

Inclusion methods for real and complex functions in one variable

D. M. Claudio

UFRGS Inst. Informática
 Porto Alegre-RS Brasil
 dalcidio@inf.ufrgs.br

S. M. Rump

T. U. Hamburg-Harburg
 Hamburg Alemanha
 rump@tu-hamburg.d400.de

Abstract

A new method is discussed for obtaining a validated inclusion of a zero of a real or complex function in one variable. The method is given by means of a sufficient criterion which can be tested on a digital computer. The numerical results show that, compared to other methods, the new one frequently yields very narrow bounds for the solution.

1. Introduction

Let \mathbb{IR} , \mathbb{IC} denote the set of real, complex intervals, resp. In \mathbb{IC} rectangular, circular or any other kind of intervals may be used. For practical applications we only require that the operations are executable on digital computers and satisfy the basic property of isotonicity:

$$\forall A, B \in \mathbb{IS} : \{ a * b \mid a \in A, b \in B \} \subseteq A * B \quad (1)$$

for $S \in \{\mathbb{IR}, \mathbb{IC}\}$ and $* \in \{+, -, \cdot, /\}$,

with $0 \notin B$ in case of division. For more details refer to [1], [2].

\mathbb{IP} always denotes the power set, $\underline{\cup}$ denotes the convex union. We use the usual definition of the evaluation of a function over a set, namely

for $f : S \rightarrow S$ and $X \subseteq S$ it is

$$f(X) := \{ f(x) \mid x \in X \} \in \mathbb{IPS},$$

$S \in \{\mathbb{IR}, \mathbb{IC}\}$. $\text{int}(X)$ denotes the interior of a set.

2. Basic results

In [4] the following sufficient criterion has been given to check whether an interval contains a zero of a nonlinear function. The theorem has been formulated for the n -dimensional case; here we only need the 1-dimensional case. We only give the complex version of it.

Theorem 1. Let a holomorphic function $f : G \rightarrow \mathbb{C}$ for some closed $G \subseteq \mathbb{C}$ be given. Let $R \in \mathbb{C}$, $\tilde{z} \in G$ and define

$$f' : \mathbb{PC} \rightarrow \mathbb{PC} \text{ by}$$

$$Z \subseteq \mathbb{C} \Rightarrow f'(Z) := \cap \{ Y \in \mathbb{IC} \mid f'(z) \in Y \text{ for all } z \in Z \}. \quad (2)$$

For some $Z \in \mathbb{IC}$ with $0 \in Z$ and $\tilde{z} + Z \subseteq G$ define

$$L(Z) := -R \cdot f(\tilde{z}) + \{1 - R \cdot f'(\tilde{z} + Z)\} \cdot Z. \quad (3)$$

If

$$L(Z) \subseteq \text{int}(Z)$$

then there exists one and only one $\hat{z} \in \tilde{z} + Z$ with $f(\hat{z}) = 0$. Furthermore

$$\hat{z} \in (\tilde{z} + Z) \cap (\tilde{z} + L(Z)).$$

The proof uses a complex version of the mean-value theorem due to Böhm (cf. [6]):

Theorem 2. Let a holomorphic function $f : G \rightarrow \mathbb{C}$ for some closed set $G \subseteq \mathbb{C}$ be given. Then for $z, \tilde{z} \in \mathbb{C}$ with $z \cup \tilde{z} \subseteq G$ here holds

$$f(z) \in f(\tilde{z}) + f'(z \cup \tilde{z}) \cdot (z - \tilde{z}). \quad (4)$$

In a practical implementation one would choose \tilde{z} s.t. $f(\tilde{z}) \approx 0$ and $R \approx f'(\tilde{z})^{-1}$.

In Theorem 1 we used the operator $L(X)$ instead of Krawczyk's operator. It turned out that computing an inclusion of the difference of the true solution \hat{z} to an approximate solution \tilde{z} yields superior computational results (cf. [3], [4] and the references mentioned in there).

In the following we show how other operators instead of $L(X)$ can be used (cf. [5]). We first derive the results for the real case.

Let $f : D \rightarrow \mathbb{R}$ for closed $D \subseteq \mathbb{R}$ be differentiable and define

$$g_R(x) := x - R \cdot f(x). \tag{5}$$

In practical computations one thinks of $R \approx f'(\tilde{x})$ s.t. g_R becomes a Newton operator. Then for every $\tilde{x} \in D$ with $x \sqcup \tilde{x} \subseteq D$ it holds

$$f(x) = f(\tilde{x}) + f'(\xi) \cdot (x - \tilde{x}) \quad \text{for some } \xi \in x \sqcup \tilde{x}$$

and therefore

$$\begin{aligned} g_R(x) &= x - R \cdot (f(\tilde{x}) + f'(\xi)(x - \tilde{x})) \\ &= \tilde{x} - R \cdot f(\tilde{x}) + \{1 - R \cdot f'(\xi)\} \cdot (x - \tilde{x}). \end{aligned} \tag{6}$$

If f is twice differentiable on D then

$$f'(\xi) = f'(\tilde{x}) + f''(\eta) \cdot (\xi - \tilde{x})$$

if $\tilde{x} \sqcup \xi \subseteq D$. Hence (6) implies

$$g_R(x) = \tilde{x} - R \cdot f(\tilde{x}) + \{1 - R \cdot f'(\tilde{x}) - R \cdot f''(\eta)(\xi - \tilde{x})\}(x - \tilde{x}). \tag{7}$$

If $f'(\tilde{x}) \neq 0$ then $g \equiv g_R$ with $R := (f'(\tilde{x}))^{-1}$ becomes

$$g(x) = \tilde{x} - f'(\tilde{x})^{-1} \cdot \{f(\tilde{x}) + f''(\eta) \cdot (\xi - \tilde{x})(x - \tilde{x})\}.$$

Thence

$$g(X) \subseteq \tilde{x} - f'(\tilde{x})^{-1} \cdot \{f(\tilde{x}) + f''(X) \cdot (X - \tilde{x})^2\} \tag{8}$$

(cf. [1], [2]). On the other hand the definition of g_R together with Brouwer's Fixed Point Theorem implies for every $R \neq 0$:

$$g_R(X) \subseteq X \Rightarrow \exists \hat{x} \in X : g_R(\tilde{x}) = \hat{x} = \hat{x} - R \cdot f(\hat{x})$$

and therefore

$$g_R(X) \subseteq X \Rightarrow \exists \hat{x} \in X : f(\hat{x}) = 0. \tag{9}$$

In other words, any methods for computing outer bounds for $g_R(X)$ for some $R \neq 0$ generates a sufficient criterion for checking the existence of some $\hat{x} \in X$ with $f(\hat{x}) = 0$. (6) gives one such formulation which yielded Theorem 1, (8) is another formulation which has also been given in [5]. In [4] (6) was used and the inclusion in the interior of X was supposed thus omitting the

requirement that $R \neq 0$. In the n -dimensional case $R \neq 0$ becomes “ R not singular” and the mentioned formulation becomes important.

Formula (8) can be developed further by using higher derivatives in an obvious way. This proves the following Theorem.

Theorem 3. Let a twice differentiable function $f : D \rightarrow \mathbb{R}$ for closed $D \subseteq \mathbb{R}$ be given. For $0 \neq R \in \mathbb{R}$, $\tilde{x} \in D$, $X \in \mathbb{IIR}$ with $0 \in X$ and $\tilde{x} + X \subseteq D$ define

$$L_2(X) := -f'(\tilde{x})^{-1} \cdot \{f(\tilde{x}) + f''(\tilde{x} + X) \cdot X^2\} \quad (10)$$

and, if f is thrice differentiable,

$$L_3(X) := -f'(\tilde{x})^{-1} \cdot \{f(\tilde{x}) + (f''(\tilde{x}) + f'''(\tilde{x} + X) \cdot X) \cdot X^2\}. \quad (11)$$

If

$$L_2(X) \subseteq X \quad \text{or} \quad L_3(X) \subseteq X$$

so, there is some $\hat{x} \in \tilde{x} + X$ with $f(\hat{x}) = 0$. Moreover,

$$\hat{x} \in (\tilde{x} + X) \cap (\tilde{x} + L_2(X)) \quad \text{or} \quad \hat{x} \in (\tilde{x} + X) \cap (\tilde{x} + L_3(X)),$$

respectively.

Proof. Obvious from the foregoing discussion. ■

Using Theorem 2 it immediately follows that Theorem 3 remains true in \mathbb{C} for holomorphic functions.

3. A new inclusion method

In [1] the following theorem is given for *refining* intervals already containing a zero of a real function.

Theorem 4 (Alefeld). Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous function and \hat{x} be a zero of f within $X \in \mathbb{IIR}$. Let $f(\underline{X}) < 0$ and $f(\overline{X}) > 0$ for $X = \{\underline{X}, \overline{X}\}$ and suppose

$$0 < m_1 \leq \frac{f(x) - f(\hat{x})}{x - \hat{x}} = \frac{f(x)}{x - \hat{x}} \leq m_2 < \infty$$

for all $\hat{x} \neq x \in X$. Then for $\tilde{x} \in X$ and $M := [m_1, m_2]$ it holds

$$\hat{x} \in X \cap \left\{ \tilde{x} - \frac{f(\tilde{x})}{M} \right\}.$$

This theorem requires the knowledge of some $X \in \text{IIR}$ containing a zero \hat{x} of f . Whereas this is simple to verify in IR it becomes difficult to verify in \mathbb{C} . As in the preceding discussion we are therefore interested in sufficient criteria for X which verify the existence of some $\hat{x} \in X$ with $f(\hat{x}) = 0$. Theorems 1 and 3 already give such criteria. Now we are going to develop another one which turns the refinement in Theorem 4 into a sufficient criterion. We develop it for the complex case.

Let $f : G \rightarrow \mathbb{C}$ be a holomorphic function for closed $G \subseteq \mathbb{C}$ and define $r : G \rightarrow \mathbb{C}$ for fixed but arbitrary $\tilde{z} \in G$ by

$$r(z) := \begin{cases} \frac{f(z) - f(\tilde{z})}{z - \tilde{z}} & \text{for } z \neq \tilde{z} \\ f'(\tilde{z}) & \text{otherwise.} \end{cases}$$

Then r is continuous. Suppose $r(z) \neq 0$ for $z \in G$ then for $g : G \rightarrow \mathbb{C}$ with

$$g(z) := z - \frac{f(z)}{r(z)} = z - \frac{f(z) \cdot (z - \tilde{z})}{f(z) - f(\tilde{z})} = \tilde{z} - \frac{f(\tilde{z})}{r(z)}. \quad (12)$$

By Theorem 2, (4) we know that for every $z, \tilde{z} \in G$ with $z \cup \tilde{z} \subseteq G$

$$r(z) \in f'(z \cup \tilde{z}).$$

Thence using (2) for the definition of f' we obtain from (12)

$$g(Z) \in \tilde{z} - \frac{f(\tilde{z})}{f'(Z)} \quad (13)$$

provided $0 \notin f'(Z)$. This leads to the following Theorem.

Theorem 5. Let $f : G \rightarrow \mathbb{C}$ be holomorphic for closed $G \subseteq \mathbb{C}$ and for closed and convex $\emptyset \neq Z \subseteq G$ suppose $0 \notin f'(Z)$. Let $\tilde{z} \in Z$ be fixed but arbitrary. So

$$\tilde{z} - \frac{f(\tilde{z})}{f'(Z)} \subseteq Z \quad (14)$$

implies the existence and uniqueness of a zero \hat{z} of f within Z . Moreover

$$\hat{z} \in Z \cap \left\{ \tilde{z} - \frac{f(\tilde{z})}{f'(Z)} \right\}. \quad (15)$$

Proof. According to the previous discussion $0 \notin f'(Z)$ implies $0 \notin r(Z)$ and using (13) implies $g(Z) \subseteq Z$. Brouwer's Fixed Point Theorem implies the existence of some $\hat{z} \in Z$ with $g(\hat{z}) = \hat{z}$. Hence by (12)

$$f(\hat{z}) / r(\hat{z}) = 0 \quad \text{and therefore} \quad f(\hat{z}) = 0.$$

Furthermore (13) proves (15). Suppose $\bar{z} \in Z$ with $f(\bar{z}) = 0$. Then by Theorem 2

$$f(\hat{z}) \in f(\bar{z}) + f'(\bar{z} \cup \hat{z}) \cdot (\hat{z} - \bar{z})$$

or $0 \in f'(\bar{z} \cup \hat{z}) \cdot (\hat{z} - \bar{z})$. $0 \notin f'(Z)$ then demonstrates the uniqueness of \hat{z} and finishes the proof. ■

Obviously, Theorem 5 immediately derives for a real function f . The main point of Theorem 5 is that there are no assumptions on the quality of \tilde{z} and, most important, no a priori assumptions on the set Z . Especially, $\hat{z} \in Z$ is not assumed *a priori*.

It will turn out from the examples that the formula in Theorem 5 is always superior to the one in Theorem 1. This, in fact, can be asserted theoretically.

Theorem 6. Let $f : G \rightarrow \mathbb{C}$ be holomorphic for closed $G \cup \mathbb{C}$ and for closed and convex $\emptyset \neq Z \subseteq G$ suppose $0 \notin f'(Z)$. Then for given $\tilde{z} \in Z$, $R \in \mathbb{C}$,

$$\tilde{z} - R \cdot f(\tilde{z}) + \{1 - R \cdot f'(Z)\} \cdot (Z - \tilde{z}) \subseteq \text{int}(Z) \tag{16}$$

implies

$$\tilde{z} - \frac{f(\tilde{z})}{f'(Z)} \subseteq Z. \tag{17}$$

Proof. (16) implies

$$R \cdot (-f(\tilde{z})) + \{1 - R \cdot f'(Z)\} \cdot Y \subseteq \text{int}(Y) \tag{18}$$

for $Y := Z - \tilde{z}$. Define $\mathcal{A} \in \Pi\mathbb{C}^{n \times n}$ with $\mathcal{A} := f'(Z)$ and $b \in \mathbb{C}^n$ by $b := f(\tilde{z})$, then (18) implies

$$\sum(\mathcal{A}, b) := \{z \in \mathbb{C}^n \mid \exists A \in \mathcal{A} : Az = b\} \subseteq Y.$$

This follows e.g. using Theorem 1; the Theorem is explicitly given in [4].
Thus

$$\forall u \in f'(Z) : u^{-1} \cdot (-f(\tilde{z})) \subseteq Z - \tilde{z}$$

which in turn implies (17). ■

It should be mentioned that in case of polynomials we can even do better than Theorem 5. Let a polynomial $p \in \mathbb{C}[z]$ of degree n be given. In this case the secant

$$r(z) = \begin{cases} \frac{p(z) - p(\tilde{z})}{z - \tilde{z}} & \text{for } z \neq \tilde{z} \\ p'(\tilde{z}) & \text{otherwise,} \end{cases} \quad (19)$$

which is a polynomial of degree $(n - 1)$, can be evaluated explicitly using Horner's scheme at \tilde{z} . Then the following corollary can be used.

Corollary 7. Let $p \in \mathbb{C}[z]$ be a polynomial of degree n and $\emptyset \neq Z \subseteq \mathbb{C}$ be closed and convex. Define $r : \mathbb{C} \rightarrow \mathbb{C}$ by (19) and assume $0 \notin r(Z)$. Let $\tilde{z} \in Z$ be fixed but arbitrary. So

$$\tilde{z} - \frac{p(\tilde{z})}{r(Z)} \subseteq Z \quad (20)$$

implies the existence and uniqueness of a zero \hat{z} of p in Z . Moreover,

$$\hat{z} \in Z \cap \left\{ \tilde{z} - \frac{p(\tilde{z})}{r(Z)} \right\}.$$

The *proof* is similar to the one of Theorem 5. ■

4. Numerical results

In the following examples we are only considering polynomials. Remember that our theorems except the last one are valid for general nonlinear functions as well. All our computations are performed on an IBM 4361 in double precision according to 14 hexadecimal or approximately 16 decimal places. For the computation we use the interactive programming environment CALCULUS (cf. [7]).

The following tests are performed. Given an approximation \tilde{x} of a zero, usually being good up to a few digits in the last place of the mantissa, we

define $X := [-\epsilon \cdot \tilde{x}, +\epsilon \cdot \tilde{x}]$ for a given $\epsilon > 0$. Then we perform the following four tests:

$$- f(\tilde{x}) / f'(\tilde{x} + X) \subseteq X \quad (21)$$

$$R := 1/f'(\tilde{x}); -R \cdot f(\tilde{x}) + \{1 - R \cdot f'(\tilde{x} + X)\} \cdot X \subseteq X \quad (22)$$

$$- \{f(\tilde{x}) + f''(\tilde{x} + X) \cdot X^2\} / f'(X) \subseteq X \quad (23)$$

$$- \{f(\tilde{x}) + (f''(\tilde{x}) + f'''(\tilde{x} + X) \cdot X) \cdot X^2\} / f'(X) \subseteq X. \quad (24)$$

Due to Theorems 5, 1, and 3 the validity of any of the conditions (21) ... (24) implies the existence of some $\hat{x} \in \tilde{x} + X$ with $f(\hat{x}) = 0$.

All calculations in (21) ... (24) are performed using interval operations over floating-point real or complex numbers except for the computation of R in (22), which is performed in pure floating-point with rounding to nearest.

For the first set of test samples we take n uniformly distributed real random numbers \tilde{x}_i in $[-1, 1]$ and form

$$f(x) = \prod_{i=1}^n (x - \tilde{x}_i). \quad (25)$$

f in (25) is computed in pure floating-point hence altering the zeros of f from \tilde{x}_i . Nevertheless we take $X_i := [-\epsilon \cdot \tilde{x}_i, +\epsilon \cdot \tilde{x}_i]$ and test (21) ... (24) with \tilde{x}_i, x_i for $i = 1 \dots n$.

All approximations \tilde{x}_i and potential inclusion intervals X_i are real, hence up to now only real operations are involved. It turned out that in all cases where (21) ... (24) was satisfied, the diameter of the left hand side of (21) was the smallest followed by that of (24), (23) and (22). Denote

$$d_i := \text{diameter (l.h.s. of (2i))}, \quad 1 \leq i \leq 4.$$

Then we display the minimum and maximum ratio of diameters

$$d_2/d_1, d_3/d_2 \text{ and } d_4/d_3,$$

the minimum and maximum taken for all zeros. In some cases some of (2i), $1 \leq i \leq 4$ were not satisfied for larger values of ϵ . This may partly be due to the fact that some zeros of f became complex because of rounding errors in the computation of (25). Therefore the number of pairs (\tilde{x}_i, X_i) as defined above for which (2i), $1 \leq i \leq 4$ was satisfied is also displayed.

n	min	min	min	# success of			
	(d_2/d_1)	(d_3/d_2)	(d_4/d_3)	(21)	(22)	(23)	(24)
	max	max	max				
5	4.7e8	4.2	1.0003	5	5	5	5
	3.5e9	163	1.0105				
10	3.7e8	*	1.0006	2	2	5	5
	1.0e9	99.6	1.0933				
15	5.8e8	26.1	1.003	3	3	3	3
	1.9e9	133	1.030				

Table 1. Random real zeros in $[-1, 1]$ with $\epsilon = 10^{-4}$

We see that in some cases methods (23) and (24) are more efficient in the sense that they allow to verify $\hat{x} \in \tilde{x} + X$ whereas (21) and (22) do not. Therefore the minimum ratio d_i/d_j is denoted by * if the tests fails in cases where (2i) succeeds in proving existence (and uniqueness) of a zero in $\tilde{x} + X$.

We are especially in the performance for wide intervals $\tilde{x} + X$. Then the first method becomes even more advantageous. The following table displays the results for $\epsilon = 10^{-2}$.

n	min	min	min	# success of			
	(d_2/d_1)	(d_3/d_2)	(d_4/d_3)	(3.1)	(3.2)	(3.3)	(3.4)
	max	max	max				
5	2.0e12	*	1.03	4	4	5	5
	1.1e13	48.3	2.10				
10	2.1e12	*	*	1	1	2	3
	2.1e12	9.3	2.4				
15	4.4e12	*	1.2	2	2	3	3
	6.5e12	51.0	5.5				

Table 2. Random real zeros in $[-1, 1]$ with $\epsilon = 10^{-2}$

So up to now methods (21) and (22) enclose the same number of zeros whereas there are cases where (24) is better than (23) whereas (23) is better than (22) and (21).

Our next example are real polynomials with uniformly distributed random coefficients within $[-1, 1]$. Those polynomials have almost only complex zeros. We used the Matlab-function `roots` to produce approximations \tilde{z}_i to the roots of the polynomial. After that we proceed as before.

Note that now complex roots are to be included. We ran the same set of examples and obtained the following results.

n	min	min	min	# success of			
	(d_2/d_1)	(d_3/d_2)	(d_4/d_3)	(21)	(22)	(23)	(24)
	max	max	max				
5	1.3e9	3.7	1.0002	3	3	3	3
	2.2e9	4.3	1.0004				
10	1.9e9	4.3	1.0006	9	9	9	9
	5.0e9	18.8	1.0060				
15	3.0e9	7.8	1.002	12	12	12	12
	8.5e9	50.8	1.027				

Table 3. Random real coefficients in $[-1, 1]$ with $\epsilon = 10^{-4}$

n	min	min	min	# success of			
	(d_2/d_1)	(d_3/d_2)	(d_4/d_3)	(21)	(22)	(23)	(24)
	max	max	max				
5	2.0e12	*	1.03	4	4	5	5
	1.1e13	48.3	2.10				
10	*	*	1.06	6	3	9	9
	1.2e13	8.3	1.56				
15	*	*	*	1	0	10	12
			3.2				

Table 4. Random real coefficients in $[-1, 1]$ with $\epsilon = 10^{-2}$

For $\epsilon = 10^{-2}$ and method (22) does not work for any pair (\tilde{z}_i, Z_i) . Therefore the first two columns only contain a *. Interestingly in this case (23) and more (24) perform much better than (21).

Finally we give two examples for higher degree, namely $n = 20$ and $n = 50$. In this ϵ has to be fairly small to allow an inclusion because of instability of zeros of polynomials of high degree. We choose $\epsilon = 10^{-8}$ and 10^{-10} .

n	ϵ	min	min	min	# success of			
		(d_2/d_1)	(d_3/d_2)	(d_4/d_3)	(21)	(22)	(23)	(24)
		max	max	max				
20	10^{-8}	39.4	10.5	1.0000	16	16	16	16
		126.2	70.9	1.0000				
50	10^{-10}	1.001	1.0009	1.0000	50	50	50	50
		1.033	1.0342	1.0000				

Table 5. Random real coefficients in $[-1, 1]$

5. Conclusion

In Theorem 5 a new method for verifying the existence and uniqueness of a real or complex zero of a real or complex nonlinear function in one variable within an interval is given. It is compared with other known methods. We were especially interested in the contraction property of the Ansatz used in Theorem 5.

The numerical results show that for wider intervals method (24) offers the highest chances to be successful in verifying existence and uniqueness of a zero within the test interval, followed by method (23), (21) and finally (22). If (21) succeeds in verification it produces a very narrow inclusion being better by several orders of magnitude than the others. Moreover, the new method (21) requires very little computational effort.

References

- [1] Alefeld, G.; Herzberger, J.: Einführung in die Intervallrechnung, B.I. Wissenschaftsverlag (1974).
- [2] Moore, R.E.: Interval Analysis, Englewood Cliffs: Prentice Hall (1966).
- [3] Rump, S.M.: Estimation of the Sensitivity of Linear and Nonlinear Algebraic Problems, Linear Algebra and its Applications 153:1–34 (1991).
- [4] Rump, S.M.: Solving Algebraic Problems with High Accuracy, in “A New Approach to Scientific Computation” (eds. U. Kulisch, W.L. Miranker), Academic Press, pp. 51–120 (1983).
- [5] Alefeld, G.: private communication.
- [6] Böhm, H.: Berechnung von Polynomnullstellen und Auswertung arithmetischer Ausdrücke mit garantierter maximaler Genauigkeit, Dissertation, University of Karlsruhe (1983).
- [7] Rump, S.M.: CALCULUS, in “Wissenschaftliches Rechnen mit Ergebnisverifikation”, ed. U. Kulisch, Vieweg und Akademie Verlag Berlin (1989).