Backward Error Analysis of GEPP

Emre Mengi

Department of Mathemtics
Koç University
Istanbul, Turkey

November 30, 2011

- View the solution $x \in \mathbb{C}^n$ of the linear system $Ax = b$ as a function of $A \in \mathbb{C}^{n \times n}$.

Sensitivity Question

How does the solution $x \in \mathbb{C}^n$ vary w.r.t. perturbations in $A \in \mathbb{C}^{n \times n}$?

- View the solution $x \in \mathbb{C}^n$ of the linear system $Ax = b$ as a function of $A \in \mathbb{C}^{n \times n}$.

Sensitivity Question

How does the solution $x \in \mathbb{C}^n$ vary w.r.t. perturbations in $A \in \mathbb{C}^{n \times n}$?

Example

Original systems

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}, \quad x(A) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

Slightly perturbed system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.05 \end{bmatrix}}_{A+\delta A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}, \quad x(A+\delta A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Sensitivity

$$\frac{\|x(A+\delta A) - x(A)\|}{\|\delta A\|} = \left\| \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \right\| / \left\| \begin{bmatrix} 0 & 0 \\ 0 & 0.05 \end{bmatrix} \right\|$$

Example

Original systems

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}, \quad x(A) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

Slightly perturbed system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.05 \end{bmatrix}}_{A+\delta A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}, \quad x(A+\delta A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Sensitivity

$$\frac{\|x(A+\delta A) - x(A)\|}{\|\delta A\|} = \left\| \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \right\| / \left\| \begin{bmatrix} 0 & 0 \\ 0 & 0.05 \end{bmatrix} \right\|$$

Example

Original systems

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.1 \end{bmatrix}}_{A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}, \quad x(A) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

Slightly perturbed system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1.05 \end{bmatrix}}_{A+\delta A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1.05 \end{bmatrix}}_{b}, \quad x(A+\delta A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Sensitivity

$$\frac{\|x(A+\delta A) - x(A)\|}{\|\delta A\|} = \left\| \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} \right\| / \left\| \begin{bmatrix} 0 & 0 \\ 0 & 0.05 \end{bmatrix} \right\|$$

Relative condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|x(A+\delta A) - x(A)\| / \|x(A)\|}{\|\delta A\| / \|A\|}$$

- Let δx be such that $x + \delta x = x(A + \delta A)$.

$$(A + \delta A)(x + \delta x) = b$$
$$\Longrightarrow$$
$$Ax + (\delta A)x + A(\delta x) + (\delta A)(\delta x) = b$$
$$\Longrightarrow$$
$$x(A + \delta A) - x(A) = \delta x = -A^{-1}(\delta A)x - A^{-1}(\delta A)(\delta x)$$

Relative condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|x(A+\delta A) - x(A)\| / \|x(A)\|}{\|\delta A\| / \|A\|}$$

- Let δx be such that $x + \delta x = x(A + \delta A)$.

$$(A + \delta A)(x + \delta x) = b$$
$$\Longrightarrow$$
$$Ax + (\delta A)x + A(\delta x) + (\delta A)(\delta x) = b$$
$$\Longrightarrow$$
$$x(A + \delta A) - x(A) = \delta x = -A^{-1}(\delta A)x - A^{-1}(\delta A)(\delta x)$$

Relative condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|x(A+\delta A) - x(A)\|/\|x(A)\|}{\|\delta A\|/\|A\|}$$

- Let δx be such that $x + \delta x = x(A + \delta A)$.

$$(A+\delta A)(x+\delta x) = b$$
$$\Longrightarrow$$
$$Ax + (\delta A)x + A(\delta x) + (\delta A)(\delta x) = b$$
$$\Longrightarrow$$
$$x(A+\delta A) - x(A) = \delta x = -A^{-1}(\delta A)x - A^{-1}(\delta A)(\delta x)$$

Relative condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|x(A+\delta A) - x(A)\| / \|x(A)\|}{\|\delta A\| / \|A\|}$$

- Let δx be such that $x + \delta x = x(A + \delta A)$.

$$(A + \delta A)(x + \delta x) = b$$
$$\Longrightarrow$$
$$Ax + (\delta A)x + A(\delta x) + (\delta A)(\delta x) = b$$
$$\Longrightarrow$$
$$x(A + \delta A) - x(A) = \delta x = -A^{-1}(\delta A)x - A^{-1}(\delta A)(\delta x)$$

Relative condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|x(A+\delta A) - x(A)\|/\|x(A)\|}{\|\delta A\|/\|A\|}$$

- Let δx be such that $x + \delta x = x(A + \delta A)$.

$$(A+\delta A)(x+\delta x) = b$$
$$\Longrightarrow$$
$$Ax + (\delta A)x + A(\delta x) + (\delta A)(\delta x) = b$$
$$\Longrightarrow$$
$$x(A+\delta A) - x(A) = \delta x = -A^{-1}(\delta A)x - A^{-1}(\delta A)(\delta x)$$

Relative condition number

$$\tilde{\kappa} = \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x + A^{-1}(\delta A)(\delta x)\|/\|x\|}{\|\delta A\|/\|A\|}$$

$$= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x\|/\|x\|}{\|\delta A\|/\|A\|} \quad (\text{since } ((\delta A)(\delta x))/\|\delta A\| \to 0 \text{ as } \delta \to 0)$$

- In general

$$\|A^{-1}(\delta A)x\| \leq \|A^{-1}\|\|\delta A\|\|x\|$$

- But for all x and A there exists δA (exercise) such that

$$\|A^{-1}(\delta A)x\| = \|A^{-1}\|\|\delta A\|\|x\|$$

Relative condition number

$$\begin{aligned}\tilde{\kappa} &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x + A^{-1}(\delta A)(\delta x)\|/\|x\|}{\|\delta A\|/\|A\|} \\ &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x\|/\|x\|}{\|\delta A\|/\|A\|} \quad (\text{since } ((\delta A)(\delta x))/\|\delta A\| \to 0 \text{ as } \delta \to 0)\end{aligned}$$

- In general

$$\|A^{-1}(\delta A)x\| \leq \|A^{-1}\|\|\delta A\|\|x\|$$

- But for all x and A there exists δA (exercise) such that

$$\|A^{-1}(\delta A)x\| = \|A^{-1}\|\|\delta A\|\|x\|$$

Relative condition number

$$\begin{aligned}\tilde{\kappa} &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x + A^{-1}(\delta A)(\delta x)\|/\|x\|}{\|\delta A\|/\|A\|} \\ &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x\|/\|x\|}{\|\delta A\|/\|A\|} \quad (\text{since } ((\delta A)(\delta x))/\|\delta A\| \to 0 \text{ as } \delta \to 0)\end{aligned}$$

- In general
$$\|A^{-1}(\delta A)x\| \leq \|A^{-1}\|\|\delta A\|\|x\|$$
- But for all x and A there exists δA (exercise) such that
$$\|A^{-1}(\delta A)x\| = \|A^{-1}\|\|\delta A\|\|x\|$$

Relative condition number

$$\begin{aligned}\tilde{\kappa} &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x + A^{-1}(\delta A)(\delta x)\|/\|x\|}{\|\delta A\|/\|A\|} \\ &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}(\delta A)x\|/\|x\|}{\|\delta A\|/\|A\|} \quad (\text{since } ((\delta A)(\delta x))/\|\delta A\| \to 0 \text{ as } \delta \to 0)\end{aligned}$$

- In general
$$\|A^{-1}(\delta A)x\| \leq \|A^{-1}\|\|\delta A\|\|x\|$$

- But for all x and A there exists δA (exercise) such that
$$\|A^{-1}(\delta A)x\| = \|A^{-1}\|\|\delta A\|\|x\|$$

Consequently

$$\begin{aligned}\tilde{\kappa} &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}\|\|\delta A\|\|x\|/\|x\|}{\|\delta A\|/\|A\|} \\ &= \|A\|\|A^{-1}\|.\end{aligned}$$

Sensitivity of solution of x of $Ax = b$ w.r.t. A

$$\tilde{\kappa} = \|A\|\|A^{-1}\|$$

Consequently

$$\begin{aligned}\tilde{\kappa} &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}\|\|\delta A\|\|x\|/\|x\|}{\|\delta A\|/\|A\|} \\ &= \|A\|\|A^{-1}\|.\end{aligned}$$

Sensitivity of solution of x of $Ax = b$ w.r.t. A

$$\tilde{\kappa} = \|A\|\|A^{-1}\|$$

Consequently

$$\begin{aligned}\tilde{\kappa} &= \lim_{\delta \to 0^+} \sup_{\|\delta A\| \leq \delta} \frac{\|A^{-1}\|\|\delta A\|\|x\|/\|x\|}{\|\delta A\|/\|A\|} \\ &= \|A\|\|A^{-1}\|.\end{aligned}$$

Sensitivity of solution of x of $Ax = b$ w.r.t. A

$$\tilde{\kappa} = \|A\|\|A^{-1}\|$$

Three stages contributing to the backward error

1. Computation of LU Factorization
2. Forward Substitution
3. Backward Substitution

Error due to LU Factorization

Theorem

The computed factors $\widetilde{L}$, $\widetilde{U}$ by GEPP satisfy

$$P(A + \delta A) = \widetilde{L}\,\widetilde{U}$$

where

$$\|\delta A\|_1 \leq \epsilon_{\text{mach}} n \;\; \|L\|_1 \;\; \|U\|_1$$

with L, U denoting the exact factors.

Error due to LU Factorization

Backward error should be in terms of A (input), but not L and U.

- Since all multipliers ℓ_{kj} are less than one,

$$\|L\|_1 = \max_j \|\ell_j\|_1 \leq n.$$

- It turns out that

$$\|U\|_1 \leq n\|A\|_1 \rho_{\max} \quad \text{where} \quad \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \text{ (growth factor)}$$

i.e.

$$\|A\|_1 \geq \max_{i,j} |a_{ij}|, \quad n \max_{i,j} |u_{ij}| \geq \|U\|_1 \implies \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \geq \frac{\|U\|_1}{n\|A\|_1}$$

Error due to LU Factorization

Backward error should be in terms of A (input), but not L and U.

- Since all multipliers ℓ_{kj} are less than one,

$$\|L\|_1 = \max_j \|\ell_j\|_1 \leq n.$$

- It turns out that

$$\|U\|_1 \leq n\|A\|_1 \rho_{\max} \quad \text{where} \quad \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \text{ (growth factor)}$$

i.e.

$$\|A\|_1 \geq \max_{i,j} |a_{ij}|, \quad n \max_{i,j} |u_{ij}| \geq \|U\|_1 \implies \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \geq \frac{\|U\|_1}{n\|A\|_1}$$

Error due to LU Factorization

Backward error should be in terms of A (input), but not L and U.

- Since all multipliers ℓ_{kj} are less than one,

$$\|L\|_1 = \max_j \|\ell_j\|_1 \leq n.$$

- It turns out that

$$\|U\|_1 \leq n\|A\|_1 \rho_{\max} \quad \text{where} \quad \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \text{ (growth factor)}$$

i.e.

$$\|A\|_1 \geq \max_{i,j} |a_{ij}|, \quad n \max_{i,j} |u_{ij}| \geq \|U\|_1 \implies \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \geq \frac{\|U\|_1}{n\|A\|_1}$$

Error due to LU Factorization

Backward error should be in terms of A (input), but not L and U.

- Since all multipliers ℓ_{kj} are less than one,

$$\|L\|_1 = \max_j \|\ell_j\|_1 \leq n.$$

- It turns out that

$$\|U\|_1 \leq n\|A\|_1 \rho_{\max} \quad \text{where} \quad \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \text{ (growth factor)}$$

i.e.

$$\|A\|_1 \geq \max_{i,j} |a_{ij}|, \quad n \max_{i,j} |u_{ij}| \geq \|U\|_1 \implies \rho_{\max} = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|} \geq \frac{\|U\|_1}{n\|A\|_1}$$

Error due to LU Factorization

Worst Case : $\rho_{\max} = 2^{n-1}$

- When processing jth column the largest entry can at most be doubled in absolute value.

A worst case example

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

Error due to LU Factorization

Worst Case : $\rho_{\max} = 2^{n-1}$

- When processing jth column the largest entry can at most be doubled in absolute value.

A worst case example

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

Error due to LU Factorization

Worst Case : $\rho_{\max} = 2^{n-1}$

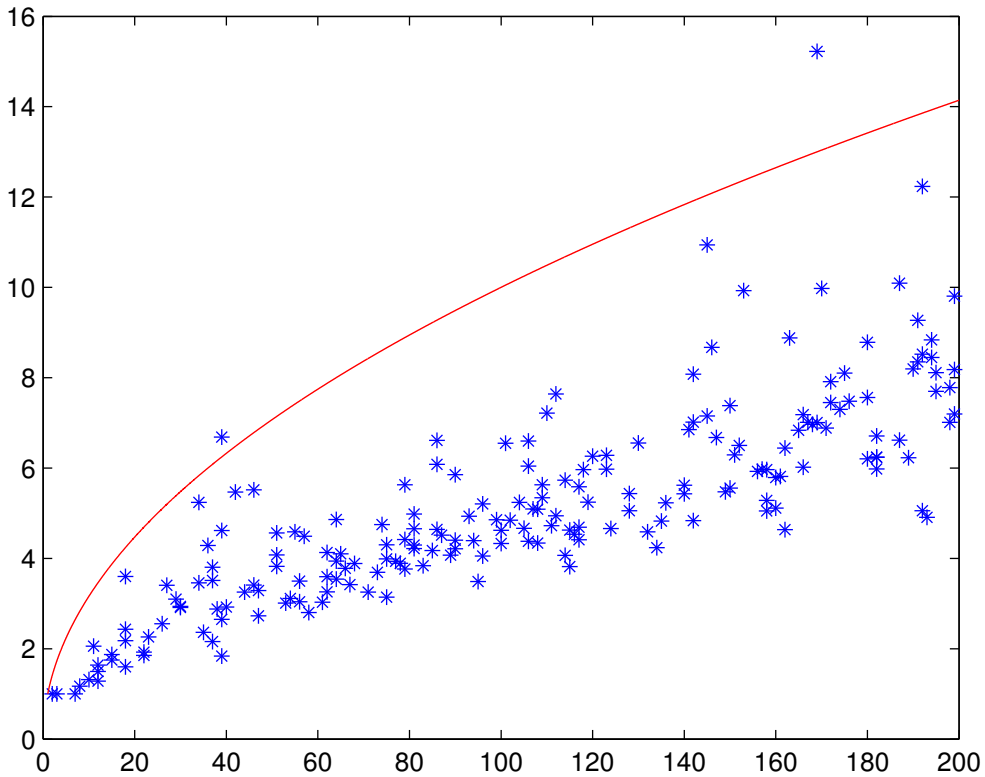
- When processing jth column the largest entry can at most be doubled in absolute value.

A worst case example

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

Error due to LU Factorization

Practice : $\rho_{\max}$ rarely exceeds $\sqrt{n}$.



Each blue asterisks represents a pair (size, growth rate) for a randomly chosen matrix of size 200 or smaller. Growth rate is calculated for 200 such matrices.

Red curve represents $f(n) = \sqrt{n}$

Error due to LU Factorization

Theorem

The computed factors $\widetilde{L}$, $\widetilde{U}$ *by GEPP satisfy*

$$P(A + \delta A) = \widetilde{L}\,\widetilde{U}$$

where $\dfrac{\|\delta A\|_1}{\|A\|_1} \leq \epsilon_{\text{mach}}\ \rho_{\text{max}}\ n^3$.

Error due to Forward Substitution

Theorem

The computed solution $\tilde{y}$ of $Ly = d$ by forward substitution satisfies

$$(L + \delta L)\tilde{y} = d$$

for some δL such that $\frac{\|\delta L\|_1}{\|L\|_1} \leq \epsilon_{\text{mach}} n.$

Error due to Back Subsitution

Theorem

The computed solution $\tilde{x}$ of $Ux = y$ by back substitution satisfies

$$(U + \delta U)\tilde{x} = y$$

for some δU such that $\frac{\|\delta U\|_1}{\|U\|_1} \leq \epsilon_{\text{mach}} n.$

Summary of Backward Errors

General Procedure (To solve $Ax = b$)	Backward Error
(i) Compute $PA \approx \widetilde{L}\widetilde{U}$.	$\|\delta A\|_1 \leq \epsilon_{\text{mach}} n^3 \rho_{\max} \|A\|_1$ $(\delta A$ s.t. $P(A + \delta A) = \widetilde{L}\widetilde{U})$
(ii) Permute rows of b, *i.e.* $d := Pb$.	None
(iii) Solve $\widehat{L}y = d$ by forward substitution where $\widehat{L} \approx \widetilde{L}$	$\|\delta L\|_1 = \|\widehat{L} - \widetilde{L}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{L}\|_1$
(iv) Solve $\widehat{U}\hat{x} = y$ by back substitution where $\widehat{U} \approx \widetilde{U}$	$\|\delta U\|_1 = \|\widehat{U} - \widetilde{U}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{U}\|_1$

Summary of Backward Errors

General Procedure (To solve $Ax = b$)	Backward Error
(i) Compute $PA \approx \widetilde{L}\widetilde{U}$.	$\|\delta A\|_1 \leq \epsilon_{\text{mach}} n^3 \rho_{\max} \|A\|_1$ (δA s.t. $P(A+\delta A) = \widetilde{L}\widetilde{U}$)
(ii) Permute rows of b, *i.e.* $d := Pb$.	None
(iii) Solve $\widehat{L}y = d$ by forward substitution where $\widehat{L} \approx \widetilde{L}$	$\|\delta L\|_1 = \|\widehat{L} - \widetilde{L}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{L}\|_1$
(iv) Solve $\widehat{U}\hat{x} = y$ by back substitution where $\widehat{U} \approx \widetilde{U}$	$\|\delta U\|_1 = \|\widehat{U} - \widetilde{U}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{U}\|_1$

Summary of Backward Errors

General Procedure (To solve $Ax = b$)	Backward Error
(i) Compute $PA \approx \widetilde{L}\widetilde{U}$.	$\|\delta A\|_1 \leq \epsilon_{\text{mach}} n^3 \rho_{\max} \|A\|_1$ $(\delta A$ s.t. $P(A + \delta A) = \widetilde{L}\widetilde{U})$
(ii) Permute rows of b, *i.e.* $d := Pb$.	None
(iii) Solve $\widehat{L}y = d$ by forward substitution where $\widehat{L} \approx \widetilde{L}$	$\|\delta L\|_1 = \|\widehat{L} - \widetilde{L}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{L}\|_1$
(iv) Solve $\widehat{U}\hat{x} = y$ by back substitution where $\widehat{U} \approx \widetilde{U}$	$\|\delta U\|_1 = \|\widehat{U} - \widetilde{U}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{U}\|_1$

Summary of Backward Errors

General Procedure (To solve $Ax = b$)	Backward Error
(i) Compute $PA \approx \widetilde{L}\widetilde{U}$.	$\|\delta A\|_1 \leq \epsilon_{\text{mach}} n^3 \rho_{\max} \|A\|_1$ $(\delta A$ s.t. $P(A + \delta A) = \widetilde{L}\widetilde{U})$
(ii) Permute rows of b, *i.e.* $d := Pb$.	None
(iii) Solve $\widehat{L}y = d$ by forward substitution where $\widehat{L} \approx \widetilde{L}$	$\|\delta L\|_1 = \|\widehat{L} - \widetilde{L}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{L}\|_1$
(iv) Solve $\widehat{U}\hat{x} = y$ by back substitution where $\widehat{U} \approx \widetilde{U}$	$\|\delta U\|_1 = \|\widehat{U} - \widetilde{U}\|_1 \leq \epsilon_{\text{mach}} n \|\widetilde{U}\|_1$

Backward Error of GEPP

Total backward error in the solution of $Ax = b$

- Let $\widetilde{x}$ be the computed solution.

$$
\begin{aligned}
& \widehat{L}\widehat{U}\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L} + \delta L)(\widetilde{U} + \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L}\widetilde{U} + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (P(A + \delta A) + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = Pb \\
\Longrightarrow \quad & (A + \underbrace{P^T(\delta A + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)}_{\widetilde{\delta A}})\widetilde{x} = b
\end{aligned}
$$

(Note: For all permutation matrices $P^T P = I$.)

Backward Error of GEPP

Total backward error in the solution of $Ax = b$

- Let $\widetilde{x}$ be the computed solution.

$$\begin{aligned}
& \widehat{L}\widehat{U}\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L} + \delta L)(\widetilde{U} + \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L}\widetilde{U} + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (P(A + \delta A) + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = Pb \\
\Longrightarrow \quad & (A + \underbrace{P^T(\delta A + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)}_{\widetilde{\delta A}})\widetilde{x} = b
\end{aligned}$$

(Note: For all permutation matrices $P^T P = I$.)

Backward Error of GEPP

Total backward error in the solution of $Ax = b$

- Let $\widetilde{x}$ be the computed solution.

$$
\begin{aligned}
& \widehat{L}\widehat{U}\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L} + \delta L)(\widetilde{U} + \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L}\widetilde{U} + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (P(A + \delta A) + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = Pb \\
\Longrightarrow \quad & (A + \underbrace{P^T(\delta A + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)}_{\widetilde{\delta A}})\widetilde{x} = b
\end{aligned}
$$

(Note: For all permutation matrices $P^T P = I$.)

Backward Error of GEPP

Total backward error in the solution of $Ax = b$

- Let $\widetilde{x}$ be the computed solution.

$$
\begin{aligned}
& \widehat{L}\widehat{U}\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L} + \delta L)(\widetilde{U} + \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L}\widetilde{U} + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (P(A + \delta A) + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = Pb \\
\Longrightarrow \quad & (A + \underbrace{P^T(\delta A + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)}_{\widetilde{\delta A}})\widetilde{x} = b
\end{aligned}
$$

(Note: For all permutation matrices $P^T P = I$.)

Backward Error of GEPP

Total backward error in the solution of $Ax = b$

- Let $\widetilde{x}$ be the computed solution.

$$
\begin{aligned}
& \widehat{L}\widehat{U}\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L} + \delta L)(\widetilde{U} + \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L}\widetilde{U} + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (P(A + \delta A) + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = Pb \\
\Longrightarrow \quad & (A + \underbrace{P^T(\delta A + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)}_{\widetilde{\delta A}})\widetilde{x} = b
\end{aligned}
$$

(Note: For all permutation matrices $P^T P = I$.)

Backward Error of GEPP

Total backward error in the solution of $Ax = b$

- Let $\widetilde{x}$ be the computed solution.

$$
\begin{aligned}
& \widehat{L}\widehat{U}\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L} + \delta L)(\widetilde{U} + \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (\widetilde{L}\widetilde{U} + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = d \\
\Longrightarrow \quad & (P(A + \delta A) + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)\widetilde{x} = Pb \\
\Longrightarrow \quad & (A + \underbrace{P^T(\delta A + \delta L\, \widetilde{U} + \widetilde{L}\, \delta U + \delta L\, \delta U)}_{\widetilde{\delta A}})\widetilde{x} = b
\end{aligned}
$$

(Note: For all permutation matrices $P^T P = I$.)

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \|\widetilde{L}\|_1 \right) \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \|L\|_1 \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\max} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \|L\|_1 \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\max} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \ \|\widetilde{L}\|_1 \right) \ \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\max} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\max} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \ \|\widetilde{L}\|_1 \right) \ \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\max} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\max} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \ \|\widetilde{L}\|_1 \right) \ \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \ \|\widetilde{L}\|_1 \right) \ \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \ \|\widetilde{L}\|_1 \right) \ \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Backward Error of GEPP

- Using the submultiplicative property and triangular inequality

$$\|\widetilde{\delta A}\|_1 \leq \underbrace{\|P^T\|_1}_{1} \left(\|\delta A\|_1 + \|\delta L\|_1 \|\widetilde{U}\|_1 + \|\widetilde{L}\|_1 \|\delta U\|_1 + \underbrace{\|\delta L\|_1 \|\delta U\|_1}_{o(\epsilon_{\text{mach}})} \right)$$

where

$$\begin{aligned} \|\delta L\|_1 \|\widetilde{U}\|_1 &\leq \left(n\epsilon_{\text{mach}} \ \|\widetilde{L}\|_1 \right) \ \|\widetilde{U}\|_1 \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

$$\begin{aligned} \|\widetilde{L}\|_1 \|\delta U\|_1 &\leq \|\widetilde{L}\|_1 \left(n\epsilon_{\text{mach}} \|\widetilde{U}\|_1 \right) \\ &\leq n\epsilon_{\text{mach}} \ \|L\|_1 \ \|U\|_1 + o(\epsilon_{\text{mach}}) \\ &\leq n^3 \epsilon_{\text{mach}} \rho_{\text{max}} \|A\|_1 + o(\epsilon_{\text{mach}}) \end{aligned}$$

Small-o Notation

Notation

$$g(h) = o(f(h)) \text{ means that } \lim_{h \to 0} \frac{g(h)}{f(h)} = 0$$

e.g. $h^2 = o(h)$ as well as $h\sqrt{h} = o(h)$.

Backward Error Analysis of GEPP

Accuracy of computed $\widetilde{x}$ for $Ax = b$

- Backward Error

$$(A + \widetilde{\delta A})\widetilde{x} = b \quad \text{where } \frac{\|\widetilde{\delta A}\|_1}{\|A\|_1} = 3n^3 \rho_{\max} \epsilon_{\text{mach}} + o(\epsilon_{\text{mach}})$$

- Relative condition number

$$\tilde{\kappa}_\delta = \kappa_1(A) = \|A\|_1 \|A^{-1}\|_1$$

- Forward Error

$$\frac{\|\widetilde{x} - x\|_1}{\|x\|_1} \leq \tilde{\kappa}_\delta \frac{\|\widetilde{\delta A}\|_1}{\|A\|_1}$$

Backward Error Analysis of GEPP

Accuracy of computed $\widetilde{x}$ for $Ax = b$

- Backward Error

$$(A + \widetilde{\delta A})\widetilde{x} = b \quad \text{where } \frac{\|\widetilde{\delta A}\|_1}{\|A\|_1} = 3n^3 \rho_{\max} \epsilon_{\text{mach}} + o(\epsilon_{\text{mach}})$$

- Relative condition number

$$\tilde{\kappa}_\delta = \kappa_1(A) = \|A\|_1 \|A^{-1}\|_1$$

- Forward Error

$$\frac{\|\widetilde{x} - x\|_1}{\|x\|_1} \leq \tilde{\kappa}_\delta \frac{\|\widetilde{\delta A}\|_1}{\|A\|_1}$$

Backward Error Analysis of GEPP

Accuracy of computed $\widetilde{x}$ for $Ax = b$

- Backward Error

$$(A + \widetilde{\delta A})\widetilde{x} = b \quad \text{where } \frac{\|\widetilde{\delta A}\|_1}{\|A\|_1} = 3n^3 \rho_{\max} \epsilon_{\text{mach}} + o(\epsilon_{\text{mach}})$$

- Relative condition number

$$\tilde{\kappa}_\delta = \kappa_1(A) = \|A\|_1 \|A^{-1}\|_1$$

- Forward Error

$$\frac{\|\widetilde{x} - x\|_1}{\|x\|_1} \leq \tilde{\kappa}_\delta \frac{\|\widetilde{\delta A}\|_1}{\|A\|_1}$$

Backward Error Analysis of GEPP

Theorem (Forward Error of GEPP)

The computed solution $\widetilde{x}$ of $Ax = b$ by GEPP satisfies

$$\frac{\|\widetilde{x}-x\|_1}{\|x\|_1} = O(\kappa_1(A) n^3 \rho_{\max} \epsilon_{\text{mach}})$$