1. 一二の慣用計算における Numerical Instability について

宇野利雄（日大理工）

次の2つの例において数値計算が不安定となる場合をしらべ，その原因の解析を試みたものを報告する。

i) 共軛勾配法による連立1次方程式の求解

ii) Frame 法による行列諸要素の計算

1　共軛勾配法が不安定となる場合

n 元連立1次方程式 $AX=K$ を解くための共軛勾配法とは出立のベクトル X_0 を任意にとり，初期剰余 $r_0=K-AX_0$，および初期方向 $p_0=A'r_0$ を計算し，以後は次の操作を n 回くりかえして，n 回目に得られた X_n が解答となるものである。

$$a_i=|A'r_i|^2/|Ap_i|^2,\quad r_{i+1}=r_i-a_iAp_i;$$

$$X_{i+1}=X_i+a_ip_i,$$

$$b_i=|A'r_{i+1}|^2/|A'r_i|^2,\qquad p_{i+1}=r_{i+1}+b_ip_i$$

若し計算途中の丸めの誤差がなければ $|A|\neq 0$ である限り，x_0 の如何にかかわらず X_n は正確な解答となるはずである。途中の丸めの誤差が入つても通例の場合はこれで差支えないのであるが，$|A|$ が小さいときには飛んでもない見当ちがいの値が出てくることを次の例で示しておく。

$$A=\begin{pmatrix} 3.2 & 1.2598 & -2.02 & 5.1398 \\ 1.0201 & -1.3501 & 3.1 & -2.1201 \\ -2.0298 & 2.55 & -1.3702 & 3.64 \\ 3.21 & 1.1102 & 2.81 & 4.54 \end{pmatrix}$$

$$K=\begin{pmatrix} -1.4402 \\ 3.54 \\ -1.94 \\ 3.7004 \end{pmatrix}$$

スタートのベクトルを（1，0，0，0）とするとX_4として

(0.31966145，0.23495757，1.0557621，−0.12185761)

またスタートを（0，1，0，0）とすると

(0.48921335，0.67483617，1.0418652，−0.34071910)

となり，この両者の間にあまりに開きがありすぎて何とも判定にくるしむ。事実正解は（1，2，1，−1）であつたので，上の両者とも第3成分をのぞいては正解とは似ても似つかないものである。

機械計算では途中の段階は一々記録されずにすむのであるから，出て来た上のような結果をそのまま答としてしまう危険性も起り得る。上の計算は筆算でやつたので実は異常とみとめられることの途中で起つたのがわかつていた。それは上の2つの場合ともr_3の計算のときこのベクトルの4成分がみなそろつて大きく桁落ちしたことである。すなわち最初の場合についていえば　$r_3=r_2-a_2Ap_2$が次のようになつた。

$$r_3^{(1)}=-0.87815214+0.87809165=-0.00006049$$
$$r_3^{(2)}=-0.4043532\ +0.40423190=-0.0001213$$
$$r_3^{(3)}=\ 1.4376088\ -1.4376693\ =-0.0000605$$
$$r_3^{(4)}=-0.2484802\ +1.2485410\ =-0.0000608$$

このような4成分の一斉桁落ちが，信用し得る有効数字の桁数を大幅に減らし，これにともなつて以後の計算が不正確になると考えられる。その理由を更にくわしく考える。

最初の方程式は$A'AX=A'K$とすれば対称形となるので，以後問題をAが対称の形で論ずることにする。このとき共軛勾配法はやや簡単になり

$$\text{スタート}\quad X_0\ ,\ r_0=p_0=K-AX_0\ ,$$

繰返し操作；

$$a_i=|r_i|^2/(p_i\ ,Ap_i),\ r_{i+1}=r_i-a_iAp_i\ ,X_{i+1}=X_i+a_ip_i; \tag{C1}$$

$$b_i = |r_{i+1}|^2 / |r_i|^2, \quad p_{i+1} = p_i + b_i p_i$$

で行われる。

Aの特有根を λ_1, λ_2, ……, λ_n とし，これらを対角要素とする対角行列Λを考えると，適当な直交変換Tがあつて$A = T^{-1} \Lambda T$である。これをつかえば原方程式は$\Lambda T X = T K$となるので，変数変換

$$y = T X, \qquad h = T K$$

を行つて原方程式を$\Lambda y = h$，あるいはあらためてyをXと書直し

$$\lambda_1 x^{(1)} = h, \quad \lambda_2 x^{(2)} = h^{(2)}, \cdots\cdots, \quad \lambda_n x^{(n)} = h^{(n)}$$

の形にして考える。もちろんこの形式にすれば各変数に分離された形となつているから，すぐに$x^{(i)} = h^{(i)} / \lambda_i$となることは明らかであるが，共軛勾配法のときの誤差の行方をたどるために，わざわざこの形式のものにその方法をほどこすわけである。ところで上にあげた共軛勾配操作の計算式はその形からわかるように，直交変換によつて形式を変えないので，そのような簡単化した形で問題を論じても差支えないわけである。

附表Ⅰに模型化した場合の一数値例が出ている（浮動小数点8桁の計算），これは特有根のうちの一つλ_1が他にくらべて非常に小さい場合である。このとき相手の$h^{(1)}$も同時に小さく，対応する$x^{(1)}$が他の$x^{(i)}$と同程度の大きさのときには，いつもこの例に示されるような経過をたどるであろう。

この場合，λ_1に対応する部分からのr，pなどは最初のうち，その他の部分にくらべて非常に小さく，ほとんど計算に寄与するところがない。それが関与しないとすれば，その残りだけをとつた次元の１つ低いsystemで$x^{(i)}$がほとんどきまつてしまうはずである。この例でも実際そのようになり，X_2の段階で$x^{(2)}$，$x^{(3)}$が既にほとんど正解に近いものに到達してしまう。

こうなると，そこでのresidualにつき，これらの部分に相当する成分がほとんど0になり，このときはじめて第Ⅰ成分からの$r_2^{(1)}$が外のものと肩をならべて計算に影響を及ぼし得ることになる。ところがこの

ときの$r_2^{(2)}$，$r_2^{(3)}$などについてはこれが0に近くなつたというのは大幅な桁落ちの結果であり，この例では1桁の数字が出てはいるものの，これは全然noise と見なさるべきものである。

ところがこのnoise が次の段階（p_2 ，Ap_2 ）のところでは，折角入つて来た第1成分からの正しいinformation と同じ程度の大きさとなつてあらわれて来て，計算全体をa_2 を通じて大きくかき乱してしまう。その経過は波線のアンダーラインをたどつて説明される通りである。

最後の結果における$x_3^{(1)}$は全然頭から信用のおけない数字である。$x_3^{(2)}$，$x_3^{(3)}$などについては前の段階でほとんど正解に近いものを得てをりながら，ここでnoise ばかりになつた補正を受けるわけである。この補正が前の段階の$x_2^{(2)}$，$x_3^{(3)}$などの約10^{-3} 程度となつているので結果は頭の3，4桁だけは信用がおける。

参考に書かれている正解と比較して我々の見込はよく験証される。

ところで上の例では各成分毎に分離した形式で問題が与えられたのであるが，変換Tによつてもとへ戻した　$A=T^{-1}\Lambda T$　の場合についていえば，T^{-1} で戻されたときに全然出たらめな$x_3^{(1)}$が各成分に入りこんで，一般にはどの成分も全部頭から信用のおけないものになる。最初に出した例はこのような場合に該当するものであろう。参考のため最初に出した例について$A'A$の特有根を出して見たら，絶対値最小の根は5.914×10^{-9}，その上の根は7.5447となつていた。

さて上のような場合に，これを矯正する方法はないものであろうか。桁落ちの悪影響からまぬかれるためには一つの方法は矢張り超桁数計算である。第2，第3成分についての計算を最初から全部超桁数でやつて行くと$r_2^{(2)}$，$r_2^{(3)}$がもつと精密な形で出せ，これにともなつてa_2 の値が正しいものに近く得られるようになる。

附表Iの例を試みに16桁計算でやると，$r_2^{(2)}$，$r_2^{(3)}$のところで桁落ちはおこるものの尚多くの有意の数字が残り

$$r_2^{(2)} = -8.74952 \times 10^{-11}$$

$$r_2^{(3)} = 2.49986 \times 10^{-11}$$

となり，次の（p_2 ，Ap_2 ）への寄与はそれぞれ

$$(Ap_2)^{(2)} \cdot p_2^{(2)} = 3.745 \times 10^{-30}$$

$$(Ap_2)^{(3)} \cdot p_2^{(3)} = 6.735 \times 10^{-31}$$

となつて，これらは第1成分からのinformation

$$(Ap_2)^{(1)},\ p_2^{(1)} = 9.999047652902649 \times 10^{-16}$$

にほとんど影響を与えない　これから計算して

$$x_3^{(1)} = 0.9999\ \ 9999\ \ 9982\ \ 8082$$

$$x_3^{(2)} = 6.6666\ \ 6666\ \ 6605\ \ 011$$

$$x_3^{(3)} = 4.2857\ \ 1428\ \ 5776\ \ 660$$

が得られるが，正解と比較し11桁まで一致が見られる。

ここに注目すべきことを書き添えよう。最初に出した例を通常の消去法でそのまま計算機にかける。

4つの答は

$$+9.9988775 \times 10^{-1}, \quad +1.9997089 \times 10^{0}$$

$$+1.0000092 \times 10^{0}, \quad -9.9985512 \times 10^{-1}$$

となり，超桁計算などの面倒な必要もなく，あつさり5桁までの一致が見られる。

2.　Frame 法の不安定性

Frame 法は行列の特有方程式，逆行列，行列式などを次の操作により求めるものである。

(1)　$A_k = AA_{k-1} - C_k E, \quad C_k = \dfrac{1}{K} tr(AA_{k-1})$

ただし$A_0 = E$　ここでC_kは特有方程式の係数となる。すなわち

$$|\lambda E - A| = \lambda^n - C_1 \lambda^{n-1},\ -C_n$$

であり，又行列式，逆行列は

$$|A| = (-1)^{n-1} C_n, \quad A^{-1} = \frac{1}{C_n} A_{n-1}$$

となる。

この方法が不安定となる1例を附表IIにあげた（要点となる対角要素のみが示してある）。ここでもC_kの計算のときに桁落ちがおこつて信用し得る数字が1桁だけになつてしまうという現象がおこつたのである。なおこのとき計算されたA_{n-1}から逆行列を出すと，

−4.1654731	124.78950	−499.65532	437.05440
31.212640	−624.54349	2810.4061	−2623.1196
−62.464154	936.77188	−4496.6393	4371.7275
36.422983	−437.20281	2185.8732	−2185.8696

となつた。正しい値は

$$\begin{pmatrix} -4 & 120 & -480 & 420 \\ 30 & -600 & 2700 & -2520 \\ -60 & 900 & -4320 & 4200 \\ 35 & -420 & 2100 & -2100 \end{pmatrix}$$

なので，やつと1桁しか合わないということである。

さてFrame 法のこのような不安定の原因をしらべる。そのために$C_k = \frac{1}{k} tr(AA_{k-1})$となる理由を帰納法的に考える。この関係が$C_k$まで成立する，すなわち

$$C_m = (-1)^{m-1} \sum \lambda_{i_1} \lambda_{i_2} \cdots\cdots \lambda_{i_m} = \frac{1}{m} t_r (AA_{m-1}),$$

$$i_1 < i_2 < \cdots\cdots < i_m$$

$(m \leq k)$とし，又$tr(A^m) = \sum_h \lambda_h^m$を考慮すると

$$tr\ (AA_k) = tr(A^{k+1}) - C_1\, tr(A^k) - C_2\, tr(A^{k-1}) - \cdots\cdots =$$

$$= \left\{ \begin{array}{ll} +\left(\sum_h \lambda_h^{k+1}\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & 〔0〕 \\ & 〔 \\ -\left(\sum_i \lambda_i\right)\left(\sum_h \lambda_h^k\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots & 〔1〕 \\ +\left(\sum_{i<j} \lambda_i \lambda_j\right)\left(\sum_h \lambda_h^{k-1}\right) \cdots\cdots\cdots\cdots\cdots\cdots & 〔2〕 \end{array} \right.$$

$$(2)\quad \left\{\begin{array}{ll} -\left(\sum_{i_1<i_2<i_3} \lambda_{i_1}\lambda_{i_2}\lambda_{i_3}\right)\left(\sum_h \lambda_h^{k-2}\right) & \cdots\cdots\cdots\cdots\quad [3] \\ +\quad \cdots\cdots\cdots\cdots & \cdots\cdots\cdots \\ +(-1)^k\left(\sum_{i_1<i_2<\cdots<i_k} \lambda_{i_1}\lambda_{i_2}\cdots\cdots\lambda_{ik}\right. & [k] \end{array}\right.$$

この右辺(2)を整頓し直すと

$$(-1)^k k \sum_{i_1<i_2<\cdots<i_{k+1}} \lambda_{i_1}\lambda_{i_2}\cdots\ \lambda_{i_{k+1}}$$

となつて，C_{k+1} が特有方程式の $k+1$ 番目の係数であることを示してくれる。なお，ここで λ_i は i 番目特有根をあらわすとした。

ところで(1)を計算するときには(2)の形での加減算が行われるのであるが，このときに $\lambda_1, \cdots, \lambda_n$ に大小の不つりあいがあり，絶対値の大きい方から順に番号をつけたとき，第1組 $\lambda_1, \cdots, \lambda_p$ まで通常の大きさであるが，第2組 $\lambda_{p+1}, \cdots, \lambda_n$ は上にくらべて非常に小さいとする。今この $\lambda_{p+1}, \cdots, \lambda_n$ を neglect して考えると，$m>p$ なる〔m〕には必ず第2組の λ_i が入つて negligible になるから，寄与する項は〔0〕，〔1〕，…，〔p〕のみである。

一方 C_k の表示においても，$k>p$ のとき必ず第2組の λ_i が因数として入つて来て，C_k が negligible になることがわかる。すなわち $k>p$ のとき，(2)の計算で通常の大きさを持つた〔0〕，〔1〕，…，〔p〕の加減算が桁落ちによつて小さい値になるということになつた。ここに附表Ⅱの例を説明する原因が求められるであろう。

〔参考〕 この行列は元来相当に singular なのであるが，かまわず消去法で逆行列を出すと，

−4.0000219	120.00034	−480.00171	420.00163
30.000075	−600.00122	2700.0064	−2520.0063
−60.000071	900.00124	−4320.0068	4200.0067
35.000017	−420.00036	2100.0021	−2100.0021

となつた。消去法の意外な優秀性がここにも示されている。

(附表Ⅰ)　　　　$AX=K$　(A 対称)

$$r_0=K-AX_0\ ,\quad p_0=r_0$$

$$r_{i+1}=r_i-a_iAp_i,\quad p_{i+1}=r_{i+1}+b_ip_i,\quad X_{i+1}=X_i+a_ip_i$$

$$a_i=|r_i|^2/(p_i,Ap_i)\ ,\ b_i=|r_{i+1}|^2/|r_i|^2$$

例

$$\begin{cases} 0.00001x_1=0.00001 \\ 0.3\qquad x_2=2 \\ 0.7\qquad x_3=3 \end{cases}$$

スタート (0, 0, 0)

$r_0=p_0$: $(10^{-5}, 2, 3)$　　　$|r_0|^2=13$

Ap_0 : $(10^{-10}, 0.6, 2.1)$

$(p_0, Ap_0)=10^{-15}+1.2+6.3=7.5$, $a_0=1.7333333$

$-a_0Ap_0$: $(-1.^{-10}7333333, -1.040, -3.6399999)$

$r_1=r_0-a_0Ap_0$: $(9.^{-6}9998267, 0.96, -0.6399999)$

$|r_1|^2=1.3311999$, $b_0=0.10239999$

b_0p_0 : $(1.^{-6}0239999, 0.20479998, 0.30719997)$

$p_1=r_1+b_0p_0$: $(1.^{-5}1023827, 1.1648000, -0.33279993)$

Ap_1 : $(1.^{-10}1023827, 0.34944000, -0.23295995)$

$(p_1Ap_1)=1.^{-15}2152476+0.40702771+0.077529055$

$=0.48455677$

$a_1=2.7472527$

$-a_1Ap_1$: $(-3.^{-10}0285238, -0.95999998, 0.63999985)$

$r_2=r_1-a_1Ap_1$: $(9.^{-6}9995239, 0.000000002, -0.00000005)$

$|r_2|^2=9.^{-11}9993378$, $b_1=7.^{-11}5115223$

b_1p_1 : $(8.^{-16}2805722, 8.^{-11}7494212, -2.^{-11}4998341)$

$p_2=r_2+b_1p_1$: $(9.^{-6}9995239, 2.^{-8}0087494, -5.^{-8}0024998)$

Ap_2 : $(9.^{-11}9995239,\ 6.^{-9}0262482,\ -3.^{-8}5017499)$

$(p_2,\ Ap_2)$: $9.^{-16}9990478+1.^{-16}2105222+1.^{-15}7517503$

$=2.^{-15}8727073$

$a_2=3.^{4}\ 4808063$

X_0	0	0	0
$a_0\ p_0$	$1.^{-5}7333333$	3.4666666	5.1999999
X_1	$1.^{-5}7333333$	3.4666666	5.1999999
$a_1\ p_1$	$3.^{-5}0285238$	3.1999999	−0.91428551
X_2	$4.^{-5}7618571$	6.6666665	4.2857144
$a_2\ p_2$	$3.^{-1}4806406$	$6.^{-4}9920676$	$-1.^{-3}7412733$
X_3	0.34811168	6.6673657	4.2839731

正解　1　, 6.6666667 , 4.2857143

（附表 II）

Frame 法

$$|\lambda E-A|=\phi(\lambda)=\lambda^{n}-C_1\lambda^{n-1}-\ \cdots\ -C_n$$

$$A_k=AA_{k-1}-C_kE,\quad C_k=\frac{1}{k}\,tr\,(AA_{k-1})$$

（ただし　$A_0=E$）

逆行列　$A^{-1}=C_n^{-1}A_{n-1}$

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix}$$ への応用

対角線要素の表

$A=AA_0$	C_1E	$A_1=AA_0-C_1E$
1	1.6761905	−0.6761905
0.33333333	〃	−1.3428572
0.2	〃	−1.4761905
0.14285714	〃	−1.5333334

AA_1	C_2E	$A_2=AA_1-C_2E$
0.40714283	0.39456347	0.01257936
0.15488094	〃	−0.23968253
0.12837301	〃	−0.26619046
0.098730159	〃	−0.29583331

AA_2	C_3E	$A_3=AA_2-C_3E$
0.004642828	0.0046454747	−0.0000026467
0.004298646	〃	−0.0003468287
0.001788355	〃	−0.0028571197
0.003256595	〃	−0.0013888797

AA_3 対角要素の計算（10^{-4}単位）

noise 部分

$$\begin{array}{r} -0.026467 \\ 0.198322 \\ -0.396891 \\ 0.231428 \\ \hline 0.006392 \end{array} (+$$

$$\begin{array}{r} 0.39645 \\ -1.3227623 \\ 1.4880388 \\ -0.55558860 \\ \hline 0.0061379 \end{array} (+$$

$$\begin{array}{r} -1.0582533 \\ 4.46426 \\ -5.7142394 \\ 2.3148034 \\ \hline 0.0065707 \end{array} (+$$

$$\begin{array}{r} 0.69425 \\ -3.333408 \\ 4.6295868 \\ -1.9841138 \\ \hline 0.0063150 \end{array} (+$$

$\mathrm{trace} = 2.54156 \times 10^{-6}$

$C_4 = 6.3539 \times 10^{-7}$

本 PDF ファイルは 1960 年発行の「第 1 回プログラミング–シンポジウム報告集」をスキャンし、項目ごとに整理して、情報処理学会電子図書館「情報学広場」に掲載するものです。

この出版物は情報処理学会への著作権譲渡がなされていませんが、情報処理学会公式 Web サイトの https://www.ipsj.or.jp/topics/Past_reports.html に下記「過去のプログラミング・シンポジウム報告集の利用許諾について」を掲載して、権利者の捜索をおこないました。そのうえで同意をいただいたもの、お申し出のなかったものを掲載しています。

過去のプログラミング・シンポジウム報告集の利用許諾について

情報処理学会発行の出版物著作権は平成 12 年から情報処理学会著作権規程に従い、学会に帰属することになっています。

プログラミング・シンポジウムの報告集は、情報処理学会と設立の事情が異なるため、この改訂がシンポジウム内部で徹底しておらず、情報処理学会の他の出版物が情報学広場 (=情報処理学会電子図書館) で公開されているにも拘らず、古い報告集には公開されていないものが少からずありました。

プログラミング・シンポジウムは昭和 59 年に情報処理学会の一部門になりましたが、それ以前の報告集も含め、この度学会の他の出版物と同様の扱いにしたいと考えます。過去のすべての報告集の論文について、著作権者（論文を執筆された故人の相続人）を探し出して利用許諾に関する同意を頂くことは困難ですので、一定期間の権利者捜索の努力をしたうえで、著作権者が見つからない場合も論文を情報学広場に掲載させていただきたいと思います。その後、著作権者が発見され、情報学広場への掲載の継続に同意が得られなかった場合には、当該論文については、掲載を停止致します。

この措置にご意見のある方は、プログラミング・シンポジウムの辻尚史運営委員長 (tsuji@math.s.chiba-u.ac.jp) までお申し出ください。

加えて、著作権者について情報をお持ちの方は事務局まで情報をお寄せくださいますようお願い申し上げます。

期間：2020 年 12 月 18 日 〜 2021 年 3 月 19 日
掲載日：2020 年 12 月 18 日

プログラミング・シンポジウム委員会

情報処理学会著作権規程
https://www.ipsj.or.jp/copyright/ronbun/copyright.html