

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なので、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, \dots, x_n から、ある値 $f(x_1, \dots, x_n)$ を求めることである。計算にともなう多くの近似により、えられた値 $\tilde{f}(x_1, \dots, x_n)$ は真の値とは差がある。その差が誤差で、そ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 $|a| \leq b$ 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 とが桁数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。それはえらんだ値 $\tilde{f}(x_1, \dots, x_n)$ 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, $|x_k - x_k^*|$ の上界として, 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

$$(1) \quad fl(a+b) = (a+b)(1+\delta), \quad |\delta| \leq C u, \quad C: \text{定数}$$

が成立する必要があるが, (1) は一般には成立しない (桁落ちを全うするとき)。しかし Wilkinson は, たとえ桁落ちを全うしたとしても, 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

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

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 の自分自身の一次変換とみるされる。 Ω にノルム $\|\cdot\|$ があるとき, それから誘導された行列のノルム

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

が定義され, つぎの性質がある。

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

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

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

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

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

定義から $\|Ax\| \leq \|A\| \cdot \|x\|$ 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,

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

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 (可能なときは) 一意である。 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

$$\|E\|_{\infty} \leq n^2 \rho \|A\|_{\infty} \cdot u$$

を示している (じつせ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

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

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, $|a_{ij}| \leq M$ のとき

$$(6) \quad \text{部分選状なら} \quad |a_{ij}^{(k)}| \leq 2^{k-1} M$$

$$(7) \quad \text{完全選状なら} \quad |a_{ij}^{(k)}| < (k \cdot 2^1 3^{\frac{1}{2}} 4^{\frac{1}{3}} \dots k^{\frac{1}{k-1}})^2 \cdot M \\ \sim 1.8 k^{(1/4) \log k} \cdot M$$

がえられてくる. (6) は $=$ が成立するという意味で「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. $|a_{ij}^{(k)}| \leq kM$ という予想もある.*)

しかし現実には (6) でつねに等号が成立し, 最終的に

$$\rho \cdot \|A\|_{\infty} = 2^{n-1} \cdot M \quad \text{となることは「めった」にない.}$$

Wilkinson の「実験」によれば, 「たいていの例」では (6) で

$$\rho \cdot \|A\|_{\infty} \leq 8 \cdot M \quad \text{であるという. — ここにつき「」なので}$$

ような, 確率論的評価の問題が生ずる.

(*) 複素係数では「反例」が作られているが, その例でも $\leq 1.05 kM$ 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, \dots, 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、 $S = a_1 + \dots + a_m$ 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」に生じないと思えるほうが自然である。丸め誤差 $\varepsilon_1, \dots, \varepsilon_m$ 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 $\varepsilon = \varepsilon_1 + \dots + \varepsilon_m$ は、平均値 0、標準偏差

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

の正規分布をなす。 $|\varepsilon| \geq 3\sigma$ となることは稀(確率 0.3% 以下)だから、 $\sigma_1 = \dots = \sigma_m = u$ ならば、事實上 $|\varepsilon| \leq 3\sqrt{m}u$ としよ。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るが、 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 $|\varepsilon| \leq 2bu$ である確率は ([4] 参照)

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

であることが計算される。 m が十分大ならば、実用上はほぼ
 $|E| \leq 2\sqrt{m} \mu$ としてよい。

Wilkinson の「実験結果」を裏づける理論として、 $|a_{ij}| \leq 1$ の範囲で、 $\{a_{ij}\}$ の分布を適当に仮定し、 $|a_{ij}^{(k)}| \leq c_k$ である確率を計算してみることが望ましい。そして $\rho \|A\|_\infty \leq 8$ (右辺の 8 は本質的でなく、10 でも 7 でも大差ないか) である確率がかたがり小さい、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。最初の枢軸選出から、 $a_{11} = 1$ とし、 a_{i1} ($i=2, \dots, n$) を一様分布としてもよからう。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か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てもよいかかもしれない。

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
 (2) $a \leq b + \varepsilon$ とか $a \leq (1 + \varepsilon)b$ という左形の評価がほ

しいことがよくある。四則演算の誤差評価 ([1], [2], [3]) には、しばしば $a \leq 1.01 n u$ という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 ^([1], [2], [3]) があって [5] で、相接近した正の数 $a_1 > a_2 > \dots > a_n > 0$ ($(a_1 - a_n)/a_1 \approx 1/30$ 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 <sup>でも十分であり、
「たとえは」</sup> $0.9 A_n$ 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, \dots, a_n の相加平均、相乗平均を A_n, G_n とすると、

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

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

であるから、 $G_n > A_n - (\Delta/n)$ 。—— しかし $a_1, \dots, a_n \rightarrow a$ のとき、(8) の中央項 $\rightarrow 2/n^2$ であり、 $G_n > A_n - (e\Delta/n^2)$ が期待される。じつせい ^{当面} の例では $\Delta = O(n^6)$ であり、 $G_n > A_n - O(n^4)$ がほしかつたので、これでは不十分だった。けっきょく a_k の具体的形式を使って展開して主要項を評価するほうがよかった。(もっとも後には梶浦蔵氏から、 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る。

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] 宇野利左衛門, 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.