Complex Interval Division with Maximum Accuracy

R. Lohner, J. Wolff v. Gudenberg

Institute for Applied Mathematics,University of Karlsruhe
7500 Karlsruhe, West Germany

1. Introduction

In complex interval arithmetic three different types of intervals are introduced: rectangles, circles and circular sectors[1,4]. The arithmetic operations for all three types of intervals should deliver the closest inclusion of the set of all possible values, i.e.

$$A \boxed{\circ} B = \square(A \circ B) = \square\{a \circ b \mid a \in A,\ b \in B\}$$

for any two intervals A,B and $\circ \in \{+,-,*,/\}$, is the rounding from the powerset of $\mathbb{C}$ into the set of complex intervals[6]. For circular and circular sector arithmetic multiplication and division are optimal, whereas addition and subtraction are not. For rectangular arithmetic addition, subtraction and multiplication are optimal. Although several different approaches have been made e.g.[1,8] to improve rectangular division none of these is optimal.
In this paper we introduce an algorithm to compute complex rectangular interval division optimally. In part 2 we derive how the maximum and the minimum of the real and imaginary part can be calculated. In part 3 we show that the computation can be carried out with maximum accuracy which means that at most one floating point number lies between the computed bounds and the exact bounds[11]. In nearly all cases the computed interval is the closest inclusion of the exact solution in the floating point system. Part 4 contains some numerical examples computed in PASCAL-SC[7].

2. Theoretical Determination of the Extrema

For two complex rectangular intervals

$$A = [\underline{u},\overline{u}] + i[\underline{v},\overline{v}] \text{ and } C = [\underline{c},\overline{c}] + i[\underline{d},\overline{d}]$$

we want to compute the optimal rectangular interval inclusion of the set of all quotients

$$f+gi := \frac{u+vi}{c+di} = \frac{uc+vd}{c^2+d^2} + i\frac{vc-ud}{c^2+d^2} ,$$

$$u+vi \in A, \quad c+di \in C, \quad (c^2+d^2>0) .$$

This means that we have to compute the maximum and the minimum of the two functions

$$f = f(u,v,c,d) = \frac{uc+vd}{c^2+d^2} \quad \text{and}$$

$$g = g(u,v,c,d) = \frac{vc-ud}{c^2+d^2}$$

with respect to all $u \in [\underline{u},\overline{u}]$, $v \in [\underline{v},\overline{v}]$, $c \in [\underline{c},\overline{c}]$, $d \in [\underline{d},\overline{d}]$.

We only discuss this for the maximum of f, since all other cases can be treated analogously.

Since f is linear in u and v the maximum is taken for $\underline{u}$ or $\overline{u}$ and $\underline{v}$ or $\overline{v}$ where c and d are fixed. This reduces the number of independent variables and we only need to consider

$$f(c,d) := \max_{\substack{u \in [\underline{u},\overline{u}] \\ v \in [\underline{v},\overline{v}]}} f(u,v,c,d) = \frac{1}{c^2+d^2} \cdot q$$

$$\text{with } q = \begin{cases} \overline{u}c+\overline{v}d \text{ , if } \underline{c} \geq 0 \text{ , } \underline{d} \geq 0 \text{ ,} \\ \overline{u}c+\underline{v}d \text{ , if } \underline{c} \geq 0 \text{ , } \overline{d} \leq 0 \text{ ,} \\ \underline{u}c+\overline{v}d \text{ , if } \overline{c} \leq 0 \text{ , } \underline{d} \geq 0 \text{ ,} \\ \underline{u}c+\underline{v}d \text{ , if } \overline{c} \leq 0 \text{ , } \overline{d} \leq 0 \text{ ,} \\ \{{\overline{u} \atop \underline{u}}\}\ c+\overline{v}d\text{,if } \underline{c} < 0 < \overline{c} \text{ , } \underline{d} > 0 \text{ ,} \\ \{{\overline{u} \atop \underline{u}}\}\ c+\underline{v}d\text{,if } \underline{c} < 0 < \overline{c} \text{ , } \overline{d} < 0 \text{ ,} \\ \overline{u}c+\{{\overline{v} \atop \underline{v}}\}d\text{,if } \underline{c} > 0 \text{ , } \underline{d} < 0 < \overline{d} \text{ ,} \\ \underline{u}c+\{{\overline{v} \atop \underline{v}}\}d\text{,if } \overline{c} < 0 \text{ , } \underline{d} < 0 < \overline{d} \text{ ,} \end{cases}$$

In the last four cases u resp. v cannot be chosen a priori since c resp. d changes sign. In these cases we have to compute the maximum for $u = \overline{u}$ as well as $u=\underline{u}$ resp. $v=\underline{v}$ as well as $v=\overline{v}$.

The real part

$$f(c,d) = \frac{uc+vd}{c^2+d^2}$$

of the analytic function $\frac{A}{C}$ (A fixed, C independent variable) is a harmonic function and therefore takes its extrema on the boundary of the rectangle

CH2146-9/85/0000/0332$01.00 © 1985 IEEE

$[\underline{c},\overline{c}] \times [\underline{d},\overline{d}]$

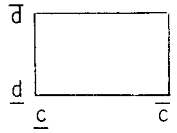


Figure 1

For each of the four edges we have to compute the maximum of f(c,d). This can be reduced to

$$\max_{x\in[\underline{x},\overline{x}]} f(x) \text{ with } f(x) = \frac{ax+by_0}{x^2+y_0^2} .$$

All four edges as well as the computation of the imaginary part can be treated by choosing a,b,x and y_0 suitably. (Note that for the imaginary part g(u,v,c,d) = f(v,-u,c,d). The first two derivatives of f are

$$f'(x) = \frac{1}{(x^2+y_0^2)^2}\{a(x^2+y_0^2) - 2x(ax+by_0))$$

$$= \frac{1}{(x^2+y_0^2)^2}(a(y_0^2-x^2) - 2by_0x)$$

$$f''(x) = \frac{1}{(x^2+y_0^2)^3}(-2(ax+by_0)(x^2+y_0^2)$$

$$- 4x(a(y_0^2-x^2) - 2by_0x)) .$$

To compute the stationary points, we have to consider three cases

(i) $a = 0$,
(ii) $a \neq 0$ and $y_0 = 0$,
(iii) $a \neq 0$ and $y_0 \neq 0$.

(i): $\boxed{a = 0}$

$$\Rightarrow f'(x) = -\frac{2by_0x}{(x^2+y_0^2)^2}$$

(α) $b=0$ or $y_0=0$ $\Rightarrow \boxed{f\equiv 0}$

(β) $b\neq 0$ and $y_0\neq 0$ $\Rightarrow f'(x)=0 \Leftrightarrow x_1 = 0$

$$\Rightarrow f''(x_1)=f''(0)$$

$$=-2\frac{b}{y_0^3}\begin{cases} <0, \text{ if } by_0 > 0 , \\ >0, \text{ if } by_0 < 0 . \end{cases}$$

For $x_1 = 0$ there is a maximum of f, if $by_0 > 0$, and a minimum if $by_c < 0$.

The extremal value is $\boxed{f(x_1) = f(0) = \frac{b}{y_0}}$

(ii): $\boxed{a \neq 0 \text{ and } y_0 = 0}$

$\Rightarrow f(x) = \frac{a}{x}$, i.e. f is a hyperbola and thus

$\boxed{f(\underline{x}) = \frac{a}{\underline{x}}}$ is a maximum of f on $[\underline{x},\overline{x}]$,if $a > 0$,

$\boxed{f(\overline{x}) = \frac{a}{\overline{x}}}$ is a maximum of f on $[\underline{x},\overline{x}]$,if $a < 0$.

(iii): $\boxed{a \neq 0 \text{ and } y_0 \neq 0}$

$$f'(x) = 0 \Leftrightarrow a(y_0^2-x^2) - 2by_0x = 0$$

$$\Leftrightarrow x^2+2\frac{by_0}{a}x - y_0^2 = 0$$

$$\Leftrightarrow x_{1,2} = -\frac{by_0}{a} \pm \sqrt{\frac{b^2y_0^2}{a^2} + y_0^2}$$

$$\Leftrightarrow \boxed{\begin{aligned} x_1 &= (-b+\sqrt{a^2+b^2})\frac{y_0}{a} , \\ x_2 &= (-b-\sqrt{a^2+b^2})\frac{y_0}{a} \end{aligned}}$$

Since

$$f''(x_1) = \frac{-2}{(x_1^2+y_0^2)^2}(ax_1+by_0) = \frac{-2y_0\sqrt{a^2+b^2}}{(x_1^2+y_0^2)^2} ,$$

and

$$f''(x_2) = \frac{2y_0\sqrt{a^2+b^2}}{(x_2^2+y_0^2)^2}$$

it follows:

$f''(x_1) < 0 \Leftrightarrow y_0 > 0$, i.e. $f(x_1)$ is a maximum ,

and

$f''(x_2) < 0 \Leftrightarrow y_0 < 0$, i.e. $f(x_2)$ is a maximum .

The value of the maximum is

$$f(x_1) = \frac{(-b+\sqrt{a^2+b^2})y_0+by_0}{(-b+\sqrt{a^2+b^2})\frac{y_0^2}{a^2} + y_0^2}$$

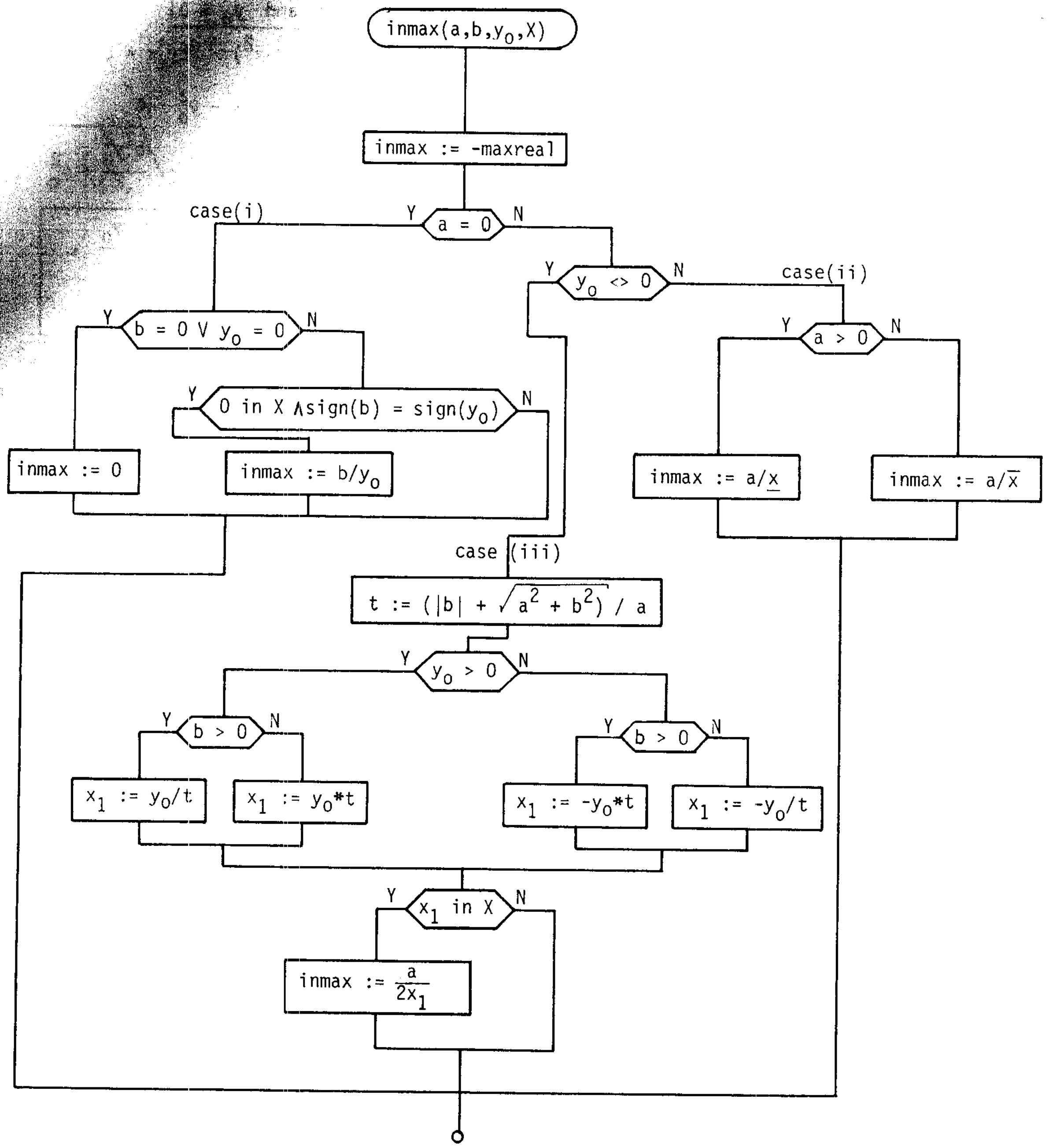
$$= \frac{a^2}{y_0}\cdot\frac{\sqrt{a^2+b^2}}{2(a^2+b^2) -2b\sqrt{a^2+b^2}}$$

$$= \frac{a^2}{2y_0}\frac{1}{-b+\sqrt{a^2+b^2}} = \frac{a}{2x_1} , \text{ if } y_0 > 0 ,$$

analogously

$$f(x_2) = \frac{a}{2x_2} , \text{ if } y_0 < 0 .$$

In cases (i) and (iii) we have to check, if the point x_1, resp. x_2, where f has a relative maximum is contained in $[\underline{x},\overline{x}]$. If this holds then this is also the absolute maximum of f on this edge, since f has at most one relative maximum and one relative minimum on $\mathbb{R}$ and vanishes asymptotically for $|x| \to \infty$. The algorithm of the computation of the maximum is displayed in the following flow chart.



To obtain the final value of the maximum we have to compare the result of this computation with f($\underline{x}$) and f($\overline{x}$).

3. Numerical computation with maximum accuracy

Let the problem be given in a floating point system with n digits. Then we want the final result to be correct to n digits. For convenience we use a decimal system for the estimations.

In case (i) and (ii) the computation of the maximum is a single division, which is carried out with maximum accuracy and directed rounding.

In case (iii) let x1 be the point where f has its relative maximum. Then we compute $\tilde{x}$1 and the maximum f($\tilde{x}$1) in a floating point system with $\ell > n$ mantissa digits. We now determine ℓ such that the results are accurate to at least n digits.

During the computation of x1 intermediate overflow or underflow can easily be avoided by an error free multiplication with a proper constant. We now give an estimation for the rounding error of $\tilde{x}$1. We only discuss the more complicated case $x1 = y_0/t$. Floating point operations are denoted in circles and the rounded values are marked with $\sim$.

$$\left|\frac{x1 - \tilde{x}1}{x1}\right| = \left|\frac{y_0/t - y_0 \oslash \tilde{t}}{y_0/t}\right|$$

$$\leq \underbrace{\left|\frac{t - \tilde{t}}{t}\right|}_{:= \gamma} \left|\frac{t}{\tilde{t}}\right| + \left|\frac{y_0/\tilde{t} - y_0 \oslash \tilde{t}}{y_0/\tilde{t}}\right| \left|\frac{\tilde{t}}{t}\right| .$$

With $w := \sqrt{a^2+b^2}$ and $\tilde{w} := \widetilde{\sqrt{}_o(a^2+b^2)}$ an approximation for w we have

$$\gamma = \left|\frac{(|b|+w)/a - (|b| \oplus \tilde{w}) \oslash a}{(|b|+w)/a}\right|$$

$$\leq \left|\frac{(|b|+w) - (|b| \oplus \tilde{w})}{|b|+w}\right| + 5_{10^{-\ell}} \left|\frac{|b| \oplus \tilde{w}}{|b|+w}\right|$$

$$\leq \left|\frac{w-\tilde{w}}{|b|+w}\right| + 5_{10^{-\ell}} \left|\frac{|b|+\tilde{w}}{|b|+w}\right| + 5_{10^{-\ell}} \left|\frac{|b| \oplus \tilde{w}}{|b|+w}\right|$$

$$\leq \left|\frac{w-\tilde{w}}{w}\right| + 5_{10^{-\ell}} \left(1+ \left|\frac{w-\tilde{w}}{w}\right|\right)$$

$$+ 5_{10^{-\ell}} (1+5_{10^{-\ell}}) \left(1+ \left|\frac{w-\tilde{w}}{w}\right| \right).$$

With an appropriate approximation $\tilde{w}$ for w we obtain

$$\left|\frac{x1-\tilde{x}1}{x1}\right| < 21_{10^{-\ell}} .$$

For the relative error ε of the computed maximum $a \oslash (2 \odot \tilde{x}1)$ we have

$$\varepsilon \leq \left|\frac{a/(2x1)-a/(2\tilde{x}1)}{a/(2x1)}\right| + \left|\frac{a/(2\tilde{x}1)-a \oslash (2 \odot \tilde{x}1)}{a/(2x1)}\right|$$

$$\leq \frac{21_{10^{-\ell}}}{1-21_{10^{-\ell}}} + \left|\frac{a/(2\tilde{x}1)-a/(2 \odot \tilde{x}1)}{a/(2\tilde{x}1)}\right|$$

$$+ \left|\frac{a/(2 \odot \tilde{x}1)-a \oslash (2 \odot \tilde{x}1)}{a/2 \odot \tilde{x}1)}\right| \left|\frac{2\tilde{x}1}{2 \odot \tilde{x}1}\right| \left|\frac{x1}{\tilde{x}1}\right|$$

$$\leq 32_{10^{-\ell}} .$$

If $x1 \in [\underline{x},\overline{x}]$ the maximum of f is taken at $\underline{x}$ or $\overline{x}$. We then have to compute $f(\underline{x})$ and $f(\overline{x})$. This computation can be done with two optimal dot products yielding ℓ mantissa digits and one final division. The error of this computation is less than $32_{10^{-\ell}}$.[10]

Of course $x1 \in [\underline{x},\overline{x}]$ does not necessarily imply $\tilde{x}1 \in [\underline{x},\overline{x}]$ and vice versa. In such cases $f(\underline{x})$ resp. $f(\overline{x})$ satisfies the above estimations.

The final result is obtained by correction of the computed value with the relative error and directed rounding to n digits, provided the computation has been carried out with $\ell \geq n+2$ digits.

Remarks: 1) In the worst case the algorithm requires the computation of 32 stationary points and evaluation of up to 32 function values of f(x).

2) It is possible to compute the closest inclusion of the exact solution without exception. But this would mean that we have to check the result, which requires the exact computation of a sum of products of 3 or 4 floating point numbers.

4. Examples and Applications

We compare

(i) $\frac{A}{B} = \frac{U+iV}{C+iD} = \frac{UC+VD}{C^2+D^2} + i\ \frac{VC-UD}{C^2+D^2}$

(ii) $\frac{A}{B} = A \cdot \frac{1}{B}$ where $\frac{1}{B}$ is computed optimally and $A \cdot X$ is the usual multiplication,[8]

(iii) our method.

1) $A = 1 + i$, $B = 1 + i[0,1]$

(i) $A / B = [0.5,2] + i[0,1]$
(ii) $A / B = [0.5,1.5] + i[0,1]$
(iii) $A / B = [1,1.20710678119] + i[0,1]$

2) $A = [1,2] + i[1,2]$ $B = [1,2] + i[1,2]$

(i) $A / B = [0.25,4] + i[-1.5,1.5]$
(ii) $A / B = [0.4,2] + i[-0.8,0.8]$
(iii) $A / B = [0.5,2] + 0.6180339887450\ i[-1,1]$

3) $A = 10^{30} + 10^{30}i$, $B = 3 + 3i$

(i) $A / B= [0.3333333333333,0.3333333333334] \cdot 10^{30} + 0 \cdot i$
(ii) $A / B= [0.3333333333332,0.3333333333334] \cdot 10^{30} + [-10^{18},10^{18}]i$
(iii) $A / B= [0.3333333333333,0.3333333333334] \cdot 10^{30} + 0 \cdot i$

4) Execution of the interval Newton method

$$Z_{k+1} := m(Z_k) - f(m(Z_k)) / f'(Z_k)$$

for the function $f(z) = z^2-2z+2$ starting with the interval $Z_0 = [0,1.5] + i[0.17,1.2]$, which contains exactly one zero $z^* = 1+i$ of f, yields the following results.

Using method (ii) to compute $f(m(Z_k)) / f'(Z_k)$ we obtain

```
k                  Z_k

1  [-9.5E-02, 2.7E+00] + i[ 2.3E-01, 3.0E+00]
2  [-2.2E+00, 2.8E+00] + i[-2.5E+00, 2.5E+00]
```

Obviously the third iteration is not defined since $0 \in f'(Z_k)$.

Our method (iii), however, yields convergence towards the exact solution (even in a finite number of steps):

```
k                                  Z_k

1  [          2.4E-01,             2.3E+00]
 +i[          5.5E-01,             2.6E+00]
2  [          1.7E-01,             1.7E+00]
 +i[          2.2E-01,             1.5E+00]
3  [          7.0E-01,             1.5E+00]
 +i[          7.9E-01,             1.6E+00]
```

```
4  [          8.9E-01,              1.1E+00]
  +i[          9.2E-01,              1.1E+00]
5  [        9.993E-01,            1.001E+00]
  +i[        9.993E-01,            1.001E+00]
6  [    9.9999993E-01,        1.0000001E+00]
  +i[    9.9999993E-01,        1.0000001E+00]
7  [9.999999999989E-01, 1.000000000001E+00]
  +i[9.999999999994E-01, 1.000000000001E+00]
8  [1.000000000000E+00, 1.000000000000E+00]
  +i[1.000000000000E+00, 1.000000000000E+00]
```

Note that the first intervals in this example do not show all digits, since they are not interesting for the result.

References

[1] G.Alefeld, J.Herzberger: Einführung in die Intervallrechnung, BI Mannheim (1974).

[2] G.Alefeld, J.Herzberger: Introduction to Interval Computations. Tr. by J. Rokne, Academic Press, New York (1983).

[3] G.Bohlender, H.Böhm, K.Grüner, E.Kaucher, R. Klatte, W.Krämer, U.W.Kulisch, W.L.Miranker, S.M.Rump, Ch.Ullrich, J.Wolff v. Gudenberg: MATRIX PASCAL in [7] p. 311-384.

[4] R. Klatte, Ch.Ullrich: Complex Sector Arithmetic, Computing 24, (1980) p. 139-148.

[5] U.Kulisch: Grundlagen des Numerischen Rechnens, BI Mannheim (1971).

[6] U.Kulisch, W.L.Miranker: Computer-Arithmetic in Theory and Practice, Academic Press, New York (1981).

[7] U.Kulisch, W.L.Miranker (ed): A New Approach to Scientific Computation. Academic Press, New York (1983).

[8] J.Rokne, P.Lancaster: Complex Interval Arithmetic, Comm. ACM 14 (1971) p.111-112.

[9] S.M. Rump: Über die komplexe Intervalldivision, Pers. Mitteilung (1983).

[10] J.Wolff v. Gudenberg: Einbettung allgemeiner Rechnerarithmetik in PASCAL mittels eines Operatorkonzeptes und Implementierung der Standardfunktionen mit maximaler Genauigkeit, Dissertation, Universität Karlsruhe (1980).

[11] J.Wolff v. Gudenberg: An Introduction to MATRIX PASCAL in [7] p. 225-246.