Program Verification in the presence of complex numbers, functions with branch cuts etc.

J.H. Davenport — J.H.Davenport@bath.ac.uk

September 24, 2012

Abstract

In considering the reliability of numerical programs, it is normal to "limit our study to the semantics dealing with numerical precision" (Martel, 2005) [Mar05]. On the other hand, there is a great deal of work on the reliability of programs that essentially ignores the numerics. The thesis of this paper is that there is a class of problems that fall between these two, which could be described as "does the low-level arithmetic implement the high-level mathematics". Many of these problems arise because mathematics, particularly the mathematics of the complex numbers, is more difficult than expected: for example the complex function log is not continuous, writing down a program to compute an inverse function is more complicated than just solving an equation, and many algebraic simplification rules are not universally valid.

The good news is that these problems are *theoretically* capable of being solved, and are *practically* close to being solved, but not yet solved, in several real-world examples. However, there is still a long way to go before implementations match the theoretical possibilities.

1 Introduction

It is customary, even though not often explicitly stated, to think that programming errors in numerical programs come in three distinct flavours, which we can categorise as follows.

blunder (of the coding variety) This is the sort of error traditionally addressed in "program verification": are all array elements properly initialised before use, "fence post" errors[1], are array bounds always respected etc.? These problems are essentially independent of the numerics of the problem, and indeed are normally taught/described in integer contexts.

[1] From the old puzzle "A farmer wishes to make a 100-metre fence with supporting posts every 10 metres — how many posts are needed", to which the answer is 11, since each end needs a post.

parallelism Issues of deadlocks or races occurring due to the parallelism of an otherwise correct sequential program. Again, these problems are essentially independent of the numerics of the problem.

numerical Do truncation and round-off errors, individually or combined, mean that the program computes approximations to the "true" answers which are out of tolerance. This is the area traditionally addressed in Numerical Analysis. There are really two subquestions here: the rounding question, i.e. does $\mathbf{R}_{IEEE}$ (or whatever other arithmetic we are using) approximate $\mathbf{R}$ sufficiently well, and the truncation error question, e.g. is the discretisation h small enough that it is the mathematical ϵ or is $\sum_1^N$ equivalent to $\sum_1^\infty$. Unfortunately the two interact; for example reducing h in $f'(x) \approx \frac{f(x+h)-f(x)}{h}$ to reduce the truncation error increases the rounding problems.

We note that [CE05], taken as a specimen of the verification literature, contains 30 papers, of which only [Mar05] deals with strictly numerical issues, four with parallelism issues, and the rest (83%) with the first kind.

It is the thesis of this paper that there is a fourth kind of error, which we can describe as follows

manipulation A piece of algebra, which is "obviously correct", turns out not to be correct when interpreted, not as abstract algebra, but as the manipulation of functions $\mathbf{R} \to \mathbf{R}$ or $\mathbf{C} \to \mathbf{C}$.

Note: throughout this paper we take the standard definitions of the branch cuts of the elementary functions from [AS64, Nat10, as tightened in [CDJW00]]. Other definitions would have different, but not fewer, problems. We also use the Anglo-Saxon convention that log etc. (and $\surd$) denote single-valued functions ($\log 1 = 0$, $\surd 4 = 2$), whereas Log etc. denote multi-valued functions ($\mathrm{Log}(1) = \{2k\pi i : k \in \mathbf{Z}\}$, $\mathrm{Sqrt}(4) = \{2, -2\}$).

The author would like to thank Russell Bradford, Acyr Locatelli, Gregory Sankaran and David Wilson of the Bath Triangular Sets seminar for their input, and the referees for their comments, but the errors and omissions are all his.

2 Examples

The problems we are going to describe arise largely (though not entirely[2]) from complex numbers, and it is sometimes said "real programs don't use complex numbers", despite the fact that the introduction of `COMPLEX` into Fortran II was probably the first instance of a language data type that did not correspond to a machine data type. The author knows of several major uses of complex numbers and complex analysis, in particular many problems which arise in fluid dynamics, where two-dimensional real space $\mathbf{R}^2 = \{(x, y)\}$ is viewed as the complex plane $\mathbf{C} = \{z = x + iy\}$. It is then normal to map this copy of $\mathbf{C}$

[2]See section 2.4 for a counter-example.

to another (in which the variable is traditionally denoted w or ζ) where the problem is easier to solve. Such an analytic map $z \mapsto w$ is termed a *conformal* map.

2.1 Kahan's example

This example comes from [Kah87, pp. 187–189], and the ultimate motivation is fluid flow in a slotted strip (z space), which we wish to transform to a more convenient shape (w space).

With the usual definitions, the necessary conformal map

$$w = g(z) := 2\,\mathrm{arccosh}\left(1 + \frac{2z}{3}\right) - \mathrm{arccosh}\left(\frac{5z+12}{3(z+4)}\right) \tag{1}$$

is only the same as the ostensibly more efficient

$$w \stackrel{?}{=} q(z) := 2\,\mathrm{arccosh}\left(2(z+3)\sqrt{\frac{z+3}{27(z+4)}}\right), \tag{2}$$

if we avoid the negative real axis and the area

$$\left\{ z = x + iy : |y| \le \sqrt{\frac{(x+3)^2(-2x-9)}{2x+5}} \wedge -9/2 \le x \le -3 \right\} \tag{3}$$

In fact, most computer algebra systems will refuse, these days, to convert one into the other, but this does not constitute a proof of difference.

Challenge 1 *Demonstrate automatically that g and q are not equal, by producing a z at which they give different results.*

The technology described in [BBDP07] will isolate the curve $y = \pm\sqrt{\frac{(x+3)^2(-2x-9)}{2x+5}}$ as a potential obstacle (it is the branch cut of q), but the geometry questions are too hard for a fully-automated solution at the moment.

However, simplification *is* possible: g can legitimately be rewritten to

$$w = h(z) := 2\,\ln\left(\frac{1}{3}\,\frac{\sqrt{3\,z+12}\left(\sqrt{z+3}+\sqrt{z}\right)^2}{2\,\sqrt{z+3}+\sqrt{z}}\right), \tag{4}$$

The technology in [BBDP07] can show this, i.e. $\forall z \in \mathbf{C} g(z) = h(z)$, but again the geometry questions are too hard for a fully-automated solution at the moment. Indeed Maple 16 currently gets this wrong: `coulditbe(g<>h);` returns `true`, which *ought* to indicate that there is a counter-example.

Challenge 2 *Demonstrate automatically that g and h are equal.*

The technology in [BBDP07], implemented in a mixture of Maple and QEPCAD, could in principle do this, but the geometry questions are too hard for a fully-automated solution at the moment. In addition, we would be left with the problem of trusting the underlying demonstration code, so there is the additional problem of translating this methodology into a tool such as MetiTarski [Pau12].

2.2 Joukowski (a)

Consider the Joukowski map [Hen74, pp. 294–298]:

$$f : z \mapsto \frac{1}{2}\left(z + \frac{1}{z}\right). \tag{5}$$

Lemma 1 *f is injective as a function from $D := \{z : |z| > 1\}$.*

If $z \mapsto \zeta$ then $1/z \mapsto \zeta$, and there are no other pre-images of ζ (since the algebraic inverse of (5) is the solution of a quadratic). If $|z| > 1$, then $|1/z| < 1$, so z is unique in D.

In fact f is a bijection from D to $\mathbf{C}^\ddagger := \mathbf{C}\setminus[-1, 1]$, and hence has an inverse.

Of course, (5) is the conformal map $\mathbf{C} \to \mathbf{C}$ that equates to the map

$$f_R : (x, y) \mapsto \left(\frac{1}{2}\,x + \frac{1}{2}\,\frac{x}{x^2 + y^2}, \frac{1}{2}\,y - \frac{1}{2}\,\frac{y}{x^2 + y^2}\right) \tag{6}$$

$\mathbf{R}^2 \to \mathbf{R}^2$. However, it is not obvious from (6) alone that f_R is a bijection $D \to \mathbf{C}^\ddagger$, i.e. that

$$\begin{aligned}\forall x_1 \forall x_2 \forall y_1 \forall y_2 \quad \Big(x_1^2 + y_1^2 > 1 \wedge x_2^2 + y_2^2 > 1 \wedge\ & x_1 + \tfrac{x_1}{x_1^2+y_1^2} = x_2 + \tfrac{x_2}{x_2^2+y_2^2} \wedge \\ & y_1 - \tfrac{y_1}{x_1^2+y_1^2} = y_2 - \tfrac{y_2}{x_2^2+y_2^2}\Big) \Rightarrow \big(x_1 = x_2 \wedge y_1 = y_2\big).\end{aligned} \tag{7}$$

Challenge 3 *Demonstrate automatically the truth of (7).*

We have been unable to do this with either the QEPCAD [Bro03] implementation of Partial Cylindrical Algebraic Decomposition [CH91] or the Maple implementation of Cylindrical Algebraic Decomposition via triangular decomposition [CMMXY09].

However, Brown [Bro12] has been able to reformulate the problem (manually) to make it amenable to QEPCAD, and indeed solved it in under 12 seconds.

Challenge 4 *Automate these techniques and transforms.*

Having established (or not) that f is a bijection $D \to \mathbf{C}^\ddagger$, we want its inverse. Formally, this is trivial, as one referee said

> The inverse of Joukowski is the solution of a quadratic with the usual sign ambiguity:

if $\zeta = \frac{1}{2}\left(z + \frac{1}{z}\right)$, then $2z\zeta = z^2 + 1$ and $z = \zeta \pm \sqrt{\zeta^2 - 1}$. This is easily within the grasp of computer algebra, as seen in Figure 1. The only challenge might be the choice implicit in the $\pm$ symbol: which do we choose?

Figure 1: Maple's `solve` on inverting Joukowski

```
> [solve(zeta = 1/2*(z+1/z), z)];
```

$$\left[\zeta + \sqrt{\zeta^2 - 1}, \zeta - \sqrt{\zeta^2 - 1}\right]$$

Figure 2: Maple's actual `solve` on inverting injective Joukowski

```
> [solve(zeta = 1/2*(z+1/z), z)] assuming abs(z) > 1
```

$$\left[\zeta + \sqrt{\zeta^2 - 1}, \zeta - \sqrt{\zeta^2 - 1}\right]$$

Unfortunately, the answer is "neither", or at least "neither, uniformly". The true answer is that, for f a bijection from $\{z : |z| > 1\}$ to $\mathbf{C} \setminus [-1, 1]$, its inverse is

$$f_1(\zeta) = \zeta \begin{cases} +\sqrt{\zeta^2 - 1} & \Im(\zeta) > 0 \\ -\sqrt{\zeta^2 - 1} & \Im\zeta) < 0 \\ +\sqrt{\zeta^2 - 1} & \Im(\zeta) = 0 \wedge \Re(\zeta) > 1 \\ -\sqrt{\zeta^2 - 1} & \Im(\zeta) = 0 \wedge \Re(\zeta) < -1 \end{cases} \tag{8}$$

In fact, a better (at least, free of case distinctions) definition is

$$f_2(\zeta) = \zeta + \sqrt{\zeta - 1}\sqrt{\zeta + 1}. \tag{9}$$

The techniques of [BBDP07] are able to **verify** (9), in the sense of showing that $f_2(f(z)) - z$ is the zero function on $\{z : |z| > 1\}$.

Challenge 5 *Derive automatically, and demonstrate the validity of, either (8) or (9). In terms of Maple, we would want to see Figure 3, rather than the actual Figure 2.*

In terms of derivation, the techniques of [CJ96] seem worthy of investigation, but the author has been unable to do this derivation satisfactorily by this route.

2.3 Joukowski (b)

Here the function is again given by (5).

Figure 3: Ideal Maple `solve` on inverting injective Joukowski

```
> solve(zeta = 1/2*(z+1/z), z) assuming abs(z) > 1
```

$$\zeta + \sqrt{\zeta - 1}\sqrt{\zeta + 1}$$

Lemma 2 *f is injective as a function from $H := \{z : \Im z > 0\}$.*

As in Lemma 1, if $z \mapsto \zeta$ then $1/z \mapsto \zeta$, and there are no other pre-images of ζ. If $\Im(z) > 0$, $\Im(1/z) < 0$, and f in therefore injective from H.

In fact, f is a bijection from H to $\mathbf{C} \setminus ((-\infty, -1] \cup [1, \infty))$, and hence has an inverse.

Again, it is not obvious from (6) alone that f_R is a bijection, now from $\{(x, y)|y > 0\}$, i.e. that

$$\forall x_1 \forall x_2 \forall y_1 \forall y_2 \quad \left(y_1 > 0 \wedge y_2 > 0 \wedge x_1 + \frac{x_1}{x_1^2+y_1^2} = x_2 + \frac{x_2}{x_2^2+y_2^2} \wedge y_1 - \frac{y_1}{x_1^2+y_1^2} = y_2 - \frac{y_2}{x_2^2+y_2^2} \right) \Rightarrow \big(x_1 = x_2 \wedge y_1 = y_2\big). \tag{10}$$

Challenge 6 *Demonstrate automatically the truth of (10).*

It is likely that the ideas of [Bro12] can do this, but again these need automation.

We have the same challenge over the inverse of f: again formally it is $f^{-1} \stackrel{?}{=} \zeta \pm \sqrt{\zeta^2 - 1}$, and the only challenge is the $\pm$ symbol: which do we choose? Here [Hen74, (5.1-13), p. 298] argues for

$$f_3(\zeta) = \zeta + \underbrace{\sqrt{\zeta - 1}}_{\arg \in (-\pi/2, \pi/2]} \underbrace{\sqrt{\zeta + 1}}_{\arg \in (0, \pi]}. \tag{11}$$

Challenge 7 *Find a way to represent functions such as* $\underbrace{\sqrt{\zeta + 1}}_{\arg \in (0,\pi]}$

Fortunately this one is soluble in this case[3], we can write $\underbrace{\sqrt{\zeta + 1}}_{\arg \in (0,\pi]} = i \underbrace{\sqrt{-\zeta - 1}}_{\arg \in (-\pi/2,\pi/2]}$, and the latter is the normal `sqrt` function of [AS64]. Hence we have an inverse function

$$f_4(\zeta) = \zeta + \sqrt{\zeta - 1} i \sqrt{-\zeta - 1}. \tag{12}$$

Challenge 8 *Demonstrate automatically that this is an inverse to f on $\{z : \Im z > 0\}$.*

2.4 A Real Example

Just in case the reader thinks that the real numbers are immune from these problems, consider the addition rule for the inverse tangent, quoted as

$$\mathrm{Arctan}(x) \pm \mathrm{Arctan}(y) = \mathrm{Arctan}\left(\frac{x \pm y}{1 \mp xy}\right). \quad [\mathrm{AS64}, (4.4.34)][\mathrm{Nat10}, (4.24.15)]$$

[3] And is probably soluble more generally, but the author knows of no general work on "alternative formulations".

Despite the caveat in [Nat10] that "The above equations are interpreted in the sense that every value of the left-hand side is a value of the right-hand side and vice versa", it is in fact the case that the 'obvious' two equations are true separately, `viz.`

$$\text{Arctan}(x) + \text{Arctan}(y) = \text{Arctan}\left(\frac{x+y}{1-xy}\right) \tag{13}$$

$$\text{Arctan}(x) - \text{Arctan}(y) = \text{Arctan}\left(\frac{x-y}{1+xy}\right) \tag{14}$$

Consider (13): This is valid for the multi-valued Arctan, but for the single-valued arctan only when $|1 - xy| < 1$, due to a "branch cut at infinity" of arctan. Nevertheless, the single-valued version of (13) is often cited as true: see for example [Ter12, (5.2.5)].

Over the reals, this is a non-challenge, the techniques of [BBDP07] do solve it easily, and produce a counterexample.

3 So why are these challenges?

3.1 Complex functions and branch cuts

These are difficult subjects, which have tended to be brushed under the carpet. The first truly algorithmic approach is ten years old ([BCD$^+$02], refined in [BBDP07]), and has various difficulties.

1. At its core is the use of Cylindrical Algebraic Decomposition of $\mathbf{R}^N$ to find the connected components of $\mathbf{C}^{N/2} \setminus \{\text{branch cuts}\}$. The complexity of this is doubly exponential in N: upper bound of $d^{O(2^N)}$ [Hon91] and lower bounds of $2^{2^{(N-1)/3}}$ [BD07, DH88]. While better algorithms are in principle known ([BRSEDS12] is $d^{O(N\sqrt{N})}$), we do not know of any accessible implementations.

   Furthermore, we are clearly limited to small values of N, at which point looking at $O(\ldots)$ complexity is of limited use. We note that the cross-over point between $2^{(N-1)/3}$ and $N\sqrt{N}$ is at $N = 21$. A more detailed comparison is given in [Hon91]. Hence there is a need for practical research on low-N Cylindrical Algebraic Decomposition.

2. While the fundamental branch cut of log is simple enough, being $\{z = x + iy | y = 0 \wedge x < 0\}$, actual branch cuts are messier. Part of the branch cut of (2) is

   $$2x^3 + 21x^2 + 72x + 2xy^2 + 5y^2 + 81 = 0 \wedge \text{other conditions}, \tag{15}$$

   whose solution accounts for the curious expression in (3). While there has been some progress in manipulating such images of half-lines (described in [PBD10, Phi11]), there is almost certainly more to be done.

3.2 Injectivity

Lemmas 1 and 2 might seem to be statements about complex functions of one variable, so why do we need to handle (or fail to handle) statements about four real variables to prove them? There are three, rather distinct, reasons for this.

1. The statements require the $|\cdot|$ function (Lemma 1) or the $\Im$ function (Lemma 2), neither of which are complex analytic functions. Hence some recourse to real analysis (and therefore twice as many variables) seems inevitable, though it would be nice to have a more formal statement and proof of this.

2. Equations (7) and (10) are the direct translations of the basic definition of injectivity. In practice, certainly if we were looking at functions $\mathbf{R} \to \mathbf{R}$, we would want to use the fact that the function concerned was continuous.

   Challenge 9 *Find a better formulation of injectivity questions* $\mathbf{R}^N \to \mathbf{R}^N$*, making use of the properties of the functions concerned (certainly continuity, possibly rationality).*

3. While equations (7) and (10) are statements from the existential theory of the reals, and so the theoretically more efficient algorithms quoted in [Hon91] are in principle applicable, the more modern developments described in [PJ09] do not seem to be directly applicable. However, we can transform then into a disjunction of statements to each of which the Weak Positivstellensatz [PJ09, Theorem 1] is applicable.

   Challenge 10 *Solve these problems using the techniques of [PJ09],*

4 Conclusions

The aim of this paper has been to demonstrate that translating mathematical problems into programs may require some algebraic manipulations whose accuracy is not as obvious as one might think, and whose verification is *currently* not as straightforward as we would like, despite the fact that their correctness is, in principle, decidable. A summary is given in Table 1.
These are, largely, concrete challenges that, we hope, will spur practical advances in this domain.

References

[AS64] M. Abramowitz and I. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. *US Government Printing Office*, 1964.

Table 1: Current state of these challenges

Challenge	State
1/2	Mathematically solved [BBDP07], geometry defeats current solvers.
3/6	Mathematically solved [Col75, etc.], geometry defeats current solvers.
4	Under development.
5/8	Mathematically solved [BBDP07], geometry defeats current solvers, and is probably significantly harder than the previous ones.
7	unknown: probably straightforward research project
9	unknown: research project
10	unknown: project for the authors of [PJ09].

[BBDP07] J.C. Beaumont, R.J. Bradford, J.H. Davenport, and N. Phisanbut. Testing Elementary Function Identities Using CAD. *AAECC*, 18:513–543, 2007.

[BCD+02] R.J. Bradford, R.M. Corless, J.H. Davenport, D.J. Jeffrey, and S.M. Watt. Reasoning about the Elementary Functions of Complex Analysis. *Annals of Mathematics and Artificial Intelligence*, 36:303–318, 2002.

[BD07] C.W. Brown and J.H. Davenport. The Complexity of Quantifier Elimination and Cylindrical Algebraic Decomposition. In C.W. Brown, editor, *Proceedings ISSAC 2007*, pages 54–60, 2007.

[Bro03] C.W. Brown. QEPCAD B: a program for computing with semi-algebraic sets using CADs. *ACM SIGSAM Bulletin 4*, 37:97–108, 2003.

[Bro12] C.W. Brown. Re: Query about QEPCAD. *Personal Commnication to David Wilson*, 2012.

[BRSEDS12] S. Basu, M.-F. Roy, M. Safey El Din, and É. Schost. A baby step-giant step roadmap algorithm for general algebraic sets. `http://arxiv.org/abs/1201.6439`, 2012.

[CDJW00] R.M. Corless, J.H. Davenport, D.J. Jeffrey, and S.M. Watt. According to Abramowitz and Stegun, or arccoth needn't be uncouth. *SIGSAM Bulletin 2*, 34:58–65, 2000.

[CE05] R. Cousot (Ed.). Verification, Model Checking, and Abstract Interpretation. *Springer Lecture Notes in Computer Science 3385*, 2005.

[CH91] G.E. Collins and H. Hong. Partial Cylindrical Algebraic Decomposition for Quantifier Elimination. *J. Symbolic Comp.*, 12:299–328, 1991.

[CJ96] R.M. Corless and D.J. Jeffrey. The Unwinding Number. *SIGSAM Bulletin 2*, 30:28–35, 1996.

[CMMXY09] C. Chen, M. Moreno Maza, B. Xia, and L. Yang. Computing Cylindrical Algebraic Decomposition via Triangular Decomposition. In J. May, editor, *Proceedings ISSAC 2009*, pages 95–102, 2009.

[Col75] G.E. Collins. Quantifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition. In *Proceedings 2nd. GI Conference Automata Theory & Formal Languages*, pages 134–183, 1975.

[DH88] J.H. Davenport and J. Heintz. Real Quantifier Elimination is Doubly Exponential. *J. Symbolic Comp.*, 5:29–35, 1988.

[Hen74] P. Henrici. Applied and Computational Complex Analysis I. *Wiley*, 1974.

[Hon91] H. Hong. Comparison of several decision algorithms for the existential theory of the reals. Technical Report 91-41, 1991.

[Kah87] W. Kahan. Branch Cuts for Complex Elementary Functions. In A. Iserles and M.J.D. Powell, editors, *Proceedings The State of Art in Numerical Analysis*, pages 165–211, 1987.

[Mar05] M. Martel. An Overview of Semantics for the Validation of Numerical Programs. In *Proceedings Verification Model Checking and Abstract Interpretation. Springer Lecture Notes in Computer Science 3385*, pages 59–77, 2005.

[Nat10] National Institute for Standards and Technology. The NIST Digital Library of Mathematical Functions. `http://dlmf.nist.gov`, 2010.

[Pau12] L.C. Paulson. MetiTarski: Past and Future. In *Proceedings Interactive Theorem Proving*, pages 1–10, 2012.

[PBD10] N. Phisanbut, R.J. Bradford, and J.H. Davenport. Geometry of Branch Cuts. *Communications in Computer Algebra*, 44:132–135, 2010.

[Phi11] N. Phisanbut. *Practical Simplification of Elementary Functions using Cylindrical Algebraic Decomposition*. PhD thesis, University of Bath, 2011.

[PJ09] G.O. Passmore and P.B. Jackson. Combined Decision Techniques for the Existential Theory of the Reals. In J. Carette *et al.*, editor, *Proceedings Intelligent Computer Mathematics*, pages 122–137, 2009.

[Ter12] D. Terr. Math is Amazingly Powerful. http://www.mathamazement.com/Lessons/Pre-Calculus/05_Analytic-Trigonometry/sum-and-difference-formulas.html, 2012.