論　文

有理式近似および連分数近似の最良化について*

浜　田　穂　積**

Abstract

A practical and efficient method to calculate minimax approximation for rational expression and continued fraction is presented. It is shown that the minimax approximation of continued fraction is more precise and more rapidly convergent in comparison with that of rational expression.

A method to decide the significant position of digit of minimax approximation coefficients is also presented. By this method, it is shown that, in the case of some favorable functions, high precision of object approximation is attainable even if the coefficients are calculated by relatively lower precision.

1. はじめに

Maehly[1]，山内[2]などにより，有理式近似の最良化は研究されている．しかし，手計算による式の展開などのため，高次の公式の計算は現実にはかなり困難である．一方多項式近似の最良化については，計算機を用いる単純で汎用的な手法が種々開発されており，高次の公式を得るにはむしろその方が現実的である．有理式の最良化に，多項式における手法をとり入れることもできるが，結果は多項式の場合ほど劇的な効果は得られない．しかしながら，連分数展開の収束する場合には，高次の非線型性の影響を弱める方法をとることにより，係数に対する補正量を計算する方法でよい結果を得ることができる．

2. 有理式近似

最良近似式は，有理式近似に限らず，次のものを定めなければならない．

（a）　近似式を適用する独立変数の区間(近似区間)

（b）　近似式の形

この章では次のように仮定する．これは特に一般性を損なうものでなく，容易に一般の場合に拡張できる．

近似区間：

$$-\rho \leqq x \leqq \rho \quad (\rho>0) \tag{1}$$

近似式の形：

$$f(x)=\frac{p_0+p_1x+\cdots+p_{n_1}x^{n_1}}{q_0+q_1x+\cdots+q_{n_2}x^{n_2}} \tag{2}$$

$f(x)$ が偶関数のとき

$$f(x)=\frac{p_0+p_1x^2+\cdots+p_{n_1}x^{2n_1}}{q_0+q_1x^2+\cdots+q_{n_2}x^{2n_2}} \tag{3}$$

$f(x)$ が奇関数のとき

$$f(x)=\frac{p_0x+p_1x^3+\cdots+p_{n_1}x^{2n_1+1}}{q_0+q_1x^2+\cdots+q_{n_2}x^{2n_2}} \tag{4}$$

近似式の形については，分母，分子の次数のほかに自由に定められる係数の個数，すなわち自由度も重要であるが．今の場合，分母，分子に同じ数を掛けても式の値は変らないため自由度は1だけ少なく

$$\text{自由度}=n_1+n_2+1 \tag{5}$$

である．

2.1 最良近似有理式の計算法

最良近似式の係数を精度よく計算する方法は

「求めようとする係数を，いくらでも正確に求められる数と，それからの偏差に分解し，後者を収束計算によって求める」

という戸田[2]の方法を，有理式の場合にあてはめるものである．ここでの偏差を0とした近似式を，基準の近似式と呼ぶことにする。こうすると，最良近似式は基準の近似式の係数を補正して得る，といえる．

---

* Minimax Approximation for Rational Expressions and Continued Fractions by Hozumi HAMADA (Systems Development Laboratory, Hitachi, Ltd.).

** (株)日立製作所システム開発研究所

ここでは基準の近似式をパデ展開にとる．被近似関数のマクローリン展開の係数が有理数であれば，計算過程から明らかなようにパデ展開の係数をすべて整数とすることができる．ここではこの基準の近似式を次の通りとする．

$$\frac{\zeta_0(x)}{\eta_0(x)}\quad\left(\begin{matrix}\zeta_0: n_1\text{ 次式}\\ \eta_0: n_2\text{ 次式}\end{matrix},\ n=n_1+n_2+1\right)\qquad(6)$$

一方求める最良近似式を次の通りとする．

$$\frac{\zeta(x)}{\eta(x)}\qquad(7)$$

ただし，$\zeta(x)$ は $\zeta_0(x)$ と，$\eta(x)$ は $\eta_0(x)$ とそれぞれ同次の多項式とする．また，自由度は係数の個数より1だけ小であるので，1つの係数を固定する必要がある．ここでは $\eta(x)$ と $\eta_0(x)$ の0次の係数は等しいものとする．またマクローリン展開の存在を前提にしているので，この0次の係数は0ではない．したがって，(6)を定めれば，(7)の各係数は一意的に定まる．

ここでの目的は次を求めることにある．

$$\left.\begin{aligned}\zeta_1(x)&=\zeta(x)-\zeta_0(x)\\ \eta_1(x)&=\eta(x)-\eta_0(x)\end{aligned}\right\}\qquad(8)$$

一方，被近似関数についても，有限の手順で計算するからには近似式であるが，目的の近似式とのバランス上，やはり整数係数のパデ展開とする方が扱いやすい．この整数係数という条件は本質的なものではないが，計算精度を少しでも損なわないようにするため，できるものはそうする方がよい．パデ展開された被近似関数の分子，分母をそれぞれ

$$Z(x),\quad H(x)\qquad(9)$$

とする．ただし，十分精度のよい公式でなければならない．

次に最良近似の条件として，ある評価式 $E(x)$ を設定し，次の2条件を満たすように近似式の係数を定めるものとする．

(A)　近似区間の両端を含む $n+1$ 個の点で極値をとる．これらの点を x_i とするとき

$$x_0(=-\rho)<x_1<x_2<\cdots<x_n(=\rho)\qquad(10)$$

(B)　ある定数 $\bar{\varepsilon}$ が存在して，次の関係を満たす．

$$\left.\begin{aligned}\bar{\varepsilon}&=(-1)^i E(x_i)\quad(i=0,1,2,\cdots,n)\\ \varepsilon&=|\bar{\varepsilon}|\end{aligned}\right\}\qquad(11)$$

ここで $E(x)$ が計算可能か調べる．(10)において，x_0, x_n は固定されているが，$x_1,\cdots,x_{n-1}$ の $n-1$ 個の極値をとるという条件式があり，(11)は $n+1$ 個の条件式で，計 $2n$ 個の式である．一方未知数は(10)に $n-1$ 個，(11)に ε と $E(x)$ に自由度 n の未知数があるので計 $2n$ 個となって，解けることがわかる．実際にはこの $2n$ 元の連立方程式は非線型の方程式のため容易には解けず，ある種の収束計算によるほかない．

$E(x)$ として何をとるかは，最良近似の目的による．絶対誤差に関して最良というのであれば，絶対誤差関数を，相対誤差に関して最良というのであれば，相対誤差関数を $E(x)$ とする．すなわち

絶対誤差近似のとき

$$E(x)=\frac{\zeta(x)}{\eta(x)}-\frac{Z(x)}{H(x)}=\frac{H(x)\zeta(x)-\eta(x)Z(x)}{\eta(x)H(x)}\qquad(12)$$

相対誤差近似のとき

$$E(x)=\left\{\frac{\zeta(x)}{\eta(x)}-\frac{Z(x)}{H(x)}\right\}\Big/\frac{Z(x)}{H(x)}$$
$$=\frac{H(x)\zeta(x)-\eta(x)Z(x)}{\eta(x)Z(x)}\qquad(13)$$

これらの評価式 (12)，(13) は，いずれも分子に多項式 $H(x)\zeta(x)-\eta(x)Z(x)$ を含む．この式の値を精度よく計算することが，最良近似を精度よく計算するためのポイントである．しかしながら，この式の値は $H(x)$，$Z(x)$，$\eta(x)$，$\zeta(x)$ の値のオーダと比べて小さいので，精度を上げる工夫が必要である．そこでこの式に(8)を代入してみると

$$\begin{aligned}H(x)\zeta(x)-\eta(x)Z(x)&\\ =\{H(x)\zeta_0(x)-\eta_0(x)Z(x)\}&\\ +\{H(x)\zeta_1(x)-\eta_1(x)Z(x)\}&\end{aligned}\qquad(14)$$

となるが，第2行は ζ_0/η_0 が被近似関数のパデ展開であることから，n 次の係数まで0で，比較的精度よく計算できる．したがって第3行の計算を工夫すれば，精度よく計算できるはずである．すなわち

$$\left.\begin{aligned}\eta_1(x)&=c_1x+c_2x^2+\cdots+c_{n_2}x^{n_2}\\ \zeta_1(x)&=c_{n_2+1}+c_{n_2+2}x+\cdots+c_nx^{n_1}\end{aligned}\right\}\qquad(15)$$

とするとき，行列を用いて表わせば

$$H(x)\zeta_1(x)-\eta_1(x)Z(x)=(1\ x\ x^2\cdots)$$
$$\times\begin{pmatrix}0 & \cdots & \cdots & H(x) & \cdots & \cdots & 0\\ -Z(x) & \ddots & & H(x) & \ddots & & \\ -Z(x) & 0 & & & \ddots & & \\ \vdots & \ddots & -Z(x) & & & H(x) & \\ & & -Z(x) & & & H(x) & \\ & & \vdots & H(x) & & \vdots & \\ -Z(x) & & & & & & \\ 0 & \cdots & -Z(x) & 0 & \cdots & 0 & H(x)\end{pmatrix}\begin{pmatrix}c_1\\ c_2\\ \vdots\\ c_n\end{pmatrix}$$
$$\underbrace{\qquad\qquad}_{n_2}\qquad\underbrace{\qquad\qquad}_{n_1+1}$$

となるが，後2項の積をとって

$$=(1\ x\ x^2\cdots)\begin{pmatrix}b_1\\ b_2\\ \vdots\end{pmatrix} \tag{16}$$

と表わす．この $b_1, b_2, \cdots$ は，$c_1, c_2, \cdots, c_n$ の一次結合で，係数は $H(x)$，$Z(x)$ のみで定まる定数である．ところが，この係数の大きさと，$b_1, b_2, \cdots$ の結果の値の大きさとは，やはりかなり開きがあるため，$c_1, c_2, \cdots$ の補正値を計算して正しい値に近づけるという通常の計算方法はあまりうまくいかない．そこで，$c_1, c_2, \cdots, c_n$ を逆に解いて，$b_1, b_2, \cdots, b_n$ の一次結合として表わし，$b_1, b_2, \cdots, b_n$ に対する補正値を計算して，正しい b_i，次に正しい c_i を計算する．すなわち (16) の行列を，上の正方行列部分 A と，その下の残りの部分 B に分けると

$$H(x)\zeta_1(x)-\eta_1(x)Z(x)$$

$$=(1\ x\ x^2\cdots)\begin{pmatrix}A\\ \cdots\cdots\\ D\end{pmatrix}\begin{pmatrix}c_1\\ c_2\\ \vdots\\ c_n\end{pmatrix} \tag{17}$$

と表わし，A を取り出して，以後列ベクトル (b_i)，(c_i) をそれぞれ B，C と表わすと

$$AC=B \tag{18}$$

この両辺に左から A^{-1} を掛けると

$$C=A^{-1}B \tag{19}$$

となるが，これを (17) に代入すると

$$H(x)\zeta_1(x)-\eta_1(x)Z(x)=(1\ x\ x^2\cdots)\begin{pmatrix}A\\ \cdots\cdots\\ D\end{pmatrix}A^{-1}B$$

$$=(1\ x\ x^2\cdots)\begin{pmatrix}I\\ \cdots\cdots\cdots\\ DA^{-1}\end{pmatrix}B \tag{20}$$

となる．この右辺の最初の2項を掛けて

$$H(x)\zeta_1(x)-\eta_1(x)Z(x)$$

$$=(p_1(x)p_2(x)\cdots p_n(x))B \tag{21}$$

とする．この $p_i(x)$ $(i=1,2,\cdots,n)$ の計算は，左辺を直接計算するのと比べて精度の点でかなり有利であることがわかる．そこで計算を B について行い，最後の結果を (19) に代入して c を求めることにする．また

$$p_0=H(x)\zeta_0(x)-\eta_0(x)Z(x) \tag{22}$$

とする．分母については，これも (12)，(13) の分母に代入することにより

$$q_0(x)+(q_1(x)q_2(x)\cdots q_n(x))B \tag{23}$$

と表わせる．これらから，$E(x)$ は一般的に次の形で表わせる．

$$E(x)=\frac{p_0(x)+\sum_{i=1}^{n}p_i(x)b_i}{q_0(x)+\sum_{i=1}^{n}q_i(x)b_i} \tag{24}$$

以上により，次の手順で計算する．

(a)　(16) の行列の各要素を計算する．

(b)　A^{-1} を計算する．

(c)　$p_i(x)$，$q_i(x)$ $(i=0,1,2,\cdots,n)$ の係数を計算する．

(d)　b_i の初期値を計算する．チェビシェフ補間によるときは

$$x_j=\rho\cos\left(\frac{2j-1}{n}\cdot\frac{\pi}{2}\right)\quad(j=1,2,\cdots,n) \tag{25}$$

において (24) の分子を0とする b_i を求める．

(e)　$E(x)$ が極値をとる x の近似値の初期値を設定する．例えば (25) に対応して次をとる．

$$x_j=\rho\cos\left(\frac{2j}{n}\cdot\frac{\pi}{2}\right)\quad(j=0,1,2,\cdots,n) \tag{26}$$

(f)　b_i の補正量を計算して補正する．すなわち

$$\sum_{i=1}^{n}\left\{\frac{\partial E(x_{j-1})}{\partial b_i}+\frac{\partial E(x_j)}{\partial b_i}\right\}db_i$$

$$=-\{E(x_{j-1})+E(x_j)\}\quad(j=1,2,\cdots,n) \tag{27}$$

を db_i について解き，b_i+db_i を新しい b_i とする．これを数回繰返す．

(g)　極値を与える x を修正する．すなわち，初期値が (e) で与えられる x_j のうち，両端を除いたものについて，$E(x)$ の真の極値を与える x を x_j の近傍で探し，x_j を新しいもので置き換える．

(h)　極値の絶対値のばらつきを調べ，収束したと見なせるようになるまで数回 (f)，(g) を繰返す．

(i)　求められた b_i を (25) に代入して c_i を求める．

以上で c_i が計算できたので，$\eta(x)$，$\zeta(x)$ も計算できたことになる．

2.2 数 値 例

正接 (tan) の場合について，次の条件で計算した．

近似区間：$-\pi/4\leq x\leq\pi/4$

近似方式：相対誤差方式

使用計算機および言語：HITAC-8700

EDOS-MSO ALGOL

計算精度：倍精度 (16進 14 桁)

計算結果は **Table 1**(次頁参照)に示す通りである．この表に記入した結果の有効桁数については，第4章を参照されたい．この場合相対誤差関数は偶関数のため，計算法に多少の変更を要するが，本質的には同様

Table 1 Constants of minimax rational approximation for tan function.

$$\tan x=\frac{a_1x+a_3x^3+\cdots+a_{i+1}x^{i+1}+\cdots}{a_0+a_2x^2+\cdots+a_ix^i+\cdots} \quad |x|\leq\frac{\pi}{4}$$

n	i	a_i	MRE
2	0	3	1.302E−03
	1	3−.0039063	
	2	−1−.0436982	
3	0	15	5.773E−06
	1	15+.0000014872	
	2	−6−.02538215	
	3	−1−.02764068	
4	0	105	1.416E−08
	1	105−.0000014872	
	2	−45−.0592563375	
	3	−10−.0591858053	
	4	1+.0202700272	
5	0	945	2.209E−11
	1	945+.000000020872	
	2	−420−.236558438117	
	3	−105−.236560008564	
	4	15+.094855112179	
	5	1+.016021150229	
6	0	10395	2.383E−14
	1	10395−.00000000024775	
	2	−4725−1.37585383397626	
	3	−1260−1.37585380684005	
	4	210+.59008651801545	
	5	21+.13146809208241	
	6	−1−.01325141140350	
7	0	135135	1.886E−17
	1	135135+.0000000000025480	
	2	−62370−10.58672355574040xx	
	3	−17325−10.58672355612337xx	
	4	3150+4.706663766826519x	
	5	378+1.177755924312213	
	6	−28−.1686256527065258	
	7	−1−.011300957111540	
8	0	2027025	1.141E−20
	1	2027025−.000000000000023120	
	2	−945945−101.6887331705855$xxxxx$	
	3	−270270−101.6887331705810$xxxxx$	
	4	51975+46.22942620394436$xxxx$	
	5	6930+12.3331818136010$4xxxx$	
	6	−630−2.057437320411063xxx	
	7	−36−.2061263400188593xx	
	8	1+.009852334635462960	

x shows the position of digit that can't be got by double precision calculation.

である．

この結果によると，最大相対誤差は $n=8$ においても正しく計算されている（次章の結果と同じ）にもかかわらず，$n=7$ 以上では補正値が必要な精度まで得られていない．この理由は，b_i は正しく計算できているにもかかわらず，(19)から c_i を求める精度が失われるためである．それは補正量が大きすぎたためである．補正量は近似区間が狭くなれば小となる．この近似区間では $n=6$ まで(少し甘くすれば $n=7$ まで)計算できたといえるが，この結果は，目的の近似式の精度と，近似式の計算精度が同程度まで計算できたということである．

3. 連分数近似

前章の結果では，パデ展開の係数の補正による最良化は，大まかにいって，目標の精度が計算精度と同程度のところまでかろうじて計算できる．という意外に期待はずれの結果となった．その理由として考えられることは，関数値の計算の精度に影響する度合の強い分母分子の低次の係数ほど，補正量の絶対値が大である（最低次を除いて）という好ましくなく傾向のためと思える．それに反して，ここで得られた結果の近似式を連分数の形に展開すると，被近似関数を連分数展開したものとの，対応する係数の差は，関数値の計算の精度に影響する度合の強い項のものほど絶対値が小となり，好都合である．そこで連分数を基にした最良化を直接実行できればそれに越したことはなさそうである．それには有理式のときと同様に，連分数展開の定数の補正量を計算するのであるが，これは次のように行う．以上において近似式を次の形とする．

$$g(x)=a_1+\frac{b_1x}{a_2+}\,\frac{b_2x}{a_3+}\,\cdots\,+\,\frac{b_{n-1}x}{a_n} \tag{28}$$

近似区間については前章と同じとする．

3.1 最良近似連分数の計算法

ここでは連分数の計算を次の漸化式で計算するものとする．まず被近似関数を

$$\left.\begin{aligned} f_m&=a_m\\ f_i&=a_i+b_ix/f_{i+1}\quad(i=m-1,m-2,\cdots,1)\end{aligned}\right\} \tag{29}$$

とし，近似式も同様に

$$\left.\begin{aligned} g_n&=a_n+d_n\\ g_i&=a_i+d_i+b_ix/g_{i+1}\quad(i=n-1,n-2,\cdots,1)\end{aligned}\right\} \tag{30}$$

とする．こうすると f_1 が被近似関数，g_1 が近似式である．ただし m は n より大で，f_1 が g_1 と比べて十分精度よく真の関数に近いように選ぶものとする．この d_i が求めようとする補正量である．

誤差の評価式 $E(x)$ はは次の通りである．

絶対誤差近似：

$$E(x)=g_1-f_1$$

$$=d_1+\frac{b_1x}{g_2}-\frac{b_1x}{f_2}=d_1-\frac{b_1x}{f_2g_2}(g_2-f_2)$$

……

$$=d_1-\frac{b_1x}{f_2g_2}\left(d_2-\frac{b_2x}{f_3}\cdot\cdots\cdot\frac{1}{g_n}\left(d_n-\frac{b_nx}{f_{n+1}}\right)\cdots\right) \quad (31)$$

相対誤差近似：

$$E(x)=\frac{1}{f_1}\left(d_1-\frac{b_1x}{f_2g_2}\right)\left(d_2-\frac{b_2x}{f_3}\cdot\cdots\cdot\frac{1}{g_n}\left(d_n-\frac{b_nx}{f_{n+1}}\right)\cdots\right) \quad (32)$$

これを漸化式で表わすと

$$\left.\begin{aligned}&s_{n+1}=-1\\&s_i=-(s_{i+1}b_ix/f_{i+1}+d_i)/g_i\\&\qquad(i=n,n-1,\cdots,1)\end{aligned}\right\} \quad (33)$$

および

絶対誤差近似：

$$E(x)=-g_1s_1 \quad (34)$$

相対誤差近似：

$$E(x)=-\frac{g_1}{f_1}s_1$$

である．この (31), (32) の特徴は，d_i について陽に一次結合になっていることである．もちろん g_i も d_i を含んではいるが，g_i の中に含まれていて $E(x)$ の計算であまり大きな影響力がなければ，陽に表わされている d_i の方が主要な働きをする．

次に d_i の補正量の計算に用いる $\partial E/\partial d_i$ は次の通りである．(以下絶対誤差の場合．相対語差も同様)

$$\frac{\partial E}{\partial d_i}=\frac{\partial E}{\partial g_1}\cdot\frac{\partial g_1}{\partial d_i}=\frac{\partial g_1}{\partial d_i}=\left(-\frac{b_1x}{g_2^2}\right)\frac{\partial g_2}{\partial d_i}$$

$$\cdots\cdots\cdots\cdots$$

$$=(-1)^{i-1}\frac{b_1b_2\cdots b_{i-1}x^{i-1}}{(g_2g_3\cdots g_i)^2}\frac{\partial g_i}{\partial d_i}$$

$$=(-1)^{i-1}\frac{b_1b_2\cdots b_{i-1}}{(g_2g_3\cdots g_i)^2}\cdot x^{i-1} \quad (35)$$

これを漸化式で表わすと次の通りである．

$$\left.\begin{aligned}&\frac{\partial E}{\partial d_1}=1\\&\frac{\partial E}{\partial d_i}=-\frac{b_{i-1}}{g_i^2}x\frac{\partial E}{\partial d_{i-1}}\quad(i=2,3,\cdots,n)\end{aligned}\right\} \quad (36)$$

相対誤差近似のときは $\partial E/\partial d_1=1/f_1$ である．

被近似関数が偶関数，あるいは奇関数のときもほとんど同様である．また (28) の逆数の形のときもあるが，これもほとんど同様である．

以上の結果を用いて，前章とほぼ同様の手順で d_i を計算する．すなわち

(a) d_i の初期値を求める．それには，例えばチェビシェフ補間を用いる．このとき g_i がまだ定まってないが，そのために (30) において $d_i=0$ とおいて求める．この仮定は，被近似関数がいわゆる収束の速いもののときはうまくゆく．

(b) $E(x)$ が極値をとる x の近似値の初期値を設定する．

(c) d_i の補正量を計算して補正する．これを数回繰返す．

(d) 極値を与える x を修正する．

(e) 極値の絶対値のバラツキを調べ，収束したと見なせるようになるまで数回 (c), (d) を繰返す．

Table 2 Constants of minimax continued fraction approximation for tan function.

$$\tan x\simeq\frac{x}{1+d_1+}\frac{-x^2}{3+d_2+}\cdots+\frac{-x^2}{2i-1+d_i+}\cdots+\frac{-x^2}{(2n-1)+d_n},\quad |x|\le\frac{\pi}{4}$$

tan(x)

n	i	d_i	MRE	
				15.38
2	1	1.3037[9]29752988941E-03	1.302E-03	2.99
	2	−1.29348[5]3229009223E-01		1.86
				5.45
3	1	−5.7731[3]83860934924E-06	5.773E-06	5.34
	2	1.369498[3]148858293E-03		4.22
	3	−1.366781[4]820018976E-01		2.63
				4.95
4	1	1.4163[8]13770077187E-08	1.416E-08	7.95
	2	−6.07588[5]4900322005E-06		6.83
	3	1.1660949[2]57125745E-03		5.23
	4	−1.4065716[0]82610192E-01		3.35
				4.81
5	1	−2.2086[9]5454085690 0E-11	2.209E-11	10.76
	2	1.500075[4]235478002E-08		9.63
	3	−4.6210355[2]25076098E-06		8.04
	4	9.9209218[5]93883622E-04		6.15
	5	−1.4316381[5]02425766E-01		4.05
				4.66
6	1	2.3833[1]73984498656E-14	2.383E-14	13.73
	2	−2.353999[7]099298057E-11		12.60
	3	1.0619908[0]11430025E-08		11.01
	4	−3.4755221[7]06645547E-06		9.12
	5	8.5748578[5]77393368E-04		7.00
	6	−1.4488807[0]57214170E-01		4.72
				4.53
7	1	−1.8855[0]76064217561E-17	1.886E-17	16.83
	2	2.554054[1]181356432E-14		15.70
	3	−1.5865284[5]83895318E-11		14.11
	4	7.3125713[8]56121008E-09		12.22
	5	−2.6719597[9]84563110E-06		10.11
	6	7.5299957[3]12680196E-04		7.82
	7	−1.4614685[9]23418696E-01		5.39
				4.48
8	1	1.1405[4]944643899 31E-20	1.141E-20	20.05
	2	−2.029938[1]982177452E-17		18.92
	3	1.6608707[2]64432188E-14		17.33
	4	−1.0225133776364552E-11		15.44
	5	5.1403248047708964E-09		13.32
	6	−2.1060008164[3]98635E-06		11.04
	7	6.7033524[5]17141616E-04		8.60
	8	−1.4710625[6]79039790E-01		6.05

MRE represents maximum relative error.
The rightmost column expresses $\log_{10}(\max_j|\partial E(x_j)/\partial d_i|/\varepsilon)$.

以上で d_i が求められた.

3.2 数 値 例

前章と同じ条件で，正接(tan) の場合について計算した．tan の連分数展開形は次のものを用いた.

$$\tan x=\frac{x}{1+}\ \frac{-x^2}{3+}\ \frac{-x^2}{5+}\ \cdots\ \frac{-x^2}{2i-1+}\ \cdots \qquad (37)$$

結果は **Table 2** に示す通りである．極値の絶対値のバラツキにおいては，その最大のもの e_1，最小のもの e_2 について，$e_2>(1-10^{-4})e_1$ となるまで繰返したが，そのとき手順（c），（d）の繰返しはいずれも2回であった．また定数の表示桁数は，上述のバラツキに見合って必要な桁数のところに印をつけてある.

この結果を見てわかることは，まず当初の予想通り関数値の結果に与える影響の小さい補正量ほど絶対値が大という好都合な傾向を持っていること，また，補正量の有対桁数は，項の番号にあまり関係ないということ．もう一つは，最も有効桁数を必要とする最終桁の，桁数の増加の度合を見ると，$n=8$ でもまだ余裕があり，$n=9$, 10 というようにもう少し先まで計算できそうだということである（最も右側の欄については4章を参照).

4. 定数の有効桁数

最良近似式の計算の目的は，被近似関数に対してある近似式の形を設定し，その近似式の中で自由に値を与え得る係数の値を調節して，誤差の絶対値の最大値を最小にすることにある．また，このときの係数を定めることが，最良近似式の計算である．ところで，この計算によって得た係数はどこまで正しいのか，あるいは何桁が有意義なのか，という点は今まで解明されていないようである．このことがわからないために，今まで行われていた計算はすべて安全側に，過大の精度で行われていたと思われる．そして，その過大の精度を得るための道具立てを作ることが容易でないために，最良近似式の計算が敬遠されてきた.

しかしながらこれは，次のように考えれば解決できる．すなわち，定数の値を δ だけ変化させたとき，$E(x)$ の値がどれだけ変化するかを見る．しかも，問題なのは極値の動きであるから，極値をとる x の値についてだけ考えればよい.

ところで，例えば連分数の場合の手順（c）を考えてみると，n 元連立方程式

$$\sum_{i=1}^{n}\left\{\frac{\partial E(x_{j-1})}{\partial d_i}+\frac{\partial E(x_j)}{\partial d_i}\right\}dd_i=-\{E(x_{j-1})+E(x_j)\}$$

$$(j=1,2,\cdots,n) \qquad (38)$$

を解いて d_i を補正するのであるが，ここに出てくる

$$\partial E(x_j)/\partial d_i\ \ (j=0,1,\cdots,n) \qquad (39)$$

は，d_i を変化させたときの x_j における $E(x)$ の値の変化する割合を示しているから，d_i を δ だけ変化させたときの $E(x)$ の変化の最大値は次で与えられる.

$$\xi=\max_j|\partial E(x_j)/\partial d_i|\delta \qquad (40)$$

これは連立方程式を作る最後の過程での値を保存すればよく，通常近似区間の端で与えられるので，この場合はさらに簡単に得られる.

このようにして得られた ξ の値が最大誤差 ε と比べてどの程度の大きさになるかが判断の目安になる．すなわち ξ が ε と等しくなる δ は

$$\delta=\varepsilon/\max_j|\partial E(x_j)/\partial d_i| \qquad (41)$$

である．そこで d_i の計算結果に添えて (41) の値も出力する．もちろん (41)の δ だけ変化するのでは大きすぎるので，誤差の絶対値の間のバラツキを考慮して，意味のある桁数を知ることができる．すなわち (41) における δ の値が $p\times10^{-q}(1\leq p<10)$ であれば，10^{-q} の桁をまず基準にとり，バラツキ $(e_1-e_2)/e_1$ が例えば 10^{-r} であればその右 r 桁程度のところまでとればよいといえる．Table 2 の $n=3$ のところを見よう．これによると，バラツキは 3.5×10^{-6} であるから，補正値は少なくともあと5桁とりたい．このようにして，Table 2 を作ることもできる Table 2 はバラツキを 10^{-4} として作成した).

5. 低精度計算による高精度公式計算の妥当性

第3章の例のように，最良近似の計算に倍精度を用いて，倍精度以上の，あるいは場合によっては4倍精度程度までの公式用の定数の計算ができるということは，一見奇異な感じを与えるかもしれない．しかしこれが実は妥当であることが，次に述べる2つの点に関する考察から確認できる.

第1は，誤差の計算を，まず近似式の値を計算し，次に真の値を計算してその差をとる，というふうにはせず，誤差関数が補正量の値に主として依存することを用いて，補正量の線型結合に近い形に式を変形し，計算途中の式の値を大きくしないようにしているからである．第3章の例について概算すると次の通りである．最終的な誤差関数は大ざっぱにいって，チェビシ

ェフの多項式を一次変換したものに近いのでこれは次と考えてよい．最大誤差を ε とするとき，

$$E(x)\sim\pm\varepsilon T_n\left(\frac{x}{\rho}\right)\ (\text{複号はいずれか一方}) \quad (42)$$

である．$n=16$ のとき，右辺の計算において ε を単位にして，上は $T_{16}(x)$ の最大の係数 2×10^5，下はバラツキの程度である 10^{-4} であるから，その比 2×10^9 は，倍精度（10 進約 16 桁）で十分計算できる．

第2の点は，近似区間の端点の座標が2進数で正しく表わせないため，低精度では誤差が出て，その影響が係数の計算結果に及ぼすかを調べる必要がある．特に端点以外の極大（小）点では，誤差関数の微係数が0であるから，その座標の誤差は問題にならないが，端点では0になるどころか，区間内で微係数の絶対値が最大になるので，この点を調べておかねばならない．ここでも (42) を用いることにして，これを微分すると

$$\left.\frac{dE(x)}{dx}\right|_{x=\rho}\sim\pm\varepsilon\left.\frac{dT_n\left(\frac{x}{\rho}\right)}{dx}\right|_{x=\rho}$$

$$=\pm\varepsilon\cdot\frac{n^2}{\rho} \quad (43)$$

$$\therefore\ dE(x)\sim\pm\varepsilon\cdot\frac{n^2}{\rho}dx \quad (44)$$

となる．これに $n=16$，$\rho=\pi/4$，$dx=7\times10^{-18}$ を代入すると，$2.28\times10^{-15}\varepsilon$ 程度となるので，バラツキの分まで考えても無視できることがわかる．

以上を考慮すれば，倍長計算によっても，それよりかなり高精度の公式の係数計算が可能である．

6. 有理式近似と連分数近似の関係

有限項で打切った連分数をほどいて行くと有理式になる．したがって有理式の方が一般的であり，連分数はその特殊な場合にすぎない．ちなみに，ここで計算した係数を用いて作成する関数ルーチンの計算法として，連分数のまま (30) の漸化式で計算する方法と，有理式にほどいて計算する方法とがあり得る．前者の方が最終的結果の精度をよくできるが，後者の方が計算時間を短くできることが多いであろう．

本論文では，有理式による近似と連分数による近似の，両方の定数の計算法を述べたが，有理式の方は期待したほどよい結果ではない．そうであれば，一般的な有理式形式の近似式の定数の計算の見通しは暗い．しかし幸いなことに，多項式近似が適さない関数を有理式近似するとしても，その場合，係数の自由度が等しいいくつかの有理式のうち，分子，分母の次数のほぼ等しいものが最も精度がよいことが経験的に知られているようである[3,4]．そこでこれを信じれば，連分数による近似の係数を計算して有理式にほどくという方法で大抵間に合うことになる．

7. おわりに

一変数の関数の近似式として，有理式および連分数の形のものの，係数の計算法を示した．この主要な用途は関数サブルーチンの作成のためであるが，このときは通常収束を速めるため，近似範囲を比較的せまくとることが多い．この条件の下では，本方式を適用することによって，直接有理式の係数を求める方法の場合，係数計算の精度と目的の近似式の精度とが，ほぼ同程度まで計算できること，直接連分数の係数を求める方法の場合は，係数計算の精度がかなり小でも計算できることが判明した．

また，計算によって得られる係数の有意な桁数を定める方法を示したが，これは逆に見れば，係数の計算に必要な計算精度を知ることができることを意味している．その他に，近似区間の端点の設定誤差が近似式に与える影響を示すことによって，目的の近似式の誤差よりも低い精度で計算するのであっても，この点に関しては十分であることを示した．

ここで述べた計算法のうち，連分数についてのものは，計算量も多くなく，精度もよいという好ましい結果であるので，この方法で手軽に最良近似式の係数が計算できれば，一松の指摘[4]した転記の誤りなどによる問題点をかなりの程度に軽減し，今後よい近似式による関数ルーチンを容易に作成できるようになると期待できよう．

参考文献

1) Maehly, H. J.: Methods for Fitting Rational Approximations, Part I: J. ACM, Vol. 7, pp. 150～162 (1960), Part II, III: J. ACM, Vol. 9, pp. 257～277 (1962).
2) 山内二郎，森口繁一，一松　信 編：電子計算機のための数値計算法II，培風館 (1967).
3) 山下真一郎：有理式の最良近似式を求めるプログラム，情報処理，Vol. 10, No. 6, pp. 442～446 (1969).
4) 一松　信：初等関数の数値計算，教育出版 (1974).

(昭和52年4月28日受付)
(昭和53年3月16日再受付)