Regular Paper

"Approximate Zero-points" of Real Univariate Polynomial with Large Error Terms

Akira Terui† and Tateaki Sasaki†

Let $P(x)$ be a given real univariate polynomial and let $\tilde{P}(x) = P(x) + \Delta(x)$, where $\Delta(x)$ is the sum of error terms, that is, a polynomial with small real unknown but bounded coefficients. We first consider specifying the "existence domain" of the values of $\tilde{P}(x)$, or the domain in which the value of $\tilde{P}(x)$ exists for any real number x, by the coefficient bounds for $\Delta(x)$, and then introduce a concept of an "approximate real zero-point" of $\tilde{P}(x)$. We present a practical method for estimating the existence domain of zero-points of $\tilde{P}(x)$ by applying Smith's celebrated theorem. We next consider counting the number of real zero-points of $\tilde{P}(x)$. If all the zero-points are sufficiently far apart from each other, the number of real zero-points of $\tilde{P}(x)$ is the same as that of $P(x)$, and we derive a condition for which we can assert that $P(x)$ and $\tilde{P}(x)$ have the same number of real zero-points. We calculate the actual number of real zero-points by Sturm's method, which encounters the so-called small leading coefficient problem. For this problem, we show that, under some conditions, small leading terms can be discarded. Furthermore, we investigate four methods for evaluating the effect of error terms on the elements of the Sturm sequence.

1. Introduction

In traditional computer algebra on polynomials, we usually assume that the coefficients of polynomials are given rigorously by integers, rational numbers, or algebraic numbers, and that manipulation on the polynomials is 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 that are much larger than ε_M in general.

If a polynomial $P(x)$ has error terms, we cannot draw the graph of function $y = P(x)$; all we can draw is the "existence domain" of $P(x)$, or the domain in which values of $P(x)$ can exist. Similarly, in such a case, the positions of its zero-points cannot be determined exactly; all we can handle is the domains in which zero-points can exist. Therefore, in this paper, we introduce a concept of an "approximate real zero-point" by defining 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 for calculating them approximately and efficiently by using Smith's theorem on the error bounds of zero-points of a polynomial [11].

Next, we consider calculation of 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 whether it is equal to zero or not. Since the sign of the leading coefficient in the Sturm sequence is essential in determining the number of real zero-points, this is a serious problem. Our answer to it is that, under some conditions, we may discard the small leading term and continue further calculation of the Sturm sequence. Shirayanagi and Sekigawa [10] also attacked this problem, and proposed an interval arithmetic method with zero rewriting. We will investigate the Sturm sequence with interval coefficients in Section 5.

In Section 2, we investigate the existence domains of the values of a real approximate poly-

† Institute of Mathematics, University of Tsukuba

nomial, then define an approximate real zero-point. In Section 3, we propose a practical method for calculating the existence domains of the zero-points of an approximate polynomial. In Section 4, on the assumption that the polynomial does not have multiple or close zero-points, we derive a sufficient condition for the number of real zero-points not to be changed by error terms. In Section 5, we propose and investigate several methods for checking the effect of the error terms of a given polynomial on the Sturm sequence.

2. Approximate Polynomials and Approximate Real Zero-points

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

$$P(x) = c_n x^n + \cdots + c_0 x^0, \tag{1}$$

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

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

where $\Delta(x)$ represents the sum of real "error terms," that is, a polynomial with unknown small real coefficients. 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

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

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

$$|\delta_i| \le \varepsilon_i, \quad i = n-1, \ldots, 0. \tag{4}$$

Throughout this paper, we write $\tilde{P}(x \mid \delta_i = \varepsilon_i' \ (i = n-1, \ldots, 0))$ to denote that the values of $\delta_{n-1}, \ldots, \delta_0$ in $\tilde{P}(x)$ are specified as $\delta_{n-1} = \varepsilon_{n-1}', \ldots, \delta_0 = \varepsilon_0'$, and so on.

2.1 Existence Domain of Values of $\tilde{P}(x)$

Supposing that the variable x is fixed to x_0 and that $\delta_{n-1}, \ldots, \delta_0$ are changed continuously under the restrictions in Eq. (4); the value of $\tilde{P}(x_0)$ moves continuously inside an interval. By changing x_0 in $\mathbf{R}$, we will have the minimal domain outside of which there is no possibility of the existence of the 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 $i = 0, \ldots, n-1$. Define $P_U(x_0)$ and $P_L(x_0)$ as

$$P_U(x_0) = \max_{\substack{\delta_i \in [-\varepsilon_i, \varepsilon_i] \\ i=0,\ldots,n-1}} \tilde{P}(x_0), \tag{5}$$

$$P_L(x_0) = \min_{\substack{\delta_i \in [-\varepsilon_i, \varepsilon_i] \\ i=0,\ldots,n-1}} \tilde{P}(x_0). \tag{6}$$

By changing the value x_0 in $\mathbf{R}$, we obtain a domain

$$\{[P_L(x), P_U(x)] \mid x \in \mathbf{R}\}. \tag{7}$$

We call this domain the "existence domain of $\tilde{P}(x)$." □

The existence domain of $\tilde{P}(x)$ can be specified rigorously by using $P(x)$.

Lemma 1 Let the value of δ_i in $\tilde{P}(x)$ be changed continuously within the range $[-\varepsilon_i, \varepsilon_i]$, while the values of δ_j's $(j \ne i)$ are fixed, and, for each real value of x, define $P_{U_i}(x)$ and $P_{L_i}(x)$ as

$$P_{U_i}(x) = \max_{\delta_i \in [-\varepsilon_i, \varepsilon_i]} \tilde{P}(x), \tag{8}$$

$$P_{L_i}(x) = \min_{\delta_i \in [-\varepsilon_i, \varepsilon_i]} \tilde{P}(x). \tag{9}$$

Then, we have

$$P_{U_i}(x) = \begin{cases} \tilde{P}(x \mid \delta_i = \varepsilon_i) \\ \quad \text{if } x \ge 0 \text{ or } i \text{ is even}, \\ \tilde{P}(x \mid \delta_i = -\varepsilon_i) \\ \quad \text{if } x \le 0 \text{ and } i \text{ is odd}, \end{cases} \tag{10}$$

$$P_{L_i}(x) = \begin{cases} \tilde{P}(x \mid \delta_i = -\varepsilon_i) \\ \quad \text{if } x \ge 0 \text{ or } i \text{ is even}, \\ \tilde{P}(x \mid \delta_i = \varepsilon_i) \\ \quad \text{if } x \le 0 \text{ and } i \text{ is odd}. \end{cases} \tag{11}$$

Furthermore, for any real value x_0, $\tilde{P}(x_0)$ moves all the points inside $[P_{L_i}(x_0), P_{U_i}(x_0)]$.

Proof. Let x_0 be any real number. We see that $-\varepsilon_i |x_0|^i \le \delta_i |x_0|^i \le \varepsilon_i |x_0|^i$, and since $\delta_i |x_0|^i$ moves all the points inside $[-\varepsilon_i |x_0|^i, \varepsilon_i |x_0|^i]$, we obtain the lemma. □

This lemma directly leads us to the following theorem:

Theorem 2 Let the polynomials $P(x)$ and $\tilde{P}(x)$ be as above; then the functions $P_U(x)$ and $P_L(x)$ in Eq. (7) are given as follows:

$$P_U(x) = \begin{cases} \tilde{P}(x \mid \delta_i = \varepsilon_i \\ \qquad (i = n-1, \ldots, 0)) \\ \qquad \text{for } x \ge 0, \\ \tilde{P}(x \mid \delta_i = (-1)^i \varepsilon_i \\ \qquad (i = n-1, \ldots, 0)) \\ \qquad \text{for } x < 0, \end{cases} \tag{12}$$

$$P_L(x) = \begin{cases} \tilde{P}(x \mid \delta_i = -\varepsilon_i \\ \qquad (i = n-1, \ldots, 0)) \\ \qquad \text{for } x \ge 0, \\ \tilde{P}(x \mid \delta_i = (-1)^{i+1} \varepsilon_i \\ \qquad (i = n-1, \ldots, 0)) \\ \qquad \text{for } x < 0. \end{cases} \tag{13}$$

Furthermore, for any real number x_0, the values of $\tilde{P}(x_0)$ move all the points inside $[P_L(x_0), P_U(x_0)]$. □

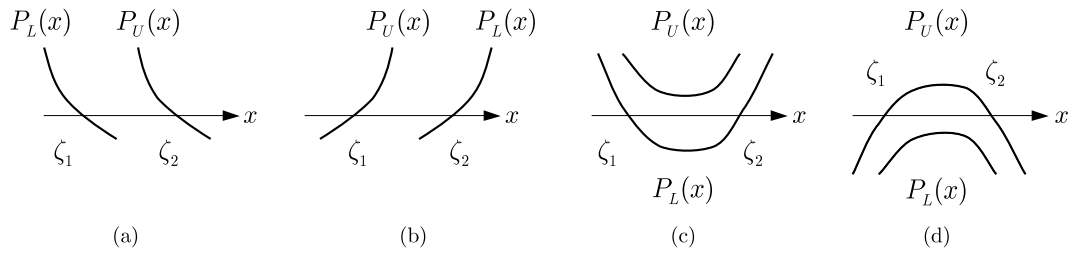


Fig. 1 Existence domain of an approximate real zero-point.

2.2 Approximate Real Zero-points and Their Existence Domains

We first define a concept of "approximate real zero-points" and their existence domains.

Definition 2 (approximate real zero-point) A real number ζ is an "approximate real zero-point of $\tilde{P}(x)$" if there exist numbers $\varepsilon'_i \in [-\varepsilon_i, \varepsilon_i]$ $(i = n-1, \ldots, 0)$ such that $\tilde{P}(\zeta \mid \delta_i = \varepsilon'_i \ (i = n-1, \ldots, 0)) = 0$. Let $[\zeta_{1,1}, \zeta_{1,2}], \ldots, [\zeta_{r,1}, \zeta_{r,2}]$, with $\zeta_{1,1} \le \zeta_{1,2} < \cdots < \zeta_{r,1} \le \zeta_{r,2}$, be the set of all the approximate real zero-points of $\tilde{P}(x)$. Then, we call each interval $[\zeta_{i,1}, \zeta_{i,2}]$, $1 \le i \le r$, an "existence domain" of the approximate real zero-point of $\tilde{P}(x)$. □

Theorem 2 tells us that the existence domains of all the approximate real zero-points can be specified rigorously by drawing graphs of $P_L(x)$ and $P_U(x)$. Suppose $[\zeta_1, \zeta_2]$ is an existence domain of an approximate real zero-point. Since ζ_1 and ζ_2 are real zero-points of $P_U(x)$ and/or $P_L(x)$, and since $P_L(x_0) < P_U(x_0)$ for any real number x_0, the graphs of $P_L(x)$ and $P_U(x)$ around this interval can be classified into one of the following four cases:

(a) $P_L(\zeta_1) = P_U(\zeta_2) = 0$, $P_L(x) < 0$ for $\zeta_1 < x \le \zeta_2$, $P_U(x) > 0$ for $\zeta_1 \le x < \zeta_2$, and there exists $\delta > 0$ such that $P_L(\zeta_1 - x) > 0$ and $P_U(\zeta_2 + x) < 0$ for any $x \in [0, \delta]$,

(b) $P_U(\zeta_1) = P_L(\zeta_2) = 0$, $P_U(x) > 0$ for $\zeta_1 < x \le \zeta_2$, $P_L(x) < 0$ for $\zeta_1 \le x < \zeta_2$, and there exists $\delta > 0$ such that $P_U(\zeta_1 - x) < 0$ and $P_L(\zeta_2 + x) > 0$ for any $x \in [0, \delta]$,

(c) $P_L(\zeta_1) = P_L(\zeta_2) = 0$, $P_U(x) > 0$ for $\zeta_1 \le x \le \zeta_2$, $P_L(x) < 0$ for $\zeta_1 < x < \zeta_2$, and there exists $\delta > 0$ such that $P_L(\zeta_1 - x) > 0$ and $P_L(\zeta_2 + x) > 0$ for any $x \in [0, \delta]$,

(d) $P_U(\zeta_1) = P_U(\zeta_2) = 0$, $P_L(x) < 0$ for $\zeta_1 \le x \le \zeta_2$, $P_U(x) > 0$ for $\zeta_1 < x < \zeta_2$, and there exists $\delta > 0$ such that $P_U(\zeta_1 - x) < 0$ and $P_U(\zeta_2 + x) < 0$ for any $x \in [0, \delta]$.

Figure 1 illustrates these four cases conceptually. Cases (a) and (b) usually correspond to a single zero-point, while Cases (c) and (d) correspond to multiple zero-points.

We now give a simple example of approximate real zero-points and their existence domains. We will see that one of the existence domains is fairly wide, which indicates that the concept of approximate zero-point is indispensable in handling polynomials with error terms.

Example 1 Let $F(x, y)$ be

$$F(x, y) = x^3 - x^2 + y^2. \tag{14}$$

We calculate a singular point of $F(x, y)$ with approximate arithmetic of precision $\varepsilon_M = 1.0 \times 10^{-6}$. First, let us calculate the discriminant $R(y)$ of $F(x, y)$ with respect to x:

$$\begin{aligned} R(y) &= \mathrm{res}(F, dF/dx) \\ &= 27y^4 - 4y^2. \end{aligned} \tag{15}$$

$R(y)$ has zero-points at $y = 0$ and $\pm 2\sqrt{3}/9$. Assume that we have calculated the value of $y = 2\sqrt{3}/9$ approximately as 0.384900. (Note that if $\deg(R) \ge 5$ then use of approximate arithmetic is necessary in general to solve $R(y) = 0$.) Let $P(x)$ and $\tilde{P}(x)$ be

$$\begin{aligned} P(x) &= x^3 - x^2 + (0.384900)^2, \\ \tilde{P}(x) &= P(x) + \delta_0, \end{aligned} \tag{16}$$

where $|\delta_0| \le 1.0 \times 10^{-6}$, and let us calculate the approximate real zero-points of $\tilde{P}(x)$. From Theorem 2, we have

$$\begin{aligned} P_U(x) &= x^3 - x^2 + 0.148149, \\ P_L(x) &= x^3 - x^2 + 0.148147. \end{aligned} \tag{17}$$

$P_U(x)$ has a real zero-point at $x \simeq -0.333334$, and $P_L(x)$ has real zero-points at $x \simeq -0.333332$, 0.665595, and 0.667738. From Definition 2, the existence domains of approximate real zero-points of $\tilde{P}(x)$ are intervals $[-0.333334, -0.333332]$, and $[0.665595, 0.667738]$. Therefore, with an approximate arithmetic of precision $\varepsilon_M = 1.0 \times 10^{-6}$, the singular point (x_0, y_0) of $F(x, y)$ can be specified only vaguely as $y_0 \in [0.384899, 0.384901]$ and $x_0 \in [0.665595, 0.667738]$. □

3. Bounding Existence Domains by Using Smith's Theorem

Although we have defined rigorously the existence domain of only real zero-points, we present in this section a method for bounding the existence domains of both real and complex zero-points by means of discs in the complex plane, because the method is common to both of them.

A key to bounding existence domains is Smith's celebrated theorem. (For the proof, see Smith [11].)

Theorem 3 (Smith) Let $P(x)$ be as above. Let $x_1, \ldots, x_n$ be n distinct numbers in $\mathbf{C}$ and $r_1, \ldots, r_n$ be defined as

$$r_j = \left| \frac{nP(x_j)}{a_n \prod_{k=1,\neq j}^{n}(x_j - x_k)} \right|, \quad j = 1, \ldots, n. \tag{18}$$

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

3.1 Single Zero-points

Without loss of generality, we assume that P and $\tilde{P}$ are monic. Let $\zeta_1, \ldots, \zeta_n$ and $\tilde{\zeta}_1, \ldots, \tilde{\zeta}_n$ be the zero-points of $P(x)$ and $\tilde{P}(x)$, respectively:

$$P(x) = (x - \zeta_1)(x - \zeta_2) \cdots (x - \zeta_n), \tag{19}$$

$$\tilde{P}(x) = (x - \tilde{\zeta}_1)(x - \tilde{\zeta}_2) \cdots (x - \tilde{\zeta}_n). \tag{20}$$

First, we consider the case in which ζ_1 is a single zero-point such that $|\zeta_1 - \zeta_j| \gg \varepsilon_M$ for $j = 2, \ldots, n$. Let $z_1, \ldots, z_n$ be approximate values for $\zeta_1, \ldots, \zeta_n$, respectively. (Actually, we may determine $z_1, \ldots, z_n$ by solving equation $P(x) = 0$ numerically, and hence approximately, with accuracy ε_M.) Using Theorem 3, we can formally calculate the domain that contains $\tilde{\zeta}_1$ in $\mathbf{C}$, as follows. Let R_1 be

$$R_1 = n \cdot \frac{|\tilde{P}(z_1)|}{\left| \prod_{j=2}^{n}(z_1 - z_j) \right|}, \tag{21}$$

then $\tilde{\zeta}_1$ is contained in the disc of radius R_1 with its center at z_1. Although we cannot calculate $\tilde{P}(z_1)$ explicitly, we have

$$\begin{aligned} |\tilde{P}(z_1)| &\leq |P(z_1)| + |\Delta(z_1)| \\ &\leq |P(z_1)| + \sum_{j=0}^{n-1} \varepsilon_j |z_1|^j. \end{aligned} \tag{22}$$

Therefore, R_1 is bounded as

$$R_1 \leq n \cdot \frac{|P(z_1)| + \sum_{j=0}^{n-1} \varepsilon_j |z_1|^j}{\left| \prod_{j=2}^{n}(z_1 - z_j) \right|}. \tag{23}$$

In ordinary numerical computation, we calculate an error bound by the above formula with $\varepsilon_j = 0$, which gives a good estimate such that the magnitude of the error bound is only several times larger than the true error. Therefore, we expect that the above formula gives a good bound.

3.2 Multiple or Close Zero-points

Next, we consider the case of multiple or close zero-points. Without loss of generality, let $\zeta_1 \simeq \cdots \simeq \zeta_m$ $(m \leq n)$ and assume that $\zeta_{m+1}, \ldots, \zeta_n$ satisfy $|\zeta_j - \zeta_1| \gg \sqrt[m]{\varepsilon_M}$ for $j = m+1, \ldots, n$. In this case, we cannot apply Eq. (23) directly, for the following reason. Let $z_1, \ldots, z_n$ be the same as above and assume that we have calculated them by a numerical method. Then $z_1, \ldots, z_m$ usually satisfy $|z_j - z_1| \simeq \sqrt[m]{\varepsilon_M}$ for $j = 2, \ldots, m$; hence, in Eq. (23), we have

$$\left| \prod_{j=2}^{n}(z_1 - z_j) \right| \simeq \varepsilon_M \cdot \left| \prod_{j=m+1}^{n}(z_1 - z_j) \right|. \tag{24}$$

Therefore, if $|\Delta(z_i)| \gg \varepsilon_M$, an upper bound calculated by Eq. (23) will be an overestimate.

We determine $z_1, \ldots, z_m$ so that the radius R_1 in Eq. (21) becomes as small as possible. (The determination method is the same as that described in the literature; for example, see Iri [5]; the only difference is that our setting of error terms is different from the conventional ones.) We express $P(x)$ as

$$P(x) = (x - \zeta_1) \cdots (x - \zeta_m) \cdot Q(x). \tag{25}$$

From our assumption, we have

$$\begin{aligned} Q(z_1) &= \prod_{j=m+1}^{n} (z_1 - \zeta_j) \\ &\simeq \prod_{j=m+1}^{n} (z_1 - z_j); \end{aligned} \tag{26}$$

hence R_1 defined by Eq. (21) can be approximated as follows:

$$R_1 = n \cdot \frac{\left| \prod_{j=1}^{n}(z_1 - \zeta_j) + \Delta(z_1) \right|}{\left| \prod_{j=2}^{n}(z_1 - z_j) \right|} \tag{27}$$

$$\simeq n \cdot \frac{\left| \prod_{j=1}^{m}(z_1 - \zeta_j) + \Delta(z_1)/Q(z_1) \right|}{\left| \prod_{j=2}^{m}(z_1 - z_j) \right|}. \tag{28}$$

If $z_1, \ldots, z_m$ are distributed equally on a disc of radius r with its center at $(\zeta_1 + \cdots + \zeta_m)/m$, we have

$$\begin{aligned} \left| \prod_{j=1}^{m} (z_1 - \zeta_j) \right| &\approx r^m, \\ \left| \prod_{j=2}^{m} (z_1 - z_j) \right| &= m r^{m-1}, \end{aligned} \tag{29}$$

and Eq. (28) can be evaluated as

$$R_1 \simeq n \cdot \frac{r^m + C}{m r^{m-1}}, \tag{30}$$

where $C = |\Delta(z_1)/Q(z_1)|$. We can almost minimize the magnitude of R_1 by setting r as

$$r = \sqrt[m]{(m-1)C}. \tag{31}$$

With the above consideration, we calculate an upper bound for R_1 as follows:

(1) Calculate r from Eq. (31).

(2) Let $\beta = (\zeta_1 + \cdots + \zeta_m)/m$ and
$$z_j = \beta + r \exp(2\pi j i/m) \tag{32}$$
for $j = 1, \ldots, m$. The approximate values $z_1, \ldots, z_m$ are distributed equally on a disc of radius r with its center at β.

(3) Substitute $z_1, \ldots, z_m$ into Eq. (23) to obtain a rigorous bound of R_1.

4. Calculating the Number of Real Zero-points of a Real Approximate Polynomial

If a real approximate polynomial has multiple or close zero-points, they may change significantly, or some real zero-points may become complex, when the coefficients are changed slightly. Therefore, it is not adequate to count the number of real zero-points of a real approximate polynomial that may have multiple or close zero-points. On the other hand, if a polynomial has only single zero-points, the number of its real zero-points rarely changes, although their positions may change considerably, when the coefficients are changed slightly. In this section, we focus on calculating the number of real zero-points of a real approximate polynomial containing only single zero-points.

4.1 Sufficient Condition for Fixing the Number of Real Zero-points

We first derive a sufficient condition for asserting that $P(x)$ and $\tilde{P}(x)$ have the same number of real zero-points.

Theorem 4 Let $P(x)$ and $\tilde{P}(x)$ be as in Eqs. (1) and (2), respectively. 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 $\mathrm{res}(\tilde{P}, d\tilde{P}/dx)$ does not become zero for any values $\delta_{n-1}, \ldots, \delta_0$ satisfying Eq. (4).

Proof. As the coefficients of $\tilde{P}(x)$ change continuously, the number of real zero-points of $\tilde{P}(x)$ changes only if there exist $\delta_i \in [-\varepsilon_i, \varepsilon_i]$ for $i = 0, \ldots, n-1$ such that $\tilde{P}(x)$ has real multiple zero-points. Its contraposition shows the validity of the theorem. □

Theorem 4 tells us that we can calculate the number of real zero-points of an unknown polynomial $\tilde{P}(x)$ by calculating the number of the real zero-points of $P(x)$, so long as the discriminant $\mathrm{res}(\tilde{P}, d\tilde{P}/dx)$ does not become zero for any values $\delta_{n-1}, \ldots, \delta_0$ satisfying Eq. (4). Therefore, we can check the definiteness of the number of real zero-points by checking whether or not $\mathrm{res}(\tilde{P}, d\tilde{P}/dx)$ becomes zero because of the error terms.

4.2 Problem of Small Leading Coefficient in the Sturm Sequence

Below, the leading coefficient and the degree of $P(x)$ are denoted as $\mathrm{lc}(P)$ and $\deg(P)$, respectively. Let $\zeta_{\max}$ be the maximum of the absolute values of real zero-points of $P(x)$.

The p-norm of $P(x)$, with $P(x)$ given in Eq. (1), is defined as

$$\|P\|_p = \left(\sum_{i=1}^{n} |c_i|^p \right)^{1/p}, \quad p = 1, 2, \ldots, \infty. \tag{33}$$

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

Assuming that $P(x)$ and $\tilde{P}(x)$ satisfy the condition in Theorem 4, $\|P\|_2 \simeq 1$, and $\|\tilde{P}\|_2 \simeq 1$, let us consider calculation of the number of real zero-points of $\tilde{P}(x)$ by application of Sturm's famous method to $P(x)$. Sturm's theorem is as follows (for the proof, see Cohen [3], for example):

Theorem 5 (Sturm) Let $P(x)$ be a real square-free polynomial of degree n, and define a polynomial sequence (the Sturm sequence)

$$(P_0(x), P_1(x), \ldots, P_n(x)) \tag{34}$$

as

$$\begin{cases} P_0 = P(x), \\ P_1 = \frac{d}{dx} P(x), \\ P_i = -\mathrm{rem}(P_{i-2}, P_{i-1}) \\ \qquad \text{for } i = 2, \ldots, n, \end{cases} \tag{35}$$

where $\mathrm{rem}(P_{i-2}, P_{i-1})$ denotes the remainder of P_{i-2} divided by P_{i-1}. For a real number x, let $N(x)$ be the number of sign changes, counting

from the left to the right without counting zeros, in the sequence (34), and let s and t be real numbers satisfying $s < t$. Then, the number of the real zero-points of P in the interval $[s, t]$ is equal to $N(s) - N(t)$. □

Note that we can calculate the number of all the real zero-points of P by putting $s = -\infty$ and $t = \infty$ in Theorem 5. In the following, the zeros of the Sturm sequence and its modifications are not counted as sign changes.

Consider calculation of the Sturm sequence of $P(x)$ by means of floating-point arithmetic. During the calculation, we may encounter the leading coefficient problem: (1) it is hard for us to decide whether or not a very small leading coefficient is equal to zero, and (2) the division by a polynomial by a small leading coefficient will cause large cancellation errors in the coefficients of the remainder polynomial.

Let P, s, and t be the same as in Theorem 5. A Sturm sequence of P with $P_n \equiv (\text{constant}) \neq 0$ has the following properties (for example, see Cohen [3]):

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 $(1 \leq j < n)$ and $x \in \mathbf{R}$, then we have $P_{j-1}(x)P_{j+1}(x) < 0$.

3° P_n has no real zero-point.

With Property 1°, we can calculate the number of sign changes by investigating each P_i separately. Let $P_j(x_j) = 0$ for some $x_j \in \mathbf{R}$; then Property 2° means that P_{j-1} and P_{j+1} have no zero-point in the neighborhood of $x = x_j$. Property 3° is trivial in our case, because $P_n = (\text{constant})$, but it is not trivial for the general Sturm sequence. The above three properties are sufficient for determining the number of real zero-points, and a sequence that has those properties is called a general Sturm sequence.

We note that the sign change of $P_j(x)$ at $x = x_j$, $j \geq 1$, does not affect the number of sign changes in the sequence (34); the value of $N(x)$ changes only when the evaluation point x passes a real zero-point of $P_0(x)$ $(= P(x))$. Furthermore, we can prove the following property of the Sturm sequence:

Lemma 6 Let $P(x)$ and $P_0, \ldots, P_n$ be the same as in Theorem 5, and assume that $P_k(x) = 0$ $(1 < k < n)$ at $x = x_{k,1}, \ldots, x_{k,l_k}$, where $l_k < \deg(P_k)$ and $|x_{k,j}| > \zeta_{\max}$ for $j = 1, \ldots, l_k$. Define $P''_k(x)$ as

$$P''_k(x) = \frac{P_k(x)}{(x - x_{k,1}) \cdots (x - x_{k,l_k})}, \quad (36)$$

and let s and t be real numbers satisfying $s < t$. For real number x, let $N(x)$ be the same as in Theorem 5, and let $N''_k(x)$ be the numbers of sign changes in the sequence

$$\begin{aligned} &(P_0(x), \ldots, P_{k-1}(x), \\ &\ P''_k(x), P_{k+1}(x), \ldots, P_n(x)). \end{aligned} \quad (37)$$

Then we have

$$N''_k(s) - N''_k(t) = N(s) - N(t). \quad (38)$$

That is, $N''_k(s) - N''_k(t)$ is equal to the number of real zero-points of $P(x)$ in the interval $[s, t]$.

Proof. Property 1° assures us that there exists a small positive number δ such that $[x_{k,j_1} - \delta, x_{k,j_1} + \delta] \cap [x_{k,j_2} - \delta, x_{k,j_2} + \delta] = \emptyset$ for $1 \leq j_1 < j_2 \leq l_k$ and $P_{k\pm1}(x) \neq 0$ for any $x \in [x_{k,j} - \delta, x_{k,j} + \delta]$. We show $N''_k(x) = N(x)$ for any $x \in [x_{k,j} - \delta, x_{k,j} + \delta]$. Consider a case in which $dP_k/dx < 0$ at $x = x_{k,1}$, $P_{k-1}(x_{k,1}) > 0$, and $P_{k+1}(x_{k,1}) < 0$. Property 2° says that the sequence of signs of polynomials $(P_{k-1}(x), P_k(x), P_{k+1}(x))$ at $x = x_{k,1} - \delta$, $x = x_{k,1}$ and $x = x_{k,1} + \delta$ are $(+, +, -)$, $(+, 0, -)$ and $(+, -, -)$, respectively; hence the number of sign changes of the sequence $(P_{k-1}(x), P_k(x), P_{k+1}(x))$ is equal to 1 for any $x \in [x_{k,1} - \delta, x_{k,1} + \delta]$. Now, assume that $P''_k(x) > 0$ for $x \in [x_{k,1} - \delta, x_{k,1} + \delta]$; then the sequence of signs of polynomials $(P_{k-1}(x), P''_k(x), P_{k+1}(x))$ is $(+, +, -)$ for any $x \in [x_{k,1} - \delta, x_{k,1} + \delta]$. Therefore, we have $N''_k(x) = N(x)$ for any $x \in [x_{k,1} - \delta, x_{k,1} + \delta]$. The other cases can be proved similarly. □

Theorem 7 Assume the same hypotheses as in Lemma 6, and define a polynomial sequence

$$(P_0(x), \ldots, P_{k-1}(x), P''_k(x), \ldots, P''_{n''}(x)) \quad (39)$$

as

$$\begin{cases} P''_k &= \dfrac{P_k(x)}{(x - x_{k,1}) \cdots (x - x_{k,l_k})}, \\ P''_{k+1} &= -\mathrm{rem}(P_{k-1}, P''_k), \\ P''_i &= -\mathrm{rem}(P''_{i-2}, P''_{i-1}) \\ &\quad \text{for } i = k+2, \ldots, n'', \end{cases} \quad (40)$$

where $\deg(P''_{n''}) = 0$. For a real number x, let $N''(x)$ be the number of sign changes in the sequence (39), and let s and t be real numbers satisfying $s < t$. Then, the number of real zero-points of $P(x)$ in the interval $[s, t]$ is equal to $N''(s) - N''(t)$.

Proof. From Lemma 6, we need not consider $x_{k,1}, \ldots, x_{k,l_k}$ for calculating the number of real zero-points of $P(x)$. Let x_k be any zero-point of P''_k; hence $P_{k-1}(x_k) \neq 0$. Then, $P_{k-1}(x_k) \cdot P''_{k+1}(x_k) < 0$ because $-P''_{k+1}(x) =$

$P_{k-1}(x) - Q''_k(x)P''_k(x)$. Repeating this argument for P''_{k+1}, P''_{k+2}, and so on, we see that the new polynomial sequence (39) satisfies Properties 1°, 2°, and 3° described above, and that the sequence (39) is a general Sturm sequence of $P(x)$. Thus, we can count all the real zero-points of $P(x)$ by using the sequence (39). □

Remark 1 Properties 1°, 2°, and 3° are enough to prove Theorem 7, and Lemma 6 is unnecessary. We introduced Lemma 6 to help the reader to understand what happens when large real zero-points of P_k are removed. □

In Theorem 7, calculating the general Sturm sequence by using P''_k in Eq. (40) is theoretically simple but not practical, because we have to calculate the real zero-points of P_k rigorously. We next show that, if a polynomial has small leading terms, these terms correspond to zero-points of large magnitudes.

Lemma 8 Let $\varepsilon_n, \ldots, \varepsilon_{n-s+1}$ be real numbers such that $0 < |\varepsilon_j| \ll 1$, and, without loss of generality, let $Q(x)$ be

$$\begin{aligned} Q(x) = {} & \varepsilon_n x^n + \cdots + \varepsilon_{n-s+1} x^{n-s+1} \\ & + b_{n-s} x^{n-s} + \cdots + b_0 x^0, \end{aligned} \tag{41}$$

where $|b_i| \geq 1$ $(i = n-s, \ldots, 0)$ for $b_i \neq 0$. Let $x_1, \ldots, x_n$ be the zero-points of $Q(x)$ such that $|x_1| < \cdots < |x_n|$. Then we have

$$\begin{aligned} & \lim_{(\varepsilon_n, \ldots, \varepsilon_{n-s+1}) \to (0, \ldots, 0)} |x_j| = \infty, \\ & j = n-s+1, \ldots, n. \end{aligned} \tag{42}$$

Proof. Define $Q_I(x)$ as

$$\begin{aligned} Q_I(x) &= x^n \cdot Q(1/x) \\ &= \bar{b}_n x^n + \cdots + \bar{b}_0 x^0, \end{aligned} \tag{43}$$

and let $\bar{x}_1, \ldots, \bar{x}_n$ be the zero-points of $Q_I(x)$ with $|\bar{x}_1| < \cdots < |\bar{x}_n|$. Then we have $\bar{b}_{n-j} = \varepsilon_j$ for $j = n, \ldots, n-s+1$ and $\bar{x}_{n-i+1} = 1/x_i$ for $i = 1, \ldots, n$. We have $|\bar{x}_i| \to 0$ $(i = n, \ldots, n-s+1)$ for $|\bar{b}_{n-j}| \to 0$ $(j = n, \ldots, n-s+1)$; hence $|x_i| \to \infty$ for $\varepsilon_j \to 0$. □

Remark 2 Although Lemma 8 is a limiting case of $(\varepsilon_n, \ldots, \varepsilon_{n-s+1}) \to (0, \ldots, 0)$ and is sufficient to prove Theorem 9, we investigate the location of zero-points of $Q_I(x)$ in the appendix. □

Theorem 7 and Lemma 8 lead us to an idea of discarding the small leading terms to calculate a general Sturm sequence in practice. Since the zero-points of $P_k(x)$ are moved slightly by discarding the small leading terms, we must be more careful than in Theorem 7.

Theorem 9 Define $P(x)$ and $\tilde{P}(x)$ as in Eqs. (1) and (2), respectively. Let $(P_0 = P(x), P_1 = dP/dx, P_2, \ldots, P_i, \ldots)$ be the Sturm sequence of $P(x)$ and assume that $P_k(x)$ has small leading terms as

$$\begin{aligned} P_k(x) = {} & \varepsilon_{k,n_k} x^{n_k} + \cdots \\ & \cdots + \varepsilon_{k,n_k-s+1} x^{n_k-s+1} \\ & + b_{k,n_k-s} x^{n_k-s} + \cdots \\ & \cdots + b_{k,0} x^0, \end{aligned} \tag{44}$$

where

$$\begin{aligned} & \max\{|\varepsilon_{k,n_k}|, \ldots, |\varepsilon_{k,n_k-s+1}|\} \\ & \ll \min_{b_{k,j} \neq 0} \{|b_{k,n_k-s}|, \ldots, |b_{k,0}|\}. \end{aligned}$$

Define a polynomial sequence

$$\begin{aligned} & (P_0(x), \ldots, P_{k-1}(x), \\ & P'_k(x), \ldots, P'_{n'}(x)) \end{aligned} \tag{45}$$

as

$$\begin{cases} P'_k &= b_{k,n_k-s} x^{n_k-s} + \cdots \\ & \qquad \cdots + b_{k,0} x^0, \\ P'_{k+1} &= -\mathrm{rem}(P_{k-1}, P'_k), \\ P'_i &= -\mathrm{rem}(P'_{i-2}, P'_{i-1}) \\ & \quad \text{for } i = k+2, \ldots, n', \end{cases} \tag{46}$$

where $\deg(P'_{n'}) = 0$. For a real number x, let $N'(x)$ be the number of sign changes in the sequence (45), and let s and t be real numbers such that $s < -\zeta_{\max}$ and $\zeta_{\max} < t$. Then, if $\tilde{P}(x)$, $P_{k-1}(x)$, and $P_k(x)$ satisfy the following two conditions, the number of real zero-points of $\tilde{P}(x)$ is equal to $N'(s) - N'(t)$:

(1) The resultant $\mathrm{res}(\tilde{P}, P_k)$ does not become zero for any values $\delta_{n-1}, \ldots, \delta_0$ satisfying Eq. (4) or when the values of $\varepsilon_{k,n_k}, \ldots, \varepsilon_{k,n_{k-s+1}}$ are changed to zero.

(2) The resultant $\mathrm{res}(P_{k-1}, P_k)$ does not become zero when the values of $\varepsilon_{k,n_k}, \ldots, \varepsilon_{k,n_{k-s+1}}$ are changed to zero.

Proof. Even if $P_k(x)$ has real zero-points whose magnitudes are larger than that of any zero-point of $\tilde{P}(x)$, Lemma 8 and Condition (1) assure us that these real zero-points will be "safely removed" from $P_k(x)$ by changing the values of $\varepsilon_{k,n}, \ldots, \varepsilon_{k,n-s+1}$ to 0. We also see that the removed zero-points do not affect the calculation of the number of real zero-points, as Theorem 7 shows. Next, changing the values of $\varepsilon_{k,j}$'s to 0 will change the values of the other zero-points of $P_k(x)$ slightly. However, Condition (2) assures us that none of the real zero-points of $P_k(x)$ passes through the real zero-points of $P_{k-1}(x)$; hence the sequence (45) is a general Sturm sequence. Therefore, as in Theorem 7, we can calculate the number of real zero-points of $\tilde{P}(x)$ by using the sequence (45). □

Theorem 9 tells us that the problem of small leading coefficients reduces to checking whether or not any resultants become zero. We will propose several methods for this in Section 5.

We explain Theorem 9 by means of an example with exact arithmetic.

Example 2 Let $P(x)$ and $\tilde{P}(x)$ be

$$\begin{aligned} P(x) &= x^5 + 4x^4 + \tfrac{6401}{1000}x^3 \\ &\quad - 20x^2 + 5x + 1, \\ \tilde{P}(x) &= P(x) + \delta_{0,4}x^4 \\ &\quad + \delta_{0,3}x^3 + \cdots + \delta_{0,0}x^0, \end{aligned} \tag{47}$$

where numbers $\delta_{0,4}, \ldots, \delta_{0,0}$ are unknown but bounded as

$$|\delta_{0,j}| \le \varepsilon = 1/10000. \tag{48}$$

We obtain $(P_0, \ldots, P_5)$, the Sturm sequence of $P(x)$, as follows:

$$\begin{aligned} P_0 &= P(x), \\ P_1 &= \tfrac{d}{dx}P(x) \\ &= 5x^4 + 16x^3 + \tfrac{19203}{1000}x^2 \\ &\quad -40x + 5, \\ P_2 &= -\tfrac{1}{2500}x^3 + \tfrac{94203}{6250}x^2 \\ &\quad - \tfrac{52}{5}x - \tfrac{1}{5}, \\ P_3 &= -\tfrac{7099837085603}{1000}x^2 \\ &\quad + 4898974540x + 94210995, \\ P_4 &= -\tfrac{1838986143841703970}{5040768664210370074987360 9}x \\ &\quad + \tfrac{5814705282399344 09}{5040768664210370074987360 9}, \\ P_5 &= -(31566508567667286525829957694414724087925197085 57) \\ &\quad / (33818700372417803243846409931137603 00900000). \end{aligned} \tag{49}$$

Therefore, we have $N(-\infty) - N(\infty) = 3$.

In Eq. (49), P_2 has a small leading coefficient. (Correspondingly, $P_2(x)$ has a real zero-point at $x \simeq 37680.5$.) The conditions in Theorem 9 are satisfied as follows. First, the existence domains of approximate zero-points of $\tilde{P}(x)$ in the neighborhood of $x = 0$ are the intervals $[-0.12992, -0.12989]$, $[0.44536, 0.44541]$, and $[0.19803, 0.19810]$, while the existence domains of approximate zero-points of $P_2(x)$ when we change the value of the leading coefficient continuously from $-1/2500$ to 0 are the intervals $[-0.01877227\cdots, -0.01877227\cdots]$, $[0.708722, 0.708735]$, and $[37680.5, \infty)$. Therefore, the existence domains of the real zero-points of $\tilde{P}(x)$ and $P_2(x)$ do not overlap; hence we have $\mathrm{res}(\tilde{P}, P_2) \neq 0$. Second, the existence domains of approximate zero-points of $P_1(x)$ are the intervals $[0.134731, 0.134738]$, and $[0.910227, 0.910260]$. Therefore, the existence domains of the real zero-points of $P_1(x)$ and $P_2(x)$ do not overlap; hence we have $\mathrm{res}(P_1, P_2) \neq 0$. Since $P(x)$, $\tilde{P}(x)$, P_1, and P_2 satisfy the conditions in Theorem 9, we can calculate $P_2', \ldots, P_4'$ as follows:

$$\begin{aligned} P_2' &= \tfrac{94203}{6250}x^2 - \tfrac{52}{5}x - \tfrac{1}{5}, \\ P_3' &= \tfrac{14367059719609325}{835976753303427}x \\ &\quad - \tfrac{181700163229606 75}{3343907013213708}, \\ P_4' &= (6544015983161815588348053 \\ &\qquad 0106785213) \\ &\quad / (3302598479789132420606890 \\ &\qquad 0312900000). \end{aligned} \tag{50}$$

We have $N'(-\infty) - N'(\infty) = 3 = N(-\infty) - N(\infty)$. □

5. Evaluating the Effects of Error Terms

Theorems 4 and 9 show that some important problems in counting the number of approximate real zero-points can be reduced to checking whether or not some resultants become zero owing to the error terms. In this section, we consider how to evaluate errors in the resultant of an approximate univariate polynomial. We investigate four methods: (1) evaluating the "subresultant determinant" by using Hadamard's inequality, (2) calculating the Sturm sequence with the coefficients of interval numbers, (3) solving a linear system on polynomial coefficients and evaluating errors in the solution by a standard method in numerical analysis, and (4) calculating the Sturm sequence with parametric error terms. The experiments were performed with GAL (General Algebraic Language/Laboratory, a LISP-based computer algebra system) on NS-LISP (Nara Standard LISP) running on a SPARC Station 5 (CPU: microSPARC II, 70 MHz) and SunOS 4.1.4.

5.1 Evaluation of the Subresultant Determinant

Except for the overall signs of polynomials, the Sturm sequence is the same as the polynomial remainder sequence (PRS) for which the subresultant theory has been developed. (For subresultant theory, see Mishra [7], for example.) With this theory, we can express the elements in the Sturm sequence by the determinants of the coefficients of two consecutive elements. Let $(P_0 = P, P_1 = dP/dx, P_2, \ldots, P_{k-1}, P_k, \ldots)$ be a Sturm sequence, and assume that

$$\begin{aligned} P_{k-1}(x) &= a_l x^l + \cdots + a_0 x^0, \\ P_k(x) &= \varepsilon_m x^m + \cdots \\ &\quad \cdots + \varepsilon_{m-s+1}x^{m-s+1} \\ &\quad + b_{m-s}x^{m-s} + \cdots + b_0 x^0, \end{aligned} \tag{51}$$

where

$$\begin{aligned} &\max\{|\varepsilon_{k,n_k}|, \ldots, |\varepsilon_{k,n_k-s+1}|\} \\ &\ll \min_{b_{k,j}\neq 0}\{|b_{k,n_k-s}|, \ldots, |b_{k,0}|\} \end{aligned}$$

as before.

Let $\mathrm{S}_i(P_{k-1}, P_k)$ be the following determinant:

$$\mathrm{S}_i(P_{k-1}, P_k) = \begin{vmatrix} a_l \cdots & \cdots & \cdots & \cdots \\ \ddots & & & \\ & a_l & \cdots & \cdots \\ \varepsilon_m \cdots & \varepsilon_{m-s+1} & b_{m-s} & \cdots \\ \ddots & & \ddots & \ddots \\ & \varepsilon_m & \cdots & \varepsilon_{m-s+1} \\ \cdots & \cdots & a_{l-2i+1} & x^{i-1}P_{k-1} \\ & & \vdots & \vdots \\ \cdots & \cdots & a_{l-i} & x^0 P_{k-1} \\ \cdots & \cdots & b_{m-2i+1} & x^i P_k \\ & & \vdots & \vdots \\ b_{m-s} \cdots & & b_{m-i+1} & x^0 P_k \end{vmatrix}. \tag{52}$$

$\mathrm{S}_i(P_{k-1}, P_k)$ is called the i-th subresultant of $P_{k-1}(x)$ and $P_k(x)$, and we have $P_{k+i}(x) = \gamma_i \mathrm{S}_i(P_{k-1}, P_k)$, with γ_i a constant. For example, if $\deg(P_{k-1}) = \deg(P_k) + 1$, we have

$$\begin{aligned} P_{k+1}(x) &= \mathrm{S}_1(P_{k-1}, P_k) \\ &= \begin{vmatrix} a_l & a_{l-1} & P_{k-1}(x) \\ \varepsilon_m & \varepsilon_{m-1} & xP_k(x) \\ & \varepsilon_m & P_k(x) \end{vmatrix}. \end{aligned} \tag{53}$$

Below, we consider only the leading coefficients of P_{k+1}, P_{k+2}, and so on. Applying Hadamard's inequality to the subresultant, we can bound the effect of $\varepsilon_m, \ldots, \varepsilon_{m-s+1}$ on $\mathrm{lc}(P_{k+i})$, as follows:

Proposition 10 Define P'_k and L as follows.

$$\begin{aligned} P'_k &= P_k - (\varepsilon_m x^m + \cdots \\ &\qquad \cdots + \varepsilon_{m-s+1} x^{m-s+1}) \\ &= b_{m-s} x^{m-s} + \cdots + b_0, \end{aligned} \tag{54}$$

$$\begin{aligned} L = {} & \|P_k\|_2^{(i-1)} \times \\ & \Big\{ (i-s)\,|a_l|^s \, \|P_{k-1}\|_2^{(i-s)} \\ & + \sum_{j=1}^{s} |a_l|^{(j-1)} \|P_{k-1}\|_2^{(i-j+1)} \Big\}. \end{aligned} \tag{55}$$

If $\mathrm{lc}(\mathrm{S}_i(P_{k-1}, P_k)) \neq 0$ and

$$\begin{aligned} & \{|\varepsilon_m| + \cdots + |\varepsilon_{m-s+1}|\} \cdot L \\ & < |{a_l}^s \cdot \mathrm{lc}(\mathrm{S}_i(P_{k-1}, P'_k))|, \\ & \quad i = s, \ldots, m, \end{aligned} \tag{56}$$

then

$$\begin{aligned} & \mathrm{lc}(\mathrm{S}_i(P_{k-1}, P_k)) \times \\ & {a_l}^s \cdot \mathrm{lc}(\mathrm{S}_i(P_{k-1}, P'_k)) > 0. \end{aligned} \tag{57}$$

Proof. Note that

$$\mathrm{lc}(\mathrm{S}_i(P_{k-1}, P_k)) = \begin{vmatrix} a_l \cdots & \cdots & \cdots & \cdots \\ \ddots & & & \\ & a_l & \cdots & \cdots \\ \varepsilon_n \cdots & \varepsilon_{m-s+1} & b_{m-s} & \cdots \\ \ddots & & \ddots & \ddots \\ & \varepsilon_m & \cdots & \varepsilon_{m-s+1} \\ & \cdots & \cdots & a_{l-2i} \\ & & & \vdots \\ & \cdots & \cdots & a_{l-i-1} \\ & \cdots & \cdots & b_{m-2i} \\ & & & \vdots \\ & b_{m-s} & \cdots & b_{m-i} \end{vmatrix}, \tag{58}$$

and

$$\mathrm{lc}(\mathrm{S}_i(P_{k-1}, P'_k)) = \begin{vmatrix} a_l & \cdots & \cdots & \cdots & a_{l-2i+s} \\ & \ddots & & & \vdots \\ & & a_l & \cdots & a_{l-i-1} \\ b_{m-s} & \cdots & \cdots & \cdots & b_{m-2i} \\ & \ddots & & & \vdots \\ & & b_{m-s} & \cdots & b_{m-i} \end{vmatrix}, \tag{59}$$

where $a_j = b_j = 0$ for $j < 0$. By expanding the determinant in Eq. (58) with respect to the $(i+1)$-th row as

$$\begin{aligned} & \begin{vmatrix} \cdots & & \cdots \\ \varepsilon_m \cdots & \varepsilon_{m-s+1} \; b_{m-s} & \cdots b_{m-2i} \\ \cdots & & \cdots \end{vmatrix} \\ & = \begin{vmatrix} \cdots & & \cdots \\ \varepsilon_m \cdots & \varepsilon_{m-s+1} \; 0 & \cdots 0 \\ \cdots & & \cdots \end{vmatrix} \\ & \quad + \begin{vmatrix} \cdots & & \cdots \\ 0 \cdots & 0 \; b_{m-s} & \cdots b_{m-2i} \\ \cdots & & \cdots \end{vmatrix}, \end{aligned} \tag{60}$$

and expanding the last determinant similarly, we finally obtain

$$\begin{aligned} & \mathrm{lc}(\mathrm{S}_i(P_{k-1}, P_k)) \\ & = {a_l}^s \cdot \mathrm{lc}(\mathrm{S}_i(P_{k-1}, P'_k)) \\ & \qquad + \sum_{j=1}^{i+1} \det(R_{i,j}), \end{aligned} \tag{61}$$

where

$$R_{i,j} = \begin{pmatrix} a_l & \cdots & \cdots & \cdots & \cdots \\ & \ddots & \cdots & \cdots & \cdots \\ & & a_l & \cdots & \cdots \\ & & & b_{m-s} & \cdots \\ & & & & \ddots \\ & & & & \\ & & & \varepsilon_m & \cdots \\ & & & & \varepsilon_m \\ & & & & \\ & & & & \end{pmatrix}$$

$$\begin{pmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ b_{m-s} & \cdots & \cdots & \cdots \\ \varepsilon_{m-s+1} & 0 & \cdots & \cdots \\ \cdots & \varepsilon_{m-s+1} & b_{m-s} & \cdots \\ \ddots & \cdots & \ddots & \ddots \\ & \varepsilon_m & \cdots & \varepsilon_{m-s+1} \end{pmatrix} \quad (62)$$

$$\begin{pmatrix} \cdots & \cdots & a_{l-2i} \\ & & \vdots \\ \cdots & \cdots & a_{l-i-1} \\ \cdots & \cdots & b_{m-2i} \\ & & \vdots \\ \cdots & \cdots & b_{m-2i+j-2} \\ \cdots & \cdots & 0 \\ \cdots & \cdots & b_{m-2i+j} \\ & & \vdots \\ b_{m-s} & \cdots & b_{m-i} \end{pmatrix}.$$

Expanding $\det(R_{i,j})$ with respect to the $(i+j)$-th row, or the row

$$(0 \cdots 0\ \varepsilon_m \cdots \varepsilon_{m-s+1}\ 0 \cdots 0), \quad (63)$$

we have

$$\begin{aligned} &\det(R_{i,j}) \\ &= (-1)^{i+2j}\varepsilon_m \det(\tilde{R}_{i,j,j}) + \cdots \\ &\quad \cdots + (-1)^{i+2j+s}\varepsilon_{m-s+1} \det(\tilde{R}_{i,j,j+s}), \end{aligned} \quad (64)$$

where $\tilde{R}_{i,j,q}$ is a $2i \times 2i$ submatrix obtained by removing the $(i+j)$-th row and the q-th column from $R_{i,j}$. After removing several top-left diagonal elements a_l's of $\tilde{R}_{i,j,q}$, and applying Hadamard's inequality to $\det(\tilde{R}_{i,j,q})$, with inequalities $|a_l|^2 + |a_{l-1}|^2 + \cdots + |a_{l-2i}|^2 \le \|P_{k-1}\|_2^2$ and $|\varepsilon_m|^2 + \cdots + |\varepsilon_{m-s+1}|^2 + |b_{m-s}|^2 + \cdots + |b_{m-2i}|^2 \le \|P_k\|_2^2$, we finally obtain the following inequality:

$$\begin{aligned} &|\det(R_{i,j})| \\ &\le M_i\{|a_l|^{(j-1)}\, \|P_{k-1}\|_2^{(i-j+1)}\}, \\ &\qquad j = 1, \ldots, s, \end{aligned} \quad (65)$$

$$\begin{aligned} &|\det(R_{i,j})| \\ &\le M_i\{|a_l|^{s}\, \|P_{k-1}\|_2^{(i-s)}\}, \\ &\qquad j = s+1, \ldots, i+1, \end{aligned} \quad (66)$$

where

$$M_i = \{|\varepsilon_m| + \cdots + |\varepsilon_{m-s+1}|\} \cdot \|P_k\|_2^i. \quad (67)$$

From the assumption (56), we have

$$\begin{aligned} &\left|\sum_{j=1}^{i+1} \det(R_{i,j})\right| \le \sum_{j=1}^{i+1} |\det(R_{i,j})| \\ &\le \{|\varepsilon_m| + \cdots |\varepsilon_{m-s+1}|\} \cdot L \\ &< |a_l{}^s \cdot \mathrm{lc}(\mathrm{S}_i(P_{k-1}, P'_k))|. \end{aligned} \quad (68)$$

Therefore, from Eqs. (61) and (68), we obtain Eq. (57). □

From the fundamental theorem of subresultants [2], we have

$$\begin{aligned} \mathrm{S}_i(P_{k-1}, P'_k) &= P'_{k+h}\, \mathrm{lc}(P'_{k+h})^{d_{k+h-1}-1} \\ &\times \prod_{l=1}^{h} \Big\{ \mathrm{lc}(P'_{k+l-1})^{(d_{k+l-2}+d_{k+l-1})} \\ &\times (-1)^{(n_{k+l-2}-n_{k+h})(n_{k+l-1}-n_{k+h})} \Big\}, \end{aligned} \quad (69)$$

where $h = i - s$, $n_{k+j} = \deg(P'_{k+j})$ and $d_j = n_j - n_{j+1}$. Therefore, we can calculate $\mathrm{lc}(\mathrm{S}_i(P_{k-1}, P'_k))$ easily from $\mathrm{lc}(P'_{k+h})$.

Proposition 10 shows that, so long as $\varepsilon_m, \ldots, \varepsilon_{m-s+1}$ satisfy the condition (56), discarding terms $\varepsilon_m x^m, \ldots, \varepsilon_{m-s+1} x^{m-s+1}$ in P_k does not change the signs of leading coefficients of the subresultants $\mathrm{S}_i(P_{k-1}, P'_k)$ for $i = 0, \ldots, m-s-1$. However, in actual calculation of the Sturm sequence, the number L in Eq. (55) seems to become too large, hence the condition (56) is not useful in practice.

5.2 Utilization of Interval Arithmetic

In this method, we transform the coefficients of the given polynomial into interval numbers each of which includes the corresponding error, and calculate the Sturm sequence by using interval arithmetic.

By observing how the widths of intervals increased during the calculation, we found that the increase of the width of each interval was about one decimal-digit for each remainder computation. In fact, the division of polynomials of degree difference 1 requires two "polynomial × number" multiplications and two polynomial subtractions. The width of an interval is increased to about twice that of the original interval by one arithmetic operation if the

operands are of almost the same widths; hence the width increases by about $2^4 = 16$ times after the polynomial division. Thus, for a polynomial of degree 10, for example, the width of an interval in the last element of the Sturm sequence may become about 10^{10} times larger than the initial widths, which shows that this method is not 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 can be reduced to solving a linear system.

Usually, the norm of vectors and matrices are defined as follows. Let $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ be a vector in $\mathbf{R}^m$. Then, the p-norm of $\boldsymbol{x}$ is defined as

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{1/p}, \quad p = 1, 2, \infty. \tag{70}$$

Let $A = (a_{ij})$ be a real (m, m)-matrix. Then, by using the norm of a vector, we define the p-norm of A as

$$\|A\|_p = \max_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|A\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}. \tag{71}$$

In this paper we use only $\|A\|_1$ and $\|A\|_\infty$.

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

$$F(x) = f_m x^m + \cdots + f_0 x^0, \quad f_m \neq 0, \tag{72}$$

$$G(x) = g_n x^n + \cdots + g_0 x^0, \quad g_n \neq 0, \tag{73}$$

where $m \geq n$. Calculation of the PRS is equivalent to eliminating the terms of higher degrees of F and G to derive R_s, a polynomial of degree s, for $0 \leq s \leq n-1$. For each R_s, there exist polynomials U_s and V_s such that

$$\begin{aligned} &U_s F + V_s G = R_s, \\ &\deg(U_s) \leq n - s - 1, \\ &\deg(V_s) \leq m - s - 1. \end{aligned} \tag{74}$$

We consider calculating $R_0 = \mathrm{res}(F, G)$. Let U_0 and V_0 be expressed as

$$U_0 = u_{n-1} x^{n-1} + \cdots + u_0 x^0, \tag{75}$$

$$V_0 = v_{m-1} x^{m-1} + \cdots + v_0 x^0. \tag{76}$$

From the relation $U_0 F + V_0 G = R_0$, we obtain a system of linear equations on the coefficients in U_0 and V_0, as follows:

$$\begin{pmatrix} f_m & & & g_n & & \\ \vdots & f_m & & \vdots & g_n & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & f_m & \vdots & \vdots & \vdots & g_n \\ f_0 & \vdots & \vdots & \vdots & g_0 & \vdots & \vdots & \vdots \\ & f_0 & \vdots & \vdots & & g_0 & \vdots & \vdots \\ & & \ddots & \vdots & & & \ddots & \vdots \\ & & & f_0 & & & & g_0 \end{pmatrix} \times \begin{pmatrix} u_{n-1} \\ u_{n-2} \\ \vdots \\ u_0 \\ v_{m-1} \\ v_{m-2} \\ \vdots \\ v_0 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ R_0 \end{pmatrix}. \tag{77}$$

U_0 and V_0 can be normalized in any way so long as U_0 and V_0 satisfy the above relation. Therefore, we normalize U_0 and V_0 as $u_{n-1} = g_n$ and $v_{m-1} = -f_m$. With this normalization, we can rewrite the relation (77) as

$$\begin{pmatrix} f_m & & & g_n & & \\ \vdots & \ddots & & \vdots & \ddots & \\ \vdots & \vdots & f_m & \vdots & \vdots & g_n \\ f_1 & \vdots & \vdots & f_m\, g_1 & \vdots & \vdots & g_n \\ f_0 & \ddots & \vdots & \vdots & g_0 & \ddots & \vdots & \vdots \\ & \ddots & \vdots & \vdots & & \ddots & \vdots & \vdots \\ & & f_1 & \vdots & & & g_1 & \vdots \\ & & f_0 & f_1 & & & g_0\, g_1 \end{pmatrix} \begin{pmatrix} u_{n-2} \\ \vdots \\ \vdots \\ u_0 \\ v_{m-2} \\ \vdots \\ \vdots \\ v_0 \end{pmatrix} = \begin{pmatrix} g_{n-1} f_m - f_{m-1} g_n \\ \vdots \\ g_{n-m} f_m - f_0 g_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{78}$$

where $g_j = 0$ for $j < 0$, and

$$R_0 = f_0 u_0 + g_0 v_0. \tag{79}$$

The linear system (78) is of the form

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{80}$$

where A is a "coefficient matrix," and $\boldsymbol{x}$ and $\boldsymbol{b}$ are vectors of unknowns and given numbers, respectively. We briefly describe a perturbation

Table 1 Condition number of the matrix in Eq. (78) computed for 10 polynomials with random-number coefficients.

Degree of $P(x)$	Condition number					
	1-norm			∞-norm		
	Maximum	Minimum	Average	Maximum	Minimum	Average
10	8.73×10^3	1.69×10^2	2.55×10^3	7.96×10^3	2.92×10^2	2.99×10^3
20	2.57×10^6	4.44×10^3	2.95×10^5	8.51×10^5	1.83×10^3	1.08×10^5
30	1.16×10^7	4.97×10^4	2.46×10^6	5.97×10^7	2.45×10^4	1.18×10^6
40	5.37×10^7	1.48×10^5	7.44×10^6	4.76×10^7	6.01×10^4	6.00×10^6
50	1.47×10^8	1.56×10^5	2.25×10^7	6.42×10^7	7.38×10^4	8.09×10^6

theory for linear system. (The theory can be found in various works on numerical analysis; see Higham [4] for example.) Assume that $\boldsymbol{b}$ has an error $\Delta\boldsymbol{b}$ that causes an error $\Delta\boldsymbol{x}_1$ in the solution $\boldsymbol{x}$. Then we have

$$A(\boldsymbol{x} + \Delta\boldsymbol{x_1}) = \boldsymbol{b} + \Delta\boldsymbol{b}. \tag{81}$$

Using Eq. (80), we can easily evaluate the magnitude of $\Delta\boldsymbol{x}_1$ as

$$\frac{\|\Delta\boldsymbol{x_1}\|}{\|\boldsymbol{x}\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}. \tag{82}$$

Furthermore, assume that A has an error ΔA and that the error of $\boldsymbol{x}$ becomes $\Delta\boldsymbol{x_1} + \Delta\boldsymbol{x_2}$, as follows:

$$(A + \Delta A)(\boldsymbol{x} + \Delta\boldsymbol{x}_1 + \Delta\boldsymbol{x}_2) = \boldsymbol{b} + \Delta\boldsymbol{b}. \tag{83}$$

Using Eq. (81), we derive the following evaluation of $\Delta\boldsymbol{x}_2$.

$$\frac{\|\Delta\boldsymbol{x}_2\|}{\|\boldsymbol{x} + \Delta\boldsymbol{x}_1 + \Delta\boldsymbol{x}_2\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\Delta A\|}{\|A\|}. \tag{84}$$

Equations (82) and (84) lead us to the following evaluation:

$$\frac{\|\Delta\boldsymbol{x}_1\| + \|\Delta\boldsymbol{x}_2\|}{\|\boldsymbol{x}\| + \|\Delta\boldsymbol{x}_1\| + \|\Delta\boldsymbol{x}_2\|} \leq \|A\|\,\|A^{-1}\|\left\{\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right\}. \tag{85}$$

The number $\|A\|\,\|A^{-1}\|$, which is called the "condition number," specifies how the initial errors are magnified in the solution.

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 ΔR, a term representing rounding errors, into A. It is known that, if we solve Eq. (80) by Gaussian elimination with pivoting, for example, the errors $\Delta\boldsymbol{x}_1 + \Delta\boldsymbol{x}_2$ in the solution $\boldsymbol{x}$ are well bounded by Eq. (85) (see Higham [4]).

Applying Eq. (85) to the linear system (78), we can bound the errors $|\delta_{u_0}|$ and $|\delta_{v_0}|$ of the solutions u_0 and v_0, due to the perturbations δ_{f_i} of f_i $(i = 0, \ldots, m)$ and δ_{g_j} of g_j $(j = 0, \ldots, n)$. Equation (79) tells us that if $|f_0u_0 + g_0v_0| \gg |f_0 \cdot \delta_{u_0}|, |g_0 \cdot \delta_{v_0}|$ then we can say definitely that $R_0 \neq 0$ for the perturbations of the coefficients of F and G. If $|f_0u_0 + g_0v_0| \ll |f_0u_0|, |g_0v_0|$ then this case corresponds to F and G having mutually close zero-points, and the above method cannot be applied to such cases. If $|f_0u_0 + g_0v_0|$ is not small, then we can apply the above method so long as $|\delta_{u_0}|$ and $|\delta_{v_0}|$ are not large. Equation (85) shows that the measure of largeness of $|\delta_{u_0}|$ and $|\delta_{v_0}|$ is the condition number. Therefore, in order to check whether or not the above method is useful, we check the largeness of the condition number for polynomials of degrees from 10 to 50. We generate a real univariate polynomial $P(x)$ with random coefficients, and construct the matrix in the left-hand-side of Eq. (78) by putting $F = P$ and $G = dP/dx$. We generate each coefficient c of $P(x)$ to satisfy $|c| \leq 10$. We set $\deg(P) = 10, 20, 30, 40, 50$, and generate 10 polynomials for each degree. We used the LAPACK library [1] linked to GAL to estimate the condition number (for estimating the condition number, see Natori [8], for example).

Table 1 shows the result of computations. For each degree of polynomial, we show the maximum, minimum, and average values of our estimates of 10 condition numbers. We see from this result that, for a polynomial of degree 10, for example, the error in $\mathrm{res}(P, dP/dx)$ may become 10^3 or 10^4 times larger than the error in the initial polynomial. Although these numbers are rather large, they are much smaller than the increase of the interval width explained above.

5.4 Calculating Error Terms Parametrically

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

For simplicity, we assume that $P(x)$ is monic in Eq. (1), and express $\tilde{P}(x)$ in Eq. (2) as

Table 2 $\|\tilde{P}_n(x, \delta_{n-1}, \ldots, \delta_0)\| / \|\tilde{P}_n(x, 0, \ldots, 0)\|$ for 10 polynomials, where $\tilde{P}_n$ is the last element of the Sturm sequence.

Degree of $\tilde{P}(x)$	Polynomial norm					
	1-norm			∞-norm		
	Maximum	Minimum	Average	Maximum	Minimum	Average
10	2.57×10^4	1.02×10^2	5.97×10^3	1.52×10^4	1.68×10^2	5.00×10^3
20	2.83×10^5	1.34×10^5	1.81×10^5	3.65×10^5	3.44×10^5	2.62×10^5
30	2.68×10^9	7.01×10^8	1.13×10^9	4.75×10^9	1.78×10^9	1.76×10^9
40	3.50×10^{13}	2.99×10^{11}	5.39×10^{12}	1.60×10^{14}	1.42×10^{11}	3.65×10^{12}
50	2.80×10^{17}	7.81×10^{15}	9.19×10^{16}	2.47×10^{17}	1.32×10^{16}	1.36×10^{17}

Table 3 Computing times for calculating Sturm sequences with and without parameterized error terms.

Degree of $\tilde{P}(x)$	Computing time (msec.)					
	With error terms			Without error terms		
	Maximum	Minimum	Average	Maximum	Minimum	Average
10	70	50	55	10	< 10	< 10
20	420	400	403	10	< 10	< 10
30	1420	1330	1357	20	10	11
40	3280	3210	3244	50	10	31
50	6080	6030	6050	50	30	37

$$\begin{aligned}&\tilde{P}(x, \delta_{n-1}, \ldots, \delta_0)\\ &= x^n + (c_{n-1} + \delta_{n-1})x^{n-1} + \cdots \\ &\qquad \cdots + (c_0 + \delta_0)x^0,\end{aligned} \tag{86}$$

where $\delta_{n-1}, \ldots, \delta_0$ are parameters representing errors in the coefficients. Exact calculation of the Sturm sequence of a parametric polynomial $\tilde{P}$ exactly is extremely time-consuming, because $\tilde{P}$ is $(n+1)$-variate. However, if we neglect all the quadratic and higher-order terms with respect to $\delta_{n-1}, \ldots, \delta_0$, then the computation cost is only $O(n)$ times larger than that of a numerical Sturm sequence. Therefore, we calculate the i-th element $\tilde{P}_i$ of the Sturm sequence as

$$\begin{aligned}&\tilde{P}_i(x, \delta_{n-1}, \ldots, \delta_0) \simeq \tilde{P}_i(x, 0, \ldots, 0)\\ &+ \tilde{P}_{i,n-1}(x, 0, \ldots, 0)\delta_{n-1} + \cdots \\ &\qquad \cdots + \tilde{P}_{i,0}(x, 0, \ldots, 0)\delta_0,\end{aligned} \tag{87}$$

where $\tilde{P}_{i,j} = \partial \tilde{P}_i / \partial \delta_j$ $(j = n-1, \ldots, 0)$. Then, by neglecting the terms of order $O(\delta^2)$, we can approximately bound the effect of error terms fairly well, as

$$\begin{aligned}&|\tilde{P}_i - P_i|\\ &\lesssim |\tilde{P}_{i,n-1}(x, 0, \ldots, 0)|\, \varepsilon_{n-1} + \cdots \\ &\qquad \cdots + |\tilde{P}_{i,0}(x, 0, \ldots, 0)|\, \varepsilon_0,\end{aligned} \tag{88}$$

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

Actually, the calculation is performed by introducing the total-degree variable t for $\delta_{n-1}, \ldots, \delta_0$ as $\delta_i \to \delta_i t$ $(i = 0, \ldots, n-1)$. We calculate the Sturm sequence only up to the total-degree 1, and substitute 1 for t after the calculation.

We calculated the Sturm sequences with and without parameterized error terms. For this experiment, we used the same polynomials as in Section 5.3.

Table 2 shows the value $\|\tilde{P}_n(x, \delta_{n-1}, \ldots, \delta_0)\| / \|\tilde{P}_n(x, 0, \ldots, 0)\|$, where $\tilde{P}_n(x, \delta_{n-1}, \ldots, \delta_0)$ is the last element of the Sturm sequence, and **Table 3** shows the computing times of Sturm sequences with and without parametric errors. In Table 2, for each degree of polynomial, we show the maximum, the minimum, and the average of 10 ratios. Note that the values in Table 2 show how the initial errors are magnified by the computation of Sturm sequence, just as the values in Table 1 show. Comparing with Table 1, we see that the numbers are too large for polynomials of higher degrees. Table 3 shows the maximum, minimum, and average values of the computation times for ten examples. We see that, very roughly speaking, the computation time for a parameterized sequence is about $\deg(P)$ times larger than that for a numerical sequence. These results indicate that we can use this method only for polynomials of low or medium degrees.

6. Discussion

In this paper we have considered the real zero-points of a real univariate polynomial with error terms whose coefficients may be much larger than ε_M. For such an approximate polynomial, we introduced the concept of an

"approximate real zero-point" and proposed a method for calculating the existence domains of zero-points fairly accurately and simply.

Next, we considered how to calculate the number of real zero-points of an approximate polynomial by Sturm's method. We gave a sufficient condition for the number of real zero-points to be definite. We also derived a sufficient condition for the small leading coefficients in the Sturm sequence to be discarded, and showed that these problems can be reduced to a problem to that of estimating the errors in the resultants of univariate polynomials.

Finally, in order to estimate the errors in the Sturm sequence, we investigated four methods: (1) evaluating the "subresultant determinant" by using Hadamard's inequality, (2) calculating the Sturm sequence with coefficients of interval numbers, (3) solving a linear system on polynomial coefficients and evaluating errors in the solution by a standard method in numerical analysis, and (4) calculating the Sturm sequence with parametric error terms. Method 1 is theoretically correct, but the calculated upper bound is too large, and with method 2 the width of each interval number grows too rapidly during the calculation of the Sturm sequence; hence methods 1 and 2 do not seem to be useful in practice. Method 3 gives a rather practical estimation, and thus seems to be useful in practice. Method 4 gives the errors rather accurately, and we have seen that calculating the resultant by PRS gives much larger errors than method 3. This means that the errors contained in the resultant depend on which method we have used to calculate the resultant, and method 3 seems to be the best for evaluating the errors.

We still have a problem in cases where $P(x)$ has multiple or close zero-points. Let us briefly mention what happens if $P(x)$ has close zero-points. Let $\|.\|$ be an appropriate norm of a polynomial defined by Eq. (33), and assume that $\|P\| = 1$ and P_k contains m close zero-points of closeness δ, $0 < \delta \ll 1$, around the origin. Then, Sasaki and Sasaki [9] tell us that $\|P_k\| = O(\delta^0)$ and $\|P_{k+1}\| = O(\delta^2)$, $\|P_{k+2}\| = O(\delta^3), \ldots, \|P_{k+m}\| = O(\delta^{m+1})$. Therefore, if these close zero-points can be separated and counted as m single zero-points, we must have $\|P_{k+m}\| \gg \varepsilon_M$ or $\delta \gg \sqrt[m+1]{\varepsilon_M}$. On the other hand, if we change coefficients of $P(x)$ slightly, the positions of these close zero-points are changed considerably. Thus, the treatment of close zero-points is not easy and remains an open problem.

Acknowledgments The authors thank to anonymous referees for their valuable comments.

References

1) Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., Croz, J.D., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S. and Sorensen, D.: *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia (1995).
2) Brown, W.S. and Traub, J.F.: On Euclid's Algorithm and the Theory of Subresultants, *J. ACM*, Vol.18, No.4, pp.505–514 (1971).
3) Cohen, H.: *A Course in Computational Algebraic Number Theory*, Graduate Texts in Mathematics, Vol.138, Springer-Verlag, Berlin (1993).
4) Higham, N.J.: *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia (1996).
5) Iri, M.: *Numerical Analysis* (in Japanese), Asakura Publishing, Tokyo (1981).
6) Mignotte, M.: *Mathematics for Computer Algebra*, Springer-Verlag (1992).
7) Mishra, B.: *Algorithmic Algebra*, Texts and Monographs in Computer Science, Springer-Verlag, New York (1993).
8) Natori, M.: *Numerical Analysis and Its Applications* (in Japanese), Corona Publishing, Tokyo (1990).
9) Sasaki, T. and Sasaki, M.: Analysis of Accuracy Decreasing in Polynomial Remainder Sequence with Floating-point Number Coefficients, *J. Inform. Process.*, Vol.12, No.4, pp.394–403 (1989).
10) Shirayanagi, K. and Sekigawa, H.: An Interval Method Based on Zero Rewriting and Its Application to Sturm's Algorithm (in Japanese), *Trans. Inst. Electronics, Inform. and Comm. Engineers A*, Vol.J80-A, No.5, pp.791–802 (1997).
11) Smith, B.T.: Error Bounds for Zeros of a Polynomial Based upon Gerschgorin's Theorems, *J. ACM*, Vol.17, No.4, pp.661–674 (1970).

Appendix: On the Zero-points of Eq. (43)

Let $0 < \varepsilon \ll 1$ and let $P(x)$ be

$$P(x) = c_n x^n + \cdots + c_{m+1}x^{m+1} + x^m + \varepsilon_{m-1}x^{m-1} + \cdots + \varepsilon_0, \quad (89)$$

where $n > m$ and $c_n, \ldots, c_{m+1}, \varepsilon_{m-1}, \ldots, \varepsilon_0$ are numbers such that

$$\max\{|c_n|, \ldots, |c_{m+1}|\} = 1,\ c_n \neq 0,\quad |\varepsilon_{m-i}| \leq (\sqrt[m]{\varepsilon})^i \ (i = 1, \ldots, m). \quad (90)$$

We choose ε to satisfy $\sqrt[m]{\varepsilon} = \max\{\sqrt[i]{|\varepsilon_{m-i}|} \mid i = 1, \ldots, m\}$. Putting $e = \sqrt[m]{\varepsilon}$, we prove the following theorem in this appendix:

Theorem 11 Let $\zeta_1, \ldots, \zeta_n$ be the zero-points of $P(x)$, where

$$\begin{aligned} &|\zeta_1| \le \cdots \le |\zeta_m| \\ &< |\zeta_{m+1}| \le \cdots \le |\zeta_n|. \end{aligned} \tag{91}$$

If $e = \sqrt[m]{\varepsilon} \le 1/9$ then $|\zeta_m|$ and $|\zeta_{m+1}|$ are bounded as

$$\begin{aligned} &|\zeta_m| \\ &< \frac{1+3e}{4}\left[1 - \sqrt{1 - \frac{16e}{(1+3e)^2}}\right], \\ &|\zeta_{m+1}| \\ &> \frac{1+3e}{4}\left[1 + \sqrt{1 - \frac{16e}{(1+3e)^2}}\right]. \end{aligned} \tag{92}$$

Furthermore, we can approximate the right-hand-side expressions of Eq. (92) as

$$\begin{aligned} &|\zeta_m| \\ &< 2e \cdot \left[\frac{1}{1+3e} + \frac{16e}{(1+3e)^3}\right], \\ &|\zeta_{m+1}| \\ &> \frac{1}{2} - \frac{e(1-9e)}{2(1+3e)} - \frac{32e^2}{(1+3e)^3}. \end{aligned} \tag{93}$$

□

Before the proof, we investigate the zero-points of $P(x)$ roughly. Put

$$\begin{aligned} P'(x) &= x^m + \varepsilon_{m-1}x^{m-1} + \cdots \\ &\qquad \cdots + \varepsilon_1 x + \varepsilon_0, \\ P''(x) &= c_n x^{n-m} + \cdots \\ &\qquad \cdots + c_{m+1}x + 1. \end{aligned} \tag{94}$$

Note that $P(x) \approx P''(x)P'(x)$. Using the following well-known theorem (see Mignotte [6] for example), we can bound the zero-points of $P'(x)$ and $P''(x)$ easily as in Corollaries 13 and 14 below.

Theorem 12 Let $A(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$, with $a_n a_0 \neq 0$, be a polynomial with complex coefficients and with zero-points $\zeta_1, \ldots, \zeta_n$. Then, we have the following bounds for the zero-points of $A(x)$:

$$\begin{aligned} &\max\{|\zeta_1|, \ldots, |\zeta_n|\} \\ &\le \frac{|a_n| + \max\{|a_{n-1}|, \ldots, |a_0|\}}{|a_n|}, \\ &\min\{|\zeta_1|, \ldots, |\zeta_n|\} \\ &\ge \frac{|a_0|}{|a_0| + \max\{|a_1|, \ldots, |a_n|\}}. \end{aligned} \tag{95}$$

□

Applying Theorem 12 to $\varepsilon^{-1}P'(\sqrt[m]{\varepsilon}x)$ and $P''(x)$, respectively, we obtain the following corollaries:

Corollary 13 Let the zero-points of $P'(x)$ be $\zeta'_1, \ldots, \zeta'_m$; then we have $\max\{|\zeta'_1|, \ldots, |\zeta'_m|\} \le 2\sqrt[m]{\varepsilon}$. □

Corollary 14 Let the zero-points of $P''(x)$ be $\zeta''_{m+1}, \ldots, \zeta''_n$; then we have $\min\{|\zeta''_{m+1}|, \ldots, |\zeta''_n|\} \ge 1/2$. □

These corollaries show that $P(x)$ has m zero-points of magnitude $\lesssim 2\sqrt[m]{\varepsilon}$ and that the other $(n-m)$ zero-points have absolute values $\gtrsim 1/2$. We now prove Theorem 11.

Proof of Theorem 11. We first consider the zero-point ζ of Eq. (89), such that $|\zeta| \lesssim 2\sqrt[m]{\varepsilon}$. Applying the transformation $\zeta = e\bar{\zeta}$ $(= \sqrt[m]{\varepsilon}\bar{\zeta})$ to $P(\zeta) = 0$, we obtain

$$\begin{aligned} &c_n e^{n-m}\bar{\zeta}^n + \cdots + c_{m+1}e\bar{\zeta}^{m+1} \\ &+ \bar{\zeta}^m + (\varepsilon_{m-1}/e)\bar{\zeta}^{m-1} + \cdots \\ &\qquad + (\varepsilon_0/e^m)\bar{\zeta}^0 = 0. \end{aligned} \tag{96}$$

We are considering the zero-point $\bar{\zeta}$ such that $|\bar{\zeta}| \lesssim 2$, hence the zero-point is determined mostly by the terms of degree $\le m$ and the terms $c_{m+j}e^j\bar{\zeta}^{m+j}$ $(j = 1, \ldots, n-m)$ contribute only as small correction terms because $e \ll 1$. (We can state this situation as follows. Consider a set of equations of degree m:

$$\left\{\begin{aligned} &a_m z^m + (c_{m-1}/e)z^{m-1} + \cdots \\ &\qquad \cdots + (c_0/e^m)z^0 = 0, \\ &a_m \in \{1 + c_{m+1}e\bar{z} + \cdots \\ &\qquad \cdots + c_n e^{n-m}\bar{z}^{n-m} \mid \\ &\qquad |\bar{z}| \le \bar{\zeta}_{\max}\}, \end{aligned}\right. \tag{97}$$

where $\bar{\zeta}_{\max}$ is an upper bound of $|\zeta_m/e|$. Obviously, $\bar{\zeta} = \zeta_m/e$ is a solution of one equation in this set. For the solution of any equation in this set, we can derive an upper bound.) Thus, rewriting the above equation as

$$\begin{aligned} &(c_n e^{n-m}\bar{\zeta}^{n-m} + \cdots + c_{m+1}e\bar{\zeta} + 1)\bar{\zeta}^m \\ &+ (\varepsilon_{m-1}/e)\bar{\zeta}^{m-1} + \cdots \\ &\qquad \cdots + (\varepsilon_0/e^m)\bar{\zeta}^0 = 0, \end{aligned} \tag{98}$$

we can regard Eq. (98) as an equation of degree m with the leading coefficient $a_m = 1 + c_{m+1}e\bar{\zeta} + \cdots + c_n e^{n-m}\bar{\zeta}^{n-m} \approx 1$. Therefore, from Theorem 12, we obtain

$$\begin{aligned} |\bar{\zeta}| &\le 1 + \max\{|\varepsilon_{m-1}/e|, \ldots, |\varepsilon_0/e^m|\}/|a_m| \\ &\le 1 + \frac{1}{1 - |e\bar{\zeta}| - \cdots - |e\bar{\zeta}|^{n-m}} \\ &< 1 + \frac{1}{1 - |e\bar{\zeta}|/(1 - |e\bar{\zeta}|)} \\ &= \frac{2 - 3|e\bar{\zeta}|}{1 - 2|e\bar{\zeta}|}, \end{aligned} \tag{99}$$

or

$$|\zeta| < \frac{2e - 3e|\zeta|}{1 - 2|\zeta|}. \tag{100}$$

Inequality (100) gives us

$$2|\zeta|^2 - (1+3e)|\zeta| + 2e > 0. \tag{101}$$

Let z_- and z_+ be the solutions of equation $2z^2 -$

$(1+3e)z+2e=0$ with $z_- \le z_+$. We see that $z_\pm$ are real if and only if $e \le 1/9$, and $z_- \simeq 2e$ and $z_+ \simeq (1-e)/2$ for $|e| \ll 1$. Therefore, we have

$$4|\zeta| < (1+3e) \times \left[1-\sqrt{1-\frac{16e}{(1+3e)^2}}\right] \quad (102)$$

for $e \le 1/9$. Using the inequality $\sqrt{1-x} > 1-x/2-x^2/2$, which is valid for $0<x<1$, and putting $x=16e/(1+3e)^2$, we obtain

$$4|\zeta| < (1+3e) \times \left[\frac{8e}{(1+3e)^2}+\frac{128e^2}{(1+3e)^4}\right], \quad (103)$$

or

$$|\zeta| < 2e\cdot\left[\frac{1}{1+3e}+\frac{16e}{(1+3e)^3}\right]. \quad (104)$$

This inequality is valid for $16e/(1+3e)^2 < 1$, or for $e < 1/9$.

Next, we consider the zero-point ζ of Eq. (89), such that $1/2 \lesssim |\zeta|$. Dividing $P(\zeta)=0$ by ζ^m, we obtain the equality

$$c_n\zeta^{n-m}+\cdots+c_{m+1}\zeta+1 + \varepsilon_{m-1}/\zeta+\varepsilon_{m-2}/\zeta^2+\cdots \cdots+\varepsilon_0/\zeta^m=0. \quad (105)$$

Since we are considering the zero-point ζ such that $1/2 \lesssim |\zeta|$, the terms $\varepsilon_{m-j}/\zeta^j$ $(j=1,\ldots,m)$ contribute only as small correction terms because $|\varepsilon_{m-j}| \ll 1$. Thus, following the same reasoning as for Eq. (98), we can regard Eq. (105) as an equation of degree $n-m$ with the constant term $a_0 = 1+\varepsilon_{m-1}/\zeta+\cdots+\varepsilon_0/\zeta^m \approx 1$. From Theorem 12, we obtain

$$\begin{aligned}|\zeta| &\ge \frac{1}{1+\max\{|c_n|,\ldots,|c_{m+1}|\}/|a_0|} \\ &\ge \frac{1}{1+1/(1-|e/\zeta|-\cdots-|e/\zeta|^m)} \\ &> \frac{1}{1+\frac{1}{1-|e/\zeta|/(1-|e/\zeta|)}} \quad (106) \\ &= \frac{1-2|e/\zeta|}{2-3|e/\zeta|}.\end{aligned}$$

Inequality Eq. (106) gives us

$$2|\zeta|-(1+3e)|\zeta|+2e>0. \quad (107)$$

Solving Eq. (107) with condition $e \le 1/9$, we obtain

$$4|\zeta| > (1+3e) \times \left[1+\sqrt{1-\frac{16e}{(1+3e)^2}}\right]. \quad (108)$$

Using the inequality $\sqrt{1-x} > 1-x/2-x^2/2$ again, we obtain

$$4|\zeta| > (1+3e) \times \left[2-\frac{8e}{(1+3e)^2}-\frac{128e^2}{(1+3e)^4}\right], \quad (109)$$

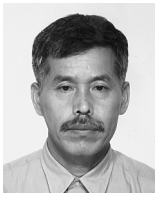
or

$$|\zeta| > \frac{1}{2}-\frac{e(1-9e)}{2(1+3e)}-\frac{32e^2}{(1+3e)^3} \quad (110)$$

for $e<1/9$. □

(Received December 11, 1998)
(Accepted January 6, 2000)

Akira Terui was born in 1971. He received his M.S. degree from Univ. Tsukuba in 1997. Since 1999 he has been in Univ. Tsukuba as a research associate. His research interests include theory and application of approximate algebraic computation: solving system of algebraic equations, calculation of singularities of algebraic functions, etc., by means of computer algebra with approximate computation. He is a member of IPSJ, JSIAM, JSSAC and ACM.



Tateaki Sasaki was born in 1946. He received M.S. and D.S. degrees from Univ. Tokyo in 1970 and 1973, respectively. He had been a researcher of RIKEN: The Institute of Physical and Chemical Research (Information Science Laboratory) in 1974–1991 and a visiting researcher of Univ. Utah (Dept. Computer Science) in 1978–1979. Since 1991, he has been a professor of Univ. Tsukuba (Institute of Mathematics). His research interests include algorithm development of computer algebra and development of formula manipulation system. In particular, he is an initiator of approximate algebra. He is a member of IPSJ, JSIAM, JSSAC, MSJ and ACM.