Floating-point arithmetic and error analysis (AFAE)

Fast verification methods for linear systems - Part II

Stef Graillat

LIP6/PEQUAN - Sorbonne University

Lecture Master 2 CCA

SCIENCES SORBONNE UNIVERSITÉ

Outline

1. Iterative refinement
   - Error analysis
   - Computation of a verified solution
   - Numerical experiments
2. Inversion of ill-conditioned matrices
   - Statement of the problem
   - Rump's algorithm
3. Conclusion

Problem statement

Let $A \in \mathbb{F}^{n\times n}$ be nonsingular, $b \in \mathbb{F}^n$, and $\widehat{x} \in \mathbb{F}^n$ an approximate solution of the linear system $Ax = b$. We have already seen four "fast" to compute an upper bound for $\|\widehat{x} - x\|_\infty$.

How can we increase the accuracy of $\widehat{x}$ if this one is not sufficient?

One possibility is to redo all the computations using extended precision.

Another possibility is to use iterative refinement to increase the accuracy of the computed solution.

Iterative refinement

- The iterative refinement is a technique to increase the accuracy of the computed solution $\widehat{x}$ of a linear system $Ax = b$:
  1. $\widehat{x}_i \leftarrow$ computed solution of $Ax = b$
  2. $\widehat{r}_i \leftarrow$ computed residual $b - A\,\widehat{x}_i$
  3. $\widehat{c}_i \leftarrow$ computed solution of $Ac_i = \widehat{r}_i$
  4. $\widehat{x}_{i+1} \leftarrow \mathrm{fl}(\widehat{x}_i + \widehat{c}_i)$
  5. go to step 2 if the stopping criterion is not satisfied
- We generally distinguish 2 cases:
  1. either the residual is computed with the working precision: this makes it possible to increase the backward error of the computed solution [Hig02, Thm 12.1 and 12.2],
  2. or the residual is computed with twice the working precision: this makes it possible to increase the forward error [Hig02, Thm 12.1].

Outline

1. Iterative refinement
   - Error analysis
   - Computation of a verified solution
   - Numerical experiments
2. Inversion of ill-conditioned matrices
   - Statement of the problem
   - Rump's algorithm
3. Conclusion

Conditioning

We introduce the following componentwise condition number:

$$\mathrm{cond}_{E,f}(A,x) := \lim_{\varepsilon\to 0} \sup_{\substack{|\Delta A|\le\varepsilon|E|\\ |\Delta b|\le\varepsilon|f|}} \left\{ \frac{\|\widehat{x}-x\|_\infty}{\varepsilon\|x\|_\infty}, (A+\Delta A)\,\widehat{x} = b+\Delta b \right\}.$$

If we take $E=|A|$ and $f=|b|$, we denote $\mathrm{cond}(A,x) := \mathrm{cond}_{A,b}(A,x)$, and we have

$$\mathrm{cond}(A,x) = \frac{\||A^{-1}||A||x|\|_\infty}{\|x\|_\infty}.$$

If we take $E=|A|$ and $f=0$, we denote $\mathrm{cond}(A) := \mathrm{cond}_{A,0}(A,x)$, and we have

$$\mathrm{cond}(A) = \||A^{-1}||A|\|_\infty.$$

Let us remark that $\mathrm{cond}(A,x) \le \mathrm{cond}(A) \le \kappa_\infty(A)$.

Error bound

We have already proved the following result:

Theorem 1 (Thm 9.4, p. 164, in [Hig02])

Let $A \in \mathbb{F}^{n\times n}$ and suppose GE produces computed LU factors $A \approx \widehat{L}\,\widehat{U}$, and a computed solution $\widehat{x}$ to $Ax = b$. Then

$$(A + \Delta A)\,\widehat{x} = b, \quad |\Delta A| \le \gamma_{3n}|\widehat{L}||\widehat{U}|.$$

This is not entirely satisfactory: in the normwise sense, even with GE, "the solution is stable only of the growth factor is small".
In this case, we can show that

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \le \gamma_{3n}\frac{\||A^{-1}||\widehat{L}||\widehat{U}||\widehat{x}|\|_\infty}{\|x\|_\infty},$$

where the second factor in the RHS is to be compared with $\mathrm{cond}(A, x)$.

Refinement with fixed precision u

We assume that for all matrix $A \in \mathbb{F}^{n\times n}$ and for all vector $b \in \mathbb{F}^n$, the method used for solving $Ax = b$ returns an approximate solution satisfying

$$(A + \Delta A)\,\widehat{x} = b, \quad \text{with} \quad |\Delta A| \le uW(A, n).$$

It is what we obtain with GE; in this case, indeed, $\widehat{x}$ satisfies

$$(A + \Delta A)\,\widehat{x} = b, \quad \text{with} \quad |\Delta A| \le \gamma_{3n}|\widehat{L}||\widehat{U}|.$$

Theorem 2 (Thm 12.2, p. 234, in [Hig02])

Let iterative refinement in fixed precision be applied to the nonsingular linear system $Ax = b$ of order n. Let $\eta = \mathbf{u}\| |A^{-1}|(|A| + W(A,n))\|_\infty$. Then, provided η is sufficiently less than 1, iterative refinement reduces the forward error by a factor approximately η at each stage, until

$$\frac{\|\widehat{x}_k - x\|_\infty}{\|x\|_\infty} \leq 2n\mathbf{u}\,\mathrm{cond}(A,x) + O(\mathbf{u}^2).$$

- If GE is used, $\eta = \mathbf{u}\| |A^{-1}|(|A| + 3n|\widehat{L}||\widehat{U}|)\|_\infty + O(\mathbf{u}^2)$: if $|\widehat{L}||\widehat{U}| \approx |A|$, then $\eta \approx 3n\mathbf{u}\,\mathrm{cond}(A)$.
- An upper bound in $\mathbf{u}\,\mathrm{cond}(A,x)$ is the best we can expect with working in precision $\mathbf{u}$.
- The theorem gives a sharper bound than the initial one

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \leq \gamma_{3n} \frac{\| |A^{-1}||\widehat{L}||\widehat{U}||\widehat{x}| \|_\infty}{\|x\|_\infty}.$$

Refinement with mixed precision

Theorem 3 ([Hig02, § 12.1])

Let iterative refinement be applied to the nonsingular linear system $Ax = b$ with residuals computed in double the working precision. Let $\eta = \mathbf{u}\| |A^{-1}|(|A| + W(A, n)) \|_\infty$. Then, provided η is sufficiently less than 1, iterative refinement reduces the forward error by a factor approximately η at each stage until

$$\frac{\| \widehat{x}_k - x \|_\infty}{\|x\|_\infty} \approx \mathbf{u}.$$

- This time, we obtain a forward error of the order of $\mathbf{u}$ of the condition number is sufficiently small
- By allowing to double the precision used to computer the residuals, we can obtain a forward error that is smaller than $\mathbf{u}\,\mathrm{cond}(A, x)$.

Outline

1. Iterative refinement
   - Error analysis
   - Computation of a verified solution
   - Numerical experiments
2. Inversion of ill-conditioned matrices
   - Statement of the problem
   - Rump's algorithm
3. Conclusion

We will use iterative refinement into those certified algorithm:

- certifLSV1 : $6 \cdot n^3 + O(n^2)$;
- certifLSV4 : $4/3 \cdot n^3 + O(n^2)$.

The algorithms will have the following form:

1. $\widehat{x}_i \leftarrow$ computed solution of $Ax = b$
2. $\widehat{r}_i \leftarrow$ computed residual $b - A\widehat{x}_i$ { with double the working precision }
3. $\widehat{c}_i \leftarrow$ computed solution of $Ac_i = \widehat{r}_i$
4. $\widehat{x}_{i+1} \leftarrow \mathrm{fl}(\widehat{x}_i + \widehat{c}_i)$
5. Computation of a verified upper bound δ for $\|\widehat{x}_{i+1} - x\|_\infty$
6. Go to step 2 if the stopping criterion is not valid

Choosing a stopping criterion

- Assume that we know an upper bound $\overline{\delta}$ such that $\|x - \widehat{x}\|_\infty \leq \overline{\delta}$.
  We have $\overline{\delta} \geq \|\widehat{x}\|_\infty - \|x\|_\infty$, so $\|x\|_\infty \geq \|\widehat{x}\|_\infty - \overline{\delta}$.
  Hence, assuming that $\overline{\delta} < \|\widehat{x}\|_\infty$, we obtain:

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \leq \frac{\overline{\delta}}{\|\widehat{x}\|_\infty - \overline{\delta}}.$$

  From algorithm `bndDelta1`, we can deduce an algorithm `bndEpsilon`, to upper bound the relative error $\|x - \widehat{x}\|_\infty / \|x\|_\infty$.

- We choose to stop the iterative refinement when:
  - either $\|x - \widehat{x}\|_\infty / \|x\|_\infty \leq \tau$, where τ is the tolerance;
  - or 3 steps of iterative refinement have already been done.

Verified solution of linear system with iterative refinement

function $[\widehat{x}, \overline{\varepsilon}]$ = certifLSV1rafit(A, b, τ)

$[L, U, P] = \mathrm{fl}\left(\mathrm{xGETRF}(A)\right)$ $\{\ 2/3 \cdot n^3 + O(n^2)\ \}$

$\widehat{x} = \mathrm{fl}\left(\mathrm{xGETRS}(P, L, U, b)\right)$ $\{\ O(n^2)\ \}$

$R = \mathrm{fl}\left(\mathrm{xGETRI}(P, L, U)\right)$ $\{\ 4/3 \cdot n^3 + O(n^2)\ \}$

$\overline{\alpha} = \mathrm{bndAlpha1}(A, R)$ $\{\ 4 \cdot n^3 + O(n^2)\ \}$

if $\overline{\alpha} \geq 1$ then error('Certification failed')

k = 0; while true do

$< m_{res}, r_{res} >= \mathrm{resLinSys2}(A, b, \widehat{x})$ $\{\ O(n^2)\ \}$

$\overline{\varepsilon} = \mathrm{bndEpsilon}(A, b, \widehat{x}, R, < m_{res}, r_{res} >)$ $\{\ O(n^2)\ \}$

if $\overline{\varepsilon} \leq \tau$ then return

if $k \geq 3$ then error('Convergence failed')

$\widehat{c} = \mathrm{fl}\left(\mathrm{xGETRS}(P, L, U, m_{res})\right)$ $\{\ O(n^2)\ \}$

$\widehat{x} = \mathrm{fl}(\widehat{x} + \widehat{c})$

$k = k + 1$

done

Cost of the algorithm: $6 \cdot n^3 + O(n^2)$.

Outline

All the numerical experiments were done the the same environment as previously:

- MATLAB, with INTLAB toolbox;
- IEEE-754 double precision;
- the ill-conditioned linear systems are generated as follows:

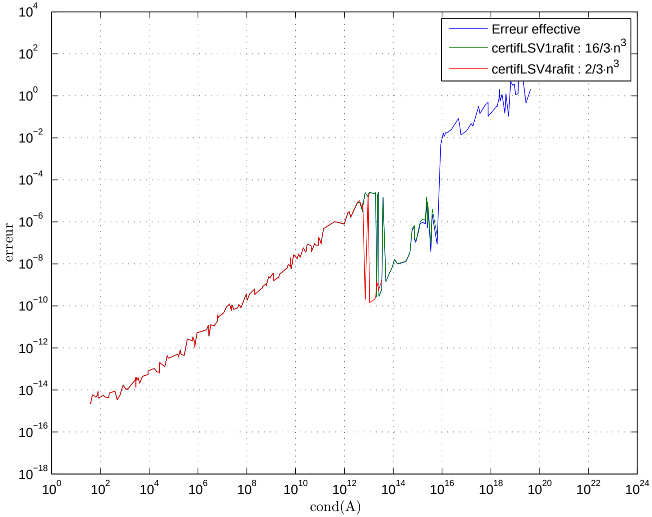
```
A = gallery('randsvd', n, 10^(k*rand));
b = A*ones(n,1);
```

  In the experiments, $n = 50$.

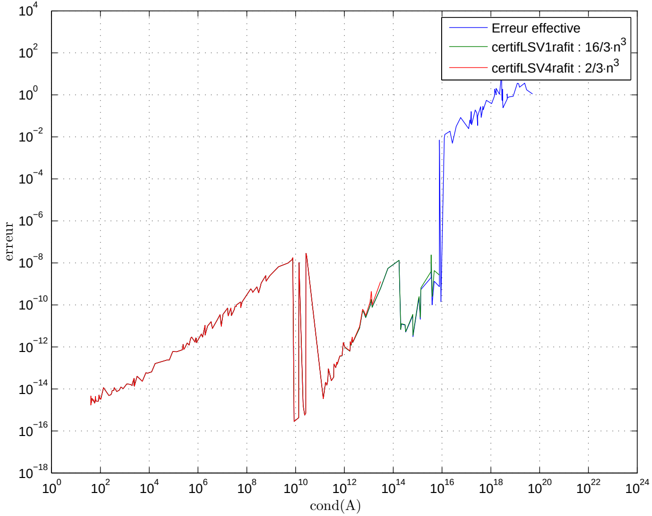
In the x-axis, we have $\kappa(A)$, in the y-axis we have the relative error. We use a logarithmic scale.

We change the tolerance τ which is the targeted relative error.

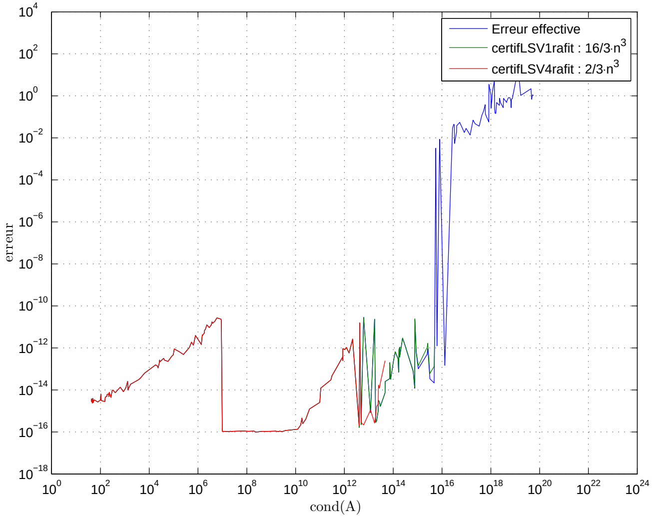
System 50×50, $\texttt{tol} = 2^{-15} \approx 3 \cdot 10^{-5}$:



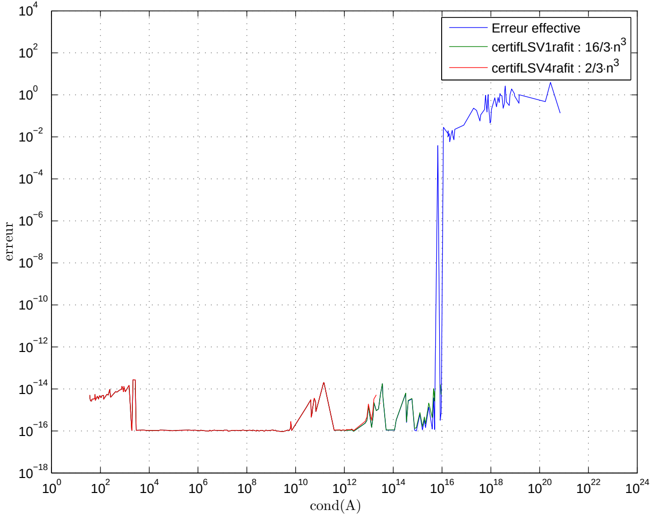
System 50×50, $\texttt{tol} = 2^{-25} \approx 3 \cdot 10^{-8}$:



System 50×50, $\mathtt{tol} = 2^{-35} \approx 3 \cdot 10^{-11}$:



System 50×50, $\mathtt{tol} = 2^{-45} \approx 3 \cdot 10^{-14}$:



Summary

- The certified error bound behave like the true relative error for the computed solution: this behavior appears in increasing the accuracy of the computed solution by iterative refinement.
- Iterative refinement combined with verified algorithms make it possible to guarantee a "small" rigorous forward error as long as the condition number is "sufficiently small" compared to $1/\mathbf{u}$. With the fastest algorithm we presented, it is still needed to double the number of floating-point operations.
- It would be interesting to provide efficient implementations of those algorithms as well as a deep experimental study in order to compare the performances with [DHK+06].

Outline

Outline

Statement of the problem

Let $A \in \mathbb{F}^{n \times n}$ be nonsingular. If we can computed an approximate inverse $R = \mathrm{inv}(A)$ in floating-point arithmetic, and that $\|RA - I\| < 1$, we can show that A is nonsingular. It will be the case if $\kappa(A)$ is sufficiently small compared to u^{-1}.

The Kahan-Gastinel theorem states that

$$\kappa(A)^{-1} = \min\left\{ \frac{\|\Delta A\|}{\|A\|}, A + \Delta A \text{ is singular} \right\}.$$

If $\kappa(A) > u^{-1}$, a perturbation of A with normwise norm of order u can lead to a singular matrix...Moreover, if one wants to prove the nonsingularity of A, the strategy consisting in verifying that $\|RA - I\| < 1$ will likely to fail.

Statement of the problem

If $\kappa(A) > \mathbf{u}^{-1}$, how can we prove the nonsingularity of A using floating-point arithmetic?

- We can still use multiprecision arithmetic
- There exists an algorithm from Rump [Rum09], in which it is only necessary to increase the precision used for matrix products.

Outline

Principle of the algorithm

Assume that we can perform some matrix operations exaclty:

$R = \mathrm{inv}(A)$ {approximate inverse, computed with work

$P = RA$ {exact}

$X = P^{-1}$ {exact}

$R' = XR$ {exact}

In this case, $X = A^{-1}R^{-1}$, so $R' = XR = A^{-1}$.

In practice, we only authorize computation with finite precision so the previous algorithm cannot be used: but we can try to use some multiplicative corrections to compute a more accurate inverse of A.

Rump's algorithm

The idea of Rump is to locally use a precision greater that the working precision. Let $A, B \in \mathbb{F}^{n\times n}$ and $k \geq 2$.

The notation $P = fl_{k,1}(AB)$ means that the product AB is computed with precision u^k, then rounded to the working precision $\mathbf{u}$:

$$\| \underset{k,1}{fl}(AB) - AB\| \leq \mathbf{u}\|AB\| + n\mathbf{u}^k\|A\|\|B\| + O(\mathbf{u}^{k+1}).$$

The notation $\{P\} = fl_{k,k}(AB)$ means that AB is computed with precision u^k and the result is represented as a non-evaluated sum $\{P\} = \sum_{i=1}^{k} P_i$ ($P_i \in \mathbb{F}^{n\times n}$):

$$\| \underset{k,1}{fl}(AB) - AB\| \leq n\mathbf{u}^k\|A\|\|B\| + O(\mathbf{u}^{k+1}).$$

Moreover, if $\{P\} = \{P_1, \ldots, P_\ell\}$,

$$\underset{k,1}{fl}(\{P\}B) = \underset{k,1}{fl}\left(\sum_{i=0}^{\ell} P_i B\right) \quad \text{and} \quad \underset{k,k}{fl}(\{P\}B) = \underset{k,k}{fl}\left(\sum_{i=0}^{\ell} P_i B\right).$$

Inversion of an ill-condition matrix

function $\{R^{(k)}\}$ = InvIllCond(A)

$R^{(0)} = \mathrm{fl}(\|A\|^{-1}) \cdot I$ {starting "approximate inverse"}

$k = 0$

repeat

$k = k + 1$

$P^{(k)} = \mathrm{fl}_{k,1}(\{R^{(k-1)}\} \cdot A)$

$X^{(k)} = \mathrm{inv}(P^{(k)})$

$\{R^{(k)}\} = \mathrm{fl}_{k,k}(X^{(k)} \cdot \{R^{(k-1)}\})$

until $\mathrm{cond}(P^{(k)}) < (100\mathbf{u})^{-1}$

If $\kappa(A) \gg \mathbf{u}^{-1}$, Rump justifies in an heuristic way [Rum09] that

$$\mathrm{cond}(\{R^{(k)}\}A) \approx \mathbf{u}^{k-1}\,\mathrm{cond}(A).$$

Moreover, we can hope that the algorithm terminates with $\|\{R^{(k)}\}A - I\| \leq 1/100$.

Let us take for example the following matrix A, such $\kappa(A_1) \approx 6.4 \cdot 10^{63}$:

$$A = \begin{bmatrix} -5046135670319638 & -3871391041510136 & -5206336348183639 & -6745986988231149 \\ -640032173419322 & 8694411469684959 & -564323984386760 & -2807912511823001 \\ -16935782447203334 & -18752427538303772 & -8188807358110413 & -14820968618548534 \\ -1069537498856711 & -14079150289610606 & 7074216604373039 & 725796028397871 \end{bmatrix}.$$

For this matrix, we can observe:

k	$\mathrm{cond}(R^{(k-1)})$	$\mathrm{cond}(R^{(k-1)}A)$	$\mathrm{cond}(P^{(k)})$	$\|I - R^{(k)}A\|$
2	$1.68 \cdot 10^{17}$	$2.73 \cdot 10^{49}$	$2.31 \cdot 10^{17}$	3.04
3	$1.96 \cdot 10^{32}$	$2.91 \cdot 10^{33}$	$2.14 \cdot 10^{17}$	5.01
4	$7.98 \cdot 10^{48}$	$1.10 \cdot 10^{17}$	$1.83 \cdot 10^{17}$	1.84
5	$6.42 \cdot 10^{64}$	8.93	8.93	$3.43 \cdot 10^{-16}$

- $\mathrm{cond}(R^{(k-1)}A)$ decreases by a factor of order u at each iteration.
- By verifying $\|I - R^{(k)}A\| < 1$, on can show that A is nonsingular.

"Final version of the algorithm"

Inversion of an ill-conditioned matrix

function {R} = InvIllCond(A)

  $\{R\} = \mathrm{fl}(\|A\|^{-1}) \cdot I$; $P = X = \infty$; $k = 0$

  repeat

    $\text{finished} = (\|P\| \cdot \|X\| < (100u)^{-1})$

    $k = k + 1$

    $P = \mathrm{fl}_{k,1}(\{R\} \cdot A)$

    $X = \mathrm{inv}(P)$

    while "inversion failed" do

      $P = P + \Delta P$ {random perturbations such that $|\Delta P| \leq u|P|$}

      $X = \mathrm{inv}(P)$

    done

    $\{R\} = \mathrm{fl}_{k,k}(X \cdot \{R\})$

  until finished

Outline

We presented four "fast" methods to compute a verified solution of linear systems with floating-point arithmetic.

- These methods require between 2 and 9 times for floating-point operations than the classic GE.
- All the methods can be implemented using the BLAS routines.
- They are efficient as long as the condition number is small compared to $\mathbf{u}^{-1}$.

When the condition number is small compared to $\mathbf{u}^{-1}$, if the accuracy of the computed solution is not sufficient, it is possible to use iterative refinement with mixed precision to obtain a normwise relative error of order $\mathbf{u}$.

If the condition number is larger than $\mathbf{u}^{-1}$, it is possible to use the Rump's inversion algorithm for preconditioning the matrix and obtain an approximate solution that can be verified.

References I

James Demmel, Yozo Hida, William Kahan, Xiaoye S. Li, Sonil Mukherjee, and E. Jason Riedy, Error bounds from extra-precise iterative refinement, ACM Trans. Math. Softw. **32** (2006), no. 2, 325–351.

Nicholas J. Higham, Accuracy and stability of numerical algorithms, second ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

S. M. Rump, Inversion of extremely ill-conditioned matrices in floating-point, Japan J. Indust. Appl. Math. (JJIAM) **26** (2009), 249–277.