MA423 Matrix Computations

Lectures 12&13: Stability Analysis of Gaussian Elimination

Rafikul Alam
Department of Mathematics
IIT Guwahati

Outline

- Stability analysis of GEPP/GECP
- Accuracy of computed solutions

GENP is unreliable and unstable

Consider $A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$ and $b := \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Note that $\mathrm{cond}_\infty(A) \approx 4$.

GENP is unreliable and unstable

Consider $A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$ and $b := \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Note that $\mathrm{cond}_\infty(A) \approx 4$.

In 3-decimal digit floating point arithmetic, the correct answer to 3-decimal places is $x = [1, 1]^t$.

GENP is unreliable and unstable

Consider $A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$ and $b := \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Note that $\mathrm{cond}_\infty(A) \approx 4$.

In 3-decimal digit floating point arithmetic, the correct answer to 3-decimal places is $x = [1, 1]^t$.

LU decomposition of A gives:

$$L = \begin{bmatrix} 1 & 0 \\ \mathrm{fl}(10^4) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix}$$

... no roundoff error.

GENP is unreliable and unstable

Consider $A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$ and $b := \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Note that $\mathrm{cond}_\infty(A) \approx 4$.

In 3-decimal digit floating point arithmetic, the correct answer to 3-decimal places is $x = [1, 1]^t$.

LU decomposition of A gives:

$$L = \begin{bmatrix} 1 & 0 \\ \mathrm{fl}(10^4) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix}$$

... no roundoff error.

$$U = \begin{bmatrix} 10^{-4} & 1 \\ 0 & \mathrm{fl}(1 - 10^4 * 1) \end{bmatrix} = \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix}$$

... roundoff error in the 4th place.

GENP is unreliable and unstable

Consider $A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$ and $b := \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Note that $\mathrm{cond}_\infty(A) \approx 4$.

In 3-decimal digit floating point arithmetic, the correct answer to 3-decimal places is $x = [1, 1]^t$.

LU decomposition of A gives:

$$L = \begin{bmatrix} 1 & 0 \\ \mathrm{fl}(10^4) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix}$$

... no roundoff error.

$$U = \begin{bmatrix} 10^{-4} & 1 \\ 0 & \mathrm{fl}(1 - 10^4 * 1) \end{bmatrix} = \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix}$$

... roundoff error in the 4th place. We get same LU for $\mathrm{fl}(a_{22} - 10^4) = -10^4$.

GENP is unreliable and unstable (cont.)

Now

$$L * U = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix} \text{ whereas } A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$$

The (2,2) entry is completely wrong.

GENP is unreliable and unstable (cont.)

Now

$$L * U = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix} \text{ whereas } A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$$

The (2,2) entry is completely wrong. Next, solving $Ly = b$ and $Ux = y$, we obtain

$$y = \begin{bmatrix} 1 \\ \mathrm{fl}(2 - 10^4 * 1) \end{bmatrix} = \begin{bmatrix} 1 \\ -10^4 \end{bmatrix}, \text{ the value 2 has been lost}$$

... rounding error in the 4th place, and

GENP is unreliable and unstable (cont.)

Now

$$L * U = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix} \text{ whereas } A := \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$$

The (2,2) entry is completely wrong. Next, solving $Ly = b$ and $Ux = y$, we obtain

$$y = \begin{bmatrix} 1 \\ \mathrm{fl}(2 - 10^4 * 1) \end{bmatrix} = \begin{bmatrix} 1 \\ -10^4 \end{bmatrix}, \text{ the value 2 has been lost}$$

... rounding error in the 4th place, and

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \mathrm{fl}((1 - x_2 * 1)/10^{-4}) \\ \mathrm{fl}(-10^4/(-10^4)) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

... no rounding error committed in either component.

Need for pivoting

There were only two floating errors in the 4th decimal place.

Need for pivoting

There were only two floating errors in the 4th decimal place. But the computed solution is completely wrong.

Need for pivoting

There were only two floating errors in the 4th decimal place. But the computed solution is completely wrong.

This phenomenon is called numerical instability, and must be eliminated to yield a reliable algorithm.

Need for pivoting

There were only two floating errors in the 4th decimal place. But the computed solution is completely wrong.

This phenomenon is called numerical instability, and must be eliminated to yield a reliable algorithm.

Partial pivoting is a standard remedy for this problem.

Need for pivoting

There were only two floating errors in the 4th decimal place. But the computed solution is completely wrong.

This phenomenon is called numerical instability, and must be eliminated to yield a reliable algorithm.

Partial pivoting is a standard remedy for this problem.

Applying GEPP to A, we obtain

$$L = \begin{bmatrix} 1 & 0 \\ \mathrm{fl}(.0001/1) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ .0001 & 1 \end{bmatrix}$$

and

$$U = \begin{bmatrix} 1 & 1 \\ 0 & \mathrm{fl}(1 - .0001 * 1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

so that $L * U$ approximates PA (row interchanged) quite accurately.

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular.

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular. The pivot growth

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}} \leq 2^{n-1}$$

plays an important role in the accuracy of computed solution.

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular. The pivot growth

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}} \leq 2^{n-1}$$

plays an important role in the accuracy of computed solution.

The partial pivoting guarantees that $\max_{ij} |L(i,j)| = 1$ and $\max_{ij} |U(i,j)| \leq 2^{n-1} \max_{ij} |A(i,j)|$.

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular. The pivot growth

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij}|U(i,j)|}{\max_{ij}|A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}} \leq 2^{n-1}$$

plays an important role in the accuracy of computed solution.

The partial pivoting guarantees that $\max_{ij}|L(i,j)| = 1$ and $\max_{ij}|U(i,j)| \leq 2^{n-1}\max_{ij}|A(i,j)|$. Indeed, we have $A \longrightarrow A^{(1)} \longrightarrow \cdots \longrightarrow A^{(n-1)} = U$.

$$A^{(1)} = \left[\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array}\right], \text{ where } a_{ij}^{(1)} = a_{ij} - \ell_{i1}a_{1j}.$$

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular. The pivot growth

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}} \leq 2^{n-1}$$

plays an important role in the accuracy of computed solution.

The partial pivoting guarantees that $\max_{ij} |L(i,j)| = 1$ and $\max_{ij} |U(i,j)| \leq 2^{n-1} \max_{ij} |A(i,j)|$. Indeed, we have $A \longrightarrow A^{(1)} \longrightarrow \cdots \longrightarrow A^{(n-1)} = U$.

$$A^{(1)} = \left[\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array}\right], \text{ where } a_{ij}^{(1)} = a_{ij} - \ell_{i1} a_{1j}.$$

Now $|a_{ij}^{(1)}| \leq |a_{ij}| + |a_{1j}| \leq 2\|A\|_{\max} \Longrightarrow \|A^{(1)}\|_{\max} \leq 2\|A\|_{\max}$.

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular. The pivot growth

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}} \leq 2^{n-1}$$

plays an important role in the accuracy of computed solution.

The partial pivoting guarantees that $\max_{ij} |L(i,j)| = 1$ and $\max_{ij} |U(i,j)| \leq 2^{n-1} \max_{ij} |A(i,j)|$. Indeed, we have $A \longrightarrow A^{(1)} \longrightarrow \cdots \longrightarrow A^{(n-1)} = U$.

$$A^{(1)} = \left[\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array}\right], \text{ where } a_{ij}^{(1)} = a_{ij} - \ell_{i1} a_{1j}.$$

Now $|a_{ij}^{(1)}| \leq |a_{ij}| + |a_{1j}| \leq 2\|A\|_{\max} \Longrightarrow \|A^{(1)}\|_{\max} \leq 2\|A\|_{\max}$. Similarly, $a_{ij}^{(2)} = a_{ij}^{(1)} - \ell_{i1} a_{1j}^{(1)}$ yields $\|A^{(2)}\|_{\max} \leq 2\|A^{(1)}\|_{\max} \leq 2^2\|A\|_{\max}$.

Pivot growth for GEPP

GEPP computes $PA = LU$, where P is a permutation matrix, L unit lower triangular and U is upper triangular. The pivot growth

$$g_{\rm pp}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}} \leq 2^{n-1}$$

plays an important role in the accuracy of computed solution.

The partial pivoting guarantees that $\max_{ij} |L(i,j)| = 1$ and $\max_{ij} |U(i,j)| \leq 2^{n-1} \max_{ij} |A(i,j)|$. Indeed, we have $A \longrightarrow A^{(1)} \longrightarrow \cdots \longrightarrow A^{(n-1)} = U$.

$$A^{(1)} = \left[\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array}\right], \text{ where } a_{ij}^{(1)} = a_{ij} - \ell_{i1} a_{1j}.$$

Now $|a_{ij}^{(1)}| \leq |a_{ij}| + |a_{1j}| \leq 2\|A\|_{\max} \Longrightarrow \|A^{(1)}\|_{\max} \leq 2\|A\|_{\max}$. Similarly, $a_{ij}^{(2)} = a_{ij}^{(1)} - \ell_{i1} a_{1j}^{(1)}$ yields $\|A^{(2)}\|_{\max} \leq 2\|A^{(1)}\|_{\max} \leq 2^2\|A\|_{\max}$. Repeating this process, $\|A^{(n-1)}\|_{\max} \leq 2^{n-1}\|A\|_{\max}$.

Wilkinson's result (1961)

Theorem: Suppose we solve $Ax = b$ using GEPP in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\hat{x}$ be the computed solution. Then

Wilkinson's result (1961)

Theorem: Suppose we solve $Ax = b$ using GEPP in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\hat{x}$ be the computed solution. Then

$$(A + \Delta A)\hat{x} = b, \ \|\Delta A\|_\infty \leq 2n^3 g_{\mathrm{pp}}(A)\|A\|_\infty \mathbf{u}$$

where $g_{\mathrm{pp}}(A)$ is the pivot growth given by

Wilkinson's result (1961)

Theorem: Suppose we solve $Ax = b$ using GEPP in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\hat{x}$ be the computed solution. Then

$$(A + \Delta A)\hat{x} = b, \ \|\Delta A\|_\infty \leq 2n^3 g_{\mathrm{pp}}(A)\|A\|_\infty \mathbf{u}$$

where $g_{\mathrm{pp}}(A)$ is the pivot growth given by

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}}$$

Thus, $\|x - \hat{x}\|_\infty / \|x\|_\infty \lesssim 2n^3 g_{\mathrm{pp}}(A)\mathrm{cond}_\infty(A)\mathbf{u}$.

Wilkinson's result (1961)

Theorem: Suppose we solve $Ax = b$ using GEPP in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\hat{x}$ be the computed solution. Then

$$(A + \Delta A)\hat{x} = b, \; \|\Delta A\|_\infty \leq 2n^3 g_{\mathrm{pp}}(A)\|A\|_\infty \mathbf{u}$$

where $g_{\mathrm{pp}}(A)$ is the pivot growth given by

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}}$$

Thus, $\|x - \hat{x}\|_\infty / \|x\|_\infty \lesssim 2n^3 g_{\mathrm{pp}}(A)\mathrm{cond}_\infty(A)\mathbf{u}$.

- Elegant way of accounting for rounding errors. Bounds backward error rather than the error.

Wilkinson's result (1961)

Theorem: Suppose we solve $Ax = b$ using GEPP in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\hat{x}$ be the computed solution. Then

$$(A + \Delta A)\hat{x} = b, \ \|\Delta A\|_\infty \leq 2n^3 g_{\rm pp}(A)\|A\|_\infty \mathbf{u}$$

where $g_{\rm pp}(A)$ is the pivot growth given by

$$g_{\rm pp}(A) := \frac{\max_{ij}|U(i,j)|}{\max_{ij}|A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}}$$

Thus, $\|x - \hat{x}\|_\infty / \|x\|_\infty \lesssim 2n^3 g_{\rm pp}(A) \mathrm{cond}_\infty(A)\mathbf{u}$.

- Elegant way of accounting for rounding errors. Bounds backward error rather than the error.
- Draws attention to pivot growth factor $g_{\rm pp}$.

Wilkinson's result (1961)

Theorem: Suppose we solve $Ax = b$ using GEPP in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\hat{x}$ be the computed solution. Then

$$(A + \Delta A)\hat{x} = b, \; \|\Delta A\|_\infty \leq 2n^3 g_{\mathrm{pp}}(A)\|A\|_\infty \mathbf{u}$$

where $g_{\mathrm{pp}}(A)$ is the pivot growth given by

$$g_{\mathrm{pp}}(A) := \frac{\max_{ij} |U(i,j)|}{\max_{ij} |A(i,j)|} = \frac{\|U\|_{\max}}{\|A\|_{\max}}$$

Thus, $\|x - \hat{x}\|_\infty / \|x\|_\infty \lesssim 2n^3 g_{\mathrm{pp}}(A)\mathrm{cond}_\infty(A)\mathbf{u}$.

- Elegant way of accounting for rounding errors. Bounds backward error rather than the error.
- Draws attention to pivot growth factor g_{pp}.
- Both $g_{\mathrm{pp}}(A)$ and $\mathrm{cond}_\infty(A)$ are easy to compute after getting L and U, costing just an extra $\mathcal{O}(n^2)$ flops.

Growth factor for GEPP

What do we know about $g_{\mathrm{pp}}(A)$?

Growth factor for GEPP

What do we know about $g_{\mathrm{pp}}(A)$?

Wilkinson (1954) proved that $g_{\mathrm{pp}}(A) \leq 2^{n-1}$. Usually $g_{\mathrm{pp}}(A) \simeq 1$ in practice. But examples exists for which $g_{\mathrm{pp}}(A) = 2^{n-1}$.

Growth factor for GEPP

What do we know about $g_{\rm pp}(A)$?

Wilkinson (1954) proved that $g_{\rm pp}(A) \leq 2^{n-1}$. Usually $g_{\rm pp}(A) \simeq 1$ in practice. But examples exists for which $g_{\rm pp}(A) = 2^{n-1}$.

Wilkinson's matrix: 5×5 Wilkinson's matrix W is given by

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 2^2 \\ 0 & 0 & 0 & 1 & 2^3 \\ 0 & 0 & 0 & 0 & 2^4 \end{bmatrix}.$$

Note that $g_{\rm pp}(W) = 2^4$.

Growth factor for GEPP

What do we know about $g_{\rm pp}(A)$?

Wilkinson (1954) proved that $g_{\rm pp}(A) \leq 2^{n-1}$. Usually $g_{\rm pp}(A) \simeq 1$ in practice. But examples exists for which $g_{\rm pp}(A) = 2^{n-1}$.

Wilkinson's matrix: 5×5 Wilkinson's matrix W is given by

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 2^2 \\ 0 & 0 & 0 & 1 & 2^3 \\ 0 & 0 & 0 & 0 & 2^4 \end{bmatrix}.$$

Note that $g_{\rm pp}(W) = 2^4$.

For an $n \times n$ Wilkinson matrix W, we have $W = LU$ with $U(n,n) = 2^{n-1}$. Hence $g_{\rm pp}(W) = 2^{n-1}$.

Growth factor for GEPP

What do we know about $g_{\mathrm{pp}}(A)$?

Wilkinson (1954) proved that $g_{\mathrm{pp}}(A) \leq 2^{n-1}$. Usually $g_{\mathrm{pp}}(A) \simeq 1$ in practice. But examples exists for which $g_{\mathrm{pp}}(A) = 2^{n-1}$.

Wilkinson's matrix: 5×5 Wilkinson's matrix W is given by

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -1 & -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 2^2 \\ 0 & 0 & 0 & 1 & 2^3 \\ 0 & 0 & 0 & 0 & 2^4 \end{bmatrix}.$$

Note that $g_{\mathrm{pp}}(W) = 2^4$.

For an $n \times n$ Wilkinson matrix W, we have $W = LU$ with $U(n, n) = 2^{n-1}$. Hence $g_{\mathrm{pp}}(W) = 2^{n-1}$. The matrix W can be generated in MATLAB as follows

```
W = tril( 2*eye(n)-ones(n) ); W(:, n) = ones(n,1);
```

Growth factor for GEPP

An $n \times n$ matrix A is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1 : n$.

Growth factor for GEPP

An $n \times n$ matrix A is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1 : n$.

An $n \times n$ matrix A is said to be banded with bandwidth ℓ if $a_{ij} = 0$ for all $|i - j| > \ell$.

Growth factor for GEPP

An $n \times n$ matrix A is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1 : n$.

An $n \times n$ matrix A is said to be banded with bandwidth ℓ if $a_{ij} = 0$ for all $|i - j| > \ell$. For example, if $\ell = 1$ then A is tridiagonal and if $\ell = 2$ then A is pentadiagonal.

Growth factor for GEPP

An $n \times n$ matrix A is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1 : n$.

An $n \times n$ matrix A is said to be banded with bandwidth ℓ if $a_{ij} = 0$ for all $|i - j| > \ell$. For example, if $\ell = 1$ then A is tridiagonal and if $\ell = 2$ then A is pentadiagonal.

An $n \times n$ matrix A is said to be Hessenberg (i.e., upper Hessenberg form) if $a_{ij} = 0$ for $i > j + 1$.

Growth factor for GEPP

An $n \times n$ matrix A is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1 : n$.

An $n \times n$ matrix A is said to be banded with bandwidth ℓ if $a_{ij} = 0$ for all $|i - j| > \ell$. For example, if $\ell = 1$ then A is tridiagonal and if $\ell = 2$ then A is pentadiagonal.

An $n \times n$ matrix A is said to be Hessenberg (i.e., upper Hessenberg form) if $a_{ij} = 0$ for $i > j + 1$.

Special matrices:

Matrix	$g_{\rm pp}(A)$
diag. dom	2
tridiagonal	2
banded (bandwidth p)	$2^{2p-1} - (p-1)2^{p-2}$
Hessenberg	n
SPD	1

Growth factor for GECP

- Wilkinson (1961) proved

$$g_{\mathrm{cp}}(A) \leq n^{1/2}(2.3^{1/2}\cdots n^{1/2})^{1/2} \sim cn^{1/2}n^{\frac{1}{4}\log n}.$$

- Usually, in practice, $g_{\mathrm{cp}}(A) \sim 1$. Determining the largest possible value of $g_{\mathrm{cp}}(A)$ is still an open problem.

Growth factor for GECP

- Wilkinson (1961) proved

$$g_{\rm cp}(A) \leq n^{1/2}(2.3^{1/2}\cdots n^{1/2})^{1/2} \sim cn^{1/2}n^{\frac{1}{4}\log n}.$$

- Usually, in practice, $g_{\rm cp}(A) \sim 1$. Determining the largest possible value of $g_{\rm cp}(A)$ is still an open problem.

Remark: There is no correlation between pivot growth of A and the condition number of A, that is, no correlation between $\mathrm{PG}(A)$ and $\mathrm{cond}(A)$. This is illustrated by Golub matrix.

```
function A = golub(n)
s = 10;
L = tril(round(s*randn(n)),-1)+eye(n);
U = triu(round(s*randn(n)),1)+eye(n);
A = L*U;
```

For $n = 10$, we have $g_{pp}(A) = 1$ and $\mathrm{cond}_\infty(A) = 2.9219 \times 10^{18}$.

Growth factor for GECP

- Wilkinson (1961) proved

$$g_{\mathrm{cp}}(A) \leq n^{1/2}(2.3^{1/2} \cdots n^{1/2})^{1/2} \sim cn^{1/2} n^{\frac{1}{4}\log n}.$$

- Usually, in practice, $g_{\mathrm{cp}}(A) \sim 1$. Determining the largest possible value of $g_{\mathrm{cp}}(A)$ is still an open problem.

Remark: There is no correlation between pivot growth of A and the condition number of A, that is, no correlation between $\mathrm{PG}(A)$ and $\mathrm{cond}(A)$. This is illustrated by Golub matrix.

```
function A = golub(n)
s = 10;
L = tril(round(s*randn(n)),-1)+eye(n);
U = triu(round(s*randn(n)),1)+eye(n);
A = L*U;
```

For $n = 10$, we have $g_{pp}(A) = 1$ and $\mathrm{cond}_\infty(A) = 2.9219 \times 10^{18}$. For Wilkinson matrix with $n = 50$, we have $g_{pp}(A) = 2^{49} = 5.6295 \times 10^{14}$ and $\mathrm{cond}(A) = 22.306$.

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Set $\hat{x}_j := x_j(1 + \delta_j)$ and $\hat{y}_j := y_j(1 + \delta_j)$ for $j = 1 : n$. Define

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Set $\hat{x}_j := x_j(1 + \delta_j)$ and $\hat{y}_j := y_j(1 + \delta_j)$ for $j = 1 : n$. Define $\hat{x} := \begin{bmatrix} \hat{x}_1 & \cdots & \hat{x}_n \end{bmatrix}^\top$, $\hat{y} := \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_n \end{bmatrix}^\top$, $|x| := \begin{bmatrix} |x_1| & \cdots & |x_n| \end{bmatrix}^\top$.

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Set $\hat{x}_j := x_j(1 + \delta_j)$ and $\hat{y}_j := y_j(1 + \delta_j)$ for $j = 1 : n$. Define $\hat{x} := \begin{bmatrix} \hat{x}_1 & \cdots & \hat{x}_n \end{bmatrix}^\top$, $\hat{y} := \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_n \end{bmatrix}^\top$, $|x| := \begin{bmatrix} |x_1| & \cdots & |x_n| \end{bmatrix}^\top$.

Define $x \leq y$ if $x_j \leq y_j$ for $j = 1 : n$. Then $\mathrm{fl}(y^\top x) = y^\top \hat{x} = \hat{y}^\top x$

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Set $\hat{x}_j := x_j(1 + \delta_j)$ and $\hat{y}_j := y_j(1 + \delta_j)$ for $j = 1 : n$. Define $\hat{x} := \begin{bmatrix} \hat{x}_1 & \cdots & \hat{x}_n \end{bmatrix}^\top$, $\hat{y} := \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_n \end{bmatrix}^\top$, $|x| := \begin{bmatrix} |x_1| & \cdots & |x_n| \end{bmatrix}^\top$.

Define $x \leq y$ if $x_j \leq y_j$ for $j = 1 : n$. Then $\mathrm{fl}(y^\top x) = y^\top \hat{x} = \hat{y}^\top x$ with $|x - \hat{x}| \lesssim n\mathbf{u}\,|x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\,|y|$. Hence ALG is (backward) stable.

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Set $\hat{x}_j := x_j(1 + \delta_j)$ and $\hat{y}_j := y_j(1 + \delta_j)$ for $j = 1 : n$. Define $\hat{x} := \begin{bmatrix} \hat{x}_1 & \cdots & \hat{x}_n \end{bmatrix}^\top$, $\hat{y} := \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_n \end{bmatrix}^\top$, $|x| := \begin{bmatrix} |x_1| & \cdots & |x_n| \end{bmatrix}^\top$.

Define $x \leq y$ if $x_j \leq y_j$ for $j = 1 : n$. Then $\mathrm{fl}(y^\top x) = y^\top \hat{x} = \hat{y}^\top x$ with $|x - \hat{x}| \lesssim n\mathbf{u}\, |x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\, |y|$. Hence ALG is (backward) stable.

Further, we have

$$\frac{|y^\top x - \mathrm{fl}(y^\top x)|}{|y|^\top |x|} \lesssim n\mathbf{u}.$$

Stability of inner product

Let $x := \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ and $y := \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top$, where x_j and y_j are floating-point numbers in $F(\beta, t, e_{\min}, e_{\max})$. Then

$$\mathrm{ALG}(x, y) := \mathrm{fl}(y^\top x) = \mathrm{fl}(\sum_{j=1}^{n} x_j y_j) = \sum_{j=1}^{n} x_j y_j (1 + \delta_j),$$

where $|\delta_j| \leq (n - j + 2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2)$, that is, $|\delta_j| \lesssim (n - j + 2)\mathbf{u}$ for $j = 1 : n$.

Set $\hat{x}_j := x_j(1 + \delta_j)$ and $\hat{y}_j := y_j(1 + \delta_j)$ for $j = 1 : n$. Define $\hat{x} := \begin{bmatrix} \hat{x}_1 & \cdots & \hat{x}_n \end{bmatrix}^\top$, $\hat{y} := \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_n \end{bmatrix}^\top$, $|x| := \begin{bmatrix} |x_1| & \cdots & |x_n| \end{bmatrix}^\top$.

Define $x \leq y$ if $x_j \leq y_j$ for $j = 1 : n$. Then $\mathrm{fl}(y^\top x) = y^\top \hat{x} = \hat{y}^\top x$ with $|x - \hat{x}| \lesssim n\mathbf{u}\, |x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\, |y|$. Hence ALG is (backward) stable.

Further, we have

$$\frac{|y^\top x - \mathrm{fl}(y^\top x)|}{|y|^\top |x|} \lesssim n\mathbf{u}.$$

This shows that if all entries of x (resp., y) have the same sign then the computed inner product is accurate.

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1 \colon n$
$\quad s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Proof

$\mathrm{ALG}(x, y)$ is given by

$$
\begin{array}{l}
s_0 = 0 \\
\text{for } j = 1\colon n \\
\qquad s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j)) \\
\text{end}
\end{array}
$$

Then

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1 \colon n$
  $s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$s_1 \quad = \quad \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1))$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1 \colon n$
$\quad s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$s_1 \quad = \quad \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1 + \epsilon_1)$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1{:}\, n$
  $s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$\begin{aligned} s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1+\epsilon_1) \\ &= x_1 y_1 (1+\eta_1)(1+\epsilon_1) \end{aligned}$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1 \colon n$
$\quad s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$\begin{aligned} s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1 + \epsilon_1) \\ &= x_1 y_1 (1 + \eta_1)(1 + \epsilon_1) \\ s_2 &= \mathrm{fl}(s_1 + \mathrm{fl}(x_2 y_2)) \end{aligned}$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1\colon n$
  $s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$\begin{aligned} s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1+\epsilon_1) \\ &= x_1 y_1 (1+\eta_1)(1+\epsilon_1) \\ s_2 &= \mathrm{fl}(s_1 + \mathrm{fl}(x_2 y_2)) = (s_1 + \mathrm{fl}(x_2 y_2))(1+\epsilon_2) \end{aligned}$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1 \colon n$
  $s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$\begin{aligned}
s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1 + \epsilon_1) \\
&= x_1 y_1 (1 + \eta_1)(1 + \epsilon_1) \\
s_2 &= \mathrm{fl}(s_1 + \mathrm{fl}(x_2 y_2)) = (s_1 + \mathrm{fl}(x_2 y_2))(1 + \epsilon_2) \\
&= (s_1 + x_2 y_2 (1 + \eta_2))(1 + \epsilon_2)
\end{aligned}$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1\colon n$
$\quad s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$
\begin{aligned}
s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1+\epsilon_1) \\
&= x_1 y_1 (1+\eta_1)(1+\epsilon_1) \\
s_2 &= \mathrm{fl}(s_1 + \mathrm{fl}(x_2 y_2)) = (s_1 + \mathrm{fl}(x_2 y_2))(1+\epsilon_2) \\
&= (s_1 + x_2 y_2 (1+\eta_2))(1+\epsilon_2) \\
&= x_1 y_1 (1+\eta_1)(1+\epsilon_1)(1+\epsilon_2) + x_2 y_2 (1+\eta_2)(1+\epsilon_2)
\end{aligned}
$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1 \colon n$
$\quad s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$
\begin{aligned}
s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1+\epsilon_1) \\
&= x_1 y_1 (1+\eta_1)(1+\epsilon_1) \\
s_2 &= \mathrm{fl}(s_1 + \mathrm{fl}(x_2 y_2)) = (s_1 + \mathrm{fl}(x_2 y_2))(1+\epsilon_2) \\
&= (s_1 + x_2 y_2 (1+\eta_2))(1+\epsilon_2) \\
&= x_1 y_1 (1+\eta_1)(1+\epsilon_1)(1+\epsilon_2) + x_2 y_2 (1+\eta_2)(1+\epsilon_2) \\
s_n &= x_1 y_1 (1+\eta_1) \prod_{j=1}^{n} (1+\epsilon_j) + x_2 y_2 (1+\eta_2) \prod_{j=2}^{n} (1+\epsilon_j) + \cdots \\
&\quad + x_n y_n (1+\eta_n)(1+\epsilon_n)
\end{aligned}
$$

Proof

$\mathrm{ALG}(x, y)$ is given by

$s_0 = 0$
for $j = 1\colon n$
  $s_j = \mathrm{fl}(s_{j-1} + \mathrm{fl}(x_j y_j))$
end

Then

$$
\begin{aligned}
s_1 &= \mathrm{fl}(s_0 + \mathrm{fl}(x_1 y_1)) = \mathrm{fl}(x_1 y_1)(1+\epsilon_1) \\
&= x_1 y_1 (1+\eta_1)(1+\epsilon_1) \\
s_2 &= \mathrm{fl}(s_1 + \mathrm{fl}(x_2 y_2)) = (s_1 + \mathrm{fl}(x_2 y_2))(1+\epsilon_2) \\
&= (s_1 + x_2 y_2 (1+\eta_2))(1+\epsilon_2) \\
&= x_1 y_1 (1+\eta_1)(1+\epsilon_1)(1+\epsilon_2) + x_2 y_2 (1+\eta_2)(1+\epsilon_2) \\
s_n &= x_1 y_1 (1+\eta_1) \prod_{j=1}^{n} (1+\epsilon_j) + x_2 y_2 (1+\eta_2) \prod_{j=2}^{n} (1+\epsilon_j) + \cdots \\
&\quad + x_n y_n (1+\eta_n)(1+\epsilon_n) = x_1 y_1 (1+\delta_1) + \cdots + x_n y_n (1+\delta_n)
\end{aligned}
$$

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$|\delta_j| \quad = \quad \left|(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1\right|$$

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$|\delta_j| \quad = \quad |(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1| \le (1+\mathbf{u})^{n-j+2} - 1$$

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$\begin{aligned} |\delta_j| &= \left|(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1\right| \leq (1+\mathbf{u})^{n-j+2} - 1 \\ &\leq (n-j+2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2) \;\text{ for } j = 1:n. \end{aligned}$$

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$\begin{aligned} |\delta_j| &= \left|(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1\right| \leq (1+\mathbf{u})^{n-j+2} - 1 \\ &\leq (n-j+2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2) \text{ for } j = 1:n. \end{aligned}$$

It is immediate that $|x - \hat{x}| \lesssim n\mathbf{u}\,|x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\,|y|$.

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$\begin{aligned} |\delta_j| &= \left|(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1\right| \leq (1+\mathbf{u})^{n-j+2} - 1 \\ &\leq (n-j+2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2) \;\text{ for } j = 1:n. \end{aligned}$$

It is immediate that $|x - \hat{x}| \lesssim n\mathbf{u}\,|x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\,|y|$. Further,

$$\mathrm{fl}(y^\top x) = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n) = y^\top x + x_1 y_1 \delta_1 + \cdots + x_n y_n \delta_n$$

Proof

Thus we have $s_n = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$\begin{aligned} |\delta_j| &= \left|(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1\right| \leq (1+\mathbf{u})^{n-j+2} - 1 \\ &\leq (n-j+2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2) \quad \text{for } j = 1:n. \end{aligned}$$

It is immediate that $|x - \hat{x}| \lesssim n\mathbf{u}\,|x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\,|y|$. Further,

$\mathrm{fl}(y^\top x) = x_1 y_1(1+\delta_1) + \cdots + x_n y_n(1+\delta_n) = y^\top x + x_1 y_1 \delta_1 + \cdots + x_n y_n \delta_n$ yields

Proof

Thus we have $s_n = x_1y_1(1+\delta_1) + \cdots + x_ny_n(1+\delta_n)$, where

$$1+\delta_j = (1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) \text{ for } j = 1:n.$$

This shows that

$$\begin{aligned} |\delta_j| &= \left|(1+\eta_j)\prod_{k=j}^{n}(1+\epsilon_k) - 1\right| \leq (1+\mathbf{u})^{n-j+2} - 1 \\ &\leq (n-j+2)\mathbf{u} + \mathcal{O}(\mathbf{u}^2) \ \text{ for } j = 1:n. \end{aligned}$$

It is immediate that $|x - \hat{x}| \lesssim n\mathbf{u}\,|x|$ and $|y - \hat{y}| \lesssim n\mathbf{u}\,|y|$. Further,

$\mathrm{fl}(y^\top x) = x_1y_1(1+\delta_1) + \cdots + x_ny_n(1+\delta_n) = y^\top x + x_1y_1\delta_1 + \cdots + x_ny_n\delta_n$ yields

$$\frac{|y^\top x - \mathrm{fl}(y^\top x)|}{|y|^\top |x|} \lesssim n\mathbf{u}. \blacksquare$$

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j.

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Theorem: Suppose that `[L, U] = GE(A)`, where `GE` $\in$ {`GENP, GEPP, GECP`}.

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Theorem: Suppose that `[L, U] = GE(A)`, where $\texttt{GE} \in \{\texttt{GENP}, \texttt{GEPP}, \texttt{GECP}\}$.
Then

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Theorem: Suppose that `[L, U] = GE(A)`, where `GE` $\in$ {`GENP, GEPP, GECP`}. Then

$$A + E = L \cdot U, \text{ where } |E| \lesssim |L| \cdot |U| n\mathbf{u}.$$

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Theorem: Suppose that `[L, U] = GE(A)`, where $\texttt{GE} \in \{\texttt{GENP, GEPP, GECP}\}$.
Then

$$A + E = L \cdot U, \text{ where } |E| \lesssim |L| \cdot |U| n\mathbf{u}.$$

Hence $\|E\| \lesssim \|L\| \cdot \|U\| n\mathbf{u}$. ■

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Theorem: Suppose that `[L, U] = GE(A)`, where `GE` $\in$ {`GENP, GEPP, GECP`}. Then

$$A + E = L \cdot U, \text{ where } |E| \lesssim |L| \cdot |U| n\mathbf{u}.$$

Hence $\|E\| \lesssim \|L\| \cdot \|U\| n\mathbf{u}$. ■

Define the pivot growth

$$\mathrm{PG}(A) := \|L\| \cdot \|U\| / \|A\|.$$

Stability of LU factorization

For $n \times n$ matrices A and B, write $A \leq B$ when $a_{ij} \leq b_{ij}$ for all i and j. Also define $|A| := [|a_{ij}|]$.

Theorem: Suppose that `[L, U] = GE(A)`, where $\texttt{GE} \in \{\texttt{GENP, GEPP, GECP}\}$. Then

$$A + E = L \cdot U, \text{ where } |E| \lesssim |L| \cdot |U| n\mathbf{u}.$$

Hence $\|E\| \lesssim \|L\| \cdot \|U\| n\mathbf{u}$. ■

Define the pivot growth

$$\mathrm{PG}(A) := \|L\| \cdot \|U\| / \|A\|.$$

Then $A + E = L \cdot U$ and

$$\|E\| / \|A\| \lesssim \mathrm{PG}(A) n\mathbf{u}.$$

Stability of triangular solver

Theorem:

Let $\hat{y}$ and $\hat{x}$ be computed solutions of $Ly = b$ and $Ux = \hat{y}$. Then

Stability of triangular solver

Theorem:
Let $\hat{y}$ and $\hat{x}$ be computed solutions of $Ly = b$ and $Ux = \hat{y}$. Then

$$(L + \Delta L)\hat{y} = b \text{ and } (U + \Delta U)\hat{x} = \hat{y}$$

with $|\Delta L| \lesssim |L| n \mathbf{u}$ and $|\Delta U| \lesssim |U| n \mathbf{u}$. ■

Stability of triangular solver

Theorem:
Let $\hat{y}$ and $\hat{x}$ be computed solutions of $Ly = b$ and $Ux = \hat{y}$. Then

$$(L + \Delta L)\hat{y} = b \text{ and } (U + \Delta U)\hat{x} = \hat{y}$$

with $|\Delta L| \lesssim |L| n\mathbf{u}$ and $|\Delta U| \lesssim |U| n\mathbf{u}$. ■

Putting these results together, we have

$$\begin{aligned} b &= (L + \Delta L)(U + \Delta U)\hat{x} \\ &= (LU + L\Delta U + \Delta L U + \Delta L \Delta U)\hat{x} \\ &= (A + E + L\Delta U + \Delta L U + \Delta L \Delta U)\hat{x} = (A + \Delta A)\hat{x} \end{aligned}$$

Stability of triangular solver

Theorem:
Let $\hat{y}$ and $\hat{x}$ be computed solutions of $Ly = b$ and $Ux = \hat{y}$. Then

$$(L + \Delta L)\hat{y} = b \text{ and } (U + \Delta U)\hat{x} = \hat{y}$$

with $|\Delta L| \lesssim |L|n\mathbf{u}$ and $|\Delta U| \lesssim |U|n\mathbf{u}$. ■

Putting these results together, we have

$$\begin{aligned} b &= (L + \Delta L)(U + \Delta U)\hat{x} \\ &= (LU + L\Delta U + \Delta L U + \Delta L \Delta U)\hat{x} \\ &= (A + E + L\Delta U + \Delta L U + \Delta L \Delta U)\hat{x} = (A + \Delta A)\hat{x} \end{aligned}$$

This gives

$$\begin{aligned} |\Delta A| &\leq |E| + |L| \cdot |\Delta U| + |\Delta L| \cdot |U| + |\Delta L| \cdot |\Delta U| \\ &\lesssim n\mathbf{u}|L| \cdot |U| + n\mathbf{u}|L| \cdot |U| + n\mathbf{u}|L| \cdot |U| = 3n\mathbf{u}|L| \cdot |U| \end{aligned}$$

Stability analysis of GE

Thus $(A + \Delta A)\hat{x} = b$ and $|\Delta A| \lesssim 3n\mathbf{u}|L| \cdot |U|$.
Taking norm, we have

$$\|\Delta A\|/\|A\| \lesssim 3n\mathbf{u}\mathrm{PG}(A).$$

Stability analysis of GE

Thus $(A + \Delta A)\hat{x} = b$ and $|\Delta A| \lesssim 3n\mathbf{u}|L| \cdot |U|$.
Taking norm, we have

$$\|\Delta A\|/\|A\| \lesssim 3n\mathbf{u}\mathrm{PG}(A).$$

For GEPP and GECP, $\|L\|_\infty \leq n$ and $\|U\|_\infty \leq n\|U\|_{\max}$. Hence

$$\mathrm{PG}(A) = \frac{\|L\|_\infty \|U\|_\infty}{\|A\|_\infty} \leq n^2 \frac{\|U\|_{\max}}{\|A\|_{\max}}.$$

Stability analysis of GE

Thus $(A + \Delta A)\hat{x} = b$ and $|\Delta A| \lesssim 3n\mathbf{u}|L| \cdot |U|$.
Taking norm, we have

$$\|\Delta A\|/\|A\| \lesssim 3n\mathbf{u}\mathrm{PG}(A).$$

For GEPP and GECP, $\|L\|_\infty \leq n$ and $\|U\|_\infty \leq n\|U\|_{\mathsf{max}}$. Hence

$$\mathrm{PG}(A) = \frac{\|L\|_\infty \|U\|_\infty}{\|A\|_\infty} \leq n^2 \frac{\|U\|_{\mathsf{max}}}{\|A\|_{\mathsf{max}}}.$$

- For GEPP, we have $\mathrm{PG}(A) \leq n^2 g_{\mathrm{pp}}(A)$.
- For GECP, we have $\mathrm{PG}(A) \leq n^2 g_{\mathrm{cp}}(A)$.

Stability analysis of GE

Theorem: Suppose we solve $Ax = b$ using GE (GENP, GEPP, GECP) and in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\mathrm{PG}(A) := \|L\| \cdot \|U\|/\|A\|$. Then

$$A + E = LU \text{ and } \|E\|/\|A\| \lesssim \mathrm{PG}(A)n\mathbf{u}.$$

Computed solution $\hat{x}$ satisfies $(A + \Delta A)\hat{x} = b$ and

$$\|\Delta A\|/\|A\| \lesssim 3\mathrm{PG}(A)n\mathbf{u}.$$

Stability analysis of GE

Theorem: Suppose we solve $Ax = b$ using GE (GENP, GEPP, GECP) and in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\mathrm{PG}(A) := \|L\| \cdot \|U\|/\|A\|$. Then

$$A + E = LU \text{ and } \|E\|/\|A\| \lesssim \mathrm{PG}(A) n\mathbf{u}.$$

Computed solution $\hat{x}$ satisfies $(A + \Delta A)\hat{x} = b$ and

$$\|\Delta A\|/\|A\| \lesssim 3\mathrm{PG}(A) n\mathbf{u}.$$

If the pivot growth $\mathrm{PG}(A)$ is not too large then $\|\Delta A\|/\|A\| = \mathcal{O}(\mathbf{u})$.

In practice, $g_{\mathrm{pp}}(A) \leq n$. The average pivot growth is like $g_{\mathrm{pp}}(A) \sim n^{2/3}$ or just $g_{\mathrm{pp}}(A) \sim n^{1/2}$. This makes GEPP algorithm of choice for most problems.

Stability analysis of GE

Theorem: Suppose we solve $Ax = b$ using GE (GENP, GEPP, GECP) and in floating point arithmetic with unit roundoff $\mathbf{u}$. Let $\mathrm{PG}(A) := \|L\| \cdot \|U\|/\|A\|$. Then

$$A + E = LU \text{ and } \|E\|/\|A\| \lesssim \mathrm{PG}(A)n\mathbf{u}.$$

Computed solution $\hat{x}$ satisfies $(A + \Delta A)\hat{x} = b$ and

$$\|\Delta A\|/\|A\| \lesssim 3\mathrm{PG}(A)n\mathbf{u}.$$

If the pivot growth $\mathrm{PG}(A)$ is not too large then $\|\Delta A\|/\|A\| = \mathcal{O}(\mathbf{u})$.

In practice, $g_{\mathrm{pp}}(A) \leq n$. The average pivot growth is like $g_{\mathrm{pp}}(A) \sim n^{2/3}$ or just $g_{\mathrm{pp}}(A) \sim n^{1/2}$. This makes GEPP algorithm of choice for most problems.

To sum up: Experience shows that GE is accurate in the sense that it is equivalent to changing the entries of A by small numbers on the order of $\|A\|\mathbf{u}$ (roundoff errors in the entries of A) and then solving this perturbed problem $(A + \delta A)\hat{x} = b$ exactly.

Pivot Growth for Cholesky factorization

Fact: Let $A = GG^\top$. Define $PG(A) := \|G\|_2 \|G^\top\|_2 / \|A\|_2$. Then $PG(A) = 1$.

Pivot Growth for Cholesky factorization

Fact: Let $A = GG^\top$. Define $PG(A) := \|G\|_2 \|G^\top\|_2 / \|A\|_2$. Then $PG(A) = 1$.

Proof: Note that $\|G^\top\|_2 = \|G\|_2$. Hence we have

$$\|G\|_2^2 = \|G^\top\|_2^2 = \lambda_{\max}(GG^\top) = \lambda_{\max}(A) = \sqrt{\lambda_{\max}(AA^\top)} = \|A\|_2.$$

Pivot Growth for Cholesky factorization

Fact: Let $A = GG^\top$. Define $PG(A) := \|G\|_2 \|G^\top\|_2 / \|A\|_2$. Then $PG(A) = 1$.

Proof: Note that $\|G^\top\|_2 = \|G\|_2$. Hence we have

$$\|G\|_2^2 = \|G^\top\|_2^2 = \lambda_{\max}(GG^\top) = \lambda_{\max}(A) = \sqrt{\lambda_{\max}(AA^\top)} = \|A\|_2.$$

This shows that $PG(A) = \|G\|_2^2 / \|A\|_2 = 1$. ■

Pivot Growth for Cholesky factorization

Fact: Let $A = GG^\top$. Define $PG(A) := \|G\|_2 \|G^\top\|_2 / \|A\|_2$. Then $PG(A) = 1$.

Proof: Note that $\|G^\top\|_2 = \|G\|_2$. Hence we have

$$\|G\|_2^2 = \|G^\top\|_2^2 = \lambda_{\max}(GG^\top) = \lambda_{\max}(A) = \sqrt{\lambda_{\max}(AA^\top)} = \|A\|_2.$$

This shows that $PG(A) = \|G\|_2^2 / \|A\|_2 = 1$. ■

Thus computation of Cholesky factroization is an unconditionally stable algorithm.

GEPP versus GECP

- GECP is more expensive ($\mathcal{O}(n^3)$ more operations) than GEPP.
- GECP is usually no more accurate than GEPP which is why GEPP is the default method for solving a linear system.

GEPP versus GECP

- GECP is more expensive ($\mathcal{O}(n^3)$ more operations) than GEPP.
- GECP is usually no more accurate than GEPP which is why GEPP is the default method for solving a linear system.
- Examples exist for which GECP does much better than GEPP.
- We still do not fully understand why GEPP and GECP work so well in the presence of roundoff errors.