

“Approximate Zero-points” of Univariate Polynomial with Large Error Terms

照井 章 (Akira TERUI) 佐々木 建昭 (Tateaki SASAKI)

University of Tsukuba*

University of Tsukuba†

1 Introduction

In traditional computer algebra on polynomials, we have assumed that the coefficients of polynomials were given rigorously by integers, rational numbers or algebraic numbers and that manipulation on the polynomials were also exact. However, in many practical applications or real world problems, the coefficients contain errors, that is, polynomials have “error terms.” In such cases, many of the traditional algorithms in computer algebra break down.

This paper considers the real zero-points of a real univariate polynomial with error terms (or “approximate polynomial”), where the coefficients of error terms can be much larger than the machine epsilon ε_M . In fact, even if the initial errors in coefficients are as small as ε_M , the errors can become much larger than ε_M after the calculation. Furthermore, in approximate algebraic calculation, we handle polynomials with perturbed terms which are much larger than ε_M in general.

If a polynomial $P(x)$ has error terms, we cannot draw the graph of function $y = P(x)$, and what we can draw is only the “existence domain” of values of $P(x)$, or the domain in which the value of $P(x)$ can exist. Similarly, in such a case, the position of its zero-points cannot be determined exactly, and what we can handle is only the domains in which the zero-points can exist. Therefore, in this paper, we introduce a concept of “approximate real zero-point” by a minimal interval outside of which no real zero-points can exist. Although the existence domains of real zero-points can be calculated rigorously, we propose methods to calculate them approximately and efficiently, by using Smith’s theorem on the error bounds of zero-points of the polynomial [3].

Next, we consider calculating the number of real zero-points of an approximate polynomial by Sturm’s method. If all the zero-points are single and well separated, the number of real zero-points is definite unless some error term is quite large, although the positions of zero-points are changed by the error terms. However, in the calculation of the Sturm sequence, the leading coefficient of some element may become too small to determine if it is equal to zero or not. Since the sign of the leading coefficient in the

*terui@math.tsukuba.ac.jp

†sasaki@math.tsukuba.ac.jp

Sturm sequence is essential in determining the number of real zero-points, this is a serious problem. For this problem, we give an answer that, under some conditions, we may discard the small leading term and continue further calculation of the Sturm sequence.

In **2**, we investigate the existence domains of real zero-points of approximate polynomial, then define approximate real zero-point. In **3**, we propose a practical method to calculate the existence domains of the zero-points of approximate polynomial. In **4**, with assumption that the polynomial does not have multiple or close zero-points, we derive a sufficient condition for that the number of the real zero-points is not changed by error terms. In **5**, we propose and investigate several methods to check the effect of error terms of a given polynomial on its discriminant or the Sturm sequence.

2 Approximate polynomials and approximate real zero-points

Let $P(x)$ be a given univariate polynomial with real coefficients such that

$$P(x) = a_n x^n + \cdots + a_0 x^0, \quad (1)$$

and let $\tilde{P}(x)$ be a univariate polynomial such that

$$\tilde{P}(x) = P(x) + \Delta(x), \quad (2)$$

where $\Delta(x)$ represents the sum of “error terms.” Hence, we know neither $\tilde{P}(x)$ nor $\Delta(x)$; what we know usually is an upper bound for each coefficient in $\Delta(x)$. Representing $\Delta(x)$ as

$$\Delta(x) = \delta_{n-1} x^{n-1} + \cdots + \delta_0 x^0, \quad (3)$$

we assume that we know upper bounds $\varepsilon_{n-1}, \dots, \varepsilon_0$ such that

$$|\delta_i| < \varepsilon_i, \quad i = n-1, \dots, 0. \quad (4)$$

We express $\tilde{P}(x)$ also as $\tilde{P}(x | \delta_i = \varepsilon'_i (i = n-1, \dots, 0))$ and so on, in the case that we specify (some of) the values of $\delta_{n-1}, \dots, \delta_0$ as $\delta_i = \varepsilon'_i (i = n-1, \dots, 0)$ and so on.

2.1 Existence domain of values of $\tilde{P}(x)$

Suppose that the variable x is fixed to x_0 and $\delta_{n-1}, \dots, \delta_0$ are changed continuously under the restrictions in (4), then the value of $y = \tilde{P}(x_0)$ moves continuously inside an interval. By changing x_0 in \mathbb{R} , we will have the minimal domain outside of which there is no possibility of existence of value of $\tilde{P}(x)$.

Definition 1 (existence domain) *Let x_0 be a real number and δ_i move continuously*

in the whole interval $[-\varepsilon_i, \varepsilon_i]$ for $i = 0, \dots, n - 1$. Let $y_U(x_0)$ and $y_L(x_0)$ be

$$y_U(x_0) = \max_{\delta_i \in [-\varepsilon_i, \varepsilon_i]} \tilde{P}(x_0), \quad (5)$$

$$y_L(x_0) = \min_{\delta_i \in [-\varepsilon_i, \varepsilon_i]} \tilde{P}(x_0). \quad (6)$$

By changing the value x_0 in \mathbb{R} , we have a domain

$$\{[y_L(x), y_U(x)] \mid x \in \mathbb{R}\}. \quad (7)$$

This domain is said to be “existence domain of values of $\tilde{P}(x)$.”

We specify the existence domain of values of $\tilde{P}(x)$ rigorously.

Theorem 2 *Let the polynomials $P(x)$ and $\tilde{P}(x)$ be defined as above. Then, existence domain of values of $\tilde{P}(x)$ is given by a set defined as*

$$\{[P_L(x), P_U(x)] \mid x \in \mathbb{R}\}, \quad (8)$$

where $P_L(x)$ and $P_U(x)$ are polynomials defined as follows.

$$P_U(x) = \begin{cases} \tilde{P}(x \mid \delta_i = \varepsilon_i \ (i = n - 1, \dots, 0)) & \text{for } x \geq 0, \\ \tilde{P}(x \mid \delta_i = (-1)^i \varepsilon_i \ (i = n - 1, \dots, 0)) & \text{for } x < 0, \end{cases} \quad (9)$$

$$P_L(x) = \begin{cases} \tilde{P}(x \mid \delta_i = -\varepsilon_i \ (i = n - 1, \dots, 0)) & \text{for } x \geq 0, \\ \tilde{P}(x \mid \delta_i = (-1)^{i+1} \varepsilon_i \ (i = n - 1, \dots, 0)) & \text{for } x < 0. \end{cases} \quad (10)$$

2.2 Approximate real zero-points and their existence domains

Using polynomials $P_L(x)$ and $P_U(x)$, we define a concept of “approximate real zero-points” as follows.

Definition 3 (approximate real zero-point) *Let ζ_U and ζ_L be real zero-points of P_U and P_L , respectively, such that a real zero-point ζ of $\tilde{P}(x)$ moves in the whole region of the interval $[\zeta_U, \zeta_L]$ if $\zeta_U < \zeta_L$ or the interval $[\zeta_L, \zeta_U]$ if $\zeta_L < \zeta_U$. Then, ζ is said to be an “approximate real zero-point of $\tilde{P}(x)$ of accuracy $|\zeta_L - \zeta_U|$.”*

3 Bounding existence domains by Smith’s theorem

Although the existence domains of approximate real zero-points can be specified rigorously, we consider specifying the existence domains approximately by using the real zero-points of $P(x)$.

A key for bounding existence domains is celebrated Smith’s theorem. (For the proof, see Smith [3].)

Theorem 4 (Smith) Let x_1, \dots, x_n be n distinct numbers in \mathbb{C} and r_1, \dots, r_n be defined as

$$r_j = \left| \frac{nP(x_j)}{a_n \prod_{k=1, \neq j}^n (x_j - x_k)} \right|, \quad j = 1, \dots, n. \quad (11)$$

Let D_j ($1 \leq j \leq n$) be a disc of radius r_j and the center at x_j . Then, the union $D_1 \cup \dots \cup D_n$ contains all the zero-points of $P(x)$. Furthermore, if a union $D_1 \cup \dots \cup D_m$ ($m \leq n$) is connected and does not intersect with D_{m+1}, \dots, D_n , then this union contains exactly m zero-points.

For further discussions, see Terui and Sasaki [4].

4 Calculating the number of real zero-points of real approximate polynomial

In the case that a real polynomial has multiple or close zero-points, its zero-points may change largely or some real zero-points may become complex, as the coefficients change slightly. Therefore, it is not adequate to count the number of real zero-points of a real approximate polynomial which may have multiple or close zero-points. On the other hand, in the case that a polynomial has only single zero-points, the number of its real zero-points rarely changes although the position of them may move considerably, as the coefficients change slightly. In this section, we focus our attention on calculating the number of real zero-points of a real approximate polynomial containing only single zero-points.

4.1 Sufficient condition for that the number of real zero-points becomes definite

We first derive a sufficient condition for error terms so that they do not change the number of real zero-points.

Theorem 5 Let $P(x)$ and $\tilde{P}(x)$ be defined as in (1) and (2), respectively. Then the number of real zero-points of $\tilde{P}(x)$ is the same as that of $P(x)$ if the discriminant of \tilde{P} , or $\text{res}(\tilde{P}, d\tilde{P}/dx)$ does not become zero as δ_i changes continuously within the range $[-\varepsilon_i, \varepsilon_i]$ for $i = 0, \dots, n-1$.

4.2 Problem of small leading coefficient in the Sturm sequence

We use famous Sturm's method to calculate the number of real zero-points. Sturm's theorem is as follows. (For the proof, see Cohen [1] for example.)

Theorem 6 (Sturm) Let $F(x)$ be a real square-free polynomial of degree m , and (P_0, P_1, \dots, P_m) be a sequence of polynomials (the Sturm sequence) defined as

$$\begin{cases} P_0 = F(x), & P_1 = \frac{d}{dx}F(x), \\ P_i = -\text{rem}(P_{i-2}, P_{i-1}) & \text{for } i = 2, \dots, m, \end{cases} \quad (12)$$

where $\text{rem}(P_{i-2}, P_{i-1})$ denotes the remainder of P_{i-2} divided by P_{i-1} . Let a and b be any real numbers satisfying $a < b$, and let s and t be the numbers of sign changes in the sequences $(P_0(a), P_1(a), \dots, P_m(a))$ and $(P_0(b), P_1(b), \dots, P_m(b))$, respectively. Then, the number of the real zero-points of F in the interval $[a, b]$ is equal to $t - s$.

Note that we can calculate the number of all the real zero-points of F by putting $a = -\infty$ and $b = \infty$ in Theorem 6.

Consider to calculate the Sturm sequence with floating-point arithmetic. During calculation of the Sturm sequence, we may encounter the leading coefficient problem: 1) we can hardly decide if a very small leading coefficient is equal to zero or not, and 2) the division by a polynomial with a small leading coefficient will cause large cancellation errors in the coefficients of the remainder polynomial.

A Sturm sequence $(P_0 = F, P_1 = dF/dx, P_2, \dots, P_i, \dots, P_k (= \text{constant}))$ satisfies the following conditions which are sufficient for determining the number of real zero-points. (For example, see Cohen[1].)

- 1° For any real number x , consecutive elements $P_{i-1}(x)$ and $P_i(x)$ do not simultaneously become zero.
- 2° If $P_j(x) = 0$ for some j ($1 \leq j \leq k$) and $x \in \mathbb{R}$, then we have $P_{j-1}(x)P_{j+1}(x) < 0$.
- 3° The sign of P_k is the same for any real number x .

With condition 1°, we can find the number of sign changes by investigating each P_i separately. Let $P_j(x_j) = 0$ for some $x_j \in \mathbb{R}$. Condition 2° means that P_{j-1} and P_{j+1} have no zero-point in the neighborhood of $x = x_j$. Condition 3° means that P_k is a nonzero constant in our case, i.e. $P_0(x)$ has no multiple zero-points. We note that the sign change of $P_j(x)$ at $x = x_j$, $j \geq 1$, does not affect the difference of the sign changes in the sequences $(P_0(a), P_1(a), \dots, P_m(a))$ and $(P_0(b), P_1(b), \dots, P_m(b))$; the number of sign change in the whole Sturm sequence occurs only when the evaluation point passes a real zero-point of $P_0(x)$.

We see that we can relax the condition 2° as follows.

- 2' If $P_j(x) = 0$ for some unknown x in the interval $[x_j, x'_j]$ then we have $P_{j-1}(x)P_{j+1}(x) < 0$ for $x = x_j - \delta$ and $x = x'_j + \delta$, where δ is an infinitesimal positive number.

These considerations lead us to an idea of discarding the small leading terms of $P_i(x)$.

Theorem 7 Let $P_0(x)$ be a real polynomial having no multiple zero-points. Let $P_i(x)$ be an element in the Sturm sequence of $P_0(x)$ and assume that $P_i(x)$ has small leading terms. Then, i) if large zero-points of $P_i(x)$ due to the small leading coefficients are sufficiently apart from the zero-points of $P_0(x)$, and ii) if the resultant $\text{res}(P_{i-1}, P_i)$ does not become zero as the small leading coefficients of P_i change continuously to zero, we can calculate the number of the real zero-points of $P_0(x)$ by calculating the Sturm sequence of $P_0(x)$ with discarding the small leading terms of $P_i(x)$.

5 Calculation of the discriminant of an approximate polynomial

We consider evaluating errors in the discriminant of approximate univariate polynomial. We investigate the four methods: 1) evaluating the “subresultant determinant” by Hadamard’s inequality, 2) calculating the Sturm sequence using interval arithmetic, 3) solving linear system on polynomial coefficients and evaluating errors in the solution by a standard method in numerical computation, and 4) calculating the Sturm sequence with parametric error terms.

5.1 Evaluation of subresultant determinant

Let $F(x)$ be a univariate polynomial such that

$$F(x) = f_m x^m + f_{m-1} x^{m-1} + \cdots + f_1 x + f_0 x^0. \quad (13)$$

Then, the p -norm of F is defined as

$$\|F\|_p = \left(\sum_{i=1}^m |f_i|^p \right)^{1/p}, \quad p = 1, 2, \dots, \infty. \quad (14)$$

In this paper, we use 2-norm for polynomials.

Let $F(x)$ be a real univariate polynomial and assume that the Sturm sequence $(P_0 = F, P_1 = dF/dx, P_2, \dots, P_k, \dots)$ has an element P_k such that

$$\begin{aligned} P_k &= \varepsilon_n x^n + \cdots + \varepsilon_{n-s+1} x^{n-s+1} + b_{n-s} x^{n-s} + \cdots + b_0 x^0, \\ |\varepsilon_i| &\ll |b_{n-s}| \quad (i = n - s + 1, \dots, n), \end{aligned} \quad (15)$$

where $s < n$. We derive a sufficient condition that the signs of “successor” elements in the Sturm sequence are unchanged if we discard the small leading terms $\varepsilon_n x^n, \dots, \varepsilon_{n-s+1} x^{n-s+1}$. (For subresultant theory, see Mishra [2] for example.)

Proposition 8 *Let (P_0, P_1, P_2, \dots) be a Sturm sequence and assume that there exist an element $P_k(x)$ in the sequence satisfying (15). Let $P_{k-1}(x)$ be*

$$P_{k-1}(x) = a_l x^l + \cdots + a_0 x^0, \quad (16)$$

where $l = n + 1$, and define \hat{P}_k and N as follows.

$$\begin{aligned} \hat{P}_k &= P_k - (\varepsilon_n x^n + \cdots + \varepsilon_{n-s+1} x^{n-s+1}) \\ &= b_{n-s} x^{n-s} + \cdots + b_0, \end{aligned} \quad (17)$$

$$N = \|P_k\|_2^{(l-1)} \left\{ (l-s) |a_l|^s \|P_{k-1}\|_2^{(n-s)} + \sum_{j=1}^s |a_l|^{(j-1)} \|P_{k-1}\|_2^{(n-j+1)} \right\}. \quad (18)$$

(Note that the norm of polynomials is 2-norm.) If $\varepsilon_n, \dots, \varepsilon_{n-s+1}$ satisfy

$$\{|\varepsilon_n| + \dots + |\varepsilon_{n-s+1}|\} \cdot N < |a_i^s \cdot \text{lc}(S_i(P_{k-1}, \hat{P}_k))|, \quad (19)$$

for $i = 0, \dots, n-s-1$, where $S_i(P_{k-1}, P_k)$ denotes the i -th subresultant of P_{k-1} and P_k , then we have

$$\text{lc}(S_i(P_{k-1}, P_k)) \cdot a_i^s \cdot \text{lc}(S_i(P_{k-1}, \hat{P}_k)) > 0. \quad (20)$$

5.2 Interval arithmetic

In this method, we first transfer the coefficients of the given polynomial into interval numbers each of which includes the corresponding error, then calculate the Sturm sequence.

We observed how the width of the intervals of coefficients increased during the calculation. We found that the increase of the width of intervals was about one decimal-digit for each remainder computation. In fact, division of polynomial of degree n by polynomial of degree $n-1$ requires twice of “polynomial \times number” multiplication and twice of polynomial subtraction. Since the width of each interval number increases about twice after one arithmetic operation if the width of each operand is almost the same, the width of interval coefficient increases about $2^4 = 16$ times after the polynomial division. Therefore, for example, the widths of intervals in the last element of the Sturm sequence of a polynomial of degree 10 may become about 10^{10} times larger than the initial intervals, which shows that this method is never useful in practice.

5.3 Standard method in numerical analysis

In numerical analysis, we have a good method of error estimation for the solution of a system of linear equations. Calculation of the resultant or the Sturm sequence can be reduced easily into a linear system. (For the norm of vectors and matrices, please consult various literature on numerical analysis. We only use 1-norm and ∞ -norm for vectors and matrices in this paper.)

Let $F(x)$ and $G(x)$ be given by

$$F(x) = f_m x^m + \dots + f_0 x^0, \quad f_m \neq 0, \quad (21)$$

$$G(x) = g_n x^n + \dots + g_0 x^0, \quad g_n \neq 0, \quad (22)$$

where $m \geq n$. Then, calculation of the resultant or the Sturm sequence is equivalent to eliminate the terms of higher degrees to derive R_s , a polynomial of degree s , for $s = 0, \dots, n-1$. For every R_s , there exist polynomials U_s and V_s such that

$$U_s F + V_s G = R_s, \quad \deg(U_s) \leq n-s-1, \quad \deg(V_s) \leq m-s-1. \quad (23)$$

We consider calculating $\text{res}(F, G)$, or the resultant of $F(x)$ and $G(x)$. Let U_0 and V_0 be

expressed as

$$U_0 = u_{n-1}x^{n-1} + \cdots + u_0x^0, \quad (24)$$

$$V_0 = v_{m-1}x^{m-1} + \cdots + v_0x^0. \quad (25)$$

By the relation $U_0F + V_0G = R_0 = \text{res}(F, G)$, we obtain a system of linear equations on the coefficients in U_0 and V_0 . With normalizing U_0 and V_0 as $u_{n-1} = g_n$ and $v_{m-1} = -f_m$, respectively, we can derive a system of linear equations of the form

$$A\mathbf{x} = \mathbf{b}. \quad (26)$$

In the system (26), A is a “coefficient matrix,” and \mathbf{x} and \mathbf{b} are vectors of unknown numbers and known numbers, respectively. We briefly describe the effect of errors in A and \mathbf{b} on \mathbf{x} (detailed analysis can be found in various literature in numerical analysis). Assume that \mathbf{b} has the error $\Delta\mathbf{b}$ which causes an error $\Delta\mathbf{x}_1$ in the solution \mathbf{x} . Then, we have

$$A(\mathbf{x} + \Delta\mathbf{x}_1) = \mathbf{b} + \Delta\mathbf{b}. \quad (27)$$

Using (26), we easily evaluate the magnitude of $\Delta\mathbf{x}_1$ as

$$\frac{\|\Delta\mathbf{x}_1\|}{\|\mathbf{x}\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \quad (28)$$

Furthermore, assume that A has the error ΔA and the error of \mathbf{x} becomes $\Delta\mathbf{x}_1 + \Delta\mathbf{x}_2$, as follows.

$$(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}_1 + \Delta\mathbf{x}_2) = \mathbf{b} + \Delta\mathbf{b}. \quad (29)$$

Using (27), we derive the following evaluation of $\Delta\mathbf{x}_2$.

$$\frac{\|\Delta\mathbf{x}_2\|}{\|\mathbf{x} + \Delta\mathbf{x}_1 + \Delta\mathbf{x}_2\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta A\|}{\|A\|}. \quad (30)$$

(28) and (30) lead us to the following evaluation.

$$\frac{\|\Delta\mathbf{x}_1\| + \|\Delta\mathbf{x}_2\|}{\|\mathbf{x}\| + \|\Delta\mathbf{x}_1\| + \|\Delta\mathbf{x}_2\|} \leq \|A\| \|A^{-1}\| \left\{ \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \right\}. \quad (31)$$

The number $\|A\| \|A^{-1}\|$ is called the “condition number,” which specifies the amount of the error in the solution.

Equation (26) can be solved by Gaussian elimination with pivoting. Although we did not consider rounding errors in floating-point arithmetic in the above evaluation, the evaluation of rounding errors can easily be included by adding ΔA_M , the error whose magnitude is about machine epsilon ε_M , into A .

5.4 Calculating error terms parametrically

The method described in this subsection gives a good estimation of errors in the Sturm sequence, although the calculated value does not give the rigorous error bound.

Let n be the degree of the given polynomial $P(x)$ and $\delta_{n-1}, \delta_{n-2}, \dots, \delta_0$ be parameters representing the errors of the coefficients. Then, the discriminant or an element in the Sturm sequence is a polynomial in x and the parameters $\delta_{n-1}, \delta_{n-2}, \dots, \delta_0$. Although it is very inefficient to calculate all the terms of the polynomial, calculation of only terms which are linear in the parameters can be executed at the cost of $O(n)$ times the cost of calculation of a polynomial with numerical coefficients.

Assume that $\tilde{P}(x)$ in (2) is monic, then $\tilde{P}(x)$ is expressed as follows.

$$\tilde{P}(x, \delta_n, \dots, \delta_0) = x^n + (c_{n-1} + \delta_{n-1})x^{n-1} + (c_{n-2} + \delta_{n-2})x^{n-2} + \dots + (c_0 + \delta_0)x^0. \quad (32)$$

Now we introduce a total degree variable t , and calculate the Sturm sequence $(\tilde{P}_0 = \tilde{P}, \tilde{P}_1 = d\tilde{P}/dx, \tilde{P}_2, \dots, \tilde{P}_k)$ of \tilde{P} up to degree 1 in t , then substitute 1 for t after the calculation. With this computation, we have

$$\tilde{P}_i(x, \delta_n, \dots, \delta_0) \simeq \tilde{P}_i(x, 0, \dots, 0) + \tilde{P}_{i,n-1}(x, 0, \dots, 0)\delta_{n-1} + \dots + \tilde{P}_{i,0}(x, 0, \dots, 0)\delta_0, \quad (33)$$

where $\tilde{P}_{i,j} = \partial\tilde{P}_i/\partial\delta_j$ for $j = n-1, \dots, 0$. By neglecting the terms of order $O(\delta^2)$, we have the following approximate inequality.

$$|\tilde{P}_i - P_i| \lesssim |\tilde{P}_{i,n-1}(x, 0, \dots, 0)|\varepsilon_{n-1} + \dots + |\tilde{P}_{i,0}(x, 0, \dots, 0)|\varepsilon_0, \quad (34)$$

where $|(\text{polynomial})|$ denotes a polynomial with the coefficients replaced by their absolute values.

References

- [1] H. Cohen. *A Course in Computational Algebraic Number Theory*, Vol. 138 of *Graduate Texts in Mathematics*. Springer-Verlag, Berlin, 1993.
- [2] B. Mishra. *Algorithmic Algebra*. Texts and Monographs in Computer Science. Springer-Verlag, New York, 1993.
- [3] B. T. Smith. Error Bounds for Zeros of a Polynomial Based Upon Gerschgorin's Theorems. *J. ACM*, Vol. 17, No. 4, pp. 661–674, 1971.
- [4] A. Terui and T. Sasaki. On Error bounds of the Zero-points of Univariate Polynomial with Incorrect Coefficients (in Japanese). In F. Kako, editor, *Formula Manipulation and Application of Mathematical Studies* (in Japanese), No. 1038 in RIMS Collection of Research Reports, pp. 106–110. Research Institute for Mathematical Sciences, Kyoto Univ., Kyoto, 1998.