2007 International Symposium on Nonlinear Theory and its Applications
NOLTA'07, Vancouver, Canada, September 16-19, 2007

Tight Error Bounds for Approximate Solutions of Linear Systems

Takeshi Ogita[†,‡], Shin'ichi Oishi[‡]

† CREST, Japan Science and Technology Agency (JST)
‡ Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan
Email: {ogita,oishi}@waseda.jp

Abstract—This paper is concerned with the problem of verifying the accuracy of an approximate solution of a linear system. A fast method is developed for calculating both lower and upper error bounds of the approximate solution, which are as tight as needed, with verifying the nonsingularity of the coefficient matrix. Numerical results are presented elucidating the performance of the proposed verification method.

1. Introduction

We are concerned with the problem of verifying the accuracy of an approximate solution $\tilde{x}$ of a linear system

$$Ax = b, \tag{1}$$

where A is a real $n \times n$ matrix and b is a real n-vector. If A is nonsingular, there exists a unique solution $x^* := A^{-1}b$. We aim on verifying the nonsingularity of A and calculating some $\underline{\epsilon}, \overline{\epsilon} \in \mathbb{R}^n$ such that

$$\mathbf{o} \le \underline{\epsilon} \le |x^* - \tilde{x}| \le \overline{\epsilon}, \tag{2}$$

where $\mathbf{o} := (0, \ldots, 0)^T \in \mathbb{R}^n$. A geometric image of the inclusion for the exact solution x^* such as (2) can be depicted as in Figure 1.

A number of fast self-validating algorithms (cf., for example, [2, 4, 8]) have been proposed to verify the nonsingularity of A and to compute $\overline{\epsilon}$ in (2). In addition, this paper also considers to compute $\underline{\epsilon}$. If $\underline{\epsilon}_i \approx \overline{\epsilon}_i$, then we can *verify* that the error bounds (and the verification) are of high quality!

A main point of this paper is to develop a method of calculating both $\underline{\epsilon}$ and $\overline{\epsilon}$ satisfying (2), which are as tight as we need. If we obtain tight error bounds, we can set an appropriate criterion for improving an approximate solution $\tilde{x}$ by the iterative refinement method.

We assume that the floating-point system used in this paper follows IEEE standard 754 for floating-point arithmetic. Moreover, we suppose that all floating-point operations are executed according to the rounding mode defined in IEEE standard 754. Under such conditions, we will propose a fast algorithm of calculating a verified solution $\tilde{x}$ of (1) in terms of (2). Numerical results are presented elucidating properties and efficiencies of the proposed verification method.

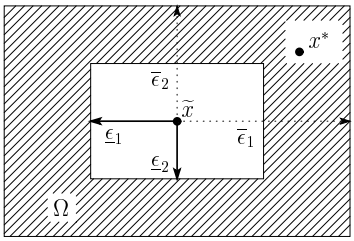

Figure 1: Inner and outer enclosure of the exact solution (two-dimensional case). The exact solution x^* exists in Ω.

2. Notation and definitions

We start by stating some well-known properties on floating-point numbers. Let $\mathbb{R}$ denote the set of real numbers. Let $\mathbb{F}$ be a set of floating-point numbers following IEEE standard 754. Let $\mathbf{u}$ be the unit roundoff. In IEEE 754 double precision arithmetic, $\mathbf{u} = 2^{-53}$. It is well-known that $\mathbb{F}$ is symmetric, *i.e.*, $f \in \mathbb{F} \Rightarrow -f \in \mathbb{F}$, so that $|f|$ is exact for $f \in \mathbb{F}$. Throughout this paper, we assume that the operations in $\mathrm{fl}(\cdots)$ is all executed by floating-point arithmetic in given rounding mode (default is round-to-nearest). Throughout the paper we assume that no overflow occurs. This usually leads to a premature stop of calculations, so we do not have to check for this. Furthermore, we assume that the floating-point system in this paper supports the gradual underflow, which is a requirement of IEEE 754 standard.

Let $\mathbb{IR}$ denote the set of interval real numbers and $\mathbb{IF}$ denote a set of interval floating-point numbers. Note that $\mathbb{IF} \subset \mathbb{IR}$. For a real matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, we denote by $|A| = (|a_{ij}|) \in \mathbb{R}^{n \times n}$ the nonnegative matrix consisting of entrywise absolute values. For real $n \times n$ matrices $A = (a_{ij}), B = (b_{ij})$, an inequality $A \le B$ is understood entrywise, *i.e.*, $a_{ij} \le b_{ij}$ for all (i, j). We express an interval matrix including A by $[A] := [\underline{A}, \overline{A}] \in \mathbb{IR}^{n \times n}$ where $\underline{A}$ and $\overline{A}$ is a lower and an upper bound of A, respectively. For real vectors, we apply these definitions similarly. The *magnitude* of an interval quantity $[a] \in \mathbb{IR}$, which is the largest absolute value in $[a]$, is defined by

$$\mathrm{mag}([a]) := \max_{a \in [\underline{a}, \overline{a}]} |a|.$$

For an interval vector and an interval matrix, it is applied

entrywise.

Throughout this paper, n-vectors $\mathbf{e}$ and $\mathbf{o}$ are defined by $\mathbf{e} := (1, \ldots, 1)^T$ and $\mathbf{o} := (0, \ldots, 0)^T$, respectively. For $p \in \{1, 2, \infty\}$ we denote p-norm of a real n-vector $x = (x_1, \ldots, x_n)^T$ and a real $m \times n$ matrix $A = (a_{ij})$ by

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \sum_{i=1}^{n} x_i^2, \quad \|x\|_\infty := \max_{1 \le i \le n} |x_i|$$

$$\|A\|_1 := \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_2 := \sigma_{\max}(A),$$

$$\|A\|_\infty := \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|,$$

where $\sigma_{\max}(A)$ is the largest singular value of A. Moreover, condition number of A is defined by

$$\mathrm{cond}_p(A) := \|A\|_p \|A^{-1}\|_p.$$

3. Verification theory

In this section, we will propose a method of calculating a tight error bound of an approximate solution $\tilde{x}$ of a linear system $Ax = b$.

We present in the following a linearized version of Yamamoto's theorem [10] to calculate the componentwise error bound of an approximate solution of a linear system.

Theorem 3.1 (Yamamoto [10]) *Let A be a real $n \times n$ matrix and b be a real n-vector. Let $\tilde{x}$ be an approximate solution of $Ax = b$ and $r := b - A\tilde{x}$. Suppose R is an approximate inverse of A and $G := I - RA$ with I denoting the $n \times n$ identity matrix. If $\|G\|_\infty < 1$, then A is nonsingular and*

$$|A^{-1}b - \tilde{x}| \le |Rr| + \frac{\|Rr\|_\infty}{1 - \|G\|_\infty} |G| e. \tag{3}$$

On the other hand, the following alternative approach for calculating the componentwise error bound is known in [2].

Theorem 3.2 (Ogita et al. [2]) *Let $A, b, \tilde{x}$ and r be as in Theorem 3.1. Let $\tilde{y}$ be an approximate solution of $Ay = r$. If A is nonsingular, then it holds that*

$$|A^{-1}b - \tilde{x}| \le |\tilde{y}| + \|A^{-1}\|_p \|r - A\tilde{y}\|_p e \tag{4}$$

for $p \in \{1, 2, \infty\}$.

The advantages of this approach are as follows:

- Although it needs an upper bound of $\|A^{-1}\|_p$, it does not necessarily need to compute an approximate inverse R of A.
- If $\tilde{y}$ is accurate enough, then the reminder term $\|A^{-1}\|_p \|r - A\tilde{y}\|_p e$ becomes almost negligible.
- It is compatible with iterative refinement and staggered correction (see Section 4).

The point of Theorem 3.2 is that $\tilde{y}$ can arbitrarily be improved for a fixed approximate solution $\tilde{x}$.

If an LU factorization with partial pivoting of A has been executed for calculating an approximate solution $\tilde{x}$ of $Ax = b$, we can compute the approximate inverse R of A by some algorithm (e.g. LAPACK's `DGETRI`) in $\frac{4}{3}n^3$ flops[1]. For example, using Matlab's notation we can proceed as follows:

$[L, U, P] = \mathtt{lu}(A);$	% LU factorization: $\frac{2}{3}n^3$ flops
$\tilde{x} = U \backslash (L \backslash (P * b));$	% forward/backward substitutions
$T = I/U;$	% solve $TU = I$ for T: $\frac{1}{3}n^3$ flops
$R = T/L;$	% solve $RL = T$ for R: n^3 flops
$R = R * P;$	% permutation, $R \approx A^{-1}$

or more simply

$$R = \mathtt{inv}(A); \quad \% \, R \approx A^{-1}\text{: } 2n^3 \text{ flops}$$
$$\tilde{x} = R * b;$$

Here, we emphasize that computing the approximate inverse R of A is a necessary measure for obtaining a rigorous error bound of the approximate solution $\tilde{x}$ of $Ax = b$, although it is widely held that computing R is not an efficient strategy for solving $Ax = b$.

After obtaining $R \in \mathbb{F}^{n \times n}$, a main part of computational effort to obtain the error bounds of $\tilde{x}$ is to calculate an upper bound of $\|I - RA\|_\infty$. To do this, a possibility is to calculate $[G] \in \mathbb{IF}^{n \times n}$ such that $I - RA \subseteq [G]$. It is known (e.g. [4]) that if $\|\mathrm{mag}([G])\|_\infty < 1$, then an upper bound ρ of $\|A^{-1}\|_\infty$ can be obtained by

$$\|A^{-1}\|_\infty \le \frac{\|R\|_\infty}{1 - \|I - RA\|_\infty} \le \frac{\|R\|_\infty}{1 - \|\mathrm{mag}([G])\|_\infty} =: \rho. \tag{5}$$

Using a usual matrix multiplication for including $I - RA$ with directed rounding requires $4n^3$ flops [4]. If an a priori estimation for $\mathrm{fl}(RA)$ is used, it requires $2n^3 + O(n^2)$ flops by calculating $\mathrm{fl}(I - RA)$ in $2n^3$ flops and $\mathrm{fl}(|R|(|A|e))$ in $O(n^2)$ flops. Faster (but less stable) methods of calculating an upper bound of $\|I - RA\|_\infty$ have also been presented in [4].

4. Iterative refinement and staggered correction

To obtain a tight enclosure of an approximate solution $\tilde{x}$ of a linear system $Ax = b$, we introduce an approach so-called "staggered correction".

Let $\mathbf{u} := 2^{-53}$. Using an iterative refinement (cf., e.g. [1]) for an approximate solution $\tilde{x} \in \mathbb{F}^n$, we may improve $\tilde{x}$ by $\tilde{x} + y$ where $y := \sum_{k=1}^{q} z^{(k)}$ with $z^{(k)} \in \mathbb{F}^n$ for $1 \le k \le q$. Then $z^{(k)}$ is called the staggered correction for $\tilde{x}$. This approach seems to be already used in [7]. If a good approximate inverse R of A has been calculated, we can obtain $\tilde{x} + y$ with

[1] addition, subtraction, multiplication or division are counted as one operation

arbitrarily high precision using the iterative refinement:

$y^{(0)} = \mathbf{o}$
for $k = 1, 2, \ldots, q$
  $r^{(k)} = \mathrm{accdot}(b - A(\tilde{x} + y^{(k-1)}))$ % accurate residual
  $z^{(k)} = \mathrm{fl}\left(Rr^{(k)}\right)$ % correction term
  $y^{(k)} = \mathrm{fl}\left(y^{(k-1)} + z^{(k)}\right)$

This makes only sense for calculating the residual $b - A(\tilde{x} + y^{(k-1)})$ when an accurate dot product is available. Fortunately, fast and portable methods for obtaining the accurate dot product have been developed in [3, 5]. We can use them for this purpose. For detail, see [3, 5].

We now assume that $\|G\|_\infty < 1$ for $G := I - RA$. Then A is nonsingular. For an arbitrary $\tilde{y} \in \mathbb{R}^n$, it holds that

$$A^{-1}b - \tilde{x} = A^{-1}b - (\tilde{x} + \tilde{y}) + \tilde{y}$$

and

$$|\tilde{y}| - \epsilon \le |A^{-1}b - \tilde{x}| \le |\tilde{y}| + \epsilon \quad \text{with} \quad \epsilon := |A^{-1}b - (\tilde{x} + \tilde{y})|.$$

By regarding $\tilde{x} + \tilde{y}$ as an approximate solution of $Ax = b$ (or $\tilde{y}$ as that of $Ay = r$ where $r := b - A\tilde{x}$), Theorem 3.1 implies

$$\begin{aligned} \epsilon &\le |R(b - A(\tilde{x} + \tilde{y}))| + \frac{\|R(b - A(\tilde{x} + \tilde{y}))\|_\infty}{1 - \|G\|_\infty}|G|e \\ &\le |R(r - A\tilde{y})| + \frac{\|G\|_\infty \|R(r - A\tilde{y})\|_\infty}{1 - \|G\|_\infty}e =: \epsilon_1. \end{aligned}$$

On the other hand, suppose $\|A^{-1}\|_p \le \rho$. Then it follows for $p \in \{1, 2, \infty\}$ that

$$\begin{aligned} \epsilon &= |A^{-1}(b - A(\tilde{x} + \tilde{y}))| = |A^{-1}(r - A\tilde{y})| \\ &\le \|A^{-1}(r - A\tilde{y})\|_\infty e \le \|A^{-1}(r - A\tilde{y})\|_p e \\ &\le \rho\|r - A\tilde{y}\|_p e =: \epsilon_2. \end{aligned}$$

From the above-mentioned discussions, we finally have the following propositions. Note that the validity of the propositions is independent of the quality of $\tilde{y}$.

Proposition 4.1 *Let $A, R, G, b, e, \tilde{x}$ and r be as in Theorem 3.1. Let $\tilde{y}$ be an approximate solution of $Ay = r$. If $\|G\|_\infty < 1$, then A is nonsingular and*

$$\max(|\tilde{y}| - \epsilon_1, \mathbf{o}) \le |A^{-1}b - \tilde{x}| \le |\tilde{y}| + \epsilon_1, \tag{6}$$

where

$$\epsilon_1 := |R(r - A\tilde{y})| + \frac{\|G\|_\infty \|R(r - A\tilde{y})\|_\infty}{1 - \|G\|_\infty}e.$$

Proposition 4.2 *Let $A, b, \tilde{x}$ and r be as in Theorem 3.1. Let $\tilde{y}$ be an approximate solution of $Ay = r$. Assume that A is nonsingular and ρ satisfies $\|A^{-1}\|_p \le \rho$ for any $p \in \{1, 2, \infty\}$. Then*

$$\max(|\tilde{y}| - \epsilon_2, \mathbf{o}) \le |A^{-1}b - \tilde{x}| \le |\tilde{y}| + \epsilon_2, \tag{7}$$

where $\epsilon_2 := \rho\|r - A\tilde{y}\|_p e$.

From Proposition 4.1, we can obtain tight componentwise lower and upper error bounds of $\tilde{x}$ by updating $\tilde{y}$ using the iterative refinement until satisfying

$$|\tilde{y}_i| \ge (\epsilon_1)_i \quad \text{for all } i,\ \tilde{y}_i \ne 0, \tag{8}$$

which becomes an appropriate stopping criterion for the iterations. Moreover, from Proposition 4.2 after obtaining an upper bound ρ of $\|A^{-1}\|_p$, we can also set an appropriate stopping criterion

$$\min_{1 \le i \le n,\ \tilde{y}_i \ne 0} |\tilde{y}_i| \ge \rho\|r - A\tilde{y}\|_p \tag{9}$$

for the iterative refinement.

5. Convergence of iterative refinement

Assume that an approximate inverse $R \in \mathbb{F}^{n \times n}$ of A is computed by some backward stable algorithm, e.g. LU factorization with partial pivoting. Let $\tilde{x} = Rb$. Then, without iterative refinement, the following is known as a rule of thumb: For $\mu := \mathrm{cond}_\infty(A) < \mathbf{u}^{-1}$ and $G := I - RA$, it holds that

$$|G_{ij}| = O(\mathbf{u}) \cdot \mu \quad \text{for all } (i, j). \tag{10}$$

Since

$$|A^{-1}b - \tilde{x}| = |A^{-1}b - Rb| = |(I - RA)A^{-1}b| \le |G||A^{-1}b|,$$

it holds that

$$|A^{-1}b - \tilde{x}| \le \|A^{-1}b\|_\infty |G|e. \tag{11}$$

After an iterative refinement, it follows by $y^{(1)} = R(b - A\tilde{x})$ that

$$\begin{aligned} |A^{-1}b - (\tilde{x} + y^{(1)})| &= |A^{-1}b - \tilde{x} - R(b - A\tilde{x})| \\ &= |(I - RA)(A^{-1}b - \tilde{x})| \\ &\le |G||A^{-1}b - \tilde{x}|. \end{aligned} \tag{12}$$

Inserting (11) into (12) yields

$$|A^{-1}b - (\tilde{x} + y^{(1)})| \le \|A^{-1}b\|_\infty |G|^2 e.$$

For $k \ge 2$, it can inductively be proven for $y^{(k)} = y^{(k-1)} + R(b - A(\tilde{x} + y^{(k-1)}))$ that

$$|A^{-1}b - (\tilde{x} + y^{(k)})| \le \|A^{-1}b\|_\infty |G|^{k+1} e$$

and

$$|A^{-1}b - (\tilde{x} + y^{(k)})| \le \alpha^{k+1}\|A^{-1}b\|_\infty e, \tag{13}$$

where $\alpha := \|G\|_\infty = O(n\mathbf{u}) \cdot \mu$. Therefore, if $\alpha < 1$, then the iterative refinement converges with the factor α for each iteration. In practice, due to the rounding error, we have $\tilde{x}^{(k)} = \mathrm{fl}\left(\tilde{x} + y^{(k)}\right)$ with $\tilde{x}^{(0)} = \tilde{x}$ and

$$|A^{-1}b - \tilde{x}^{(k)}| \le \mathbf{u}|A^{-1}b| + O(\alpha^{k+1})\|A^{-1}b\|_\infty e + \mathbf{u}_{\mathrm{N}} e,$$

where $\mathbf{u}_{\mathrm{N}}$ denotes the smallest positive normalized floating-point number. This is a *componentwise* error bound and explains behavior of the iterative refinement.

Table 1: History of iterative refinement for $k = 0, 1, 2$.

i	$\tilde{x}^{(0)}$
1	-171.1885408678072
2	$1.021301738732815 \cdot 10^3$
3	$1.000055792398647 \cdot 10^6$
4	$1.000000083648967 \cdot 10^9$

i	$\tilde{x}^{(1)}$
1	0.999935694692795
2	$1.000000011437893 \cdot 10^3$
3	$0.999999999979213 \cdot 10^6$
4	$0.999999999999980 \cdot 10^9$

i	$\tilde{x}^{(2)}$
1	1.000000000002729
2	$1.000000000000000 \cdot 10^3$
3	$1.000000000000000 \cdot 10^6$
4	$1.000000000000000 \cdot 10^9$

6. Numerical example

To confirm our discussions, let us consider the case where $A \in \mathbb{F}^{5\times5}$ with $\mathrm{cond}_\infty(A) \approx 10^{10}$ and $A^{-1}b = (1, 10^3, 10^6, 10^9, 134217728)^T$. Note that it is the *exact* solution, which does not include rounding errors. The last component 134217728 is not important but necessary only for generating a part of the exact solution $(1, 10^3, 10^6, 10^9)^T$, so that we omit to consider it. All computations are done in double precision arithmetic on Matlab, so that $\mathbf{u} = 2^{-53} \approx 10^{-16}$. An approximate inverse $R \in \mathbb{F}^{5\times5}$ of A is computed by a Matlab function `inv`, which uses LAPACK routines. Then, $\alpha = \|I - RA\|_\infty \approx 10^{-6}$ (because $\mathbf{u}\cdot\mathrm{cond}_\infty(A) \approx 10^{-6}$). An initial approximate solution $\tilde{x}^{(0)}$ is computed by $\tilde{x}^{(0)} = \mathrm{fl}\,(Rb)$.

The result of the iterative refiment is displayed in Table 1. As expected, each component is gradually improved with the factor α, in this case about 6 decimal digits, for each iteration until having the maximum accuracy.

Therefore, we can observe the following tendency of the iterative refinement: Let $x^* := A^{-1}b$, $x_{\max} := \max_{1\le i\le n} |x_i^*|$ and $x_{\min} := \min_{1\le i\le n} |x_i^*|$. Suppose $x_{\min} \neq 0$. If $x_{\max}/x_{\min}$ is very large, then it is likely that the component of $\tilde{x}$ corresponding to $x_{\min}$ is relatively less accurate than that to $x_{\max}$. An extreme case is that $x_{\min} = 0$. In such a case, the iterative refinement generally does not converge until entering underflow range.

Using our verification method, tight error bounds $\underline{\epsilon}$ and $\overline{\epsilon}$ in (2) for approximate solutions $\tilde{x}^{(k)}, k = 0, 1, 2$ can be obtained. For example, the verification result for $\tilde{x}^{(1)}$ is displayed in Table 2. From this, we can confirm that the proposed verification method provides very tight error bounds for the approximate solution of the linear system.

Table 2: Verification result for $\tilde{x}^{(1)}$.

i	$\underline{\epsilon}$
1	$2.034100276659818 \cdot 10^{-5}$
2	$7.057380003060988 \cdot 10^{-6}$
3	$3.3146924643425 79 \cdot 10^{-6}$
4	$1.502037048339663 \cdot 10^{-5}$

i	$\overline{\epsilon}$
1	$2.034100406248809 \cdot 10^{-5}$
2	$7.057380547965357 \cdot 10^{-6}$
3	$3.3146934583794 01 \cdot 10^{-6}$
4	$1.502037291962698 \cdot 10^{-5}$

References

[1] G. H. Golub, C. F. Van Loan, *Matrix Computations*, Third Edition, The Johns Hopkins University Press, Baltimore and London, 1996.

[2] T. Ogita, S. Oishi, Y. Ushiro, Computation of sharp rigorous componentwise error bounds for the approximate solutions of systems of linear equations, *Reliable Computing*, 9:3 (2003), 229–239.

[3] T. Ogita, S. M. Rump, S. Oishi, Accurate Sum and Dot Product, *SIAM J. Sci. Comput.*, 26:6 (2005), 1955–1988.

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

[5] S. M. Rump, T. Ogita, S. Oishi Accurate floating-point summation Part I and II, submitted for publication.

[6] S. M. Rump, INTLAB – INTerval LABoratory, Developments in Reliable Computing (T. Csendes ed.), Kluwer Academic Publishers, Dordrecht, 1999, 77–104.

[7] S. M. Rump, *Kleine Fehlerschranken bei Matrixproblemen*, Universität Karlsruhe, PhD thesis, 1980.

[8] S. M. Rump, Verification methods for dense and sparse systems of equations, Topics in Validated Computations – Studies in Computational Mathematics (J. Herzberger ed.), Elsevier, Amsterdam, 63–136, 1994.

[9] The MathWorks Inc., *MATLAB Users Guide, Version 7*, 2004.

[10] T. Yamamoto, Error bounds for approximate solutions of systems of equations, *Japan J. Appl. Math.*, 1:1 (1984), 157–171.