Zeros of Polynomial and an Estimation of its Accuracy

MASAO IGARASHI*

YAMASHITA, S. and SATAKE, S. show that the upper bound of the calculation errors of $f(x)=\sum_{k=0}^{n} a_k x^k$ is $\sum_{k=0}^{n} |a_k x^k| p^{-L}$, where L is the number of the digits in the mantissa based on P radix. We also show that near the zero of $f(x)$, it is $\sum_{k=0}^{n} |a_k x^k| P^{-L}/2$. Furthermore by using Newton-Raphson's iteration method, we propose a method to estimate the accurate significant digits of the numerical result and give some numerical examples.

1. Introduction

We consider a convergence criterion of real zeros of the polynomial $f(x)$ with real exact coefficients a_k

$$f(x)=\sum_{k=0}^{n} a_k x^k \quad (a_n a_0 \neq 0), \tag{1}$$

by using a Newton-Raphson's iteration methods

$$x_{i+1}=x_i-f(x_i)/f'(x_i). \tag{2}$$

Beginning with a suitable starting value x_0, successive approximations x_i $(i=1, 2, 3, \cdots)$ converge to a zero of $f(x)$. It is well known that the iteration can be broken off when $-f(x_i)/f'(x_i)$ takes the value less than its calculation errors. And if successive approximation x_i comes close to a zero of $f(x)$, then the accurate significant digits of $f(x)$ are less than that of $f'(x)$. Hence the accurate significant digits of $-f(x_i)/f'(x_i)$ are nearly equal to that of $f(x_i)$.

Now suppose that we use the floating-point arithmetic with L digits in the mantissa based on P radix such as P is 2 for binary and we calculate $f(x)$ according to Horner's scheme as follows:

$$b_n=a_n,$$
$$b_k=a_k+xb_{k+1} \quad (k=n-1, n-2, n-3, \cdots, 1),$$
$$b_0=f(x).$$

From the above assumptions, YAMASHITA, S. and SATAKE, S. [1] show that the upper bound of the calculation errors $\Delta f(x)$ of $f(x)$ is

$$|\Delta f(x)| \leqq \sum_{k=0}^{n} |a_k x^k| P^{-L}. \tag{3}$$

They used this estimation as a convergence criterion of (2), and obtained good results. However this estimation is an over-estimation when x_i comes close to a zero of $f(x)$, because the accurate significant digits of $f(x_i)$ and of $\Delta f(x_i)$ decrease if x_i comes close to a zero of $f(x)$.

Considering the above mentioned fact, we show that the upper bound of $\Delta f(x)$ near a zero of (1) is estimated as follows:

$$|\Delta f(x)| \leqq \sum_{k=0}^{n} |a_k x^k| P^{-L}/2.$$

Furthermore, using this estimation and changing the calculation process of (2) we attempt to estimate the accurate significant digits of the numerical solution of (1).

2. Calculation Errors of $f(x)$

Let $\tilde{x}$ be an approximate value for a quantity whose exact value is x, and put $\Delta x=\tilde{x}-x$. Then for two values $\tilde{x}=\Delta x+x$ and $\tilde{y}=\Delta y+y$, we have $\Delta(x+y)=\Delta x+\Delta y$. Note that if Δx and Δy are nonnegative or nonpositive then the following relation is derived.

$$|\Delta(x-y)|=|\Delta x-\Delta y| \leqq \max(|\Delta x|, |\Delta y|) \tag{4}$$

We define $f|x|=\sum_{k=0}^{n} |a_k x^k|$. Then for any given x, $(f|x|+f(x))/2$ is sum of all nonnegative terms of $a_k x^k$ $(k=0, 1, 2, \cdots n)$ while $-(f|x|-f(x))/2$ is that of the negative terms of $a_k x^k$ $(k=0, 1, 2, \cdots n)$. If we put

$$f^+(x)=(f|x|+f(x))/2, \quad f^-(x)=(f|x|-f(x))/2,$$

then we have $f(x)=f^+(x)-f^-(x)$. We use Horner's scheme to calculate $f^+(x)$ and $f^-(x)$, and denote the calculation errors of $f^+(x)$ and $f^-(x)$ by $\Delta f^+(x)$ and $\Delta f^-(x)$, respectively. From (3), the upper bound of their calculation errors is as follows:

$$|\Delta f^+(x)| \leqq (f|x|+f(x))P^{-L}/2,$$
$$|\Delta f^-(x)| \leqq (f|x|-f(x))P^{-L}/2.$$

If we use the chopped arithmetic (cf. [2]) to obtain the value of $f^+(x)$ and $f^-(x)$, then $\Delta f^+(x) \geqq 0$ and $\Delta f^-(x) \geqq 0$. Hence from (4), we obtain the following result.

$$|\Delta f^+(x)-\Delta f^-(x)| \leqq \max((f|x|+f(x))P^{-L}/2, (f|x|-f(x))P^{-L}/2)) \tag{5}$$

If x comes close to a zero of $f(x)$ then the value $|f(x)|$ is negligible for the value of $f|x|$. Therefore the estimation (5) is nearly equivalent to

$$|\Delta f^+(x)-\Delta f^-(x)| \leqq f|x|P^{-L}/2.$$

When we calculate $f^+(x)$ and $f^-(x)$ by using floating-point arithmetic, the phenomenon of the disappearance of the leading digits called cancellation (cf. [3]) does not occur. Hence, near the zero of $f(x)$, the numerical result

*College of Agriculture and Veterinary Medicine, Nihon University.

$f^+(x)-f^-(x)$ contains less effective information about $f(x)$ than the numerical result $f(x)$. It means that

$$|\Delta f(x)| \leqq |\Delta f^+(x)-\Delta f^-(x)|.$$

Concluding the above discussion, we have

THEOREM. If we use the chopped floating-point arithmetic with L digits in the mantissa based on P radix, and carry out Horner's scheme for the value of $f(x)$, and furthermore if x is near a zero of $f(x)$, then the calculation errors $\Delta f(x)$ are estimated as follows:

$$|\Delta f(x)| \leqq \sum_{k=0}^{n} |a_k x^k| P^{-L}/2. \tag{6}$$

3. Estimation of the Accurate Significant Digits

We consider the following two calculation processes to estimate the accurate significant digits of the numerical solution of (1) by using Newton-Raphson's iteration methods.

Case I. First, we calculate $f(x_i)$ and $f'(x_i)$, next $f(x_i)/f'(x_i)$ and finally $x_i-f(x_i)/f'(x_i)$.

Case II. We put

$$g(x)=xf'(x)-f(x)$$
$$=(n-1)a_n x^n+(n-2)a_{n-1}x^{n-1}+\cdots+a_2x^2-a_0.$$

First, we calculate $g(x_i)$ and $f'(x_i)$ and then $g(x_i)/f'(x_i)$. Hereafter we put $\hat{x}_{i+1}=g(x_i)/f'(x_i)$.

If x_i comes close to a zero of $f(x)$, then the leading digits of $f(x_i)$ disappear in the floating-point arithmetic. This tendency becomes more pronounced in the case where x_i comes toward multiple zeros of $f(x)$ or toward a nest of zeros of $f(x)$. In any case the following relation comes near the zero of $f(x)$

$$|x_i f'(x_i)| > |f(x_i)|. \tag{7}$$

This relation means that the numerical result $g(x_i)$ (case II) contains less effective information about $f(x_i)$ than the numerical result $f(x_i)$ (case I). This is due to the fact that in case I, $f(x_i)$ is calculated by using full significant digits, whereas in case II, $f(x_i)$ is calculated by losing a few significant digits (see Fig. 1). On the other hand, their exponents may agree with each other. Hence it is reasonable to show that near the zero of $f(x)$, the lower digits of the mantissa of x_{i+1} do not agree with one or more of those of $\hat{x}_{i+1}$ with respect to the same exponent.

It is well known that if $f(x_i)$ has a small degree of accuracy then the succeeding numerical result x_{i+1} will be more accurate than that of x_i. Here we assume that if (6) is satisfied, then $f(x_i)$ has no accuracy. That is, if (6) is solved, then the succeeding x_{i+1} and $\hat{x}_{i+1}$ are less accurate than the preceding ones respectively, and the accurate significant digits in their calculation nearly agree with the accurate significant digits of the numerical solution.

Considering the above mentioned facts we are able to conclude the following:

PROPOSITION. If $|\Delta f(x_i)| \leqq \sum_{k=0}^{n} (a_k x^k) P^{-L}/2$ is satisfied, then the leading agreement digits of x_i and

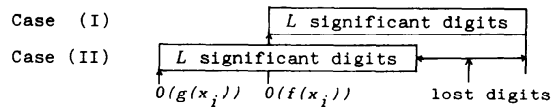

Fig. 1 A comparison between case (I) and (II) with respect to the amount of effective information about $f(x_i)$.

$\hat{x}_i$ are nearly equal to the accurate significant digits of the numerical solution.

4. Numerical Examples and Remarks

We conducted some numerical experiments to illustrate whether the above conclusion is suitable or not. The numerical examples are listed in Table 1, their results are shown in Table 2, and some of them in detail in Table 3.

As far as the above numerical examples are concerned our conclusion is appropriate for the purpose of estimating the accuracy of numerical solutions. The following well known facts are obtained accurately.

1) If x_i is an approximated solution of M-ple zeros of (1), then its accurate significant digits are about L/M.

2) If x_i comes towards a nest of zeros of $f(x)$, then a few accurate significant digits of x_i are lost depending on its density.

Remarks I. When we actually use the convergence criterion (5), we must consider that for some number system machines, except binary machines, a few leading bits are sometimes lost in calculation. For example, in the hexadecimal number system machine, the maximum number of bits lost is 3. In this case, (5) is suitable for 2^{-22} (single precision) and 2^{-54} (double precision).

Table 1 The Lists of Numerical Examples.

1. $(x-1.2340)(x-1.2342)(x-1.2345)$
   $=x^3-3.7027x^2+4.5699957x-1.8801469566$
2. $(x-12.5)^3=x^3-37.5x^2+468.75x-1953.125$
3. $(x-123)(x-0.5)(x-1)(x+1)=x^4-123.5x^3+60.5x^2$
   $+123.5x-61.5$
4. $(x+1.25)^4=x^4+5x^3+9.375x^2+7.8125x+2.44140625$
5. $(3x+2)^5=243x^5+810x^4+1080x^3+720x^2+240x+32$
6. $(x-1)(x-2)(x-3)(x-4)(x-5)(x-6)$
   $=x^6-21x^5+175x^4-735x^3+1624x^2-1764x+720$
7. $(x-1.20)(x-1.21)(x-1.22)(x-1.23)(x-1.24)(x-1.25)$
   $(x-1.26)$
   $=x^7-8.61x^6+31.7695x^5-65.121735x^4+80.08914424x^3$
   $-59.0953690404x^2+24.22376210088x-4.2553354536$
8. $(x-1.25)^{10}$
   $=x^{10}-12.5x^9+70.3125x^8-234.375x^7+512.6953125x^6$
   $-769.04296875x^5+801.08642578125x^4-572.20458984375x^3$
   $+268.220901489258x^2-74.5058059692383x$
   $+9.31322574615479$
9. $(x-0.5(1-\cos\pi/25))(x-0.5(1-\cos 3\pi/25))\cdots(x-0.5$
   $(1-\cos 23\pi/25))$
   $=x^{12}-78x^{11}+1001x^{10}-5005x^9+12870x^8-19448x^7$
   $+18564x^6-11628x^5+4845x^4-1330x^3+231x^2-23x+1$
10. $2^{19}(x-\cos\pi/40)(x-\cos 3\pi/40)(x-\cos 5\pi/40)\cdots(x-\cos$
    $39\pi/40)$
    $=524288x^{20}-2621440x^{18}+5570560x^{16}-6553600x^{14}$
    $+4659200x^{12}-2050048x^{10}+549120x^8-84480x^6+6600x^4-$
    $-200x^2+1$

II. In addition to case I and case II, we perform the following calculation process.

First, we calculate $f^+(x_i)$ and $f^-(x_i)$, next $(f^+(x_i)-f^-(x_i))/f'(x_i)$ and finally $x_i-(f^+(x_i)-f^-(x_i))/f'(x_i)$.

Hereafter we put $\tilde{x}_{i+1}=x_i+(f^+(x_i)-f^-(x_i))/f'(x_i)$. By using this calculation process, we have almost the same results (see Table 3). However this process is more complicated than that of case II.

Table 2 The Results of the Numerical Solutions.

No.	x_0	Solutions	$f(x)$	$f\|x\|*2^{-54}$
1	1.0	$x=$ 0.123400000007801+01	0.0	0.83−15
		$\tilde{x}=$ 0.123400000169573+01	0.22−15	
	1.2348	$x=$ 0.123450000072771+01	0.22−15	0.83−15
		$\tilde{x}=$ 0.123450000095943+01	0.22−15	
	1.2341	$x=$ 0.123450000002748+01	−0.22−15	0.83−15
		$\tilde{x}=$ 0.123449999981272+01	−0.22−15	
2	10.0	$x=$ 0.124998940056445+02	−0.62−12	0.86−12
		$\tilde{x}=$ 0.124998861863364+02	−0.79−12	
	5.0	$x=$ 0.124998955753597+02	−0.11−12	0.86−12
		$\tilde{x}=$ 0.124998890578041+02	−0.56−12	
3	3.0	$x=$ 0.100000000000000+01	0.0	0.20−13
		$\tilde{x}=$ 0.100000000000000+01	0.0	
	154.123	$x=$ 0.123000000000000+03	0.0	0.25−07
		$\tilde{x}=$ 0.123000000000000+03	0.19−07	
4	−1.0	$x=$ −0.124981474919127+01	0.15−14	0.26−14
		$\tilde{x}=$ −0.124981961064475+01	0.11−14	
	−2.0	$x=$ −0.125017169627768+01	0.88−15	0.21−14
		$\tilde{x}=$ −0.125016208626706+01	0.88−15	
5	3.0	$x=$ −0.665908236665141+00	0.46−13	0.56−13
		$\tilde{x}=$ −0.665869782645817+00	0.60−13	
	5.0	$x=$ −0.666078205970774+00	0.21−13	0.56−13
		$\tilde{x}=$ −0.666008959044368+00	0.42−13	
6	1.5	$x=$ 0.200000000000000+01	0.39−12	0.11−10
		$\tilde{x}=$ 0.199999999999997+01	0.79−12	
7	0.1205	$x=$ 0.120974224662827+01	0.21−13	0.28−13
		$\tilde{x}=$ 0.120906036897733+01	0.10−12	
	1.265	$x=$ 0.126002058270954+01	−0.88−15	0.32−13
		$\tilde{x}=$ 0.126008417791960+01	0.52−12	
8	1.0	$x=$ 0.120069396067796+01	−0.21−13	0.43−12
		$\tilde{x}=$ 0.119384829264054+01	0.52−12	
	2.0	$x=$ 0.130933103463042+01	0.52−12	0.66−12
		$\tilde{x}=$ 0.131317985013408+01	0.97−12	
9	2.0	$x=$ 0.137902118690489+01	−0.68−11	0.37−10
		$\tilde{x}=$ 0.137902118690500+01	−0.24−11	
	0.38	$x=$ 0.381966011240331+00	0.37−14	0.22−13
		$\tilde{x}=$ 0.381966011180631+00	0.45−13	
	7.88	$x=$ 0.712012217452314+01	0.50−05	0.48−04
		$\tilde{x}=$ 0.712012217452315+01	−0.20−04	
10	0.972	$x=$ 0.972369920400331+00	−0.13−09	0.84−09
		$\tilde{x}=$ 0.972369920372271+00	0.21−08	
	0.38	$x=$ 0.382683432365089+00	0.13−13	0.49−13
		$\tilde{x}=$ 0.382683432365098+00	0.26−13	
	−0.64	$x=$ −0.649448048330018+00	0.28−11	0.56−11
		$\tilde{x}=$ −0.649448048328451+00	0.44−11	

(HITAC L340, DOUBLE P.)

† x_0 is initial value

Table 3 The Detailed Results of the Numerical Examples.
No. 7, $f(x)=(x-1.20)(x-1.21)\cdots(x-1.26)$
Initial value $x_0=1.205$

Iteration times	Solutions	$f(x)$	$xf'(x)$	$f\|x\|*2^{-57}$	Max
1	x=0.121065297107832+01	−0.90−13	−0.12−09	0.35−14	0.19−14
	$\hat{x}$=0.121043309353861+01	−0.57−13	−0.12−09		
	$\tilde{x}$=0.121075258290349+01	−0.90−13	−0.11−09		
2	x=0.120974224662827+01	0.21−13	0.15−09	0.35−14	0.19−14
	$\hat{x}$=0.120906036897733+01	0.10−12	−0.17−09		
	$\tilde{x}$=0.120951010118022+01	0.61−13	−0.16−09		
3	x=0.120990722558912+01	0.44−14	−0.14−09	0.35−14	0.19−14
	$\hat{x}$=0.120966220991922+01	0.38−13	−0.15−09		
	$\tilde{x}$=0.121018682151222+01	−0.28−13	−0.13−09		
4	x=0.120994337175699+01	−0.53−14	−0.14−09	0.35−14	0.19−14
	$\hat{x}$=0.120973994639523+01	0.14−13	0.15−09		
	$\tilde{x}$=0.120990722558912+01	0.44−14	0.15−09		
5	x=0.120989961326739+01	−0.88−15	−0.14−09	0.35−14	0.19−14
	$\hat{x}$=0.120944636394300+01	0.67−13	0.16−09		
	$\tilde{x}$=0.120994337175699+01	−0.53−14	−0.14−09		
		No. 8, $f(x)=(x-1.25)^{10}$ Initial value $x_0=1.0$			
1	x=0.102500000000000+01	0.33−06	−0.15−04	0.25−13	0.12−13
	$\hat{x}$=0.102500000000000+01	0.33−06	−0.15−04		
	$\tilde{x}$=0.102500000000000+01	0.33−06	−0.15−04		
	.	.	.	.	.
	.	.	.	.	.
13	x=0.118684684380199+01	0.11−11	−0.15−09	0.51−13	0.25−13
	$\hat{x}$=0.118598560482934+01	0.11−11	−0.21−09		
	$\tilde{x}$=0.118742523040777+01	0.87−12	−0.17−09		
14	x=0.119323240884895+01	0.45−12	−0.61−10	0.52−13	0.25−13
	$\hat{x}$=0.119017119191207+01	0.61−12	−0.11−09		
	$\tilde{x}$=0.119500794799051+01	0.32−12	−0.55−10		
15	x=0.120069396067796+01	−0.21−13	−0.17−10	0.54−13	0.26−13
	$\hat{x}$=0.119384829264054+01	0.52−12	−0.65−10		
	$\tilde{x}$=0.119972008516025+01	0.14−12	−0.25−10		

Here, Max=Max $(|a_k x^k|*2^{-56})$
(HITAC L340, DOUBLE P.)

References

1. Yamashita, S. and Satake, S. On the calculation limit of roots of the algebraic equation, JOHO SHORI, 7, 4 (1966) (in Japanese), 197–201.
2. Sterbenz, Pat H. Floating-point Computation. Prentice-Hall (1 974).
3. Stoer, J. and Bulirsch, R. Introduction to Numerical Analysis. Springer-Verlag, New York (1980).

(Received November 19, 1981)