Automatic Verified Numerical Computations for Linear Systems

Katsuhisa Ozaki (Shibaura Institute of Technology)
joint work with
Takeshi Ogita (Tokyo Woman's Christian University)
Shin'ichi Oishi (Waseda University)

SCAN2014, Würzburg, Germany

Sep. 23rd, 2014

Introduction

This talk is concerned with linear systems

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

Let $\tilde{x}$ be an approximate solution of the system.

Our concern is to obtain a bound for $\|\tilde{x} - x\|_\infty$.

We don't assume that A has special structure.

Introduction

Let $\mathbb{F}$ be a set of floating-point numbers as defined by IEEE 754 ($A \in \mathbb{F}^{n\times n}$, $b \in \mathbb{F}^n$).

For $R \in \mathbb{F}^{n\times n}$ and the identity matrix I,

$$\|RA - I\|_\infty < 1 \Longrightarrow \|x - \tilde{x}\|_\infty \leq \frac{\|R(A\tilde{x} - b)\|_\infty}{1 - \|RA - I\|_\infty}.$$

We focus on evaluation (upper bound) of $\|RA - I\|_\infty$.

Works

Table 1: Methods for Verification ($PA \approx LU, R \approx A^{-1}$)

	Year	Authors	R	Cost
A	2002	Oishi-Rump	$U^{-1}L^{-1}P$	$4/3n^3$
B	2011	Ozaki-Ogita-Oishi	$U^{-1}L^{-1}P$	$7/3n^3$
C	2005	Ogita-Oishi	$U^{-1}L^{-1}P$	$10/3n^3$
D	2006	Ogita-Rump-Oishi	A^{-1}	$4n^3$
E	2002	Oishi-Rump	A^{-1}	$6n^3$
F	2011	Ozaki-Ogita-Oishi	A^{-1}	$8n^3$
G	2011	Ozaki-Ogita-Oishi	A^{-1}	$12n^3$

Introduction

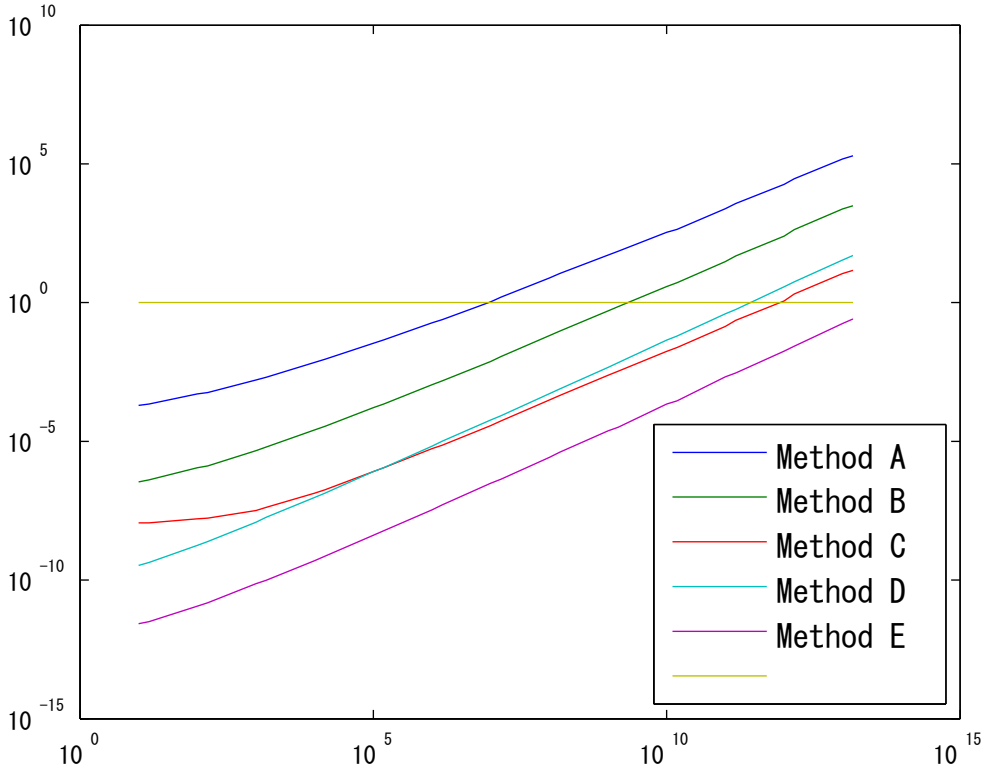


Figure 1: Treatable range of problems for each method

Introduction

We developed an algorithm which automatically selects a suitable method:

K. Ozaki, T. Ogita, S. Oishi: An Algorithm for Automatically Selecting a Suitable Verification Method for Linear Systems, Numerical Algorithms, Volume 56, Number 3 (2011), pp. 363-382.

Cost for selection is $O(n^2)$ flops.

Introduction

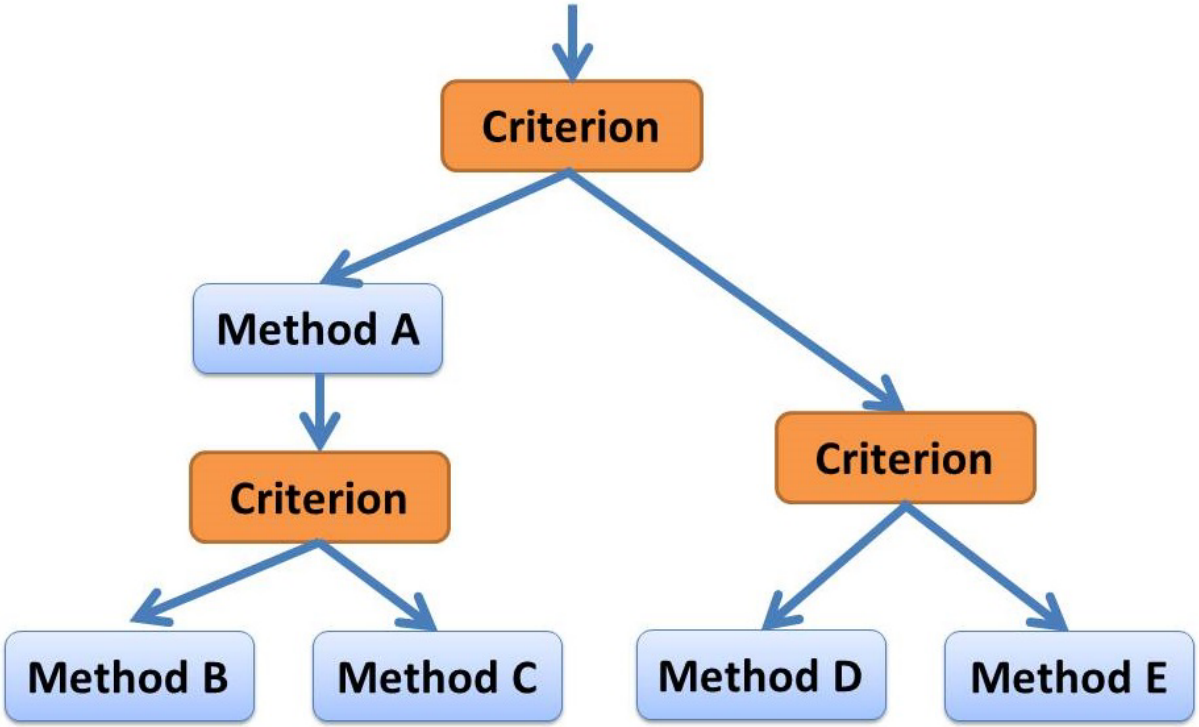
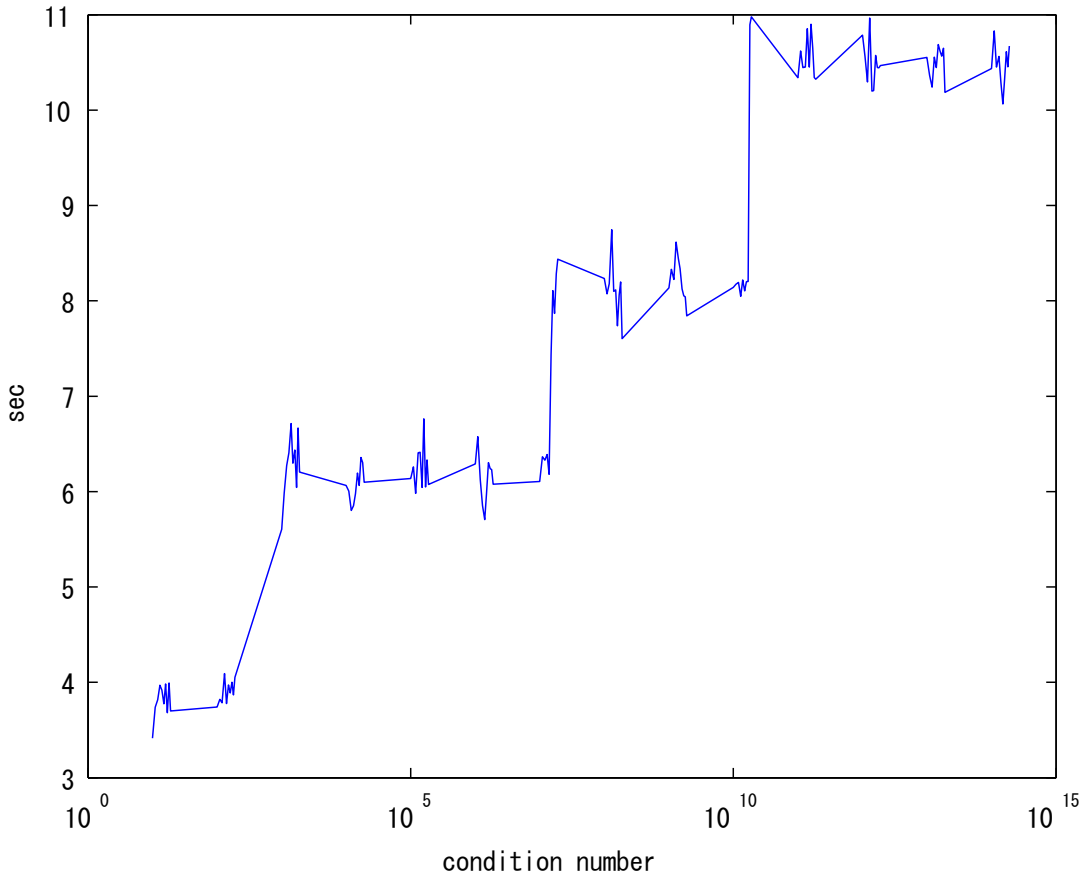


Figure 2: Flow of our Automatic Verification

Introduction



Introduction

We improve treatable range

- Method B (Ozaki-Ogita-Oishi, 2011)
- Method D (Ogita-Rump-Oishi, 2006)

by automatically using block computations.

Method D

fl(·): each operation in the parenthesis is evaluated by floating-point arithmetic. $\gamma_n = n\mathbf{u}/(1-n\mathbf{u})$, $n < \mathbf{u}^{-1}$ where $\mathbf{u}$ is the unit roundoff. We know

$$|RA - I - \mathrm{fl}(RA - I)| \leq (n+1)\mathbf{u}(|R||A| + I) \leq \gamma_{n+1}(|R||A| + I).$$

Then

$$|RA - I| \leq \mathrm{fl}(|RA - I|) + (n+1)\mathbf{u}(|R||A| + I).$$

Method D

$\mathrm{fl}_{\triangle}(\cdot)$: floating-point evaluation by rounding upwards mode.

$$|RA - I| \leq \mathrm{fl}_{\triangle}(\mathrm{fl}(|RA - I|) + (n+1)\mathbf{u}(|R||A| + I)).$$

By taking maximum norm

$$||RA - I||_{\infty} \leq \mathrm{fl}_{\triangle}(||\mathrm{fl}(|RA - I|)e + (n+1)\mathbf{u}(|R|(|A|e) + e)||_{\infty}),$$

where all elements in the vector e are ones.

Error for Dot Product

For $x, y \in \mathbb{F}^n$,

$$|\mathrm{fl}(x^T y) - x^T y| \leq n\mathbf{u}|x^T||y| \leq \gamma_n |x^T||y|$$

This inequality is satisfied for any order of computations.

But, assume that Winograd-type computations is not used.

Error for Dot Product

The following is dot product with pair wise order:

$$((x_1y_1 + x_2y_2) + (x_3y_3 + x_4y_4)) + ((x_5y_5 + x_6y_6) + (x_7y_7 + x_8y_8))$$

Let $\mathrm{fl}_p(x^T y)$ be a computed result by such pairwise computation,

$$|\mathrm{fl}_p(x^T y) - x^T y| \leq \gamma_{1+\lceil \log_2 \mathbf{n} \rceil} |x^T||y|$$

For Performance

- Generally, BLAS is used for matrix multiplication
- The detail in gemm is sometimes unknown
- We prefer small error constant
- Pair wise implementation will be slow

We apply simple block computations by using BLAS.

Block Computations

Both A and B are square matrices and n is dividable by s.

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ A_{21} & A_{22} & \cdots & A_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ A_{s1} & A_{s2} & \cdots & A_{ss} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1s} \\ B_{21} & B_{22} & \cdots & B_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ B_{s1} & B_{s2} & \cdots & B_{ss} \end{pmatrix}$$

Let C be a result of block computations, then

$$|C - AB| \leq \gamma_{n/s+s-1}|A||B|.$$

Block Computations

If we apply such block computations, then

$$\|RA - I\|_\infty \leq \|\mathrm{fl}_\triangle(\mathrm{fl}(|RA - I|)e + \gamma_{\mathbf{n/s+s}}(|R|(|A|e) + e))\|_\infty.$$

Optimal block size is approximately $\sqrt{n}$.

The constant $n\mathbf{u} \approx \gamma_n$ is approximately reduced to $\gamma_{2\sqrt{n}}$.

For example, γ_{199} is obtained for $n = 10000$.

SuperBlock Family

2-grouped computations:

$$C_{ij} = (((A_{i1}B_{1j}+A_{i2}B_{2j})+(A_{i3}B_{3j}+A_{i4}B_{4j}))+(A_{i5}B_{5j}+A_{i6}B_{6j})+\ldots$$

3-grouped computations:

$$C_{ij} = (((A_{i1}B_{1j}+A_{i2}B_{2j}+A_{i3}B_{3j})+(A_{i4}B_{4j}+A_{i5}B_{5j}+A_{i6}B_{6j}))+\ldots)$$

SuperBlock Family

For a result C by α-grouped computations,

$$|C - AB| \leq \gamma_{n'+\alpha+\lceil s/\alpha \rceil - 2}|A||B|.$$

The constant $n\mathbf{u} \approx \gamma_n$ approximately is reduced to $\gamma_{3\sqrt[3]{n}}$.

For example, γ_{58} is obtained for $n = 8000$.

For $n = 4,900$, we set $n' = 70$ and get γ_{139} for usual block computation.

However, setting $n' = 120$ and $\alpha = 8$, we get γ_{134}.

SuperBlock Family

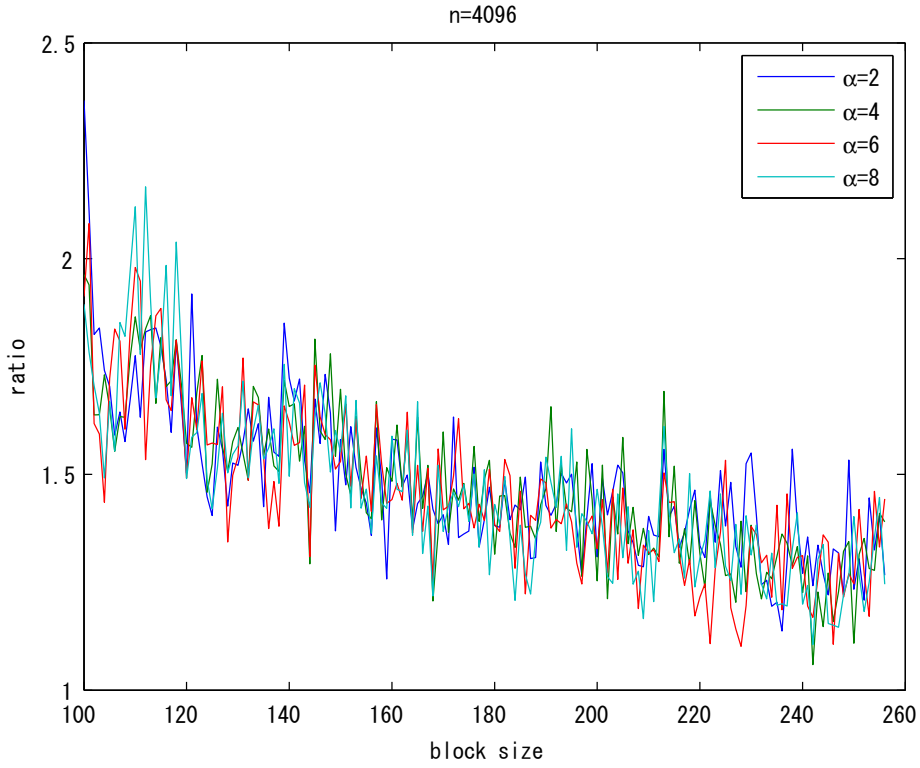


Figure 3: Each block is computed by multi-threads.

SuperBlock Family

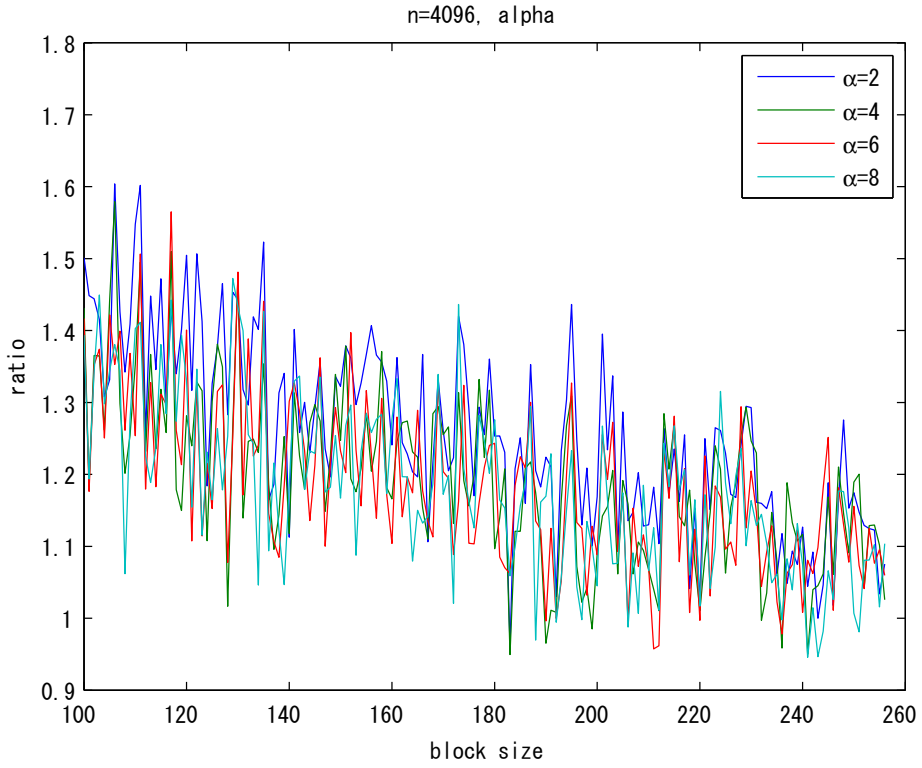


Figure 4: Some blocks are obtained in parallel.

Strategy

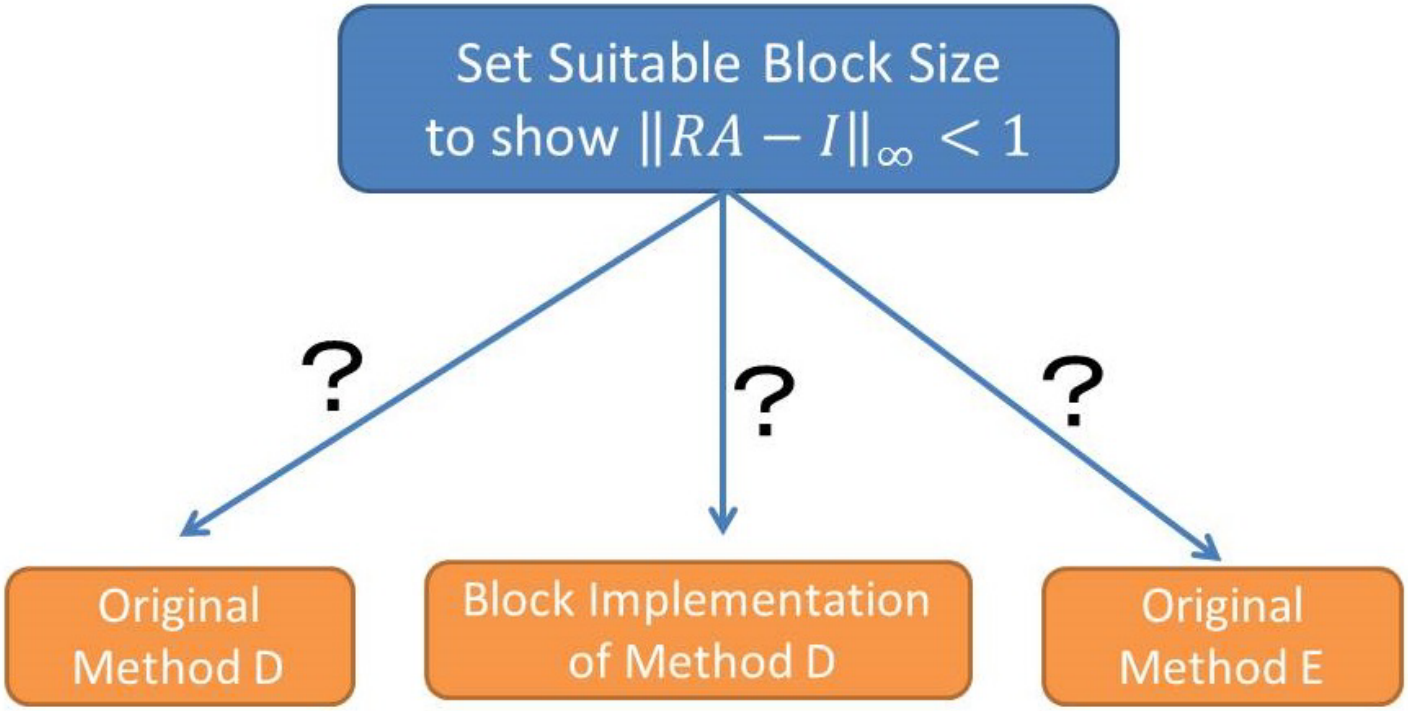


Figure 5: Choice

Modification of Method D

$$||RA-I||_\infty \leq \mathrm{fl}_\triangle(||\mathrm{fl}(|RA-I|)e+(n+1)\mathbf{u}(|R|(|A|e)+e)||_\infty). \quad (1)$$

From numerical experiments,

$$\mathrm{fl}(|RA-I|) \ll (n+1)\mathbf{u}(|R||A|+I) \leq \gamma_{n+1}(|R||A|+I).$$

First, we partially evaluate

$$\mathrm{fl}_\triangle(|||R|(|A|e)+e||_\infty) = p.$$

Modification of Method D

After that, we find k such that

$$\gamma_k p < c < 1$$

and set n' from $k = n' + \alpha + \lceil s/\alpha \rceil - 2$.

If k is too small, then we directly use Method E. If $k \geq n$, then we directly use original Method D.

Computational order is decided later

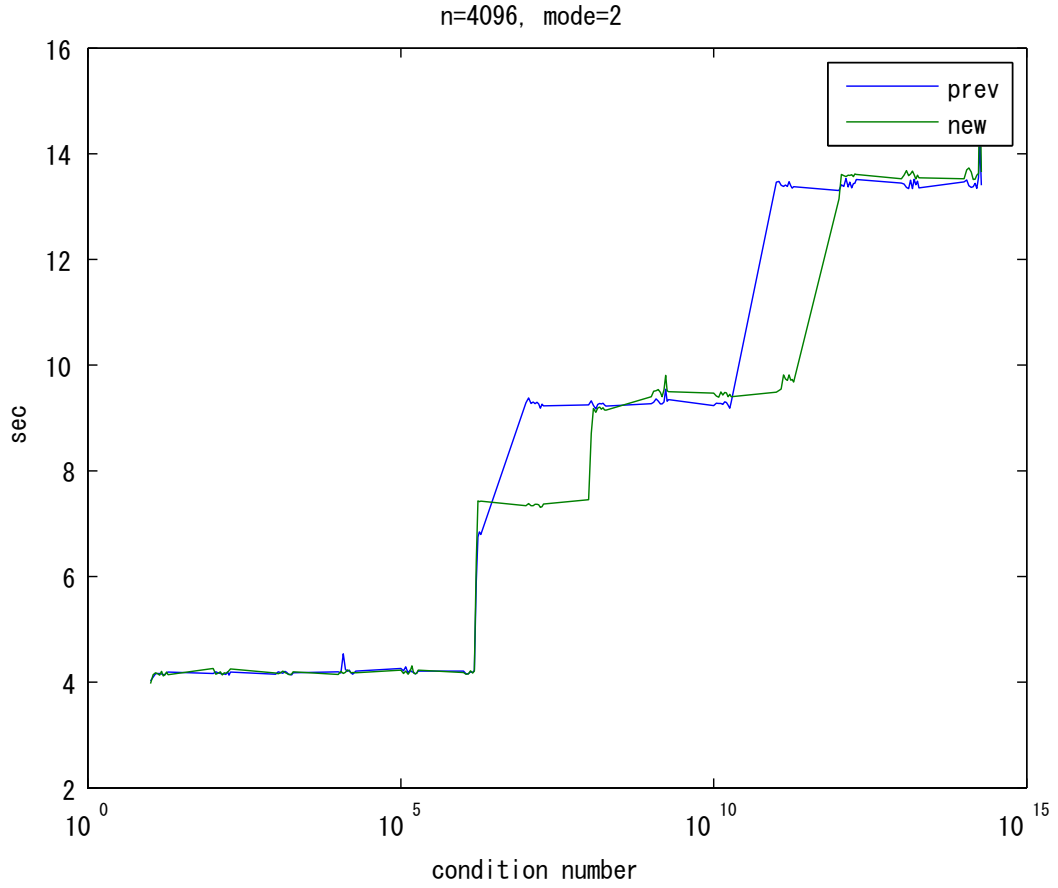
Modification of Method B

Let $PA \approx LU$ for LU decomposition, $X_U \approx U^{-1}$, $X_L \approx L^{-1}$, $e = (1, 1, \ldots, 1)^T$ and $B = PA$.

$$
\begin{aligned}
& \|RA - I\|_\infty \\
\leq\ & \| |X_U|(\mathrm{fl}(|X_L B - U|) + \mathbf{(n+1)u}(|X_L||B| + |U|) + n\mathbf{u}|U|) \|_\infty
\end{aligned}
$$

$(n+1)\mathbf{u}$ is reduced to $\gamma_{n'+\alpha+s/\alpha}$ by similar block computations.

Numerical Example (Computing time)



Conclusion

We improve two methods by automatically using block computations.

Block Computations efficiently reduce the error bound.

The performance of the grouped block computations is comparable to the original routines.