LECTURE 17

BACKWARD ERROR ANALYSIS

$$f : V \longrightarrow W$$

V, W are vector spaces

$\tilde{f}(x)$: computed value of $f\left(fl(x)\right)$ by the numerical algorithm.

Forward Error

Absolute error

$$\left\| \tilde{f}(x) - f(x) \right\|$$

Relative error

$$\frac{\left\| \tilde{f}(x) - f(x) \right\|}{\| f(x) \|}$$

Accurate algorithm (for all x)

$$(*) \quad \frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} = O(\epsilon_{mach})$$

<u>REMARK</u>

$(*)$ means there exists a constant c (independent of x)

$$\frac{\| \tilde{f}(x) - f(x) \|}{\| f(x) \|} \leqslant c\, \epsilon_{mach}$$

for all $\epsilon_{mach} > 0$ close to zero.

Forward error directly is difficult to analyze.

<u>Backward Error</u>

<u>DEFN</u>

Suppose there exists a $\delta x \in V$ s.t.

$$\tilde{f}(x) = f(x + \delta x).$$

Then

(i) $\| \delta x \|$ is the absolute backward error.

(ii) $\| \delta x \| / \| x \|$ is the relative backward error.

EXAMPLE

Matrix-vector product

$$b = f(A) = Ax \quad (A \in \mathbb{C}^{n \times n})$$

$\tilde{b}$: computed product

Find the backward error in an IEEE floating point arithmetic.

SOLN

$$\tilde{b}_j = \bigoplus_{k=1}^{n} (a_{jk} \otimes x_k)$$

$$= \bigoplus_{k=1}^{n} (a_{jk} \times x_k)(1+\epsilon_k) \quad \left(\begin{array}{c}\text{where} \\ |\epsilon_k| \leq \epsilon_{mach}\end{array}\right)$$

$$= (a_{j1} \times x_1)(1+\epsilon_1) \oplus \left(\bigoplus_{k=2}^{n} (a_{jk} \times x_k)(1+\epsilon_k) \right)$$

$$= (a_{j1} \times x_1)(1+\epsilon_1)(1+\tilde{\epsilon}_1) \quad \left(\begin{array}{c}\text{where} \\ |\tilde{\epsilon}_1| \leq \epsilon_{mach}\end{array}\right)$$

$$+$$

$$(**) \qquad \bigoplus_{k=2}^{n} (a_{jk} \times x_k)(1+\epsilon_k)(1+\tilde{\epsilon}_1)$$

Here note that (for $k=1,\ldots,n$)

$$(1+\epsilon_k)(1+\tilde{\epsilon}_1) = 1 + \epsilon_k + \tilde{\epsilon}_1 + \epsilon_k \tilde{\epsilon}_1$$

$$= 1 + \epsilon_k + \tilde{\epsilon}_1 + O(\epsilon_{mach}^2)$$

Repeat (**) by replacing $\oplus$ with $+$.

$$\tilde{b}_j = (a_{j1} \times X_1)(1+\epsilon_1)(1+\tilde{\epsilon}_1)$$
$$+$$
$$(a_{j2} \times X_2)(1+\epsilon_1)(1+\tilde{\epsilon}_1)(1+\tilde{\epsilon}_2)$$
$$+$$
$$\vdots$$
$$+$$
$$(a_{jn-1} \times X_{n-1})(1+\epsilon_1)(1+\tilde{\epsilon}_1)\cdots(1+\tilde{\epsilon}_{n-1})$$
$$+$$
$$(a_{jn} \times X_n)(1+\epsilon_1)(1+\tilde{\epsilon}_1)\cdots(1+\tilde{\epsilon}_{n-1})$$

Above note

$$(1+\epsilon_1)(1+\tilde{\epsilon}_1)\cdots(1+\tilde{\epsilon}_{\ell-1}) = (1+\epsilon_1+\tilde{\epsilon}_1+\dots+\tilde{\epsilon}_{\ell-1}+O(\epsilon_{mach}^2))$$
$$\leq (1+\ell\,\epsilon_{mach}+O(\epsilon_{mach}^2))$$

Consequently

$$\cancel{\tilde{b}_j = \sum_{k=1}^{n-1} x_k \times a_{jk} \prod (1+\epsilon_1)(1+}$$

$$\tilde{b}_j = \sum_{k=1}^{n-1} x_k \cdot a_{jk}(1+\epsilon_1)\prod_{\ell=1}^{k}(1+\tilde{\epsilon}_\ell)$$
$$+$$
$$x_n \cdot a_{jn}(1+\epsilon_1)\prod_{\ell=1}^{n-1}(1+\tilde{\epsilon}_\ell)$$

$$= \sum_{k=1}^{n} x_k \cdot a_{jk} (1 + \delta \epsilon_k) \left(\begin{array}{l} \text{where} \\ |\delta e_k| \leqslant n\, \epsilon_{mach} \\ + O(\epsilon_{mach}^2) \end{array} \right.$$

$$= \sum_{k=1}^{n} x_k \cdot (a_{jk} + \delta a_{jk})$$

where $|\delta a_{jk}| \leqslant n\, \epsilon_{mach} |a_{jk}| + O(\epsilon_{mach}^2)$.

Summary

$$\tilde{b} = (A + \delta A)\, x$$

where

* $|\delta a_{jk}| \leqslant n\, \epsilon_{mach} |a_{jk}| + O(\epsilon_{mach})$
* equivalently

$$\| \delta a_k \|_1 = \sum_{j=1}^{n} | \delta a_{jk} |$$
$$\leq n\, \epsilon_{mach} \sum_{j=1}^{n} |a_{jk}|$$
$$= n\, \epsilon_{mach} \| a_k \|_1$$

* equivalently

$$\| \delta A \|_1 = \max_{k=1,\dots,n} \| \delta a_k \|_1$$
$$\leq \max_{k=1,\dots,n} n\, \epsilon_{mach} \| a_k \|_1$$
$$= n\, \epsilon_{mach} \| A \|_1$$

Backward error

$$\frac{\|\delta A\|_1}{\|A\|_1} \leqslant n\, \epsilon_{mach}$$

EXAMPLE

Least squares problem, $\hat{x}$ is s.t.

$$\|A\hat{x}-b\|_2 = \underset{x\in\mathbb{C}^n}{\text{minimize}} \|Ax-b\|_2$$

$(A \in \mathbb{C}^{m\times n},\ b \in \mathbb{C}^m$ with $m \geqslant n)$

$\tilde{x}$: computed value of $\hat{x}$ using the QR factorization by Householder reflectors.

Assuming A is full-rank it can be shown that

$$\|A\tilde{x}-b\|_2 = \underset{x\in\mathbb{C}^n}{\text{minimize}} \|(A+\delta A)x-b\|_2$$

where

$$\frac{\|\delta A\|}{\|A\|} = O(\epsilon_{mach})$$

DEFN (Backward Stability)

An algorithm is called backward stable if for all x there exist δx s.t.

$$\tilde{f}(x) = f(x+\delta x)$$

with

$$\|\delta x\| / \|x\| = O(\epsilon_{mach}).$$

EXAMPLES

* Matrix-vector product is backward stable.
* Least squares problem using QR factor. by Householder reflectors is backward stable.

Backward Error Analysis

Relates the backward error with forward error.

THM

Suppose there exists δx s.t.

$$\tilde{f}(x) = f(x+\delta x).$$

Then

$$\tilde{\kappa}_\delta \ \|\delta x\| / \|x\|$$

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq \cancel{O\left(\tilde{\kappa} \cdot \frac{\|\delta x\|}{\|x\|}\right)}$$

where

$$\widetilde{\kappa}_\delta := \sup_{\|\widehat{\delta x}\| \leqslant \delta} \frac{\left(\|f(x+\widehat{\delta x}) - f(x)\| / \|f(x)\|\right)}{\left(\|\widehat{\delta x}\| / \|x\|\right)}$$

and $\delta := \|\delta x\|$.

PROOF

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = \frac{\|f(x+\delta x) - f(x)\|}{\|f(x)\|}$$

$$= \frac{\|f(x+\delta x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|} \cdot \frac{\|\delta x\|}{\|x\|}$$

$$\leqslant \left(\sup_{\|\widehat{\delta x}\| \leqslant \delta} \frac{\|f(x+\widehat{\delta x}) - f(x)\| / \|f(x)\|}{\|\widehat{\delta x}\| / \|x\|} \right) \cdot \frac{\|\delta x\|}{\|x\|}$$

$\square$

EXAMPLES

① Backward Error Analysis of Matrix-Vector Product

$$b = f(A) = Ax$$

$\hat{b}$: computed product

δ-relative condition number

$$\tilde{\kappa}_\delta = \sup_{\|\delta A\|_1 \leq \delta} \frac{\|f(A+\delta A) - f(A)\|_1 / \|f(A)\|_1}{\|\delta A\|_1 / \|A\|_1}$$

$$\leq \|A\|_1 \, \|A^{-1}\|_1$$

Backward stability

$$\tilde{b} = f(A+\delta A) = (A+\delta A)\, x$$

for some δA s.t.

$$\frac{\|\delta A\|_1}{\|A\|_1} \leq n\, \epsilon_{mach}$$

Forward error

$$\frac{\|\tilde{b} - b\|_1}{\|b\|_1} = \frac{\|f(A+\delta A) - f(A)\|}{\|\delta A\|}$$

$$\left(\begin{smallmatrix}\text{THM}\\ \text{ABOVE}\end{smallmatrix}\right) \leq \tilde{\kappa}_\delta \, \frac{\|\delta A\|}{\|A\|}$$

$$= \underbrace{\|A\|_1 \, \|A^{-1}\|_1}_{\kappa_1(A)} \, n\, \epsilon_{mach}$$

② Backward Error Analysis of LSP

$$\|A\hat{x} - b\|_2 = \underset{x \in \mathbb{C}^n}{\text{minimize}} \|Ax - b\|_2$$

$\tilde{x}$: computed value using
QR factor by HH reflectors

δ - relative condition number

$$\tilde{\kappa}_\delta = \sup_{\|\delta A\| \leqslant \delta} \frac{\|\hat{x}(A+\delta A) - \hat{x}(A)\| / \|\hat{x}(A)\|}{\|\delta A\| / \|A\|}$$

$$= \kappa(A) + \frac{\kappa(A) \tan^2\theta}{\eta}$$

where

$$\kappa(A) = \|A\| \, \|A^+\|$$

$$\theta = \arccos\left(\frac{y}{b}\right)$$ (where y is the orthogonal proj. of b onto Range(A))

$$\eta = \frac{\|\hat{x}\| \, \|A\|}{\|A\hat{x}\|}$$

Backward stability (assuming A is full rank)

$$\tilde{x}(A) = \hat{x}(A + \delta A)$$

for some δA s.t.

$$\frac{\|\delta A\|}{\|A\|} = O(\epsilon_{mach})$$

Forward Error

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\hat{x}\|} \leqslant \tilde{\kappa}_\delta \frac{\|\delta A\|}{\|A\|}$$

$$\leqslant c \left(\kappa(A) + \kappa(A)^2 \frac{\tan\theta}{\eta} \right) \epsilon_{mach}$$

for some constant c.