LECTURE 16

BACKWARD ERROR ANALYSIS (PART II)

A backward stable algorithm is accurate if and only if the problem is well-conditioned.

THM (Forward and Backward Error) Suppose there exists a δx such that

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

Then

$$\underbrace{\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|}}_{\text{RELATIVE FORWARD ERROR}} \leq \tilde{K}_\delta \underbrace{\frac{\|\delta x\|}{\|x\|}}_{\text{RELATIVE BACKWARD ERROR}}$$

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

PROOF

$$\frac{\|\tilde{f}(x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|} \overset{\left(\text{since } \tilde{f}(x) = f(x+\delta x)\right)}{=} \frac{\|f(x+\delta x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|} \quad (1)$$

$$\leq \underbrace{\sup_{\|\tilde{\delta x}\| \leq \delta} \frac{\|f(x+\tilde{\delta x}) - f(x)\| / \|f(x)\|}{\|\tilde{\delta x}\| / \|x\|}}_{\tilde{K}_\delta}$$

$$\Longrightarrow$$

$$\|\tilde{f}(x) - f(x)\| / \|f(x)\| \leq \tilde{K}_\delta \left(\|\delta x\| / \|x\|\right) \quad \square$$

COROLLARY

Suppose $\tilde{f}$ is a backward stable algorithm and $\tilde{K} > 0$. Then

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

PROOF

From the previous thm

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq \tilde{K}_\delta \underbrace{c\,\epsilon_{mach}}_{\left(\text{since } \frac{\|\delta x\|}{\|x\|} = O(\epsilon_{mach})\right)}$$

where $\delta := \|\delta x\| \leq c\,\epsilon_{mach} \|x\|$.

But since $\tilde{\kappa} = \lim_{\delta \to 0} \tilde{\kappa}_\delta$, for small δ we have $\tilde{\kappa}_\delta \leq d\tilde{\kappa}$ where $d > 1$ is a constant. Therefore for ϵ_{mach} small enough (ensuring δ is small)

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq cd\,\tilde{\kappa}\,\epsilon_{mach}$$

□

BACKWARD ERROR ANALYSIS FOR MATRIX-VECTOR PRODUCT

View the product Ax as a function of $A \in \mathbb{C}^{n \times n}$ for a fixed $x \in \mathbb{C}^n$.

$$f : \mathbb{C}^{n \times n} \longrightarrow \mathbb{C}^n, \quad f(A) = Ax$$

Recall that

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

THM (Backward Error of Matrix-Vector Product)
Let $\tilde{b} = \tilde{f}(A)$ be the computed value of $b = f(A) = Ax$ in a floating point arithmetic.

Then

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

for some δA such that

$$\frac{\|\delta A\|_1}{\|A\|_1} \leq n\,\epsilon_{mach} + O(\epsilon_{mach}^2)$$

PROOF

Consider the jth entry of $\tilde{b}$

$$\tilde{b}_j = (a_{j1} \otimes x_1) \oplus (a_{j2} \otimes x_2) \oplus \ldots \oplus (a_{jn} \otimes x_n)$$

$$= \left((a_{j1}x_1)(1+\epsilon_1)\right) \oplus \left((a_{j2}x_2)(1+\epsilon_2)\right) \oplus \ldots \oplus \left((a_{jn}x_n)(1+\epsilon_n)\right)$$

$$= \left(\left(a_{j1}x_1(1+\epsilon_1) + a_{j2}x_2(1+\epsilon_2)\right)(1+\tilde{\epsilon}_1) \oplus \ldots \oplus \left(a_{jn}x_n(1+\epsilon_n)\right)\right.$$

$$= \left(a_{j1}x_1\left(1+\epsilon_1+\tilde{\epsilon}_1+O(\epsilon_{mach}^2)\right) + a_{j2}x_2\left(1+\epsilon_2+\tilde{\epsilon}_1+O(\epsilon_{mach}^2)\right)\right.$$
$$\oplus \ldots \oplus \left(a_{jn}x_n(1+\epsilon_n)\right)$$

$$= \left(a_{j1}\left(1+\epsilon_{11}+\epsilon_{12}+\ldots+\epsilon_{1n}+O(\epsilon_{mach}^2)\right)\right)x_1$$
$$+ \ldots\ldots +$$
$$\left(a_{jn}\left(1+\epsilon_{n1}+\epsilon_{n2}+\ldots+\epsilon_{nn}+O(\epsilon_{mach}^2)\right)\right)x_n$$

where ϵ_{ij} and $\epsilon_k, \tilde{\epsilon}_k$ are at most ϵ_{mach} in absolute value.

Consequently

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

with

$$\delta a_{jk} = a_{jk} \left(\epsilon_{k1} + \ldots + \epsilon_{kn} + O(\epsilon_{mach}^2) \right)$$

$$\Longrightarrow$$

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

It follows that

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

where

$$\|\delta A\|_1 = \max_{1 \leq k \leq n} \|\delta a_k\|_1$$

$$= \max_{1 \leq k \leq n} \sum_{j=1}^{n} |\delta a_{jk}|$$

$$\leq \max_{1 \leq k \leq n} \sum_{j=1}^{n} \left(|a_{jk}| \, n \, \epsilon_{mach} + O(\epsilon_{mach}^2) \right)$$

$$= n \, \epsilon_{mach} \max_{1 \leq k \leq n} \sum_{j=1}^{n} |a_{jk}| + O(\epsilon_{mach}^2)$$

$$= n \, \epsilon_{mach} \|A\|_1 + O(\epsilon_{mach}^2)$$

□

Forward error of the matrix-vector product

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

FORWARD RELATIVE ERROR

BACKWARD RELATIVE ERROR

$$\leq \|A\|_1 \|A^{-1}\|_1 \left(n\epsilon_{mach} + O(\epsilon_{mach}^2)\right)$$

$$\Longrightarrow$$

$$\boxed{\frac{\|\tilde{b}-b\|_1}{\|b\|_1} = O(\kappa_1(A)\,\epsilon_{mach})}$$

BACKWARD ERROR ANALYSIS FOR LEAST-SQUARES PROBLEM

Least squares problem

Given $A \in \mathbb{C}^{m\times n}$ (with $m \geq n$) and $b \in \mathbb{C}^m$. Find $x \in \mathbb{C}^n$ such that $\|Ax-b\|_2$ is as small as possible.

Relative condition number of x w.r.t. A

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

where

$$\theta = \arccos\left(\frac{\|Ax\|_2}{\|b\|_2}\right) \quad \text{and} \quad \eta = \frac{\|A\|\,\|x\|}{\|Ax\|}$$

THM (Backward Stability of QR based algorithm)
Suppose the LSP solved using the QR factorization of A computed by Householder reflectors. Then the computed solution $\hat{x}$ solves

Find $\hat{x}$ so that $\|(A+\delta A)\hat{x} - b\|_2$ is as small as possible

where

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

Forward error of LSP

$$\frac{\|\hat{x} - x\|}{\|\hat{x}\|} = O(\tilde{\kappa}\,\epsilon_{mach})$$

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