Numerical Stability of Linear System Solution Made Easy

Ilse C.F. Ipsen

North Carolina State University
Raleigh, NC, USA

The problem

Given:

Nonsingular matrix $A \in \mathbb{R}^{n \times n}$
Right hand side vector $b \in \mathbb{R}^n$

Solve: $Ax = b$ in floating point arithmetic

Numerical stability:

Quantifies amplification of roundoff errors by algorithm

Overview:

1. Forward error: Perturbation bound (algorithm independent)
2. Direct methods for solving $Ax = b$
3. Backward error: Roundoff error bounds (algorithm dependent)
4. Perturbation bound for numerical stability of direct methods

Forward error: Perturbation bound
(algorithm independent)

Vector p-norms

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p} \quad \text{for} \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad p \geq 1$$

- $p = 1$: One norm

$$\|x\|_1 = |x_1| + \cdots + |x_n|$$

- $p = 2$: Two (Euclidean) norm

$$\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$$

- $p = \infty$: Infinity (max) norm

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$$

Induced matrix p-norms

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

- $p = 1$: Largest absolute column sum $\|A\|_1 = \max_j \sum_i |A_{ij}|$
- $p = \infty$: Largest absolute row sum $\|A\|_\infty = \max_i \sum_j |A_{ij}|$
- $p = 2$: Largest singular value $\|A\|_2 = \max_j \sqrt{|\lambda_j(A^T A)|}$

Condition number with respect to inversion of nonsingular A

$$\kappa_p(A) = \|A^{-1}\|_p \|A\|_p$$

Perturbation bound for forward error

Input: Nonsingular $A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$
Want: Solution to $Ax = b$
Computed solution: $z \neq 0$ with residual $r = Az - b$
How close is z to x?

$$\underbrace{\frac{\|z - x\|_p}{\|z\|_p}}_{\text{Relative error of } z} \leq \underbrace{\kappa_p(A)}_{\text{Conditioning}} \underbrace{\frac{\|r\|_p}{\|A\|_p \|z\|_p}}_{\text{Stability}}$$

Problem sensitivity (conditioning):

A well-conditioned if $1 \leq \kappa(A) \lesssim n$
A numerically singular if $\kappa_p(A) \gtrsim 10^{15}$ {IEEE double precision}

Algorithm: Backward stable if $\frac{\|r\|_p}{\|A\|_p \|z\|_p} \lesssim 10^{-16}$

Derivation of perturbation bound

1. Residual

$$r = Az - b = Az - Ax = A(z - x)$$

2. A is invertible

$$z - x = A^{-1}\, r$$

3. Take norms

$$\|z - x\|_p \leq \|A^{-1}\|_p\, \|r\|_p = \underbrace{\|A^{-1}\|_p \|A\|_p}_{\kappa_p(A)} \frac{\|r\|_p}{\|A\|_p}$$

4. Divide by $\|z\|_p$

Direct methods for solving $Ax = b$

Popular direct methods

Gaussian elimination without pivoting (if it exists)

1. Factor $A = LU$ where L is unit ◺ and U is ◹
2. Solve ◺ system $Ly = b$ $\{y = L^{-1}b\}$
3. Solve ◹ system $Ux = y$ $\{x = U^{-1}y = U^{-1}L^{-1}b = A^{-1}b\}$

Popular direct methods

Gaussian elimination without pivoting (if it exists)

1. Factor $A = LU$ where L is unit ◺ and U is ◹
2. Solve ◺ system $Ly = b$ $\{y = L^{-1}b\}$
3. Solve ◹ system $Ux = y$ $\{x = U^{-1}y = U^{-1}L^{-1}b = A^{-1}b\}$

Gaussian elimination with partial pivoting (GEPP)

Factor $A = (P^T L)U$ where permutation P reorders the rows

Cholesky decomposition (for symmetric positive definite A)

Factor $A = LL^T$ where L is ◺

QR decomposition

Factor $A = QR$ where $Q^T = Q^{-1}$ and R is ◹

Example: Worst case GEPP $A = (P^T L)\, U$

$n = 4$:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{pmatrix}}_{P^T L} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}}_{U}$$

Elements of L are bounded: $\|P^T L\|_\infty \le n$

Growth factor for elements of U:

$$\rho_n \equiv \frac{\max_{i,j,k} |A_{i,j}^{(k)}|}{\max_{i,j} |A_{ij}|} = 2^{n-1}$$

(Largest element in factorization / largest element of A)

Question: Why is element growth bad?
Answer: See roundoff error analysis, next

Backward error: Roundoff error bounds

Roundoff error analysis for direct methods

James H. Wilkinson
Rounding Errors in Algebraic Processes
(1963)

Nicholas J. Higham
Accuracy and Stability of Numerical Algorithms
Second edition (2002)

Roundoff error for elementary operations $\mathrm{op} \in \{+, -, *, \}$

$$\mathrm{fl}(\alpha \,\mathrm{op}\, \beta) = (\alpha \,\mathrm{op}\, \beta)(1 + \delta)$$

where $|\delta| \leq u \approx 10^{-16}$ in IEEE double precision

Lastly, suppose that $i \geqslant k+1$ and $j \geqslant k+1$. Corresponding to (3.8) and (3.9) we have

$$-\bar{m}_{ik}\bar{a}_{kj}^{(k)} \simeq -\overline{m_{ik}a_{kj}^{(k)}};\quad \mathrm{rp}\,(\gamma),$$

and

$$-\bar{m}_{ik}\bar{a}_{kj}^{(k)} \simeq -\overline{m_{ik}a_{kj}^{(k)}};\quad \mathrm{ap}\,(|\overline{m_{ik}a_{kj}^{(k)}}|\gamma\, e^{\gamma}).$$

Addition of $\bar{a}_{ij}^{(k)}$ to both sides followed by abbreviation of the right member to rp (γ) yields

$$-\bar{m}_{ik}\bar{a}_{kj}^{(k)}+\bar{a}_{ij}^{(k)} \simeq \bar{a}_{ij}^{(k+1)};\quad \mathrm{ap}\,\{(|\overline{m_{ik}a_{kj}^{(k)}}|+|\bar{a}_{ij}^{(k+1)}|)\gamma\, e^{\gamma}\}; \tag{3.13}$$

compare the first of (2.6). A further application of the exponential rule yields

$$|\overline{m_{ik}a_{kj}^{(k)}}|+|\bar{a}_{ij}^{(k+1)}| \leqslant |\overline{m_{ik}a_{kj}^{(k)}}|\, e^{\gamma_l}+|\bar{a}_{ij}^{(k+1)}|e^{\gamma_l} \leqslant \overline{\phi a_{ij}^{(k+1)}}\, e^{2\gamma_l}, \tag{3.14}$$

where

$$\phi a_{ij}^{(k+1)} = |m_{ik}a_{kj}^{(k)}|+|a_{ij}^{(k+1)}|,\quad i,j \geqslant k+1. \tag{3.15}$$

As indicated in (3.14), in computing $\overline{\phi a_{ij}^{(k+1)}}$ we assume that the product $|m_{ik}a_{kj}^{(k)}|$ and the term $|a_{ij}^{(k+1)}|$ are abbreviated separately from $\mathscr{M}$ to $\mathscr{L}$ and then added.† Using the inequality $\gamma \leqslant \gamma_l$ and substituting (3.14) in (3.13), we obtain

$$-\bar{m}_{ik}\bar{a}_{kj}^{(k)}+\bar{a}_{ij}^{(k)} \simeq \bar{a}_{ij}^{(k+1)};\quad \mathrm{ap}\,(\overline{\phi a_{ij}^{(k+1)}}\,\gamma\, e^{3\gamma_l}). \tag{3.16}$$

Then by use of (3.6) we may recast this relation in the form

$$\bar{m}_{ik}\bar{a}_{kj}^{(k+1)}+\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)}-\Delta a_{ij}^{(k+1)},\quad i,j \geqslant k+1, \tag{3.17}$$

where

$$|\Delta a_{ij}^{(k+1)}| \leqslant \overline{\phi a_{ij}^{(k+1)}}\,\gamma\, e^{3\gamma_l}. \tag{3.18}$$

In a similar manner we derive

$$\bar{m}_{ik}\bar{b}_{k}^{(k+1)}+\bar{b}_{i}^{(k+1)} = \bar{b}_{i}^{(k)}-\Delta b_{i}^{(k+1)},\quad i \geqslant k+1, \tag{3.19}$$

where

$$|\Delta b_{i}^{(k+1)}| \leqslant \overline{\phi b_{i}^{(k+1)}}\,\gamma\, e^{3\gamma_l}, \tag{3.20}$$

and

$$\phi b_{i}^{(k+1)} = |m_{ik}b_{k}^{(k)}|+|b_{i}^{(k+1)}|,\quad i \geqslant k+1. \tag{3.21}$$

Equations (3.6), (3.11), (3.17) and (3.19) may be combined into matrix form:

$$\bar{L}^{(k)}\bar{A}^{(k+1)} = \bar{A}^{(k)}-\Delta A^{(k+1)},\qquad \bar{L}^{(k)}\bar{b}^{(k+1)} = \bar{b}^{(k)}-\Delta b^{(k+1)}, \tag{3.22}$$

where

$$\Delta A^{(k+1)} = \left[\begin{array}{c|ccc} O & & 0 & \\ \hline & \Delta a_{k+1,k}^{(k+1)} & \cdots & \Delta a_{k+1,n}^{(k+1)} \\ O & \vdots & \vdots & \vdots \\ & \Delta a_{nk}^{(k+1)} & \cdots & \Delta a_{nn}^{(k+1)} \end{array}\right],\qquad \Delta b^{(k+1)} = \left[\begin{array}{c} 0 \\ \hline \Delta b_{k+1}^{(k+1)} \\ \vdots \\ \Delta b_{n}^{(k+1)} \end{array}\right]. \tag{3.23}$$

† It is essential that the value of the product $m_{ik}a_{kj}^{(k)}$ be extracted from the main computations. However, the needed relation (3.16) would remain valid if $\overline{\phi a_{ij}^{(k+1)}}$ were to be computed by adding $|m_{ik}a_{kj}^{(k)}|$ and $|a_{ij}^{(k+1)}|$ in $\mathscr{M}$ and then abbreviating the result to $\mathscr{L}$.

Gaussian elimination with partial pivoting in floating point arithmetic [Higham, Wilkinson]

Solve: $Ax = b$ where $A \in \mathbb{R}^{n\times n}$ nonsingular

Perturbation bound for computed solution z:

$$\frac{\|z-x\|_\infty}{\|z\|_\infty} \quad \leq \quad \kappa_\infty(A)\,\frac{\|r\|_\infty}{\|A\|_\infty\|z\|_\infty}$$

Backward error for GEPP in floating point $(u \approx 10^{-16})$

$$\frac{\|r\|_\infty}{\|A\|_\infty\|z\|_\infty} \lesssim 3\,n^3\,u\,\rho_n$$

Growth factor: $\rho_n \;\equiv\; \frac{\max_{i,j,k}|A_{ij}^{(k)}|}{\max_{i,j}|A_{ij}|}$

Large growth factor $\Longrightarrow$ GEPP backward unstable

Perturbation bound for numerical stability of direct methods

General view of direct methods

Exact arithmetic:

1. Factor $A = S_1 S_2$ where square S_1 and S_2 are "simple" to solve
2. Solve $S_1 y = b$ $\{y = S_1^{-1} b\}$
3. Solve $S_2 x = y$ $\{x = S_2^{-1} y = S_2^{-1} S_1^{-1} b = A^{-1} b\}$

General view of direct methods

Exact arithmetic:

1. Factor $A = S_1 S_2$ where square S_1 and S_2 are "simple" to solve
2. Solve $S_1 y = b$ $\{y = S_1^{-1} b\}$
3. Solve $S_2 x = y$ $\{x = S_2^{-1} y = S_2^{-1} S_1^{-1} b = A^{-1} b\}$

Perturbation model for floating point arithmetic:

1. Factor $A + E = S_1 S_2$ where $\epsilon_A \equiv \frac{\|E\|_p}{\|A\|_p}$
2. Solve $S_1 y = b + r_1$ where $\epsilon_1 \equiv \frac{\|r_1\|_p}{\|S_1\|_p \|y\|_p}$
3. Solve $S_2 z = y + r_2$ where $\epsilon_2 \equiv \frac{\|r_2\|_p}{\|S_1\|_p \|z\|_p}$

Splits backward error into 3 major steps

Perturbation bound for numerical stability

Model:

$$\begin{aligned} A + E &= S_1 S_2 \qquad \epsilon_A \equiv \frac{\|E\|_p}{\|A\|_p} \\ S_1 y &= b + r_1 \qquad \epsilon_1 \equiv \frac{\|r_1\|_p}{\|S_1\|_p \|y\|_p} \\ S_2 z &= y + r_2 \qquad \epsilon_2 \equiv \frac{\|r_2\|_p}{\|S_1\|_p \|z\|_p} \end{aligned}$$

Perturbation bound for computed solution z:

$$\frac{\|z - x\|_p}{\|z\|_p} \leq \kappa_p(A) \frac{\|r\|_p}{\|A\|_p \|z\|_p}$$

Stability of direct method:

$$\frac{\|r\|_p}{\|A\|_p \|z\|_p} \leq \epsilon_A + \underbrace{\frac{\|S_1\|_p \|S_2\|_p}{\|A\|_p}}_{\text{Stability Factor}} (\epsilon_2 + \epsilon_1 (1 + \epsilon_2))$$

Easy derivation of numerical stability bound

1. Determine residual

$$r = Az - b = -Ez + r_1 + S_1 r_2$$

Follows from

$$\begin{aligned}
\underbrace{(A+E)}_{\text{Factorization}} z &= S_1 S_2 z \\
&= S_1 \underbrace{S_2 z}_{\text{2. system}} = S_1(y + r_2) = S_1 y + S_1 r_2 \\
&= \underbrace{S_1 y}_{\text{1. system}} + S_1 r_2 = b + r_1 + S_1 r_2
\end{aligned}$$

Easy derivation of numerical stability bound

1. Determine residual

$$r = Az - b = -Ez + r_1 + S_1 r_2$$

Follows from

$$\begin{aligned}
\underbrace{(A+E)}_{\text{Factorization}}\, z &= S_1 S_2\, z \\
&= S_1 \underbrace{S_2 z}_{\text{2. system}} = S_1(y + r_2) = S_1 y + S_1 r_2 \\
&= \underbrace{S_1 y}_{\text{1. system}} + S_1 r_2 = b + r_1 + S_1 r_2
\end{aligned}$$

2. Bound relative residual norm

$$\frac{\|r\|_p}{\|A\|_p \|z\|_p} \leq \epsilon_A + \underbrace{\frac{\|S_1\|_p \|S_2\|_p}{\|A\|_p}}_{\text{Stability Factor}} (\epsilon_2 + \epsilon_1(1+\epsilon))$$

Stability factors for popular direct methods

- Gaussian elimination with partial pivoting $A = (P^T L)U$

$$\frac{\|P^T L\|_\infty \|U\|_\infty}{\|A\|_\infty} \leq n \frac{\|U\|_\infty}{\|A\|_\infty}$$

- Cholesky decomposition (for spd A) $A = LL^T$

$$\frac{\|L\|_2 \|L^T\|_2}{\|A\|_2} = 1$$

- QR decomposition $A = QR$

$$\frac{\|Q\|_2 \|R\|_2}{\|A\|_2} = 1$$

Example: Stability factor captures growth

$n = 4$:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{pmatrix}}_{P^T L} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}}_{U}$$

Traditional growth factor (from roundoff error analysis)

$$\rho_n \equiv \frac{\max_{i,j,k} |A_{i,j}^{(k)}|}{\max_{i,j} |A_{ij}|} = 2^{n-1}$$

Our stability factor (from perturbation bound)

$$\frac{\|P^T L\|_\infty \|U\|_\infty}{\|A\|_\infty} = \frac{n}{n}\, 2^{n-1} = \rho_n$$

Stability factor equal to growth factor

Summary

Solving systems of linear equations $Ax = b$

Contribution: Easy and intuitive perturbation bound for numerical stability of direct methods $A = S_1 S_2$

- Model: Splits backward error into 3 major steps
  (factorization $A = S_1 S_2$, solution of systems with S_1 and S_2)
- Individual backward errors amplified by stability factor

$$\|S_1\|_p \|S_2\|_p / \|A\|_p$$

- Captures instability due to element growth
- General: Applies to any factorization, in any p-norm

Ilse C.F. Ipsen: *Numerical Matrix Analysis*, SIAM, 2009