A|v)$ for some nonnegative vector v in (iii) and (iv). Although the computational cost for Method (b) is almost half of that for Method (a), Method (b) provides weaker bounds than Method (a) and is valid only for limited $n(\leq 1/(2u))$. Moreover, we can combine Methods (a) and (b); if Method (b) fails to verify that A is nonsingular in (iii), then we evaluate $\bigtriangleup(RA)$ and use Method (a). This strategy is proposed by the third author (S. M. Rump) and is used in INTLAB [9]. In this paper, on the basis of (2), we try to derive error bounds adapted to Method (b).

The remainder of the paper is organized as follows. In Section 2, we collect notation and definitions used in this paper. We also introduce some theorems on an H-matrix. In Section 3, we propose error bounds for approximate solutions of linear systems using the property of an H-matrix. In Section 4, we present numerical results showing the effectiveness of the proposed verification methods.

2. Notation and Preliminaries

Let I denote the $n \times n$ identity matrix, O the $n \times n$ matrix of all zeros, and $\mathbf{0}$ 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. The spectral radius of A, which is the largest magnitude of the eigenvalues of A, is denoted by $\rho(A)$.

Definition 2.1 (interval matrix). *An interval matrix* $\mathbf{A}$ *is defined as*

$$\mathbf{A} := [\underline{A}, \overline{A}] = \{A \in \mathbb{R}^{n\times n} : \underline{A} \leq A \leq \overline{A}\}$$

Definition 2.2 (midpoint and radius of an interval matrix). *The midpoint and the radius of* $\mathbf{A}$ *are defined as*

$$\mathrm{mid}(\mathbf{A}) := \frac{1}{2}(\underline{A} + \overline{A}), \quad \mathrm{rad}(\mathbf{A}) = \frac{1}{2}(\overline{A} - \underline{A}).$$

Definition 2.3 (mignitude of an interval vector). *The mignitude of interval vector* $\mathbf{x}$ *is defined as*

$$\mathrm{mig}(\mathbf{x}) = \min\{|x| \mid x \in \mathbf{x}\}.$$

Definition 2.4 (magnitude of an interval vector). *The magnitude of interval vector* $\mathbf{x}$ *is defined as*

$$\mathrm{mag}(\mathbf{x}) = \max\{|x| \mid x \in \mathbf{x}\}.$$

Definition 2.5 (hull of a set). *The hull of a set of* $\mathbf{X} \subseteq \mathbb{R}^n$ *is defined to be an the smallest interval vector enclosing* $\mathbf{X}$.

Definition 2.6 (comparison matrix). *The comparison matrix* $\langle A \rangle = (\widehat{a}_{ij})$ *of* A *is defined as*

$$\widehat{a}_{ij} = \begin{cases} |a_{ij}| & (i = j) \\ -|a_{ij}| & (i \neq j) \end{cases}.$$

Definition 2.7 (comparison matrix of an interval matrix). *The comparison matrix* $\langle \mathbf{A} \rangle = (\widehat{a}_{ij})$ *of* $\mathbf{A}$ *is defined as*

$$\widehat{a}_{ij} = \begin{cases} \mathrm{mig}(\mathbf{a_{ij}}) & (i = j) \\ -\mathrm{mag}(\mathbf{a_{ij}}) & (i \neq j) \end{cases}.$$

Definition 2.8 (monotone). *A real* $n \times n$ *matrix* A *is called monotone if* $Ax \geq \mathbf{0}$ *implies* $x \geq \mathbf{0}$ *for* $x \in \mathbb{R}^n$.

Definition 2.9 (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 2.10 (M-matrix). *If a Z-matrix* A *is monotone, then* A *is called an M-matrix.*

Definition 2.11 (H-matrix). *A real matrix* A *is called an H-matrix if* $\langle A \rangle$ *is an M-matrix.*

Definition 2.12 ((interval) H-matrix). *An interval matrix* $\mathbf{A}$ *is called an H-matrix if any* $A \in \mathbf{A}$ *is an H-matrix, i.e.,* $\langle \mathbf{A} \rangle$ *is an M-matrix.*

Definition 2.13 (nonsingular). *An interval matrix* $\mathbf{A}$ *is called nonsingular if any* $\hat{A} \in \mathbf{A}$ *is nonsingular.*

Definition 2.14 (strongly nonsingular). *An interval matrix* $\mathbf{A}$ *is called strongly nonsingular if* $\mathrm{mid}(\mathbf{A})^{-1}\mathbf{A}$ *is nonsingular.*

Theorem 2.1 (Fiedler-Pták [10]). *Let a Z-matrix* $A \in \mathbb{R}^{n \times n}$ *be given. Then the following conditions are equivalent:*

1. *A is nonsingular, and $A^{-1} \geq O$(, i.e. A is an M-matrix).*
2. *There exists $v \in \mathbb{R}^n$ with $v > \mathbf{0}$ satisfying $Av > \mathbf{0}$.*

Remark 1. *If there exists v with $v > \mathbf{0}$ satisfying $\langle A \rangle v > \mathbf{0}$, then A is an H-matrix.*

Theorem 2.2 (Neumaier [11]). *For an interval matrix* $\mathbf{A}$ *the following conditions are equivalent:*

(i) $\mathbf{A}$ *is an H-matrix.*

(ii) $\langle \mathbf{A} \rangle$ *is nonsingular, and* $\langle \mathbf{A} \rangle^{-1} e > \mathbf{0}$, *where* $e = (1, 1, \cdots, 1)^T$.

Theorem 2.3 (Neumaier [11]). *Let an interval matrix* $\mathbf{A}$ *be an H-matrix. Then, any* $\hat{A} \in \mathbf{A}$ *is nonsingular and satisfies*

$$|\hat{A}^{-1}| \leq \langle \mathbf{A} \rangle^{-1}.$$

Remark 2. *Let* $\mathbf{C}$ *be an interval matrix enclosing* RA *and* $\mathbf{c}$ *be an interval vector enclosing* $R(b - A\tilde{x})$. *If* $\mathbf{C}$ *is an H-matrix, then*

$$|A^{-1}b - \tilde{x}| \leq |(RA)^{-1}||R(b - A\tilde{x})| \leq \langle \mathbf{C} \rangle^{-1} \mathtt{mag}(\mathbf{c}).$$

3. Rigorous error bounds for linear systems using H-matrices

In this section, we introduce rigorous error bounds for linear systems using the property of an H-matrix. Let a linear system $Ax = b$ with $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$ be given. Let $R \in \mathbb{R}^{n\times n}$ be an approximate inverse of A. Let $\langle RA\rangle = D - E$ denote the splitting of $\langle RA\rangle$ into the diagonal part D and the off-diagonal part $-E$. If A is not ill-conditioned, then RA is close to the identity matrix. In such cases, RA is likely to be an H-matrix. If we find a positive vector v satisfying $\langle RA\rangle v > 0$, then A and R are proved to be nonsingular. There are several choices to find such vector v. For example, we choose v as a solution of $\langle RA\rangle z = e$, $e = (1, \cdots, 1)^T$ or the Perron vector of ED^{-1}. The function `verifylss` in INTLAB [6] adopts the algorithm based on the Perron iteration of ED^{-1} in [3].

3.1. Conventional error bounds

We introduce some previous works.

Theorem 3.1 (Rump [3]). *Let $A \in \mathbb{R}^{n\times n}$ and $b, \tilde{x} \in \mathbb{R}^n$ be given. Let R be an approximate inverse of A. Assume $v \in \mathbb{R}^n$ with $v > \mathbf{0}$ satisfies $u := \langle RA\rangle v > \mathbf{0}$. Let $\langle RA\rangle = D - E$ denote the splitting of $\langle RA\rangle$ into the diagonal part D and the off-diagonal part $-E$, and define $w \in \mathbb{R}^n$ by*

$$w_k := \max_{1\le i\le n} \frac{G_{ik}}{u_i} \quad \text{for } 1 \le k \le n,$$

where $G := I - \langle RA\rangle D^{-1} = ED^{-1} \ge O$. Then A is nonsingular,

$$|(RA)^{-1}| \le D^{-1} + vw^T,$$

and

$$|A^{-1}b - \tilde{x}| \le (D^{-1} + vw^T)|R(b - A\tilde{x})|.$$

Theorem 3.2 (Minamihata et al. [5]). *Under the same assumption as in Theorem 3.1, define $\Delta := uw^T - (I - \langle RA\rangle D^{-1})$. Then it holds for any $m \in \mathbb{N}$ that*

$$|(RA)^{-1}| \le (D^{-1} + vw^T)(I - \sum_{k=1}^{m}(\Delta^{2k-1} - \Delta^{2k})).$$

Moreover,

$$|A^{-1}b - \tilde{x}| \le (D^{-1} + vw^T)(I - \sum_{k=1}^{m}(\Delta^{2k-1} - \Delta^{2k}))|R(b - A\tilde{x})|. \tag{3}$$

Both Theorems 3.1 and 3.2 give upper bounds of $|(RA)^{-1}|$. When n is small, Theorems 3.1 and 3.2 work very well even if A is moderately ill-conditioned as long as RA is proved to be an H-matrix (see [3] and [5]). However, in the case of A being large and ill-conditioned, both Theorems 3.1 and 3.2 cause overestimation; It is due to the rank-1 matrix vw^T for obtaining an upper bound of $|(RA)^{-1}|$.

3.2. Proposed error bound for Method (a)

In order to reduce the overestimation, we propose the following error bound for a linear system using the property of an H-matrix.

Theorem 3.3. *Under the same assumption as in Theorem 3.1, we have*

$$|A^{-1}b - \tilde{x}| \le D^{-1}|R(b - A\tilde{x})| + \max_{1\le i\le n} \frac{(ED^{-1}|R(b - A\tilde{x})|)_i}{u_i} v. \tag{4}$$

Moreover, it holds for any $m \in \mathbb{N}$ that

$$|A^{-1}b - \tilde{x}| \leq \min_{1 \leq k \leq m} \epsilon^{(k)} , \tag{5}$$

where

$$\epsilon^{(k)} := D^{-1}(c^{(0)} + c^{(1)} + \cdots + c^{(k-1)}) + \max_{1 \leq i \leq n} \frac{c_i^{(k)}}{u_i} v$$

with

$$c^{(0)} := |R(b - A\tilde{x})|, \quad c^{(k)} := ED^{-1}c^{(k-1)}$$

for $1 \leq k \leq m$.

Proof: From Theorem 2.1,

$$\langle RA \rangle^{-1} \geq O. \tag{6}$$

For $1 \leq k \leq m$,

$$c^{(k)} \leq \max_{1 \leq i \leq n} \frac{c_i^{(k)}}{u_i} u. \tag{7}$$

From (6) and (7), we obtain

$$\langle RA \rangle^{-1} c^{(k)} \leq \max_{1 \leq i \leq n} \frac{c_i^{(k)}}{u_i} \langle RA \rangle^{-1} u = \max_{1 \leq i \leq n} \frac{c_i^{(k)}}{u_i} v, \quad 1 \leq k \leq m. \tag{8}$$

Since $\langle RA \rangle^{-1} = D^{-1} + \langle RA \rangle^{-1} ED^{-1}$, it holds for any $m \in \mathbb{N}$ that

$$\begin{aligned} \langle RA \rangle^{-1} |R(b - A\tilde{x})| &= D^{-1}c^{(0)} + \langle RA \rangle^{-1} c^{(1)} \\ &= D^{-1}(c^{(0)} + c^{(1)}) + \langle RA \rangle^{-1} c^{(2)} \\ &= D^{-1}(c^{(0)} + c^{(1)} + \cdots + c^{(m-1)}) + \langle RA \rangle^{-1} c^{(m)}. \end{aligned}$$

It can be seen from the inequality (8) that $\epsilon^{(0)}, \epsilon^{(1)}, \cdots, \epsilon^{(m)}$ are upper bounds of $|A^{-1}b - \tilde{x}|$. Thus, (5) is proved. □

Corollary 3.1. *Theorem 3.3 with $m = 1$ always gives a tighter error bound than Theorem 3.1.*

Proof: From the definition of Δ in Theorem 3.2,

$$\langle RA \rangle^{-1} + \langle RA \rangle^{-1} \Delta = D^{-1} + vw^T.$$

Then,

$$\begin{aligned} \langle RA \rangle^{-1} c^{(0)} &= (D^{-1} + vw^T)c^{(0)} - \langle RA \rangle^{-1} \Delta c^{(0)} \\ &= (D^{-1} + vw^T)c^{(0)} - \langle RA \rangle^{-1} \left((w^T c^{(0)}) u - ED^{-1} c^{(0)} \right) \\ &\leq (D^{-1} + vw^T)c^{(0)} - \langle RA \rangle^{-1} \left(w^T c^{(0)} - \max_{1 \leq i \leq n} \frac{(ED^{-1}c^{(0)})_i}{u_i} \right) u \\ &= D^{-1}c^{(0)} + \max_{1 \leq i \leq n} \frac{(ED^{-1}c^{(0)})_i}{u_i} v. \end{aligned} \tag{9}$$

Here, $ED^{-1} \leq uw^T$ holds by the definition of w, and

$$ED^{-1}c^{(0)} \leq uw^T c^{(0)} = (w^T c^{(0)})u. \tag{10}$$

Hence,

$$\max_{1 \leq i \leq n} \frac{(ED^{-1}c^{(0)})_i}{u_i} \leq w^T c^{(0)}.$$

From (10),

$$D^{-1}c^{(0)} + \max_{1\le i\le n} \frac{(ED^{-1}c^{(0)})_i}{u_i} v \le (D^{-1} + vw^T)c^{(0)},$$

which together with (9) completes the proof. □

Remark 3. *If b is an $n\times p$ matrix with $p \gg 1$, we may use Theorem 3.1 from the viewpoint of computational cost, because $(D^{-1} + vw^T)C$ for any $C \in \mathbb{R}^{n\times n}$ can be computed in $O(n^2)$ operations.*

The above theorems (Theorems 3.1, 3.2, and 3.3) are easy to apply to verification methods for linear systems. Let $\mathbf{C}$ be an inclusion of RA. We replace $\langle RA\rangle$ by $\langle \mathbf{C}\rangle$ and apply the above theorems with Theorem 2.3.

3.3. Proposed error bound for Method (b)

If we adopt Method (b) introduced in Section 1, $\mathbf{C} = [\bigtriangledown(RA), \bigtriangledown(RA) + 2nu|R||A|]$, and

$$\langle \mathbf{C}\rangle = \langle \bigtriangledown(RA) + nu|R||A|\rangle - nu|R||A|.$$

Then, for obtaining $\langle \mathbf{C}\rangle$, it is necessary to compute $|R||A|$ explicitly, which requires $O(n^3)$ operations. To avoid computing $|R||A|$ explicitly, we construct an inclusion $\mathbf{C}'$ of $\mathbf{C}$ such that $\langle \mathbf{C}'\rangle v$ can be computed in $O(n^2)$ operations. For example,

$$[\bigtriangledown(RA), \bigtriangledown(RA) + 2nu|R||A|] \subseteq [\bigtriangledown(RA), |\bigtriangledown(RA)| + 2nu|R||A|] =: \mathbf{C}'.$$

If the diagonal part of $\bigtriangledown(RA)$ is positive, then the diagonal part of $\langle \mathbf{C}'\rangle$ equals to that of $\bigtriangledown(RA)$, and the off-diagonal part of $\langle \mathbf{C}'\rangle$ equals to that of $\langle \bigtriangledown(RA)\rangle - 2nu|R||A|$. For $\mathbf{C}'$, $\mathrm{mid}(\mathbf{C}') \approx I$ is not satisfied in some ill-conditioned cases.

Theorem 3.4 (Neumaier [11]). *Let an interval matrix $\mathbf{A}$ be strongly nonsingular. Let $\mathbf{A}^H\mathbf{b}$ be the hull of a solution set of $\hat{A}x = \hat{b}$ ($\hat{A} \in \mathbf{A}$, $\hat{b} \in \mathbf{b}$). Then, for any interval vector $\mathbf{b}$ and any $C_1, C_2 \in \mathbb{R}^{n\times n}$ such that $C_1\mathbf{A}C_2$ is an H-matrix, we have*

$$\mathrm{mag}(\mathbf{A}^H\mathbf{b}) \le |C_2|\langle C_1\mathbf{A}C_2\rangle^{-1}\mathrm{mag}(C_1\mathbf{b}), \tag{11}$$

i.e., the bounds (11) becomes minimal for the choice $C_1 = \mathrm{mid}(\mathbf{A}^{-1})$ and $C_2 = I$.

From Theorem 3.4, we should choose $C_1 = \mathrm{mid}(\mathbf{C}')^{-1}$ and $C_2 = I$ for $\mathbf{C}'$ to minimize the error bounds. However, the computational costs of $\langle \mathrm{mid}(\mathbf{C}')^{-1}\mathbf{C}'\rangle v$ for any $v > \mathbf{0}$ is $O(n^3)$ operations because $\mathrm{mid}(\mathbf{C}')^{-1}$ is not a nonnegative matrix. Hence, we construct an inclusion $\mathbf{B}$ of $\mathrm{mid}(\mathbf{C}')^{-1}\mathbf{C}'$ such that $\langle \mathbf{B}\rangle v$ can be computed in $O(n^2)$ operations.

Theorem 3.5. *Let $A, R \in \mathbb{R}^{n\times n}$ and $b, \tilde{x} \in \mathbb{R}^n$ be given. Let $\mathbf{C}'$ be an inclusion of RA, i.e., $(RA)_{ij} \in \mathbf{C}'_{ij}$ for all (i, j). Assume $v \in \mathbb{R}^n$ with $v > \mathbf{0}$ satisfies $u := \langle \mathbf{C}'\rangle v > \mathbf{0}$. Then A is nonsingular, and*

$$|A^{-1}b - \tilde{x}| \le \langle \mathbf{C}'\rangle^{-1}\langle \mathrm{mid}(\mathbf{C}')\rangle|\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x})|. \tag{12}$$

Proof: From the assumption of v and Theorems 2.1 and 2.2, $\mathbf{C}'$ is a nonsingular H-matrix. Therefore, A is nonsingular, and $\mathrm{mid}(\mathbf{C}')$ is an H-matrix. From $\langle \mathbf{C}'\rangle \le \langle \mathrm{mid}(\mathbf{C}')\rangle$ and $\langle \mathbf{C}'\rangle^{-1} \ge O$,

$$I = \langle \mathbf{C}'\rangle^{-1}\langle \mathbf{C}'\rangle \le \langle \mathbf{C}'\rangle^{-1}\langle \mathrm{mid}(\mathbf{C}')\rangle. \tag{13}$$

From $\langle \mathbf{C}'\rangle \le \langle \mathrm{mid}(\mathbf{C}')\rangle$ and $\langle \mathrm{mid}(\mathbf{C}')\rangle^{-1} \ge O$,

$$\langle \mathrm{mid}(\mathbf{C}')\rangle^{-1}\langle \mathbf{C}'\rangle \le \mathrm{mid}(\mathbf{C}')\rangle^{-1}\mathrm{mid}(\mathbf{C}') = I. \tag{14}$$

From (13) and (14), the inverse of $\langle \mathrm{mid}(\mathbf{C}')\rangle^{-1}\langle \mathbf{C}'\rangle$ is nonnegative, and $\langle \mathrm{mid}(\mathbf{C}')\rangle^{-1}\langle \mathbf{C}'\rangle$ is a Z-matrix. Therefore $\langle \mathrm{mid}(\mathbf{C}')\rangle^{-1}\langle \mathbf{C}'\rangle$ is an M-matrix from Theorem 2.1.

Next, we consider an inclusion of $\mathrm{mid}(\mathbf{C}')^{-1}RA$ as follows.

$$\begin{aligned}\mathrm{mid}(\mathbf{C}')^{-1}RA &\subseteq \mathrm{mid}(\mathbf{C}')^{-1}\mathbf{C}'\\ &\subseteq [I - |\mathrm{mid}(\mathbf{C}')^{-1}|\mathrm{rad}(\mathbf{C}'), I + |\mathrm{mid}(\mathbf{C}')^{-1}|\mathrm{rad}(\mathbf{C}')]\\ &\subseteq [I - \langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}\mathrm{rad}(\mathbf{C}'), I + \langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}\mathrm{rad}(\mathbf{C}')] =: \mathbf{B}.\end{aligned}$$

Since $\langle\mathbf{C}'\rangle$ is an M-matrix, $0 \notin \mathbf{C}'_{ii}$ and

$$\langle\mathbf{C}'\rangle = \langle\mathrm{mid}(\mathbf{C}')\rangle - \mathrm{rad}(\mathbf{C}').$$

Moreover, since $\langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}\langle\mathbf{C}'\rangle = I - \langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}\mathrm{rad}(\mathbf{C}')$ is an M-matrix, $0 \notin \mathbf{B}_{ii}$ and

$$\langle\mathbf{B}\rangle = I - \langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}\mathrm{rad}(\mathbf{C}') = \langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}\langle\mathbf{C}'\rangle.$$

Thus, $\mathbf{B}$ is an H-matrix, and

$$\begin{aligned}|A^{-1}b - \tilde{x}| &\le |(\mathrm{mid}(\mathbf{C}')^{-1}RA)^{-1}|\ |\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x})|\\ &\le \langle\mathbf{B}\rangle^{-1}|\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x})| \qquad (15)\\ &= \langle\mathbf{C}'\rangle^{-1}\langle\mathrm{mid}(\mathbf{C}')\rangle|\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x})|.\end{aligned}$$

□

Remark 4. *The estimation of (15) is better than the estimation of (2) with applying* $|(RA)^{-1}| \le \langle\mathbf{C}'\rangle^{-1}$. *It is proved as follows.*

$$\begin{aligned}\langle\mathbf{B}\rangle^{-1}|\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x})| &\le \langle\mathbf{B}\rangle^{-1}|\mathrm{mid}(\mathbf{C}')^{-1}||R(b - A\tilde{x})|\\ &\le \langle\mathbf{B}\rangle^{-1}\langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}|R(b - A\tilde{x})|\\ &= \langle\mathbf{C}'\rangle^{-1}|R(b - A\tilde{x})|.\end{aligned}$$

Corollary 3.2. *Under the same assumption as in Theorem 3.5, we have*

$$|A^{-1}b - \tilde{x}| \le D^{-1}r + \alpha v, \qquad (16)$$

where $r := \langle\mathrm{mid}(\mathbf{C}')\rangle|\tilde{y}| + |\mathrm{mid}(\mathbf{C}')\tilde{y} - R(b - A\tilde{x})|$ *and* $\alpha := \max_{1\le i\le n} \dfrac{\{ED^{-1}r\}_i}{u_i}$.

Proof: Since $\mathrm{mid}(\mathbf{C}')$ is an H-matrix,

$$|\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x}) - \tilde{y}| \le \langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}|\mathrm{mid}(\mathbf{C}')\tilde{y} - R(b - A\tilde{x})|. \qquad (17)$$

From (15) and (17), we have

$$\begin{aligned}|A^{-1}b - \tilde{x}| &\le \langle\mathbf{B}\rangle^{-1}|\mathrm{mid}(\mathbf{C}')^{-1}R(b - A\tilde{x})|\\ &\le \langle\mathbf{B}\rangle^{-1}|\tilde{y}| + \langle\mathbf{B}\rangle^{-1}\langle\mathrm{mid}(\mathbf{C}')\rangle^{-1}|\mathrm{mid}(\mathbf{C}')\tilde{y} - R(b - A\tilde{x})|\\ &= \langle\mathbf{C}'\rangle^{-1}\langle\mathrm{mid}(\mathbf{C}')\rangle|\tilde{y}| + \langle\mathbf{C}'\rangle^{-1}|\mathrm{mid}(\mathbf{C}')\tilde{y} - R(b - A\tilde{x})|\\ &= \langle\mathbf{C}'\rangle^{-1}(\langle\mathrm{mid}(\mathbf{C}')\rangle|\tilde{y}| + |\mathrm{mid}(\mathbf{C}')\tilde{y} - R(b - A\tilde{x})|)\\ &= \langle\mathbf{C}'\rangle^{-1}r. \qquad (18)\end{aligned}$$

From the definition of α,

$$ED^{-1}r \le \alpha u. \qquad (19)$$

The assumption of v implies that $\langle\mathbf{C}'\rangle$ is an M-matrix. Then, $\langle\mathbf{C}'\rangle^{-1} \ge O$ and

$$\langle\mathbf{C}'\rangle^{-1}ED^{-1}r \le \alpha\langle\mathbf{C}'\rangle^{-1}u = \alpha v. \qquad (20)$$

Therefore, we have

$$\langle\mathbf{C}'\rangle^{-1}r = D^{-1}r + \langle\mathbf{C}'\rangle^{-1}ED^{-1}r \le D^{-1}r + \alpha v.$$

□

Table 1: Error bounds with the inclusion $\mathbf{C} = [\triangledown(RA), \triangle(RA)]$

I	Error bound (3) in Theorem 3.1
II	Error bound (3) in Theorem 3.2 ($m = 1$)
III	Proposed error bound (4) in Theorem 3.3 ($m = 1$)
IV	Error estimate by a precise approximation of $\langle \mathbf{C} \rangle^{-1} \lvert R(b - A\tilde{x}) \rvert$

Remark 5. *Corollary 3.2 can also be applicable to the verification method using Method (a). However, we expect that Corollary 3.2 with Method (b) works more efficiently than that with Method (a), since* $\mathrm{mid}(\mathbf{C}')$ *is often far from the identity matrix, especially for large* n *and ill-conditioned cases.*

4. Numerical results

We present some numerical results showing the performance of the proposed error bounds. To this end, we compare the proposed error bound with the conventional error bounds. All computations are performed with MATLAB R2018a and INTLAB Ver. 11 on a PC with Intel Core i7-7820X 3.6GHz CPU and 64 GB of main memory.

We generate the following random matrices as test problems, where we can specify n as the matrix dimension and `cnd` as the condition number $\kappa_2(A) := \|A\|_2 \|A^{-1}\|_2$.

1. `randmat`: Rump's test matrices using the INTLAB function `randmat`
   - MATLAB command: $A =$ `randmat`$(n,$ `cnd`$)$;
2. `randsvd`: Higham's test matrices using the MATLAB function `gallery` with `randsvd`
   - MATLAB command: $A =$ `gallery`$('$`randsvd`$', n,$ `cnd`$,$ `mode`$, n, n, 1)$;

The right-hands side vector b is generated by `randn`$(n, 1)$ in all cases. We fix $n = 10000$ and vary `cnd` up to the limit of the applicability of the verification methods. We also vary `mode` between 1 and 5 for `randsvd` in order to show results for various kinds of singular value distribution.

We first investigate the performance of the proposed verification method using Method (a), i.e., an inclusion of RA as $RA \in \mathbf{C} = [\triangledown(RA), \triangle(RA)]$. We compare the error bounds listed in Table 1. All the error bounds in Table 1 require computing R, $\tilde{x}$, and v, where R is an approximate inverse of A obtained by the MATLAB function `inv`, $\tilde{x}$ is an approximate solution computed by $R * b$ with iterative refinement to attain the maximum accuracy using the algorithm `Dot2` [12], and v is an approximate Perron vector of ED^{-1} obtained by the INTLAB function `MVector` [13, 3]. Therefore, $\tilde{x}$ is maximally accurate in double precision (binary64), i.e., all the elements of $\tilde{x}$ have almost maximum accuracy. The inclusion of $R(b - A\tilde{x})$ is computed by the algorithm `Dot2Err` [12]. In order to measure how obtained error bounds are overestimated, we also compute a precise approximation of $\langle \mathbf{C} \rangle^{-1} \lvert R(b - A\tilde{x}) \rvert$ as the error estimate IV.

In Figs. 1–6, we display the medians of the componentwise ratios of the error bounds I, II, and III to the error estimate IV. As can be seen, the proposed error bound III gives better results than the error bounds I and II except in the case of `randsvd` with `mode` $= 1$ (Fig. 2) and is almost the same as the estimation IV in all the cases. Note that the error bounds I and II overestimate the actual error for ill-conditioned cases due to the estimation using the rank-1 matrix vw^T in Theorem 3.1 and Δ such that $\rho(\Delta) > 1$ in Theorem 3.2, respectively. From these results, we can see that the proposed error bound III is more effective than the conventional error bounds I and II for a certain range of the condition number, for example, $\kappa_2(A) \in [10^{14}, 4.2 \times 10^{14}]$ in the case of `randmat` (Fig. 1) and $\kappa_2(A) \in [10^{12}, 6 \times 10^{12}]$ in the case of `randsvd` with `mode` $= 3$ (Fig. 4).

Next, we investigate the performance of the proposed verification method using Method (b), i.e., an inclusion of RA as $RA \in \mathbf{C}' = [\triangledown(RA), \lvert \triangledown(RA) \rvert + 2nu\lvert R \rvert \lvert A \rvert]$. In the same way as the previous example, the test matrix A is generated using `randmat`, and the right-hands side vector b is generated by `randn`$(n, 1)$. We fix $n = 10000$ and vary `cnd` up to 7×10^{11}. The computations of $R, \tilde{x}, v, R(b - A\tilde{x})$ are done as well.

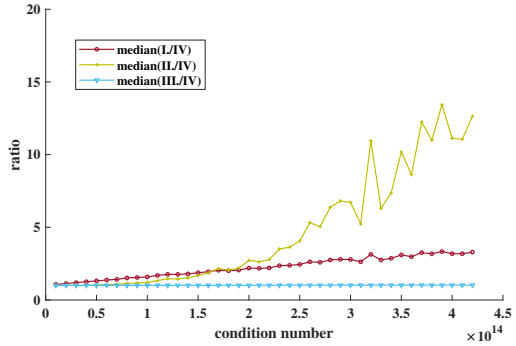

Figure 1: Ratios of the error bounds I, II, and III to the error estimate IV ($n = 10000$, `randmat`)

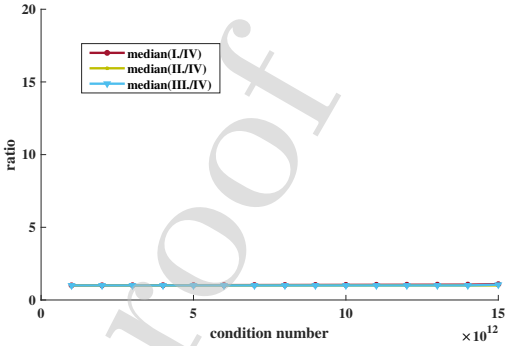

Figure 2: Ratios of the error bounds I, II, and III to the error estimate IV ($n = 10000$, `randsvd` with `mode` $= 1$)

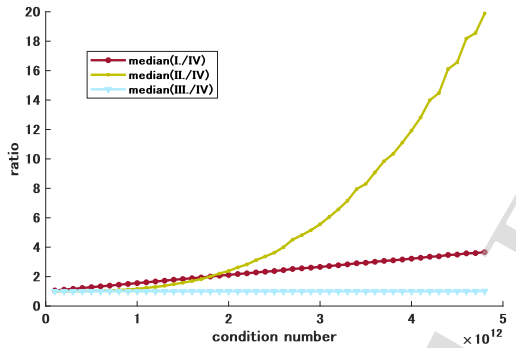

Figure 3: Ratios of the error bounds I, II, and III to the error estimate IV ($n = 10000$, `randsvd` with `mode` $= 2$)

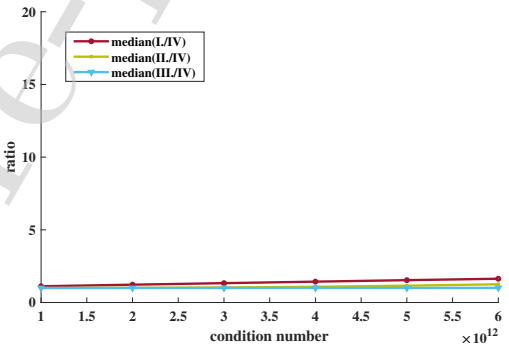

Figure 4: Ratios of the error bounds I, II, and III to the error estimate IV ($n = 10000$, `randsvd` with `mode` $= 3$)

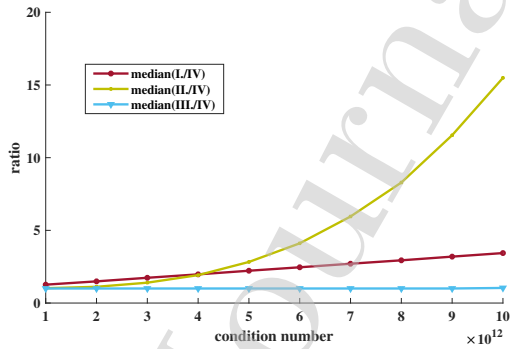

Figure 5: Ratios of the error bounds I, II, and III to the error estimate IV ($n = 10000$, `randsvd` with `mode` $= 4$)

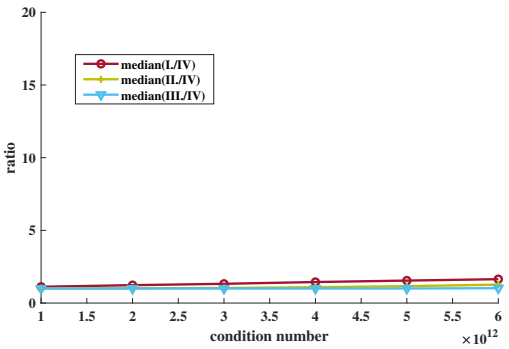

Figure 6: Ratios of the error bounds I, II, and III to the error estimate IV ($n = 10000$, `randsvd` with `mode` $= 5$)

Table 2: Error bounds with the inclusion $\mathbf{C}' = [\nabla(RA), |\triangledown(RA)| + 2nu|R||A|]$

Ⅳ	Error estimate by a precise approximation of $\langle \mathbf{C}' \rangle^{-1} |R(b - A\tilde{x})|$
Ⅴ	Proposed error bound (16) in Corollary 3.2
Ⅵ	Error estimate by a precise approximation of $\langle \mathbf{C}' \rangle^{-1} \langle \mathrm{mid}(\mathbf{C}') \rangle |\mathrm{mid}(\mathbf{C}')^{-1} R(b - A\tilde{x})|$

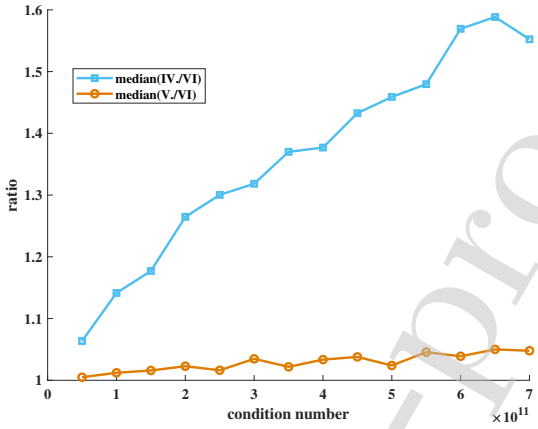

Figure 7: Ratios of the error estimate Ⅳ and the error bound Ⅴ to the error estimate Ⅵ ($n = 10000$, `randmat`)

Let $\mathbf{g}$ be an inclusion of $R(b - A\tilde{x})$ by the algorithm `Dot2Err`. We put $\tilde{y} := \mathtt{mid}(\mathbf{g})./\mathtt{diag}(\mathtt{mid}(\mathbf{C}'))$ for the proposed error bound Ⅴ. The error bounds listed in Table 2 are compared.

In Fig. 7, we display the medians of the componentwise ratios of the error estimate Ⅳ and the proposed error bound Ⅴ to the error estimate Ⅵ. It can be seen that the error bound Ⅴ gives better results than the error estimate Ⅳ and is almost the same as the error estimate Ⅵ. This also implies the error bound Ⅴ gives better results than the error bounds Ⅰ, Ⅱ, and Ⅲ.

Finally, we show computing times of the proposed verification method using Methods (a) and (b). In the same way as the previous example, the test matrix A is generated using `randmat`, and the right-hands side vector b is generated by $\mathtt{randn}(n, 1)$. We fix $n = 10000$ and $\mathrm{cnd} = 7 \times 10^{11}$. The CPU times are listed in Table 3. Note that Dot2 and Dot2Err are implemented in C. As can be seen, main computational costs in these verification methods are to compute R and inclusion of RA, and the verification method with Method (b) is faster than that with Method (a).

Appendix

Comparison between the Hansen-Bliek-Rohn-Ning-Kearfott-Neumaier (HBRNKN) method and the proposed methods

In this appendix, we compare the proposed methods with the HBRNKN method [4]. We use the same variables as in the numerical results in Section 4. Let $\mathbf{Y}$ be an inclusion of the solution of $\mathbf{C}y = \mathbf{g}$ using the HBRNKN method, i.e., $A^{-1}b \in \tilde{x} + \mathbf{Y}$. In fairness, we adopt $\tilde{x} + \mathtt{mid}(Y)$ as an approximate solution in the proposed methods.

We compared the following error bounds:

- ϵ_P: Proposed error bound (4) in Theorem 3.3 ($m = 1$)
- ϵ_N: Error bound using HBRNKN method

Table 3: Computing times (sec.) of the proposed verification methods using Methods (a) and (b)

	Verification method using (a)	Verification method using (b)
Total time	22.73	17.54
Computing R	8.01	8.01
Inclusion of RA	11.56	5.77
Iterative refinement with Dot2	1.24	1.24
Inclusion of $R(b - A\tilde{x})$ with Dot2Err	0.49	0.36
Computing v	0.08	0.82
Computing error bounds	0.02	0.48

Table 4: Results of comparison between the HBRNKN method and the proposed method for `randsvd` matrices ($n = 10000$)

`mode`	`cnd`	Ratio 1	Iter.	Ratio 2	$\rho(ED^{-1})$
1	1.5×10^{13}	1.019	2	0.0054	0.9948
2	5.0×10^{12}	1.007	3	0.0088	0.9913
3	6.0×10^{12}	1.006	2	0.0271	0.9730
4	1.0×10^{13}	1.015	3	0.0041	0.9960
5	6.0×10^{12}	1.033	2	0.0065	0.9936

On computational cost, the HBRNKN method needs to compute an approximate inverse X_C of $\langle \mathbf{C} \rangle$ and an interval matrix multiplication $X_C \langle \mathbf{C} \rangle$ explicitly, while our proposed method does not. Thus, it is clear that the proposed method is faster than the HBRNKN method.

In Table 4, we show the following results:

- Ratio 1: Median of ratios of error bounds $(\epsilon_P)_i/(\epsilon_N)_i$, $i = 1, 2, \ldots, n$
- Iter.: Number of iterations in `MVector` function for obtaining v
- Ratio 2: Median of ratios of error terms $(D^{-1}c^{(0)})_i/(\langle \mathbf{C} \rangle^{-1}c^{(1)})_i$, $i = 1, 2, \ldots, n$
- $\rho(ED^{-1})$: Spectral radius of ED^{-1}

It can be seen from the results of "Ratio 1" that ϵ_N is almost the same as ϵ_P, and the proposed method requires at most three iterations for obtaining v in these numerical results. As can be seen from the results of "Ratio 2", the main error term is not $D^{-1}c^{(0)}$ but $\langle \mathbf{C} \rangle^{-1}c^{(1)}$. Thus, even if we increase m in Theorem 3.3, it does not converge efficiently. This can also be explained from the results as $\rho(ED^{-1}) \approx 1$. Therefore, it is important to estimate $\langle \mathbf{C} \rangle^{-1}c^{(1)}$ as tight as possible, which is achieved by computing an appropriate v with negligible cost.

References

[1] E. Hansen, R. Smith, Interval arithmetic in matrix computations, Part II, SIAM Journal on Numerical Analysis 4 (1) (1967) 1–9.

[2] S. Oishi, S. Rump, Fast verification of solutions of matrix equations, Numer. Math. 90 (4) (2002) 755–773.

[3] S. Rump, Accurate solution of dense linear systems, Part II: Algorithms using directed rounding, J. Comp. Appl. Math. 242 (2013) 185–212.

[4] A. Neumaier, A simple derivation of the Hansen-Bliek-Rohn-Ning-Kearfott enclosure for linear interval equations, Reliable Computing 5 (2) (1999) 131–136.

[5] A. Minamihata, K. Sekine, T. Ogita, S. Rump, S. Oishi, Improved error bounds for linear systems with H-matrices, Nonlinear Theory and Its Applications 6 (3) (2015) 377–382.

[6] S. M. Rump, Fast and parallel interval arithmetic, BIT Numerical Mathematics 39.3 (1999) 534–554. doi:10.1007/s11075-011-9524-z.

[7] S. M. Rump, Fast interval matrix multiplication, Numerical Algorithms 61.1 (2012) 1–34. doi:10.1007/s11075-011-9524-z.

[8] M. Lange, S. Rump, Sharp estimates for perturbation errors in summations, Math. Comp. 88 (315) (2019) 349–368.

[9] S. Rump, INTLAB - INTerval LABoratory, in: T. Csendes (Ed.), Developments in Reliable Computing, Kluwer Academic Publishers, Dordrecht, 1999, pp. 77–104, http://www.ti3.tuhh.de/rump/.

[10] M. Fiedler, V. Pták, On matrices with non-positive off-diagonal elements and positive principal minors, Czech. Math. J. 12 (3) (1962) 382–400.

[11] A. Neumaier, Interval Methods for Systems of Equations, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 1990.

[12] T. Ogita, S. Rump, S. Oishi, Accurate sum and dot product, SIAM Journal on Scientific Computing 26 (6) (2005) 1955–1988.

[13] S. Rump, Fast and parallel interval arithmetic, BIT 39 (3) (1999) 534–554.