Rounding Errors and Stability of Algorithms

Backward Error Analysis of Gaussian Elimination
&
Cholesky Decomposition Methods

Backward error analysis of backward and forward substitution

Some notation:

- For any two matrices $C = [c_{ij}]_{n\times m}$ and $F = [f_{ij}]_{n\times m}$, we write $C \leq F$ if $c_{ij} \leq f_{ij}$ for all $1 \leq i \leq n$, $1 \leq j \leq m$.
- For any $C = [c_{ij}]_{n\times m}$, $|C| = [|c_{ij}|]_{n\times m}$.

Backward error analysis of backward and forward substitution

Some notation:

- For any two matrices $C = [c_{ij}]_{n\times m}$ and $F = [f_{ij}]_{n\times m}$, we write $C \leq F$ if $c_{ij} \leq f_{ij}$ for all $1 \leq i \leq n$, $1 \leq j \leq m$.
- For any $C = [c_{ij}]_{n\times m}$, $|C| = [|c_{ij}|]_{n\times m}$.

Theorem Let G be any nonsingular lower (upper) triangular $n \times n$ matrix and b be any nonzero column vector of length n. If y_c be the computed solution of the system $Gy = b$ using any variant of forward (backward) substitution in floating point arithmetic, then y_c satisfies

$$(G + \delta G)y_c = b \tag{6}$$

where δG is lower (upper) triangular such that

$$|\delta G| \leq 2nu|G| + O(u^2) \tag{7}$$

Backward error analysis of backward and forward substitution

Some notation:

- For any two matrices $C = [c_{ij}]_{n\times m}$ and $F = [f_{ij}]_{n\times m}$, we write $C \leq F$ if $c_{ij} \leq f_{ij}$ for all $1 \leq i \leq n$, $1 \leq j \leq m$.
- For any $C = [c_{ij}]_{n\times m}$, $|C| = [|c_{ij}|]_{n\times m}$.

Theorem Let G be any nonsingular lower (upper) triangular $n \times n$ matrix and b be any nonzero column vector of length n. If y_c be the computed solution of the system $Gy = b$ using any variant of forward (backward) substitution in floating point arithmetic, then y_c satisfies

$$(G + \delta G)y_c = b \tag{6}$$

where δG is lower (upper) triangular such that

$$|\delta G| \leq 2nu|G| + O(u^2) \tag{7}$$

Here (7) means $|\delta g_{ij}| \leq 2nu|g_{ij}| + O(u^2)$ for all $1 \leq i, j \leq n$ and $O(u^2)$ stands for a matrix whose entries are $O(u^2)$.

Backward error analysis of backward and forward substitution

Some notation:

- For any two matrices $C = [c_{ij}]_{n \times m}$ and $F = [f_{ij}]_{n \times m}$, we write $C \leq F$ if $c_{ij} \leq f_{ij}$ for all $1 \leq i \leq n$, $1 \leq j \leq m$.
- For any $C = [c_{ij}]_{n \times m}$, $|C| = [|c_{ij}|]_{n \times m}$.

Theorem Let G be any nonsingular lower (upper) triangular $n \times n$ matrix and b be any nonzero column vector of length n. If y_c be the computed solution of the system $Gy = b$ using any variant of forward (backward) substitution in floating point arithmetic, then y_c satisfies

$$(G + \delta G)y_c = b \tag{6}$$

where δG is lower (upper) triangular such that

$$|\delta G| \leq 2nu|G| + O(u^2) \tag{7}$$

Here (7) means $|\delta g_{ij}| \leq 2nu|g_{ij}| + O(u^2)$ for all $1 \leq i, j \leq n$ and $O(u^2)$ stands for a matrix whose entries are $O(u^2)$.

For a proof see pp. 159-161 of Fundamentals of Matrix Computations, 2nd Edition.

Backward error analysis of backward and forward substitution

Exercise Given any two matrices $C = [c_{ij}]_{n\times m}$ and $F = [f_{ij}]_{n\times m}$, show that

1. $|C| \le |F| \Rightarrow \|C\|_p \le \|F\|_p$ where $p = 1, F, \infty$.
2. $\|C\|_p = \||C|\|_p$ for $p = 1, F, \infty$.

Backward error analysis of backward and forward substitution

Exercise Given any two matrices $C = [c_{ij}]_{n\times m}$ and $F = [f_{ij}]_{n\times m}$, show that

1. $|C| \leq |F| \Rightarrow \|C\|_p \leq \|F\|_p$ where $p = 1, F, \infty$.
2. $\|C\|_p = \||C|\|_p$ for $p = 1, F, \infty$.

Corollary Let G be any nonsingular lower (upper) triangular $n \times n$ matrix and b be any nonzero column vector of length n. If y_c be the computed solution of the system $Gy = b$ using any variant of forward (backward) substitution in floating point arithmetic, then y_c satisfies

$$(G + \delta G)y_c = b$$

where δG is lower (upper) triangular such that

$$\|\delta G\|_\infty \leq 2nu\|G\|_\infty + O(u^2) \tag{8}$$

Backward error analysis of backward and forward substitution

Exercise Given any two matrices $C = [c_{ij}]_{n\times m}$ and $F = [f_{ij}]_{n\times m}$, show that

1. $|C| \leq |F| \Rightarrow \|C\|_p \leq \|F\|_p$ where $p = 1, F, \infty$.
2. $\|C\|_p = \||C|\|_p$ for $p = 1, F, \infty$.

Corollary Let G be any nonsingular lower (upper) triangular $n \times n$ matrix and b be any nonzero column vector of length n. If y_c be the computed solution of the system $Gy = b$ using any variant of forward (backward) substitution in floating point arithmetic, then y_c satisfies

$$(G + \delta G)y_c = b$$

where δG is lower (upper) triangular such that

$$\|\delta G\|_\infty \leq 2nu\|G\|_\infty + O(u^2) \tag{8}$$

Therefore backward and forward substitution processes are unconditionally backward stable.

Backward error analysis of Gaussian Elimination

Theorem Let A be an $n \times n$ matrix. Let L_c and U_c be the LU factors of A computed via Gaussian Elimination in floating point arithmetic. If no zero pivots were encountered in the process, then

$$A + \delta A = L_c U_c$$

where

$$|\delta A| \leq 2nu|L_c||U_c| + O(u^2) \tag{9}$$

and consequently,

$$\|\delta A\|_\infty \leq 2nu\|L_c\|_\infty\|U_c\|_\infty + O(u^2) \tag{10}$$

Backward error analysis of Gaussian Elimination

Theorem Let A be an $n \times n$ matrix. Let L_c and U_c be the LU factors of A computed via Gaussian Elimination in floating point arithmetic. If no zero pivots were encountered in the process, then

$$A + \delta A = L_c U_c$$

where

$$|\delta A| \leq 2nu|L_c||U_c| + O(u^2) \tag{9}$$

and consequently,

$$\|\delta A\|_\infty \leq 2nu\|L_c\|_\infty\|U_c\|_\infty + O(u^2) \tag{10}$$

Further if x_c be the computed solution of $Ax = b$, $b \neq 0$, obtained by solving lower and upper triangular systems with L_c and U_c as coefficient matrices via forward and backward substitution respectively, then

$$(A + E)x_c = b$$

where

$$|E| \leq 6nu|L_c||U_c| + O(u^2) \tag{11}$$

and consequently,

$$\|E\|_\infty \leq 6nu\|L_c\|_\infty\|U_c\|_\infty + O(u^2) \tag{12}$$

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

In GEPP and GECP, $\|L_c\|_\infty \leq n$. Therefore,

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty} \leq n \frac{\|U_c\|_\infty}{\|A\|_\infty}.$$

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

In GEPP and GECP, $\|L_c\|_\infty \le n$. Therefore,

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty} \le n \frac{\|U_c\|_\infty}{\|A\|_\infty}.$$

For GEPP it is observed that for almost all matrices in practice, $\frac{\|U_c\|_\infty}{\|A\|_\infty} \approx \sqrt{n}$.

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

In GEPP and GECP, $\|L_c\|_\infty \leq n$. Therefore,

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty} \leq n \frac{\|U_c\|_\infty}{\|A\|_\infty}.$$

For GEPP it is observed that for almost all matrices in practice, $\frac{\|U_c\|_\infty}{\|A\|_\infty} \approx \sqrt{n}$.

However, $\frac{\|U_c\|_\infty}{\|A\|_\infty} = \frac{2^{n-1}}{n}$ when A is the $n \times n$ Wilkinson matrix in GEPP.

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

In GEPP and GECP, $\|L_c\|_\infty \leq n$. Therefore,

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty} \leq n \frac{\|U_c\|_\infty}{\|A\|_\infty}.$$

For GEPP it is observed that for almost all matrices in practice, $\frac{\|U_c\|_\infty}{\|A\|_\infty} \approx \sqrt{n}$.

However, $\frac{\|U_c\|_\infty}{\|A\|_\infty} = \frac{2^{n-1}}{n}$ when A is the $n \times n$ Wilkinson matrix in GEPP.

Therefore GEPP is conditionally backward stable as stabiity is subject to $\frac{\|U_c\|_\infty}{\|A\|_\infty}$ being small.

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

In GEPP and GECP, $\|L_c\|_\infty \leq n$. Therefore,

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty} \leq n \frac{\|U_c\|_\infty}{\|A\|_\infty}.$$

For GEPP it is observed that for almost all matrices in practice, $\frac{\|U_c\|_\infty}{\|A\|_\infty} \approx \sqrt{n}$.

However, $\frac{\|U_c\|_\infty}{\|A\|_\infty} = \frac{2^{n-1}}{n}$ when A is the $n \times n$ Wilkinson matrix in GEPP.

Therefore GEPP is conditionally backward stable as stabiity is subject to $\frac{\|U_c\|_\infty}{\|A\|_\infty}$ *being small.*

For GECP it can be proved that $\frac{\|U_c\|_\infty}{\|A\|_\infty} = O(n)$.

Backward error analysis of Gaussian Elimination

Therefore the process is backward stable if

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty}$$

is small.

In GENP, the presence of small pivots can make $\|L_c\|_\infty$ very large. Therefore, GENP is unstable.

In GEPP and GECP, $\|L_c\|_\infty \leq n$. Therefore,

$$\frac{\|L_c\|_\infty \|U_c\|_\infty}{\|A\|_\infty} \leq n \frac{\|U_c\|_\infty}{\|A\|_\infty}.$$

For GEPP it is observed that for almost all matrices in practice, $\frac{\|U_c\|_\infty}{\|A\|_\infty} \approx \sqrt{n}$.

However, $\frac{\|U_c\|_\infty}{\|A\|_\infty} = \frac{2^{n-1}}{n}$ when A is the $n \times n$ Wilkinson matrix in GEPP.

Therefore GEPP is conditionally backward stable as stabiity is subject to $\frac{\|U_c\|_\infty}{\|A\|_\infty}$ being small.

For GECP it can be proved that $\frac{\|U_c\|_\infty}{\|A\|_\infty} = O(n)$.

Hence GECP is unconditionally backward stable.

GEPP: The algorithm of choice for solving square systems

However despite the verdict of conditional backward stability on GEPP, it remains the algorithm of choice when it comes to direct methods for solving square system of equations that are not too large and the coefficient matrix is not positive definite and not sparse, that is most of its entries are nonzero.

This is because it is cheaper than GECP and it has excellent stability properties for almost all matrices.

Therefore the MATLAB command $\texttt{A} \setminus \texttt{b}$ runs GEPP to solve $Ax = b$ when A is not too large, not sparse and not positive definite.

A complete understanding of the reasons for the backward stability of GEPP in almost all cases is an open problem.

Backward error analysis of Cholesky factorization

Theorem Let A be an $n \times n$ positive definite matrix and G_c be the Cholesky factor computed in floating point arithmetic via some version of Cholesky factorization algorithm. Then

$$A + \delta A = G_c^T G_c$$

where $|\delta A| \leq 2nu|G_c^T||G_c| + O(u^2)$.

Backward error analysis of Cholesky factorization

Theorem Let A be an $n \times n$ positive definite matrix and G_c be the Cholesky factor computed in floating point arithmetic via some version of Cholesky factorization algorithm. Then

$$A + \delta A = G_c^T G_c$$

where $|\delta A| \leq 2nu|G_c^T||G_c| + O(u^2)$.

Consequently,

$$\frac{\|\delta A\|_F}{\|A\|_F} \leq \frac{2n^{3/2}u}{1 - 2n^{3/2}u} + O(u^2).$$

Backward error analysis of Cholesky factorization

Theorem Let A be an $n \times n$ positive definite matrix and G_c be the Cholesky factor computed in floating point arithmetic via some version of Cholesky factorization algorithm. Then

$$A + \delta A = G_c^T G_c$$

where $|\delta A| \leq 2nu|G_c^T||G_c| + O(u^2)$.

Consequently,

$$\frac{\|\delta A\|_F}{\|A\|_F} \leq \frac{2n^{3/2}u}{1 - 2n^{3/2}u} + O(u^2).$$

Thus, the algorithms for finding a Cholesky factor of a positive definite matrix and using it to solve a system of equations is unconditionally backward stable.

Backward error analysis of Cholesky factorization

Theorem Let A be an $n \times n$ positive definite matrix and G_c be the Cholesky factor computed in floating point arithmetic via some version of Cholesky factorization algorithm. Then

$$A + \delta A = G_c^T G_c$$

where $|\delta A| \leq 2nu|G_c^T||G_c| + O(u^2)$.

Consequently,

$$\frac{\|\delta A\|_F}{\|A\|_F} \leq \frac{2n^{3/2}u}{1 - 2n^{3/2}u} + O(u^2).$$

Thus, the algorithms for finding a Cholesky factor of a positive definite matrix and using it to solve a system of equations is unconditionally backward stable.

Consequently, the solutions of a system of equations with a positive definite coefficient matrix is unconditionaly backward stable.