Reasoning about the elementary functions of complex analysis

Robert M. Corless
David J. Jeffrey
Stephen M. Watt
Ontario Research Centre
for Computer Algebra
www.orcca.on.ca

Russell Bradford
James H. Davenport*
Dept. Computer Science
University of Bath
Bath BA2 7AY, England
R.J.Bradford,J.H.Davenport@bath.ac.uk

Submitted to Annals of Mathematics and Artificial Intelligence,
August 2001

Abstract

There are many problems with the simplification of elementary functions, particularly over the complex plane. Systems tend to make "howlers" or not to simplify enough. In this paper we outline the "unwinding number" approach to such problems, and show how it can be used to prevent errors and to systematise such simplification, even though we have not yet reduced the simplification process to a complete algorithm. The unsolved problems are probably more amenable to the techniques of artificial intelligence and theorem proving than the original problem of complex-variable analysis.

Keywords: Elementary functions; Branch cuts; Complex identities.

1 Introduction

The elementary functions are traditionally thought of as log, exp and the trigonometric and hyperbolic functions (and their inverses). This should include pow-

*Much of this work was performed while this author held the Ontario Research Chair in Computer Algebra at the University of Western Ontario. We are grateful to Mrs A. Davenport for her translation of [3].

ering (to non-integral powers) and also the n-th root. These functions are built in, to a greater or lesser extent, into many computer algebra systems (not to mention other programming languages [12, 17]), and are heavily used. As abstract algebraic solutions to integrals and/or differential equations, they satisfy well-known properties [16]. However, reasoning with them as functions $\mathbb{C} \to \mathbb{C}$ is more difficult than is usually acknowledged, and all algebra systems have one, sometimes both, of the following defects:

- they make mistakes, be it the traditional schoolchild one

$$1 = \sqrt{1} = \sqrt{(-1)^2} = -1 \tag{1}$$

  or more subtle ones;

- they fail to perform obvious simplifications, leaving the user with an impossible mess when there "ought" to be a simpler answer. In fact, there are two possibilities here: maybe there is a simpler equivalent that the system has failed to find, but maybe there isn't, and the simplification that the user wants is not actually valid, or is only valid outside an exceptional set. In general, the user is informed neither what the simplification might have been nor what the exceptional set is.

Faced with these problems, the user of the algebra system is not convinced that the result is correct, or that the algebra system in use understands the functions with which it is reasoning. An ideal algebra system would never generate incorrect results, and would simplify the results as much as practicable, even though perfect simplification is impossible, and not even totally well-defined: is $1 + x + \cdots + x^{1000}$ "simpler" than $(x^{1001} - 1)/(x - 1)$?

Throughout this paper, z and its decorations indicate a complex variable, while x, y and t indicate real variables. The symbol $\Im$ denotes the imaginary part, and $\Re$ the real part, of a complex number. For the purposes of this paper, the precise definitions of the inverse elementary functions in terms of log are those of [5]: these are reproduced in Appendix A for ease of reference.

2 The Problem

The fundamental problem is that log is multi-valued: since $\exp(2\pi i) = 1$, its inverse is only valid up to adding any multiple of $2\pi i$. This ambiguity is traditionally resolved by making a *branch cut*: usually [1, p. 67] the branch cut $(-\infty, 0]$, and the rule (4.1.2) that

$$-\pi < \Im \log z \leq \pi. \tag{2}$$

This then completely specifies the behaviour of log: on the branch cut it is continuous with the positive imaginary side of the cut, i.e. counter-clockwise continuous in the sense of [14].

What are the consequences of this definition[1]? From the existence of branch cuts, we get the problem of a lack of continuity:

$$\lim_{y \to 0^-} \log(x+iy) \neq \log(x) : \tag{3}$$

for $x < 0$ the limit is $\log(x) - 2\pi i$. Related to this is the fact that

$$\log \overline{z} \neq \overline{\log z} \tag{4}$$

on the branch cut: instead $\log \overline{z} = \overline{\log z} + 2\pi i$ on the cut. Similarly,

$$\log(\frac{1}{z}) \neq -\log z \tag{5}$$

on the branch cut: instead $\log(\frac{1}{z}) = -\log z + 2\pi i$ on the cut.

Although not normally explained this way, the problem with (1) is a consequence of the multi-valued nature of log: if we define (as for the purposes of this paper we do)

$$\sqrt{z} = \exp\left(\frac{1}{2}\log z\right), \tag{6}$$

then $\frac{-\pi}{2} < \Im\sqrt{z} \leq \frac{\pi}{2}$. On the real line, this leads to the traditional resolution of (1), namely that $\sqrt{x^2} = |x|$.

Four families of solutions have been proposed to these problems.

2.1 Signed Zero

[14] points out that the concept of a "signed zero"[2] [13] (for clarity, we write the positive zero as 0^+ and the negative one as 0^-) can be used to solve the above problems, if we say that, for $x < 0$, $\log(x + 0^+ i) = \log(|x|) + \pi i$ whereas $\log(x + 0^- i) = \log(|x|) - \pi i$. Equation (3) then becomes an equality for all x, interpreting the x on the right as $x + 0^- i$. Similarly, (4), where $\overline{0^+ i} = 0^- i$, and (5) become equalities throughout. Attractive though this proposal is, it does not answer the fundamental question as far as the designer of a computer algebra system is concerned: what to do if the user types $\log(-1)$.

[1] Which we do not contest: it seems that few people today would support the rule one of us (JHD) was taught, viz. that $0 \leq \Im \log z < 2\pi$. The placement of the branch cut is "merely" a notational convention, but an important one. If we wanted a function that behaves like log but with this cut, we could consider $\underbrace{\log}_{[0,2\pi)}(z) = \log(-1) - \log(-1/z)$ instead. We note that, until 1925, astronomers placed the branch cut between one day and the next at noon [9, vol. 15 p. 417].

[2] One could ask why zero should be special and have two values. The answer seems to be that all the branch cuts we need to consider are on either the real or imaginary axes, so the side to which the branch cut adheres depends on the sign of the imaginary or real part, including the sign of zero. To handle other points similarly would require the arithmetic of non-standard analysis.

1. $z = \ln e^z + 2\pi i \mathcal{K}(z)$.
2. $\mathcal{K}(a \ln z) = 0\, \forall z \in \mathbb{C}$ if and only if $-1 < a \leq 1$.
3. $\ln z_1 + \ln z_2 = \ln(z_1 z_2) + 2\pi i \mathcal{K}(\ln z_1 + \ln z_2)$.
4. $a \ln z = \ln z^a + 2\pi i \mathcal{K}(a \ln z)$.
5. $z^{ab} = (z^a)^b e^{2\pi i b \mathcal{K}(a \ln z)}$.

5′. $e^{ab} = (e^a)^b \; (-\pi < \Im a \leq \pi)$.

Table 1: Some correct identities for logarithms and powers using $\mathcal{K}$.

2.2 Uniformly valid transformations

The authors of [6] point out that most "equalities" do not hold for the complex logarithm, e.g. $\log(z^2) \neq 2 \log z$ (try $z = -1$), and its generalisation

$$\log(z_1 z_2) \neq \log(z_1) + \log(z_2). \tag{7}$$

The most fundamental of all non-equalities is $z \stackrel{?}{=} \log \exp z$, whose most obvious violation is at $z = 2\pi i$. (A similar point was made in [2], where the correction term is called the "adjustment".) They therefore propose to formalise the violation of this equality by introducing the *unwinding number* $\mathcal{K}$, defined[3] by

$$\mathcal{K}(z) = \frac{z - \log \exp z}{2\pi i} = \left\lceil \frac{\Im z - \pi}{2\pi} \right\rceil \in \mathbf{Z} \tag{8}$$

(note that the apparently equivalent definition $\lfloor \frac{\Im z + \pi}{2\pi} \rfloor$ differs precisely on the branch cut for log as applied to $\exp z$).

This definition has several attractive features: $\mathcal{K}(z)$ is integer-valued, and familiar in the sense that "everyone knows" that the multivalued logarithm can be written as the principal branch "plus $2\pi i k$ for some integer k"; it is single-valued; and it can be computed by a formula not involving logarithms. It does have a numerical difficulty, namely that you must decide if the imaginary part is an odd integer multiple of π or not, and this can be hard (or impossible in some exact arithmetic contexts), but the difficulty is inherent in the problem and cannot be repaired e.g. by putting the branch cuts elsewhere.

Some correct identities for elementary functions using $\mathcal{K}$ are given in Table 1.

(7) can then be rescued as

$$\log(z_1 z_2) = \log(z_1) + \log(z_2) - 2\pi i \mathcal{K}(\log(z_1) + \log(z_2)). \tag{9}$$

[3]Note that the sign convention here is the opposite to that of [6], which defined $\mathcal{K}(z)$ as $\lfloor \frac{\pi - \Im z}{2\pi} \rfloor$: the authors of [6] recanted later to keep the number of -1s occurring in formulae to a minimum. We could also change "unwinding" to "winding" when we make that sign change; but "winding number" is in wide use for other contexts, and it seems best to keep the existing terminology.

Similarly (4) can be rescued as

$$\log \overline{z} = \overline{\log z} - 2\pi i \mathcal{K}(\overline{\log z}). \tag{10}$$

Note that, as part of the algebra of $\mathcal{K}$, $\mathcal{K}(\overline{\log z}) = \mathcal{K}(-\log z) \neq \mathcal{K}(\log \frac{1}{z})$. $\mathcal{K}(z)$ depends only on the imaginary part of z. $\mathcal{K}(\log z) = 0$ for all z.

$$\mathcal{K}(-\log z) = \begin{cases} -1 & z \text{ real negative} \\ 0 & \text{elsewhere} \end{cases}. \tag{11}$$

2.3 Multi-valued "functions"

Although not formally proposed in the same way in the computational community, one possible solution, often found in texts in complex analysis, is to accept the multi-valued nature of these functions (we adopt the common convention of using capital letters, e.g. Ln, to denote the multi-valued function), defining, for example

$$\operatorname{Arcsin} z = \{y \mid \sin y = z\}.$$

This leads to $\sqrt{z^2} = \{\pm z\}$, which has the advantage that it is valid throughout $\mathbb{C}$. Equation 7 is then rewritten as

$$\operatorname{Ln}(z_1 z_2) = \operatorname{Ln}(z_1) + \operatorname{Ln}(z_2), \tag{12}$$

where addition is addition of sets ($A + B = \{a + b : a \in A, b \in B\}$) and equality is set equality[4].

However, it seems to lead in practice to very large and confusing formulae. More fundamentally, this approach does not say what will happen when the multi-valued functions are replaced by the single-valued ones of numerical programming languages.

A further problem that has not been stressed in the past is that this approach suffers from the same aliasing problem that naïve interval arithmetic does [8]. For example,

$$\operatorname{Ln}(z^2) = \operatorname{Ln}(z) + \operatorname{Ln}(z) \neq 2\operatorname{Ln}(z),$$

since $2\operatorname{Ln}(z) = \{2\ln(z) + 4k\pi i : k \in \mathbb{Z}\}$, but $\operatorname{Ln}(z) + \operatorname{Ln}(z) = \{2\ln(z) + 2k\pi i : k \in \mathbb{Z}\}$: indeed if $z = -1$, $\ln(z^2) \notin 2\operatorname{Ln}(z)$. Hence this method is unduly pessimistic: it may fail to prove some identities that are true.

2.4 Riemann surfaces

A long standing treatment of multi-valued functions is based on Riemann surfaces. Riemann surfaces give a very pictorial way of visualizing multi-valuedness [18, 7], but the question here is whether they can be used computationally to

[4] "The equation merely states that the sum of one of the (infinitely many) logarithms of z_1 and one of the (infinitely many) logarithms of z_2 can be found among the (infinitely many) logarithms of $z_1 z_2$, and conversely every logarithm of $z_1 z_2$ can be represented as a sum of this kind (with a suitable choice of [elements of] $\operatorname{Ln} z_1$ and $\operatorname{Ln} z_2$)." [3, pp. 259–260] (our notation).

study elementary functions (they have been used successfully for algebraic functions [10]). In order to answer this, a computational interpretation of a Riemann surface must first be given, because the standard textbook descriptions do not offer one.

The essence of the Riemann surface approach is geometric. We associate paths with each quantity that we compute with. Consider the pair of functions $z = e^w$ and $w = \ln z$ as an example. The approaches above have concentrated on the values taken by w, or in other words on the behaviour of $\ln z$ in the w-plane; in the Riemann-surface treatment, attention is shifted to the z-plane, by considering the mapping $z = e^w$. Starting from a path, say a straight vertical segment, joining two points w_1 and $w_2 = w_1 + 2\pi i$, we consider the path it maps to in the z plane. Let $z_1 = e^{w_1}$ and $z_2 = e^{w_2} = e^{w_1+2\pi i}$ be the endpoints of the path; usually these are considered to be the same point, but in the Riemann approach, we continue to distinguish between them. In order to create a distinction, we label the two z-points with the property that makes them different, namely the value of the imaginary part of w. In [18, 7], a 3-dimensional surface is constructed by plotting $\Im w$ against complex z. This can be shown to be isomorphic to any faithful representation of the Riemann surface for $\ln z$.

Therefore, in the Riemann approach, the complex number z now becomes a number and an index: z_{w_1} and z_{w_2}. Using the index, we can decide which value to assign $\ln z$. Thus $w_1 = \ln(z_{w_1})$ and $w_2 = \ln(z_{w_2})$. This is effectively how students in a first course on complex numbers compute the n values of $z^{1/n}$. They are taught to start with the equivalence $z = re^{i\theta} = re^{i\theta+2\pi ik}$ and then they are mysteriously ordered to apply the rule $(e^{i\phi})^{1/n} = e^{i\phi/n}$ to the second form of z rather than the first. Thus they are replacing z with z_k, an equivalent point on the kth Riemann sheet, and then computing $(z_k)^{1/n}$.

Taking the point of view of a computer algebra system, we see that a complex number z does not reveal its full significance until we know what Riemann sheet it is on. Thus, to compute logs correctly in the Riemann approach, a system would have to create a new data structure, so that the correct sheet of the Riemann surface could be recorded along with the value of z. This would have to be equivalent to a path from some reference point. This does not complete the solution to the problem, however, because the Riemann surface depends upon the function being considered: the Riemann sheet for log being different from that for $\arcsin z$, for instance.

Further problems arise when we consider combining functions. Thus, when we write $\ln(u + v)$, are u and v on the same sheet? More importantly, what sheet is the sum $u + v$ on? Products are easier: we can write $u = re^{i\theta}$ and record its sheet as a multiple of 2π in θ, and likewise for v; then the index of the sheet of the product uv is just the sum of the indexes of the sheets of u and v. But sums are awkward. Moreover, if we write the expression $(z - 1)^{1/2} + \ln z$, the Riemann surface for the combined function is different from either of the component Riemann surfaces. How do we label z?

We have to distinguish between a conceptual scheme and a computational scheme. Computer systems are about computation. Often computation assists in conception, but computers must be able to compute. Riemann surfaces are a beautiful conceptual scheme, but at the moment they are not computational schemes.

3 The rôle of the Unwinding Number

We claim that the unwinding number provides a convenient formalism for reasoning about these problems. Inserting the unwinding number systematically allows one to make "simplifying" transformations that *are* mathematically valid. The unwinding number can be evaluated at any point, either symbolically or via guaranteed arithmetic: since we know it is an integer, in practice little accuracy is necessary. Conversely, removing unwinding numbers lets us genuinely "simplify" a result. We describe insertion and removal as separate steps, but in practice every unwinding number, once inserted by a "simplification" rule, should be eliminated as soon as possible. We have thus defined a concrete goal for mathematically valid simplification.[5]

The following section gives examples of reasoning with unwinding numbers. Having motivated the use of unwinding numbers, the subsequent sections deal with their insertion (to preserve correctness) and their elimination (to simplify results).

4 Examples of Unwinding Numbers

This section gives certain examples of the use of unwinding numbers. We should emphasise our view that an ideal computer algebra system should do this manipulation for the user: certainly inserting the unwinding numbers where necessary, and preferably also removing/simplifying them where it can.

4.1 Forms of arccos

The following example is taken from [5], showing that two alternative definitions of arccos are in fact equal:

Theorem 1

$$\frac{2}{i}\ln\left(\sqrt{\frac{1+z}{2}}+i\sqrt{\frac{1-z}{2}}\right)=-i\ln\left(z+i\sqrt{1-z^2}\right). \tag{13}$$

First we prove the correct (and therefore containing unwinding numbers) version of $\sqrt{z_1 z_2}\overset{?}{=}\sqrt{z_1}\sqrt{z_2}$.

Lemma 1

$$\sqrt{z_1 z_2}=\sqrt{z_1}\sqrt{z_2}(-1)^{\mathcal{K}(\ln(z_1)+\ln(z_2))}. \tag{14}$$

[5] Just to remove the terms with unwinding numbers, as is done implicitly in some software systems, could be called "over-simplification."

Proof.

$$\begin{aligned}
\sqrt{z_1 z_2} &= \exp\left(\frac{1}{2}\left(\ln(z_1 z_2)\right)\right) \\
&= \exp\left(\frac{1}{2}\left(\ln(z_1) + \ln(z_2) - 2\pi i \mathcal{K}(\ln(z_1) + \ln(z_2))\right)\right) \\
&= \sqrt{z_1}\sqrt{z_2}\exp\left(-\pi i \mathcal{K}(\ln(z_1) + \ln(z_2))\right) \\
&= \sqrt{z_1}\sqrt{z_2}(-1)^{\mathcal{K}(\ln(z_1)+\ln(z_2))}
\end{aligned}$$

Lemma 2 *Whatever the value of z,*

$$\sqrt{1-z}\sqrt{1+z} = \sqrt{1-z^2}.$$

This is a classic example of a result that is "obvious" — the naïf just squares both sides, but in fact that loses information, and the identity requires proof. To show this, consider the apparently similar "result":

$$\sqrt{z-1}\sqrt{1+z} \stackrel{?}{=} \sqrt{z^2-1}.$$

If we take $z = -2$, the left-hand side becomes $\sqrt{-3}\sqrt{-1}$: the inputs to the square roots[6] have $\arg = +\pi$, so the square roots themselves have $\arg = \frac{\pi}{2}$, and the product has $\arg = +\pi$, and therefore is $-\sqrt{3}$. However, the right-hand side is $\sqrt{4-1} = \sqrt{3}$. The true formula is

$$\sqrt{z^2-1} = \operatorname{csgn}(z)\sqrt{z-1}\sqrt{z+1}$$

where [7]

$$\operatorname{csgn}(z) = (-1)^{\mathcal{K}(2\ln(z))} = \begin{cases} +1 & \Re(z) > 0 \text{ or } \Re(z) = 0;\ \Im(z) \geq 0 \\ -1 & \Re(z) < 0 \text{ or } \Re(z) = 0;\ \Im(z) < 0 \end{cases}.$$

Proof of lemma. It is sufficient to show that the unwinding number term in lemma 1 is zero. Whatever the value of z, $1+z$ and $1-z$ have imaginary parts of opposite signs. Without loss of generality, assume $\Im z \geq 0$. Then $0 \leq \arg(1+z) \leq \pi$ and $-\pi < \arg(1-z) \leq 0$. Therefore their sum, which is the imaginary part of $\ln(1+z) + \ln(1-z)$, is in $(-\pi, \pi]$. Hence the unwinding number is indeed zero.

Proof of Theorem 1. Now

$$\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right)^2 = z + i\sqrt{1-z}\sqrt{1+z} = z + i\sqrt{1-z^2}$$

by the previous lemma. Also $2\ln a = \ln(a^2)$ if $\mathcal{K}(2\ln a) = 0$, so we need only show this last stipulation i.e. that

$$-\frac{\pi}{2} < \arg\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right) \leq \frac{\pi}{2}.$$

[6]One is tempted to say "arguments of the square root", but this is easily confused with the function arg; we use 'inputs' instead.

[7]Maple defines $\operatorname{csgn}(0) = 0$, but we are assuming $\operatorname{csgn}(0) = 1$.

This is trivially true at $z = 0$. If it is false at any point, say z_0, then a path from z_0 to 0 must pass through a z where $\left|\arg\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right)\right| = \frac{\pi}{2}$, i.e. $\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}} = it$ for $t \in \mathbb{R}$. Squaring because, first, arg is continuous for $|z| \leq \pi/2$, and indeed for $|z| < \pi$, and, second, that the inputs to arg are themselves discontinuous only on $z > 1$ and $z < -1$, and on these half-lines, the arguments in question are 0 and $\pi/2$, which are acceptable. Coming back to the continuity along the path, we find that by squaring both sides, $z + i\sqrt{1-z^2} = -t^2$, i.e. $(z+t^2)^2 = -(1-z^2)$. Hence $2zt^2 + t^4 = -1$, so $z = \frac{-(1+t^4)}{2t^2} \leq -1$, and in particular is real. On this half-line, the argument in question is $+\pi/2$, which is acceptable. Hence the argument never leaves the desired range, and the theorem is proved.

4.2 arccos and arccosh

$\cos(z) = \cosh(iz)$, so we can ask whether the corresponding relation for the inverse functions, $\operatorname{arccosh}(z) = i\arccos(z)$ holds. This is known in [5] as the "couthness" of the arccos/arccosh definitions. The problem reduces, using equations (22) and (28), to

$$2\ln\left(\sqrt{\frac{z+1}{2}} + \sqrt{\frac{z-1}{2}}\right) \stackrel{?}{=} i\left(\frac{2}{i}\ln\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right)\right),$$

i.e.

$$\ln\left(\sqrt{\frac{z+1}{2}} + \sqrt{\frac{z-1}{2}}\right) \stackrel{?}{=} \ln\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right).$$

Since $\ln(a) = \ln(b)$ implies $a = b$, this reduces to

$$\sqrt{\frac{z-1}{2}} \stackrel{?}{=} i\sqrt{\frac{1-z}{2}} = \sqrt{-1}\sqrt{\frac{1-z}{2}}.$$

By lemma 1, the right-hand side reduces to $\sqrt{\frac{z-1}{2}}(-1)^{\mathcal{K}(\ln(-1)+\ln(\frac{z-1}{2}))}$. Hence the two are equal if, and only if, the unwinding number is even (and therefore zero). This will happen if, and only if, $\arg\left(\frac{z-1}{2}\right) \leq 0$, i.e. $\Im z < 0$ or $\Im z = 0$ and $z > 1$.

4.3 arcsin and arctan

The aim of this section is to prove that

$$\arcsin z = \arctan\frac{z}{\sqrt{1-z^2}} + \pi\mathcal{K}(-\ln(1+z)) - \pi\mathcal{K}(-\ln(1-z)). \qquad (15)$$

We start from equations (21) and (23). Then

$$2i\ \arctan\frac{z}{\sqrt{1-z^2}} \quad = \quad \ln\left(1 + i\frac{z}{\sqrt{1-z^2}}\right) - \ln\left(1 - i\frac{z}{\sqrt{1-z^2}}\right)$$

$$
\begin{aligned}
&= \ln\left([1+i\frac{z}{\sqrt{1-z^2}}]/[1-i\frac{z}{\sqrt{1-z^2}}]\right) \\
&\quad +2\pi i\mathcal{K}\left(\ln(1+i\frac{z}{\sqrt{1-z^2}})-\ln(1-i\frac{z}{\sqrt{1-z^2}})\right) \\
&= \ln[iz+\sqrt{1-z^2}]^2 \\
&\quad +2\pi i\mathcal{K}(\ln(1+i\frac{z}{\sqrt{1-z^2}})-\ln(1-i\frac{z}{\sqrt{1-z^2}})) \\
&= 2i\arcsin(z) \\
&\quad -2\pi i\mathcal{K}\left(2\ln(iz+\sqrt{1-z^2})\right) \\
&\quad +2\pi i\mathcal{K}\left(\ln(1+i\frac{z}{\sqrt{1-z^2}})-\ln(1-i\frac{z}{\sqrt{1-z^2}})\right)
\end{aligned}
$$

The tendency for $\mathcal{K}$ factors to proliferate is clear. To simplify we proceed as follows. Consider first the term

$$\mathcal{K}(2\ln(iz+\sqrt{1-z^2})) \ .$$

For $|z| < 1$, the real part of the input to the logarithm is positive and hence $\mathcal{K} = 0$. For $|z| > 1$, we solve for the critical case in which the input to $\mathcal{K}$ is $-i\pi$ and find only $z = r\exp(i\pi)$, with $r > 1$. Therefore

$$\mathcal{K}(2\ln(iz+\sqrt{1-z^2})) = \mathcal{K}(-\ln(1+z)) \ .$$

Repeating the procedure with

$$\mathcal{K}(\ln(1+iz/\sqrt{1-z^2})-\ln(1-iz/\sqrt{1-z^2}))$$

shows that $\mathcal{K} \neq 0$ only for $z > 1$. Therefore

$$\mathcal{K}(\ln(1+iz/\sqrt{1-z^2})-\ln(1-iz/\sqrt{1-z^2})) = \mathcal{K}(-\ln(1-z))$$

and so finally we get

$$\arctan\frac{z}{\sqrt{1-z^2}} = \arcsin(z) - \pi\mathcal{K}(-\ln(1+z)) + \pi\mathcal{K}(-\ln(1-z)) \ , \tag{16}$$

and this cannot be simplified further as an unconditional formula. Using equation (11), we can write

$$-\pi\mathcal{K}(-\ln(1+z)) + \pi\mathcal{K}(-\ln(1-z)) = \begin{cases} +\pi & \text{if } z < -1 \\ -\pi & \text{if } z > 1 \\ 0 & \text{elsewhere on } \mathbb{C}. \end{cases}$$

5 Unwinding Number: Insertion Rules

Unwinding numbers are normally inserted by use of equation (9) and its converse:

$$\log\left(\frac{z_1}{z_2}\right) = \log(z_1) - \log(z_2) - 2\pi\mathcal{K}\left(\log(z_1) - \log(z_2)\right). \tag{17}$$

Equation (10) may also be used, as may its close relative (also a special case of (17))

$$\log\left(\frac{1}{z}\right) = -\log(z) - 2\pi\mathcal{K}\left(-\log(z)\right). \tag{18}$$

In practice, results such as lemma 1 would also be built in to a simplifier.

The definition of $\mathcal{K}$ gives us

$$\log(\exp(z)) = z - 2\pi i K(z), \tag{19}$$

which is another mechanism for inserting unwinding numbers while "simplifying". The formulae for other inverse functions are given in appendix 2 of [4].

Many other "identities" among inverse functions require unwinding numbers. For example,

$$\arctan x + \arctan y = \arctan\left(\frac{x+y}{1-xy}\right) + \pi\mathcal{K}\left(2i(\arctan x + \arctan y)\right).$$

6 The Unwinding Number: Removal

It is clearly easier to insert unwinding numbers than to remove them. There are various possibilities for the values of unwinding numbers.

- An unwinding number may be identically zero. This is the case in lemma 2 and theorem 1. The aim is then to prove this.
- An unwinding number may be zero everywhere except on certain branch cuts in the complex plane. This is the case in equation (10), and its relative $\log\frac{1}{z} = -\log z 2\pi i\mathcal{K}(-\log(z))$. A less trivial case of this can be seen in equation (15). Derive has a different definition of arctan to eliminate this, so that, for Derive, $\arcsin(z) = \underbrace{\arctan}_{\text{Derive}} \frac{z}{\sqrt{1-z^2}}$. This definition can be related to ours either via unwinding numbers or via $\underbrace{\arctan}_{\text{Derive}}(z) = \overline{\arctan\overline{z}}$.

  It is often possible to disguise this sort of unwinding number, which is often of the form $\mathcal{K}(-\ln(\ldots))$ or $\mathcal{K}(\overline{\ln z})$ by resorting to such a "double conjugate" expression, though as yet we have no algorithm for this. Equally, we have no algorithm as yet for the sort of simplification we see in section 15.
- An unwinding number may divide the complex plane into two regions, one where it is non-zero and one where it is zero. A typical case of this is given in section 4.2. Here the proof methodology consists in examining the critical case, i.e. when the input to $\mathcal{K}$ has imaginary part $\pm\pi$, and examining when the functions contained in the input to $\mathcal{K}$ themselves have discontinuities.
- An unwinding number may correspond to the usual $+n\pi$: $n \in \mathbb{Z}$ of many trigonometric identities: examples of this are given in appendix 2 of [4] — see also [15].

7 More Automation

Here we look at a potential approach to mechanizing the reasoning required. We reconsider theorem 1: in particular we look at the branch cuts of the left-hand and right-hand sides, in the complex plane, which we will regard as the (x, y) plane.

LHS The question is the branch cut of $\ln\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right)$. Hence we have to solve

$$\Im\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right) = \pi.$$

In terms of (x, y)-coordinates, and cancelling the $\sqrt{2}$ in the denominators for simplicity[8] this means that

$$\sqrt{1+x+iy} + i\sqrt{1-x-iy} = \hat{x} + 0i : \qquad \hat{x} \in (-\infty, 0).$$

Isolating the first square root[9],

$$\begin{aligned} 1+x+iy &= \left(\hat{x} - i\sqrt{1-x-iy}\right)^2 \\ &= \hat{x}^2 - 2i\hat{x}\sqrt{1-x-iy} - (1-x-iy). \end{aligned}$$

Hence

$$(2-\hat{x}^2)^2 = 4 - 4\hat{x}^2 + \hat{x}^4 = -4\hat{x}^2(1-x-iy).$$

Equating imaginary parts, $0 = 4\hat{x}^2 y$, so $y = 0$. Equating real parts, $\frac{4+\hat{x}^4}{4\hat{x}^2} = x$, so $x \in [\sqrt{2}, \infty)$.

Hence the branch cut for the logarithm term is $(\sqrt{2} \leq x < \infty, y = 0)$. We also need to add the branch cuts for the individual square roots, which are $(x < -1, y = 0)$ and $(x > 1, y = 0)$.

RHS Reasoning as above, we reach

$$x + iy + i\sqrt{1-(x+iy)^2} = \hat{x} + 0i.$$

Hence

$$\begin{aligned} (x - \hat{x} + iy)^2 &= -(1-(x+iy)^2) \\ x^2 - 2x\hat{x} + 2ixy + \hat{x}^2 - 2i\hat{x}y - y^2 &= -1 + x^2 + 2ixy - y^2. \end{aligned}$$

[8] A mechanized approach would probably not do this, and so we will not bother to justify it.

[9] The reader may complain that we are naïvely squaring an equation, something we clearly, from the examples given previously, should not be doing. If we were aiming for a precise description of the branch cuts, the reader would be correct. We may introduce spurious branch cuts, but the process cannot *lose* a solution.

Table 2: (20) at $z = 1 + i$

$\sqrt{\frac{1+z}{2}}$	$\sqrt{\frac{1-z}{2}}$	$\sqrt{\frac{1-z^2}{2}}$	value
+	+	+	0
−	+	+	$0.73 - 1.13i$
+	−	+	$0.73 - 1.13i$
−	−	+	0
+	+	−	$-0.73 + 1.13i$
−	+	−	0
+	−	−	0
−	−	−	$-0.73 + 1.13i$

Equating imaginary parts, $-2\hat{x}y = 0$, so $y = 0$. Using this, the real parts simplify to

$$-2x\hat{x} + \hat{x}^2 = -1,$$

i.e. $x = \frac{1+\hat{x}^2}{2\hat{x}}$ or $x \in (-\infty, -1]$.

We need also need to consider the branch cuts of the square root, viz. $1 - (x + iy)^2 = \hat{x} + 0i$ with $\hat{x} \in (-\infty, 0)$. The imaginary part of this is $2xy = 0$, so either $x = 0$ or $y = 0$. Substituting this in gives two critical lines: $(y = 0, x^2 = 1 - \hat{x})$ and $(x = 0, y^2 = \hat{x} - 1)$. given the range of $\hat{x}$, the latter has no solution, while the latter is $(y = 0, |x| > 1)$.

Lemma 3 *The difference between the left and right hand sides is locally constant, i.e. its derivative is zero.*

The derivative of the difference is

$$2\frac{\frac{1}{\sqrt{2+2\,z}} - \frac{i}{\sqrt{2-2\,z}}}{\sqrt{2+2\,z} + i\sqrt{2-2\,z}} - \frac{1 - \frac{iz}{\sqrt{1-z^2}}}{z + i\sqrt{1-z^2}}. \tag{20}$$

In one sense, this is "obviously" zero, but this is precisely the problem we are trying to solve. However, this or a variant with alternative signs for some of the square roots must be zero, since it does simplify "symbolically" to zero. Furthermore, the branch cuts of expression (20) are those of the square roots involved, which we have already computed as $(x < -1, y = 0)$ and $(x > 1, y = 0)$. In this case, the complex plane is divided into these two half-lines, and their complement, i.e. one two-dimensional region and two one-dimensional ones. On each of these, the expression (or one of its variants) is identically zero, so we need only find a point at which the expression is zero, and its relatives are not. The evaluation can, in this case, be done exactly, since we have exact algebraic numbers — see [11]. Unfortunately, in this case, four of the eight possibilities evaluate to 0 at $z = 1 + i$, as in the following table. Possibility 4 $(-, -, +)$ is in fact exactly equal to possibility 1, since effectively we multiply numerator and denominator by -1. Unfortunately, to deduce that possibilities 6 and 7 are equal to possibility 1, we need to use Lemma 2. Once we do this, it is then clear that possibility 1 is zero, and the remaining ones non-zero, at the sample point, and hence in the two-dimensional region. In fact, the same behaviour also takes place at sample points for the two one-dimensional cells, e.g. $z = 2$ and $z = -2$.

Proof (2) of Theorem 1 continued. Since the difference is locally constant, we need merely evaluate at sample points for the decomposition. There is only one two-dimensional component, viz. $\mathbb{C} \setminus \{\text{branch cuts}\}$. A possible sample point would be $z = 0$, when the equation reduces to

$$2 \ln \left(\frac{1+i}{\sqrt{2}} \right) = \frac{i\pi}{2},$$

which is obviously true.

The various branch cuts from the analysis of the left and right hand sides are all on the real line, viz.

$$(-\infty, -1), \quad (-\infty, -1], \quad (1, \infty), \quad (\sqrt{2}, \infty).$$

Note that the zero-dimensional point $z = -1$ is the difference of two of these intervals, so must be considered as a special case. Suitable sample points would be -2, -1, 1.25 and 2.

In theory, there is a problem of precision, but since we are asking for the equality of two logarithms (up to a factor of 2), we can consider their inputs instead. For this to work, we must be able to apply $2 \ln w = \ln(w^2)$ to the left-hand side, i.e. w must satisfy $\frac{-\pi}{2} < \Im w \leq \frac{\pi}{2}$. The evaluation of this input at the four sample points are $\frac{i}{2}(\sqrt{2} + \sqrt{6})$, i, $\frac{1}{2}\sqrt{2}$ and $\frac{1}{2}(\sqrt{6} - \sqrt{2})$, all of which satisfy these criteria.

8 Conclusion

The thesis of this paper is that, of the four methods outlined in section 2, unwinding numbers are the only one amenable to computer algebra. Unwinding number insertion allows combination of logarithms, square roots etc., as well as canceling functions and their inverses, while retaining mathematical correctness. This can be done completely algorithmically, and we claim this is one way, the only way we have seen, of guaranteeing mathematical correctness while "simplifying".

Unwinding number removal, where it is possible, then simplifies these results to the expected form. This is not a process that can currently be done algorithmically, but it is much better suited to current artificial intelligence techniques than the general problems of complex analysis.

When the unwinding numbers cannot be eliminated, they can often be converted into a case analysis that, while not ideal, is at least comprehensible while being mathematically correct.

We have also sketched an approach via decomposing the complex plane, though this is some way from being an algorithm.

More generally, we have reduced the analytic difficulties of simplifying these functions to more algebraic ones, in areas where we hope that artificial intelligence and theorem proving stand a better chance of contributing to the problem.

References

[1] Abramowitz,M. & Stegun,I., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. US Government Printing Office, 1964. 10th Printing December 1972.

[2] Bradford,R.J., Algebraic Simplification of Multiple-Valued Functions. Proc. DISCO '92 (Springer Lecture Notes in Computer Science 721, ed. J.P. Fitch), Springer, 1993, pp. 13–21.

[3] Carathéodory,C., *Theory of functions of a complex variable* (trans. F. Steinhardt), 2nd. ed., Chelsea Publ., New York, 1958.

[4] Corless,R.M., Davenport,J.H., Jeffrey,D.J., Litt,G. & Watt,S.M., Reasoning about the Elementary Functions of Complex Analysis. Artificial Intelligence and Symbolic Computation (ed. John A. Campbell & Eugenio Roanes-Lozano), Springer Lecture Notes in Artificial Intelligence Vol. 1930, Springer-Verlag 2001, pp. 115–126.

[5] Corless,R.M., Davenport,J.H., Jeffrey,D.J. & Watt,S.M., "According to Abramowitz and Stegun". SIGSAM Bulletin 34(2000) 2, pp. 58–65.

[6] Corless,R.M. & Jeffrey,D.J., The Unwinding Number. SIGSAM Bulletin **30** (1996) 2, issue 116, pp. 28–35.

[7] Corless,R.M. & Jeffrey,D.J., "Graphing Elementary Riemann Surfaces", SIGSAM Bulletin **32** (1998) 1, issue 123, pp. 11–17.

[8] Davenport,J.H. & Fischer,H.-C., Manipulation of Expressions. *Improving Floating-Point Programming* (ed. P.J.L. Wallis), Wiley, 1990, pp. 149–167.

[9] Encyclopedia Britannica, 15th. edition. Encyclopedia Britannica Inc., Chicago etc., 15th ed., 1995 printing.

[10] Gianni,P., Seppälä,M., Silhol,R. & Trager,B.M., Riemann surfaces, plane algebraic curves and their period matrices. *J. Symbolic Comp.* **26** (1998) pp. 789–803.

[11] Hur,N. & Davenport,J.H., An Exact Real Arithmetic with Equality Determination. Proc. ISSAC 2000 (ed. C. Traverso), ACM, New York, 2000, pp. 169–174.

[12] IEEE Standard Pascal Computer Programming Language. IEEE Inc., 1983.

[13] IEEE Standard 754 for Binary Floating-Point Arithmetic. IEEE Inc., 1985.

[14] Kahan,W., Branch Cuts for Complex Elementary Functions. *The State of Art in Numerical Analysis* (ed. A. Iserles & M.J.D. Powell), Clarendon Press, Oxford, 1987, pp. 165–211.

[15] Litt,G., Unwinding numbers for the Logarithmic, Inverse Trigonometric and Inverse Hyperbolic Functions. M.Sc. project, Department of Applied Mathematics, University of Western Ontario, December 1999.

[16] Risch,R.H., Algebraic Properties of the Elementary Functions of Analysis. *Amer. J. Math.* **101** (1979) pp. 743–759.

[17] Steele,G.L.,Jr., *Common LISP: The Language*, 2nd. edition. Digital Press, 1990.

[18] Trott,M., Visualization of Riemann Surfaces of Algebraic functions. *Mathematica in Education and Research* **6** (1997), no. 4, pp. 15–36,

A Definition of the Elementary Inverse Functions

These definitions are taken from [5]. They agree with [1, ninth printing], but are more precise on the branch cuts, and agree with Maple with the exception of arccot, for the reasons explained in [5].

$$\arcsin z = -i \ln\left(\sqrt{1-z^2} + iz\right). \tag{21}$$

$$\arccos(z) = \frac{\pi}{2} - \arcsin(z) = \frac{2}{i} \ln\left(\sqrt{\frac{1+z}{2}} + i\sqrt{\frac{1-z}{2}}\right). \tag{22}$$

$$\arctan(z) = \frac{1}{2i}\left(\ln(1+iz) - \ln(1-iz)\right). \tag{23}$$

$$\operatorname{arccot} z = \frac{1}{2i} \ln\left(\frac{z+i}{z-i}\right) = \arctan\left(\frac{1}{z}\right). \tag{24}$$

$$\operatorname{arcsec}(z) = \arccos(1/z) = -i\ln(1/z + i\sqrt{1-1/z^2}), \tag{25}$$

with $\operatorname{arcsec}(0) = \frac{\pi}{2}$.

$$\operatorname{arccsc}(z) = \arcsin(1/z) = -i\ln(i/z + \sqrt{1-1/z^2}), \tag{26}$$

with $\operatorname{arccsc}(0) = 0$.

$$\operatorname{arcsinh}(z) = \ln\left(z + \sqrt{1+z^2}\right). \tag{27}$$

$$\operatorname{arccosh}(z) = 2\ln\left(\sqrt{\frac{z+1}{2}} + \sqrt{\frac{z-1}{2}}\right). \tag{28}$$

$$\operatorname{arctanh}(z) = \frac{1}{2}\left(\ln(1+z) - \ln(1-z)\right). \tag{29}$$

$$\operatorname{arccoth}(z) = \frac{1}{2}\left(\ln(-1-z) - \ln(1-z)\right). \tag{30}$$

$$\operatorname{arcsech}(z) = 2\ln\left(\sqrt{\frac{z+1}{2z}} + \sqrt{\frac{1-z}{2z}}\right). \tag{31}$$

$$\operatorname{arccsch}(z) = \ln\left(\frac{1}{z} + \sqrt{1+\left(\frac{1}{z}\right)^{2}}\right), \tag{32}$$