# Lower and upper error bounds of approximate solutions of linear systems

Takeshi OGITA (CREST, JST/Waseda University)

joint work with
Prof. Shin'ichi OISHI (Waseda University)

CIA'07, Sep. 9–15, 2007; Noszvaj, Hungary

# Outline

Purpose Let us consider a linear system $Ax = b$ where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. The purpose is

- to verify the nonsingularity of $A$, and then
- to verify the accuracy of an approximate solution $\widetilde{x}$ of the linear system.

# Why not compute $x^* = A^{-1}b$ ?

To solve large (e.g. 1 million unknowns) linear system $Ax = b$ on computer, we have to use floating-point arithmetic in practice.

floating-point arithmetic $\approx$ approximate computation

$\Longrightarrow$ We cannot compute the exact inverse $A^{-1}$ of large $A$.

$\Longrightarrow$ The approximation sometimes causes serious problems!

$\Longrightarrow$ Let's see what happens... (on Matlab)

# (Usual) verified computation

Notation: For $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $|x| = (|x_1|, \dots, |x_n|)^T$.

Given an approximate solution $\widetilde{x}$ of $Ax = b$, the usual verified computation gives an upper bound of the error or its norm:

$$|\widetilde{x} - A^{-1}b| \le \epsilon \in \mathbb{R}^n \quad \text{or} \quad \|\widetilde{x} - A^{-1}b\|_\infty \le \max_{1 \le i \le n} \epsilon_i = \varepsilon \in \mathbb{R}$$

$\Longrightarrow$ At least, $\widetilde{x}_i$ has correct digits (accuracy) corresponding to $\epsilon_i$.

$\Longrightarrow$ However, $\epsilon_i$ may be overestimated (too pessimistic).

$\Longrightarrow$ The quality of the verification is still not known!

# Quality of the verification

How (and whether) can we know it?

# Why compute both lower and upper error bounds

If both $\underline{\epsilon}$ and $\overline{\epsilon}$ s.t. $\underline{\epsilon} \le |\widetilde{x} - A^{-1}b| \le \overline{\epsilon}$ and $\overline{\epsilon} \approx \underline{\epsilon}$ are obtained, then the quality of the verification (evaluation) can be confirmed!

Question: Is it possible to obtain such $\underline{\epsilon}$ and $\overline{\epsilon}$ without much computational cost?

Answer: Yes. It is not so difficult! Let's see how to do it.

# Nonsingularity of $A$ and upper bound of $\|A^{-1}\|$

It needs some effort in terms of computational cost. For example,

- Let $R$ be an approximate inverse of $A$. If $\|I - RA\| < 1$, then $A$ is proved to be nonsingular and

$$\|A^{-1}\| \le \frac{\|R\|}{1 - \|I - RA\|}.$$

- computing a lower bound $\underline{\sigma}$ of the smallest singular value of $A$
$\Longrightarrow$ If $\underline{\sigma} > 0$, then $\|A^{-1}\|_2 \le 1/\underline{\sigma}$.

# Fundamental theorem

**Theorem 1. [Ogita et al., 2003]** *Let $A$ be a real $n \times n$ matrix and $b$ be a real $n$-vector. Let $\widetilde{x}$ be an approximate solution of $Ax = b$ and $r := b - A\widetilde{x}$. Let $\widetilde{y}$ be an approximate solution of $Ay = r$. If $A$ is nonsingular, then it holds for $p \in \{1, 2, \infty\}$ that*

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

*where $e := (1, \dots, 1)^T \in \mathbb{R}^n$.*

# Tight enclosure of the solution

For an arbitrary $y \in \mathbb{R}^n$, we have

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

It follows that

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

Using this and Theorem 1, we have the following proposition.

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

$$\max(|\widetilde{y}| - \epsilon, \mathbf{o}) \le |A^{-1}b - \widetilde{x}| \le |\widetilde{y}| + \epsilon, \tag{2}$$

*where* $\epsilon := \rho\|r - A\widetilde{y}\|_p e$ *and* $\mathbf{o} = (0, \ldots, 0)^T \in \mathbb{R}^n$.

$\Longrightarrow$ If $|\widetilde{y}_i| \gg \epsilon_i$, the error bounds are very tight!

$\Longrightarrow$ Such $|\widetilde{y}|$ can be obtained by the iterative refinement method.

# Iterative refinement and staggered correction

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

$\mathbb{F}$: a set of floaing-point numbers

Using iterative refinements, we can obtain $\widetilde{x} + y$ with arbitrarily higher precision: For $R \approx A^{-1}$

$$y^{(\ell+1)} = R * (b - A(\widetilde{x} + y^{(\ell)})),$$

where $y^{(\ell)} = \sum_{k=1}^{M} y_k^{(\ell)}$ with $y_k^{(\ell)} \in \mathbb{F}^n$. $\Longrightarrow$ The correction term $y$ can be expressed by the sum of floating-point vectors.

This makes only sense for calculating the residual $b - A(\widetilde{x} + y^{(\ell)})$ when an accurate dot product is available (Fortunately, we have it!).

[1] O., Rump, Oishi: *Accurate sum and dot product*, SIAM J. Sci. Comput., 26:6 (2005), 1955–1988.

[2] Rump, O., Oishi: *Accurate floating-point summation: Part I / Part II*, submitted for SISC.

On the other hand, to obtain tight error bounds, we need to compute

$$\epsilon_i = \rho\|r - A\widetilde{y}\|_p = \rho\|b - A(\widetilde{x} + \widetilde{y})\|_p.$$

This is compatible with the iterative refinements!

# Numerical experiments

(Matlab demo)

# Thanks!