数理解析研究所講究録
第 359 巻 1979 年 152-162

中大型機の基本外部関数の誤差

群馬大　工学部　　春海佳三郎
渡辺　成良
職業訓練大　　　　室伏　誠

1. はじめに

中大型機では、ミニコンに比べて基本外部関数も注意深く作られていて誤差も小さいと考えられているが我々の調査結果では、中大型機でも相当大きい誤差を持ったものもあることが判明した。ミニコンの場合と異なり、未だ解析も十分でなく、機種も数少なく大体の傾向しか報告できないが、特に古い機種の場合誤差は大きいようであり、また各社によって誤差の大きさも差があること、また社によっては最近精度が向上したことも判明したので、ここに報告する。

2. 検定の方法

基本外部関数の誤差を検定するために、次の方法を採用した。

a) 単精度基本外部関数

基本外部関数の単精度と倍精度の差を求め、それを用いて誤差又は、相対誤差を計算する。

b) 倍精度基本外部関数

単精度の場合と同様に倍精度と4倍精度の基本外部関数の差を求める。4倍精度が無い場合は誤差の検定がなされている東大大型計算機センターの4倍精度関数の結果（約35桁の精度）を28桁の結果に縮めて出力したカードを入力したものと、倍精度基本外部関数とを比較して誤差を求める。
28桁としたのは48 bitの倍精度（仮数部88 bit）の検定も可能なようにするためである。

c) これらの誤差を演算レヂスターの最終桁を単位として表示する。

ミニコンの場合と違ってバイトマシンの場合は相対誤差と最終桁の何倍かとは単純には比例しない。このため次のような方法をとった。バイトマシンの場合は、最終桁の大きさ（Last Bit L.B.と書く）は、関数値xが $1 > x \leq 1/16$ $L.B = 2^{-N}$、$16^{-1} > x \geq 16^{-2}$ では L.B. $2^{-N}/16$、$16 > x \geq 1$ では L.B. 16×2^{-N} となる。但しNは仮数部の二進の桁数従って関数値の大きさがわかればL.Bは求まるから誤差をL.Bで割れば誤差がL.Bの何倍かが得られる。

3. 結果の解析

ミニコンと同様に $1/128 \leq x \leq 201/128$ ($\fallingdotseq \pi/2$) 及び $1 \leq x \leq 100$ の区間について前者は 1/128 毎に、後

者は1/2毎に相対誤差を求め．表1に相対誤差の絶対値の最大値を各関数毎に表示した．また図1に誤差曲線の中で明瞭に誤差が表われるものを5種類示す．また表2に誤差を最終桁（L. B）の何倍かで表わしたものの各関数毎の最大値を示す．表1ではC社のみが36 bitのワードマシンで仮数部28 bitでそれ以外は，単精度が24 bit，倍精度が56 bitのバイトマシンである．従って相対誤差は$16 \times 2^{-24} \doteqdot 9.5367$ E-7，倍長で$16 \times 2^{-56} \doteqdot 2.2204$ E-16の数倍迄は許容誤差と考えてよい．

表1ではA社の二種類のコンパイラA1，A2，B社の2種類のコンパイラB1，B2，C社の2種類のコンパイラC1，C3は何れも許容誤差の範囲内に入っている．これに対してD社及びE社のコンパイラは何れも許容誤差以上の誤差となっている．D社ではCOS，TANH，ALOG，DCOS，DTANH，DLOG，が許容誤差以上となっている．E社ではSIN，COS，ALOG，DSIN，DCOS，等が許容誤差を越えている．

図1にこれらの相対誤差の誤差曲線の中の顕著なものを示す．

1) D社及びE社のSIN，COS，DSIN，DCOSの相対誤差は何れも関数値が小さいところで大きくなっている．これはこれらの関数の近似が絶対誤差型の近似式を用いて

いるため 図1−aに示されるように関数値の小さいところで相対誤差が大きくなっている。これを表2で見るとD社及びE社のSIN, COS, DSIN, DCOSの内誤差の大きいものは最終桁の100〜6000倍になっている。これは他の関数の誤差に比べて異常に大きい値となっている。相対誤差型の関数近似を使用することによってこれらの誤差を小さくすることができる。これはミニコンの基本外部関数の誤差の場合にも指摘されたことである。

2) 次にSIN, COS, DSIN, DCOSで下段の $1 \leq x \leq 100$ B2のDSIN, DCOS, DのSIN, DSIN, DCOS, EのDSIN等は上段の $0 \leq x \leq \pi/2$ の値より大きくなっている。これは図1−bに示されるように11の整数倍の22, 44, … 又は11, 33 …等の点で誤差が大きくなっていることがわかる。これは $22 - 7\pi = 0.008851$ 等と4桁桁落ちするためでこれを防止するためにはこの部分を倍長計算をする必要がある。また1)で述べたように argument の小さい点で相対誤差が大きいこともこれらの11の整数倍の点で誤差が大きくなる一因である。

3) D及びEのEXP及びDEXPは2)と同様に下段の $1 \leq x \leq 100$ の相対誤差が上段の $0 \leq x \leq \pi/2$ に比べて大きくなっている。これもミニコンの場合と同様に

argument　を $0<x<\ln 2$ 又は $0<x<\frac{1}{4}\ln 2$ にする場合 $x=N\ln 2+y$ とし $0<y<\ln 2$ 又は $0<y<\frac{1}{4}\ln 2$ とする場合 $y=x-N\ln 2$ の計算を倍長計算しないと桁落ちを生じて図1－Cのように $\ln 2$ 毎に誤差が大きくなっていくからである．

4）DのTANHが 0.33062E－4 となっているのは TANHを作るのに（EXP（2X）－1）／（EXP（2X）＋1）としたためxが小さいところで桁落ちを生じて精度が落ちるためである．xの小さいところは Taylor 展開等を用いるべきである．

5）EのTANHは誤差に記入されていないが，これは DTANHとの混合演算ができないために誤差として大きい値が出たものでコンパイラのミスのためである．

6）表2のC社の三種のコンパイラC1，C2，C3の結果から見るとC社では最初C1では最終桁（L.B.）の15倍程度有った誤差が最近のC3では何れもL.B.の1倍以下で誤差0と考えてよい程度に改善されている．

7）6）で述べたC3の程度に誤差を小さくすることができるがこのためには関数近似の計算の全部又は一部に倍長計算を行なう必要がある．

8）この調査中にコンパイラの訂正版がATANが10^{-7}程度

の誤差を持つことを発見した．これはコンパイラの不良によるものであるが訂正前は10^{-5}程度の誤差であったのでコンパイラを訂正するには十分な注意をはらう必要があることを示している．

4．結論

以上の調査によって中大型の基本外部関数中にも相当大きい誤差を持つものがあることが判明した．又同一社のものでもコンパイラによって誤差が違い最近は最終桁の1倍以下に迄誤差が小さくなるよう改善された良いコンパイラが出現したことが判明した．このように非常に良いコンパイラ（誤差＜1．L.B.）から悪いコンパイラ（誤差＞1000 L.B）迄相当バラツキが有るのが現状である．

この報告ではまだ一部のコンパイラしかチェックされていないが次のように要約される．

1）この論文では誤差のチェック方法を示した．
　この方法の誤差は1 L.B以下である．

2）更に各コンパイラについて誤差の検定を行なう必要がある．

3）許容誤差は単精度で4 L.B．以下倍精度で15 L.B．以下と考えてよいのではないか．

4）C社のように部分的又は全面的に倍精度計算を用いるこ

とによって誤差を1L.B.以下にすることができる.

5) A.B.C 三社の最近のコンパイラは何れもこの許容誤差以内である. D.E両社のコンパイラは最大約1600L.B. という大きい誤差を持つものもあった.

6) 誤差の原因の一つはSIN, COS の関数近似に絶対誤差型の近似を用いるためで相対誤差型の近似を用いること.

7) 他の一つの原因はTANHでxの argument が小さい部分で$(\exp(2x)-1)/(\exp(2x)+1)$の型の近似を用いるために桁落ちによる誤差が生じるものである.

8) argument が大きいところで argument reduction を行なう部分を倍長演算で行なう必要がある.

9) コンパイラのミスのため大きい誤差を生じたと考えられるものが2種類発見された. 従ってコンパイラの修正を安易に行なうべきでない. 又コンパイラは新しい間は虫(bug)が有るのは当然であることを user は知る必要がある.

5 謝辞

色々御協力, 討論, 御教示いただいた永坂, 平野, 山下, 西見の各先生外諸先生方に厚く御礼申し上げます. 又, 東京水産大学の細川君達, 日産化学の阿部秀夫様, 並びに, 各計

算センター及び電々公社DEMOS担当の方々に厚く御礼申し上げます.

参考文献

1) 春海, 桧山, 小竹 : ミニコンによる数値計算のおとしあな(1)~(16)
"bit" vol 8 No. 3 4~9 (1976)

2) 春海, 小竹, 桧山 : ミニコンの基本外部関数の誤差曲線について ; 京大数理解析研講究録 310, 83 (1978)

表 1

	A1	A2	B1	B2	D	E
SQRT	0.90321E-6	0.44476E-6	0.90321E-6	0.54354E-6	0.16710E-5	0.92082E-6
EXP	0.88345E-6	0.44097E-6	0.48689E-6	0.44097E-6	0.33302E-5	0.85737E-6
SIN	0.44545E-6	0.44545E-6	0.10522E-5	0.31793E-6	0.10522E-5	0.46095E-5
COS	0.78083E-6	0.78083E-6	0.11052E-5	0.72108E-6	0.64926E-3	0.32745E-4
ATAN	0.63362E-6	0.11241E-5	0.11258E-5	0.63362E-6	0.63362E-6	0.12708E-5
TANH	0.26671E-6	0.26671E-6	0.71041E-6	0.71041E-6	0.33062E-4	
ALOG	0.78961E-6	0.81769E-6	0.10021E-5	0.70891E-6	0.11641E-3	0.23437E-1
SQRT	0.49973E-6	0.27000E-6	0.90321E-6	0.27000E-6	0.95367E-6	0.49973E-6
EXP	0.77373E-6	0.40097E-6	0.43188E-6	0.42929E-6	0.34385E-4	0.51701E-5
SIN	0.11226E-5	0.11226E-5	0.77492E-6	0.71663E-6	0.37140E-4	0.89417E-5
COS	0.95367E-6	0.95367E-6	0.11052E-5	0.46255E-6	0.33578E-3	0.12001E-4
ATAN		0.80692E-6	0.80692E-6	0.80692E-6	0.11906E-5	0.64560E-6
TANH		0.47034E-7	0.58174E-7	0.53731E-7	0.18877E-5	
ALOG		0.57387E-6	0.57387E-6	0.42769E-6	0.21521E-5	0.69068E-6
DSQRT	0.21541E-15	0.2195E-15	0.2067E-15	0.2195E-15	0.4246E-15	0.2203E-15
DEXP	0.21188E-15	0.2119E-15	0.1839E-15	0.2119E-15	0.5585E-15	0.4216E-15
DSIN	0.14825E-15	0.1483E-15	0.1210E-15	0.2221E-15	0.2504E-15	0.2265E-13
DCOS	0.14749E-15	0.1475E-15	0.2483E-15	0.5602E-13	0.1265E-12	0.4164E-12
DATAN	0.17800E-15	0.2212E-15	0.2243E-15	0.2221E-15	0.3017E-15	0.2221E-15
DTANH	0.22007E-15	0.1619E-15	0.2137E-15	0.3331E-15	0.2776E-13	
DLOG	0.22268E-15		0.2375E-15		0.4943E-13	
DSQRT	0.11102E-15	0.1570E-15	0.8865E-16	0.1570E-15	0.2514E-15	0.2220E-15
DEXP	0.24970E-15	0.2497E-15	0.1811E-15	0.2497E-15	0.3781E-14	0.3908E-14
DSIN	0.20925E-15	0.2093E-15	0.2690E-15	0.1245E-12	0.2010E-12	0.2721E-12
DCOS	0.19598E-15	0.1960E-15	0.2693E-15	0.1539E-12	0.2011E-12	0.3060E-12
DATAN	0.17777E-15	0.1866E-15	0.1734E-15	0.1483E-15	0.2090E-14	0.3731E-15
DTANH	0.86374E-16	0.1533E-16	0.8063E-16	0.1822E-16	0.4505E-15	
DLOG	0.20211E-15		0.8257E-16		0.3765E-15	

	C1	C3
SQRT	0.68986E-8	0.27647E-16
EXP	0.68158E-8	0.13794E-8
SIN	0.95792E-8	0.51273E-9
COS	0.48755E-7	0.42891E-9
ATAN	0.45929E-7	0.12326E-10
TANH		0.68313E-8
ALOG		0.31068E-8
SQRT		0.27647E-16
EXP		0.13501E-8
SIN		0.37399E-9
COS		0.50217E-9
ATAN		0.30356E-11
TANH		0.36988E-8
ALOG		0.29846E-9

相対誤差の最大値：上段は $0 \le x \le 1/128$，下段は $1 \le x \le 100$

表 2

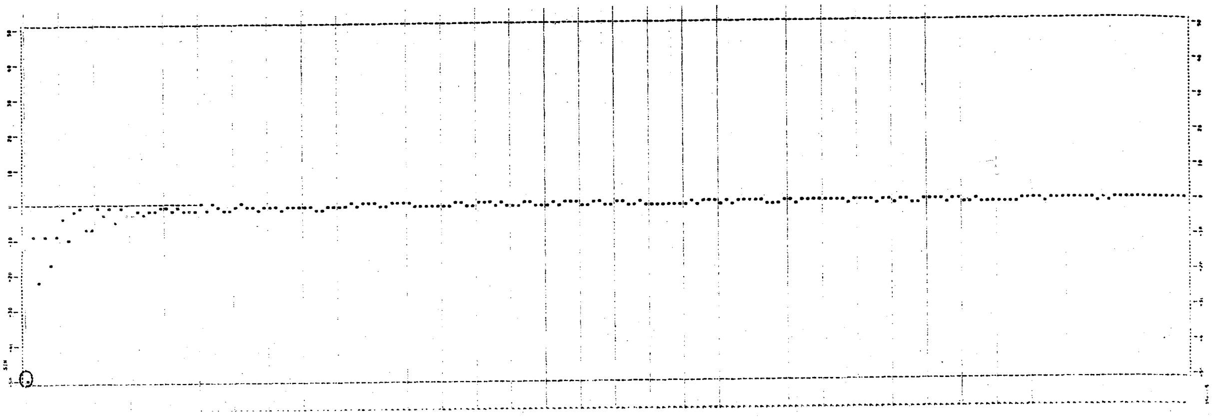
	A 1		A 2		B 1		B 2
	Single	Double	Single	Double	Single	Double	Single
SQRT	.962	.994	1		7.931	12	.5
EXP	.991	1.026	1	1	.775	6	1.0
SIN	4.523	3.682	1		5.466	14	
COS	5.832	3.249	1	1	8.141		
ATAN	2.458	3.277	2.35	2	5.930	11	.7
ALOG	1.048	2.310	1		3.193		.8
TANH	.528	6.216			2.872		

	C 1	C 2	C 3	D		E	
	Single	Single	Single	Single	Double	Single	Double
SQRT	2.0	1.74	0.0	6.0	2	2	10
EXP	1.2	1.0	0.5	11.6	7		7
SIN	6.9	1.62	0.5	138.7	7		208
COS	10.4	1.91	0.5	1349.0	1850	67	6111
ATAN	1.0	0.928	0.5	4.0	19	2	4
ALOG	14.88		0.5		457		
TANH			0.5	67.4			

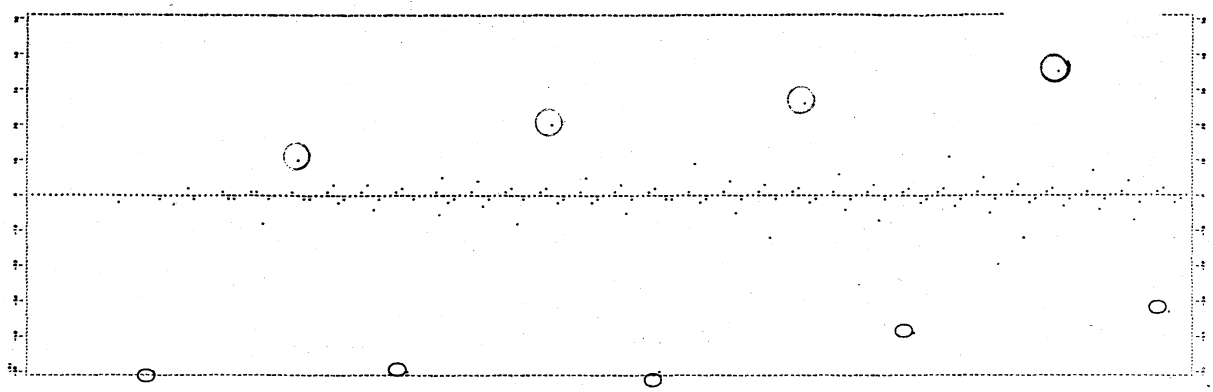
最終桁(L.B.)単位の誤差の最大値

図 1

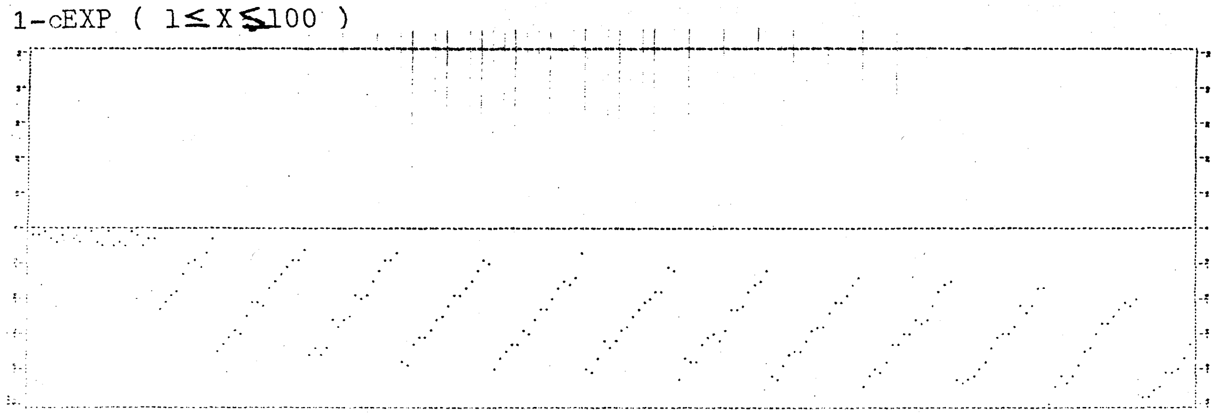
1-a SIN (0≤X≤1/128)



1-b COS (1≤X≤100)



1-cEXP (1≤X≤100)



相対誤差の誤差曲線