数理解析研究所講究録
第191巻 1973年 34-41

誤差評価と不等式

京大 数研 一松 信

0. はじめに

数値解析の中心的課題は，誤差評価であり，それは本質的に不等式にほかならない。しかし誤差評価の立場では，在来の不等式論とは，いくぶん相異なるセンスが必要な場合がある．この話は，そういった類の問題提起にすぎず，まとまった話ではないが，以前から気にかかっている問題をのべて，これからの研究者のために何かの参考になることを期待する。

1. 後退誤差解析

数値計算とは，抽象化すれば，与えられたデータ$x_1, \cdots, x_n$から，ある値$f(x_1, \cdots, x_n)$をもとめることである。計算にともなう多くの近似により，えられた値$\tilde{f}(x_1, \cdots, x_n)$は真の値とは差がある。その差が誤差で，それの上限を求めるのが誤差評価である．しかしこのような形の前進誤差解析は，一般にきわめて難しいか，またはえられた結果$|a| \leqq b$が非現実的な評価であることが多い．──非現実的とは，実際の$|a|$

とbとが数桁以上も違うことをいう.

この意味で, Wilkinsonらの後退誤差解析は, コロンブスの卵かもしれないが, 興味深い考え方である. それはえられた値 $\tilde{f}(x_1,\dots,x_n)$ は, 摂動をうけたデータ$(x_1^*,\dots,x_n^*)$に対する真の値である: $\tilde{f}(x_1,\dots,x_n)=f(x_1^*,\dots,x_n^*)$ とみなして, $|x_k-x_k^*|$ を評価しようというものである. この方法によると, 多くの場合, $|x_k-x_k^*|$ の上限として, 使用した桁数の最小単位 u の定数倍(比較的小)という「現実的」な評価がえられる.

その大きな理由は, 浮動小数点演算における加減法の誤差評価にある. 浮動小数点表示では, 相対誤差がほぼ一定とみなされるから, $a+b$ の演算結果 $fl(a+b)$ について

(1) $\quad fl(a+b)=(a+b)(1+\delta), \qquad |\delta|\leqq Cu$, C:定数

が成立する必要があるが, (1)は一般には成立しない(桁落ちを生ずるとき). しかしWilkinsonは, たとえ桁落ちを生じたとしても, 被演算数が, ある桁以降は0が無限に続いた正確な数であるとみなせば, 後退誤差解析的な評価

(2) $\quad fl(a+b)=a(1+\delta_1)+b(1+\delta_2);\ |\delta_1|,|\delta_2|\leqq u$

は正しいことを証明した. 乗除法については, (1)のような評価が成立するので, これによって後退誤差解析が可能になるのである(実例は [1], [2] に詳しい).

2. 行列のノルムと消去法の誤差解析

$n \times n$ 行列 A は，n 次元線型空間 Ω の自分自身への一次変換とみなされる。Ω にノルム $\|\ \|$ があるとき，それから誘導された行列のノルム

(3) $$\|A\| = \sup_{x \neq 0} \|Ax\| / \|x\| = \max\{\|Ax\| ; \|x\| = 1\}$$

が定義され，つぎの性質がある。

$$\|A\| \geqq 0, \quad \|A\| = 0 \leftrightarrow A = O, \quad \|cA\| = |c|\,\|A\|$$

$$\|A + B\| \leqq \|A\| + \|B\|, \quad \|AB\| \leqq \|A\| \cdot \|B\|, \quad \|I\| = 1.$$

$\|x\|$ が l_1, l_2, l_∞ ノルムであるとき，$\|A\|$ は具体的にはつぎのようになる：　$A = [a_{ij}]$

$$\|A\|_1 = \max_j \left(\sum_{i=1}^{n} |a_{ij}| \right), \qquad \|A\|_\infty = \max_i \left(\sum_{j=1}^{n} |a_{ij}| \right)$$

$$\|A\|_2 = (A^* \cdot A \text{ の最大固有値})^{1/2}.$$

定義から $\|Ax\| \leqq \|A\| \cdot \|x\|$ であるが，「多くの場合」，この両辺はかなり近いことは重要な注意である。([2], [3] 参照)

さて，連立一次方程式 $Ax = b$ を消去法で解くことは，係数行列 A を，容易に逆行列が求められる行列の積に分解することと解釈される。たとえば Gauss の消去法は，

(4) $$A = LU, \quad L \text{ は下三角行列，} U \text{ は上三角行列}$$

という分解である。L の対角要素を 1 とすれば，この分解は（可能ならば）一意的である。Wilkinson は注意深く後退誤差解析を行ない，枢軸の選択を加えた Gauss の消去法によ

ってえられた L, U が，$L \cdot U = A + E$（E は誤差行列）としたとき，たとえば

$$\|E\|_\infty \leqq n^2 \rho \, \|A\|_\infty \cdot u$$

を示している（じっさいには成分ごとの もっと詳しい評価がえられる）．ここに ρ は，消去の途中に生ずる係数 $a_{ij}^{(k)}$ から

(5) $$\rho = \max_{i,j,k} |a_{ij}^{(k)}| / \|A\|_\infty$$

としてえられる量である．各 A について「後天的」には容易にえられる．「先天的」な評価としては，$|a_{ij}| \leqq M$ のとき

(6) 部分選択なら $|a_{ij}^{(k)}| \leqq 2^{k-1} M$

(7) 完全選択なら $|a_{ij}^{(k)}| < \left(k \cdot 2^1 3^{\frac{1}{2}} 4^{\frac{1}{3}} \cdots k^{\frac{1}{k-1}}\right)^2 \cdot M$

$$\sim 1.8 \, k^{(1/4) \log k} \cdot M$$

がえられている．(6) は $=$ が成立しうるという意味で「最良」であるが，(7) は最良ではない．（(7) の最良評価は未解決のように思われる．$|a_{ij}^{(k)}| \leqq kM$ という予想もある．(*)）

しかし現実に (6) でつねに等号が成立し，最終的に $\rho \cdot \|A\|_\infty = 2^{n-1} \cdot M$ となることは「めったに」ない．Wilkinson の「実験」によれば，「たいていの例」では (6) で $\rho \cdot \|A\|_\infty \leqq 8 \cdot M$ であるという．――ここにつぎにのべるような，<u>確率論的評価</u>の問題が生ずる．

(*) 複素係数では「反例」が作られているが，その例でも $\leqq 1.05 kM$ である．

3. 確率的誤差限界の一例

たとえばある桁で丸めた数 $a_1, \cdots, a_m$ の和を求めるとする．個々の数の誤差の上限が u ならば，$s = a_1 + \cdots + a_m$ の丸め誤差の上限は mu で，これは最良の評価である．

しかしすべての丸めが皆一斉に同符号で最大値である，といった「特殊」な事態は，「めった」に生じないと考えるほうが自然である．丸め誤差 $\varepsilon_1, \cdots, \varepsilon_m$ はもちろん決定的な量であるが，それらを独立な確率変数であるかのように考えるのは，実用上有用な仮定である．もしも ε_k が平均値 0，標準偏差 σ_k の正規分布に従うと仮定すれば，s の誤差 $\varepsilon = \varepsilon_1 + \cdots + \varepsilon_m$ は，平均値 0，標準偏差

$$\sigma = (\sigma_1^2 + \cdots + \sigma_m^2)^{1/2}$$

の正規分布をなす．$|\varepsilon| \geqq 3\sigma$ となることは稀（確率 0.3% 以下）だから，$\sigma_1 = \cdots = \sigma_m = u$ なら，事実上 $|\varepsilon| \leqq 3\sqrt{m}\,u$ としてよい．これは mu より，だいぶ小さい．

実際には，ε_k は $[-u, +u]$ の一様分布とするほうがよい近似であろうが，m が十分大ならば，ε の分布は中心極限定理により，正規分布に十分近くなる．正確に計算すると，このとき $|\varepsilon| \leqq 2bu$ である確率は（[4] 参照）

$$\frac{1}{m!}\left[\sum_{k=0}^{\leqq m/2+2b} (-1)^k \binom{m}{k}\left(2b+\frac{m}{2}-k\right)^m - \sum_{k=0}^{\leqq m/2+2b} (-1)^k \binom{m}{k}\left(-2b+\frac{m}{2}-k\right)^m\right]$$

であることが計算される．mが十分大ならば，実用上ほぼ

$|\varepsilon| \leqq 2\sqrt{m}\ u$ 　としてよい．

Wilkinson の「実験結果」を裏づける理論として，$|a_{ij}| \leqq 1$ の範囲で，$\{a_{ij}\}$ の分布を適当に仮定し，$|a_{ij}^{(k)}| \leqq c_k$ である確率を計算してみることが望ましい．そうして $\rho\|A\|_\infty \leqq 8$（右辺の 8 は本質的でなく，10でも 7でも大差ないが）である確率がかなり小さい，というような結果が示されれば大いに有意義である。

ただしここで $\{a_{ij}\}$ の分布をどう仮定するかが問題であろう．すべて独立に一様分布とするのは非現実的である．最初の枢軸選びから，$a_{11}=1$ とし，a_{i1} $(i=2,\cdots,n)$ を一様分布としてもよかろう．それ以外の要素を一様分布としてよいか ―― よくわからない．一様分布より正規分布のほうが計算しやすいかもしれないが，そうすると特異行列に近いものが生じやすくなるかもしれない．―― とにかくやってみる必要があるだろう．計算しやすいように，適当な仮定を付加していってもよいかもしれない．

4. 近似評価での逆向き不等式

誤差評価においては，本来 $a > b$ であるのに，多少修正して $a \leqq b + \varepsilon$ とか $a \leqq (1+\varepsilon)\, b$ といった形の評価がほ

しいことがよくある．四則演算の誤差評価([1],[2],[3])に，しばしば $a \leqq 1.01\,nu$ といった形の式が現われるのが，その一例である．1.01 は 1 より僅かに大きいある定数という意味の量である．

この種の例として，かつて[5]で，相接近した正の数 $a_1 > a_2 > \cdots > a_n > 0$ （$(a_1 - a_n)/a_1 \fallingdotseq 1/30$ 程度）の相乗平均を下から評価する必要があった．数値的には，それより大きい相加平均 A_n でも十分であり，たとえば $0.9A_n$ とでもすればすんだが，これでは論理的でなく，一般性もない．Koberの不等式[6]も利用した：

正数 $a_1, \ldots, a_n$ の相加平均，相乗平均を A_n, G_n とすると，

$$(8) \qquad \frac{1}{n(n-1)} < \frac{A_n - G_n}{\Delta} < \frac{1}{n}$$

$$\Delta = \sum_{1 \leqq i < k \leqq n} \left[(a_i)^{1/2} - (a_k)^{1/2}\right]^2$$

であるから，$G_n > A_n - (\Delta/n)$．── しかし $a_1, \ldots, a_n \to a$ のとき，(8)の中央項 $\to 2/n^2$ であり，$G_n > A_n - (c\Delta/n^2)$ が期待される．じっさい当面の例では $\Delta = O(n^6)$ であり，$G_n > A_n - O(n^4)$ がほしかったので，これでは不十分だった．けっきょく a_k の具体的な式を使って展開して主要項を評価するほうがよかった．（もっとも後に泉脩蔵氏から，A_n と G_n との逆向き不等式に関するよい結果を御教示いただいた）．しかしこの種の逆向き不等式 $A_n - G_n$ の上からの評価は研究課題になろう．

参考文献

[1] J. H. Wilkinson, Rounding errors in algebraic processes. Prentice-Hall, 1963. (他の版，およびドイツ語訳がある)

[2] J.H.Wilkinson, The algebraic eigenvalue problem. Oxford Univ. Press, 1965.

[3] G.E.Forsythe & C.B.Moler, Computer solution of linear algebraic systems, Prentice-Hall, 1967 : 日本語訳，培風館，1969.

[4] 宇野利雄，数値計算論，岩波，1941 (とくに p.65)

[5] S.Hitotumatu, On the numerical computation of Bessel functions through continued fraction, Comm. Math. Univ. St. Paul, 16 (1968), 89-113.

[6] H.Kober, On the arithmetic and geometric means and on Hölder inequality, Proc. Amer. Math. Soc., 9 (1958) 452-459.