# MKM from book to computer: a case study

James H. Davenport*

Department of Computer Science, University of Bath, Bath BA2 7AY, England
J.H.Davenport@bath.ac.uk

**Abstract.** [2] is one of the great mathematical knowledge repositories. Nevertheless, it was written for a different era, and for human readership. In this paper, we describe the sorts of knowledge in one chapter (elementary transcendental functions) and the difficulties in making this sort of knowledge formal. This makes us ask questions about the nature of a Mathematical Knowledge Repository, and whether a database is enough, or whether more "intelligence" is required.

## 1 Introduction

It is a widely-held belief, though probably more among computer scientists and philosophers than among mathematicians themselves, that mathematics is a completely formal subject with its own, totally precise, language. Mathematicians know that what they write is in a "mathematical vernacular" [7], which could, *in principle* be rendered utterly formal, though very few mathematicians do so, or even see the point of doing so. In practice the mathematical vernacular is intended for communicating between human beings, or more precisely, mathematicians, or, more precisely still, mathematicians in that subject. The reader is meant to apply that nebulous quality of "common sense" when reading the mathematical vernacular.

It turns out to be remarkably hard to write "correct" mathematics in the mathematical vernacular. The problem is often with "obvious" special cases that are not stated explicitly[1], but which the knowledgeable reader will (and must) infer. There are even errors in [2] of this sort — see [11] and equation (23). The ninth printing of [2] contained corrections on 132 pages, and the Dover reprint of that corrected a further nine pages.

In this paper, we will explore the problems of representing a small part of one of the most well-known sources of mathematical knowledge: [2]. In particular, we consider the problems of translating the relevant content of chapter 4 — *Elementary Transcendental Functions* — into OpenMath [1, 13]. In this paper we

* The author was partially supported by the European OpenMath Thematic Network and the Mathematical Knowledge Management Network.

[1] The author recently had a problem with this, having set an examination question "if $\alpha$ is algebraic and $\beta$ is transcendental, is $\alpha\beta$ always transcendental"? One student answered in the negative, quoting the case of $\alpha = 0$.

will be concerned with the semantic problems, rather than with the OpenMath problems. It should be noted that there is concern within the computer algebra community about the treatment of these functions in computer algebra [3, 17].

It should be emphasised that this paper is in no way a criticism of [2]. The author is one of, probably literally, millions of people who have benefited from this enormous compendium of knowledge. Rather, the point is to illustrate that a book produced for other human beings to read, in days before the advent of (general-purpose) computer algebra systems or theorem provers, implicitly assumes knowledge in the reader that it is notoriously difficult to imbue such systems with. We therefore ask how we can make such knowledge explicit.

## 2 The "Elementary Transcendental Functions"

These days[2] these functions are normally considered to be exp and its inverse ln, the six trigonometric functions and their inverses, and the six hyperbolic functions and their inverses. For the purposes of this paper, we will class exp, the six trigonometric functions and the six hyperbolic functions together as the *forward functions*, and the remainder as the *inverse functions*.

The forward functions present comparatively little difficulty. They are continuous, arbitrarily-differentiable, many–to–one functions defined from $\mathbf{C}$ (possibly less a countable number of singularities) to $\mathbf{C}$. While it is possible to extend them to run from from the whole of $\mathbf{C}$ to $\mathbf{C} \cup \{\infty\}$, [2] sensibly chooses not to. The concept of $\infty$ is a difficult one to formalise (but see [4]), and, while $\mathbf{R} \subset \mathbf{C}$, it is not the case for their natural completions: $\mathbf{R} \cup \{-\infty, +\infty\} \not\subset \mathbf{C} \cup \{\infty\}$.

The problem lies rather with the inverse functions. They are continuous, arbitrarily-differentiable, one–to–many functions defined from $\mathbf{C}$ (possibly less a countable number of singularities) to an appropriate Riemann surface. The problem comes when we wish to consider them as functions from $\mathbf{C}$ (possibly less a countable number of singularities) to $\mathbf{C}$. The solution is to introduce "branch cuts", i.e. curves (though in practice we will only be considering lines in this paper) in $\mathbf{C}$ across which the inverse function is not continuous.

Provided that they satisfy appropriate mathematical conditions, any line or curve can be chosen as the branch cut. For example, ln, as one makes a complete counter-clockwise circle round the origin, increases in value by $2\pi i$. Therefore any simple curve from the origin to infinity will serve as a branch cut. The normal choice today[3], as in [2], is to choose the negative real axis.

It is also important to specify what the value of the function is on the branch cut. It clearly makes sense to have it continuous with one side or the other, and the common choice, as in [2], is to choose the value of ln on the branch cut to be continuous with the upper half-plane, so that $-\pi < \Im \ln z \leq \pi$. However,

[2] Other trigonometric variants such as versine have disappeared. However, see section 5.

[3] Though the author was taught at school to use the positive real axis, with $0 \leq \Im \ln z < 2\pi$.

this choice is essentially arbitrary, and [16] would like to make the function two-valued on the branch cut: $\ln(-1) = \pm\pi i$. This has the drawback of not fitting readily with numerical evaluation.

One still might wish to "have one's cake and eat it". [15] points out that the concept of a "signed zero"[4] [14] (for clarity, we write the positive zero as $0^+$ and the negative one as $0^-$) can be used to solve this dilemma, if we say that, for $x < 0$, $\ln(x + 0^+ i) = \ln|x| + \pi i$ whereas $\ln(x + 0^- i) = \ln|x| - \pi i$. However, this is no use to computer algebra systems, and little use to theorem provers.

The serious problem with branch cuts is that they make many "obvious" relations false. For example, exp takes complex conjugates to complex conjugates, as $\exp\overline{z} = \overline{\exp z}$, so one might expect the same, i.e.

$$\log\overline{z} \stackrel{?}{=} \overline{\log z}, \tag{1}$$

to be true of its inverse. Unfortunately, this is true everywhere except on the branch cut, where $z = \overline{z}$, and therefore $\log\overline{z} = \log z$. These complications mean that it is not a simple matter to codify knowledge about the inverse functions.

### 2.1 Encoding branch cut information

[10] points out that most 'equalities' do not hold for the complex logarithm, e.g. $\ln(z^2) \neq 2\ln z$ (try $z = -1$), and its generalisation

$$\ln(z_1 z_2) \neq \ln z_1 + \ln z_2. \tag{2}$$

The most fundamental of all non-equalities is $z = \ln\exp z$, with an obvious violation at $z = 2\pi i$. They therefore propose to introduce the *unwinding number* $\mathcal{K}$, defined[5] by

$$\mathcal{K}(z) = \frac{z - \ln\exp z}{2\pi i} = \left\lceil \frac{\Im z - \pi}{2\pi} \right\rceil \in \mathbf{Z} \tag{3}$$

We can then rescue equation (1) as

$$\ln\overline{z} = \overline{\ln z} - 2\pi i\mathcal{K}(\overline{\ln z}). \tag{4}$$

Since we know that $-\pi < \Im\ln z \leq \pi$, $-\pi \leq \Im\overline{\ln z} < \pi$. So the only places where the $\mathcal{K}$ term is non-zero is when $\Im\overline{\ln z} = -\pi$, i.e. $\Im\ln z = \pi$. Hence this equation implicitly encodes the region of invalidity of equation (1).

[4] One could ask why zero should be special and have two values (or four in the Cartesian complex plane). The answer is that all the branch cuts for the basic elementary functions (this is not true for, e.g. $\ln(i + \ln z)$, whose branch cut is $z \in \{e^t(\cos 1 + i\sin 1) \mid t \in (\infty, 0]\}$ 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. With sufficient care, this technique can be used for other branch cuts as long as they are parallel with the axes, e.g. $\ln(z + i)$.

[5] Note that the sign convention here is the opposite to that of [10], which defined $\mathcal{K}(z)$ as $\lfloor\frac{\pi - \Im z}{2\pi}\rfloor$: the authors of [10] recanted later to keep the number of $-1$s occurring in formulae to a minimum.

## 3 Codifying ln

[2, p. 67] gives the branch cut $(-\infty, 0]$, and the rule [2, (4.1.2)] that

$$-\pi < \Im \ln z \leq \pi. \tag{5}$$

OpenMath has chosen to adopt equation (5) as the definition of the branch cut, rather than words, since it also conveys the necessary information about the value on the branch cut, which the form of words does not. From equation (5), one can deduce that the branch cut is $\{z \mid \Im \ln z = \pi\}$, which should be the same as $\{z \mid \Im \ln z = -\pi\}$. However, it takes a certain subtlety to convert this to $z \in (-\infty, 0]$, and maybe the branch cut should be stated explicitly, either instead of equation (5) (but then how does one specify the value on the branch cut?) or as well as it (in which case, how does one ensure coherence between the two?). However, despite the discussion in the previous section, precisely what formal semantics can one give to the phrase "branch cut"? Does it depend on one's semantic model for $\mathbf{C}$ and functions $\mathbf{C} \to \mathbf{C}$?

Currently, OpenMath does not encode equations such as equation (1) (since they are false). There are various options.

1. Encode them with unwinding numbers, as in equation (4).
2. Encode them as conditional equations, e.g.

$$z \notin \text{branch cut} \Rightarrow \log \overline{z} = \overline{\log z}, \tag{6}$$

3. Encode them via multivalued functions (see section 6)

The unwinding number approach is attractive, and it could be used in the "unwinding number approach" to simplification [8]. However, it would be useless to a system that did not support the semantics of unwinding numbers, though an "intelligent" database might be able to convert such an encoding into the conditional one. The conditional equation approach might be helpful to theorem provers, but the proof obligations that would build up might be unmanageable. In this form, it does not say what happens when $z$ is on the branch cut, but an "else clause" could be added.

To state them in the "unwinding number" formalism, the following equations seem to be a suitable "knowledge base" for ln, in addition to equation (4).

$$\ln(z_1 z_2) = \ln z_1 + \ln z_2 - 2\pi i \mathcal{K}(\ln z_1 + \ln z_2). \tag{7}$$

$$\ln(z_1/z_2) = \ln z_1 - \ln z_2 - 2\pi i \mathcal{K}(\ln z_1 - \ln z_2). \tag{8}$$

The following is a re-writing of equation (3):

$$\ln \exp z = z - 2\pi i \mathcal{K}(z), \tag{9}$$

and we always have

$$\exp \ln z = z. \tag{10}$$

It is harder to write equations (7) and (8) in a "conditional" formalism, since what matters is not so much being on the branch cut as having crossed the branch cut. A direct formalism would be

$$(-\pi < \Im(\ln z_1 + \ln z_2)) \wedge (\Im(\ln z_1 + \ln z_2) \leq \pi) \Rightarrow \ln(z_1 z_2) = \ln z_1 + \ln z_2,$$

but, unlike equation (6), there is an input space of measure 0.5 on which this does not define the answer. One is really forced to go to something like

$$\begin{aligned} -\pi < \Im(\ln z_1 + \ln z_2) \leq \pi &\Rightarrow \ln(z_1 z_2) = \ln z_1 + \ln z_2 \\ \Im(\ln z_1 + \ln z_2) > \pi &\Rightarrow \ln(z_1 z_2) = \ln z_1 + \ln z_2 - 2\pi i \\ \Im(\ln z_1 + \ln z_2) \leq -\pi &\Rightarrow \ln(z_1 z_2) = \ln z_1 + \ln z_2 + 2\pi i \end{aligned}$$

which is essentially equation (7) unwrapped.

### 3.1 Square roots

It is possible to define $\sqrt{z} = \exp\left(\frac{1}{2}\ln x\right)$. This means that $\sqrt{\ }$ inherits the branch cut of ln. Since this definition is possible, and causes no significant problems, Occam's Razor tells us to use it. Equation (7) then implies

$$\sqrt{z_1 z_2} = \sqrt{z_1}\sqrt{z_2}\,(-1)^{\mathcal{K}(\ln z_1 + \ln z_2)}, \tag{11}$$

and the same discussion about alternative forms of equation (7) applies here. It is also possible to use the complex sign[6] function to reduce this to

$$\sqrt{z_1 z_2} = \operatorname{csgn}\left(\sqrt{z_1}\sqrt{z_2}\right)\sqrt{z_1}\sqrt{z_2}. \tag{12}$$

## 4 Other inverse functions

All the other forward functions can be defined in terms of exp. Hence one might wish to define all the other inverse functions in terms of ln. This is in fact principle 2 of [9] (and very close to the "Principal Expression" rule of [15]).

> All these functions should be mathematically[7] defined in terms of ln, thus inheriting their branch cuts from the chosen branch cut for ln (equation 5).

[6] The csgn function was first defined in Maple. There is some uncertainty about csgn(0): is it 0 or 1, but for the reasons given in [6], we choose csgn(0) = 1.

[7] This does not imply that it is always right to compute them this way. There may be reasons of efficiency, numerical stability or plain economy (it is wasteful to compute a real arcsin in terms of complex logarithms and square roots) why a numerical, or even symbolic, implementation should be different, but the *semantics* should be those of this definition in terms of logarithms, possibly augmented by exceptional values when the logarithm formula is ill-defined.

In fact, it is not just the branch cut itself, but also the definition of the function on the branch cut, that follows from this principle, since we know the definition of ln on the branch cut.

[2] does not quite adhere to this principle. It does give definitions in terms of ln, but these are secondary to the main definitions, and, as in the case of [2, 4.4.26]

$$\operatorname{Arcsin} x = -i \operatorname{Ln}\left(\sqrt{1-x^2} + ix\right) \qquad |x^2| \le 1, \tag{13}$$

the range of applicability is limited. [15] suggested, and [9] followed, that equation (13) be adopted as the definition throughout $\mathbf{C}$. This has the consequence that

$$\arcsin(-z) = -\arcsin(z) \tag{14}$$

is valid throughout $\mathbf{C}$. No choice of values on the branch cut (compatible with $\sin\arcsin z = z$) can make $\overline{\arcsin(z)} = \arcsin(\overline{z})$ valid on the branch cut: it has to be rescued as

$$\begin{aligned}\overline{\arcsin z} = (-1)^{\mathcal{K}(-\ln(1-z^2))} \arcsin \overline{z} \\ + \pi\mathcal{K}(-\ln(1+z)) - \pi\mathcal{K}(-\ln(1-z)).\end{aligned} \tag{15}$$

Here we have a fairly complicated formula, and the conditional form

$$(z \notin \mathbf{R}) \vee (z^2 \le 1) \Rightarrow \overline{\arcsin z} = \arcsin \overline{z} \tag{16}$$

(which does not tell what happens on the branch cuts, but there $\overline{z} = z$) might be simpler.

For real variables, the addition rule for arctan can be written out conditionally [6]:

$$\begin{aligned}\arctan(z_1) + \arctan(z_2) = \arctan\left(\tfrac{z_1+z_2}{1-z_1z_2}\right) \\ + \begin{cases} \pi & z_1 > 0, z_1z_2 > 1 \\ 0 & z_1 \ge 0, z_1z_2 \le 1 \\ -\pi & z_1 < 0, z_1z_2 \ge 1 \end{cases}\end{aligned} \tag{17}$$

For both real and complex variables, there is a representation [8] in terms of unwinding numbers:

$$\begin{aligned}\arctan(z_1) + \arctan(z_2) = \arctan\left(\tfrac{z_1+z_2}{1-z_1z_2}\right) + \\ \pi\mathcal{K}\left(2i(\arctan(z_1) + \arctan(z_2))\right).\end{aligned} \tag{18}$$

It is also possible to write the law for addition of real arcsin of real arguments in a conditional form:

$$\sqrt{(1-z_1^2)(1-z_2^2)} - z_1z_2 \ge 0 \Rightarrow \arcsin(z_1) + \arcsin(z_2) = A \tag{19}$$

$$(\sqrt{(1-z_1^2)(1-z_2^2)} - z_1z_2 < 0) \wedge (z_1 > 0) \Rightarrow \arcsin(z_1) + \arcsin(z_2) = \pi - A$$

$$(\sqrt{(1-z_1^2)(1-z_2^2)} - z_1z_2 < 0) \wedge (z_1 < 0) \Rightarrow \arcsin(z_1) + \arcsin(z_2) = -\pi - A,$$

where $A = \arcsin\left(z_1\sqrt{1-z_2^2} + z_2\sqrt{1-z_1^2}\right)$, but we have yet to find[8] an unwinding number formalism in terms of arcsin — there clearly is one in terms of (complex) lns, which works out to be $\arcsin(z_1) + \arcsin(z_2) =$

$$-i\Bigg[\ln\Bigg(\qquad iz_1\sqrt{1-z_2^2} + iz_2\sqrt{1-z_1^2} + \\ (-1)^{K(c_2)}\sqrt{1-\left(z_1\sqrt{1-z_2^2} + z_2\sqrt{1-z_1^2}\right)^2}\Bigg) + 2\pi i\mathcal{K}(c_1)\Bigg],$$

where the correction terms are $c_1 = i(\arcsin(z_1) + \arcsin(z_2))$ and

$$c_2 = 2\ln\left(\sqrt{1-z_1^2}\sqrt{1-z_2^2} - z_1z_2\right).$$

When $\mathcal{K}(c_2) = 0$, the main ln is recognisably $\arcsin\left(z_1\sqrt{1-z_2^2} + z_2\sqrt{1-z_1^2}\right)$, as required, but otherwise it is $\pm\pi - \arcsin\left(z_1\sqrt{1-z_2^2} + z_2\sqrt{1-z_1^2}\right)$.

It is also possible to state correct relations between the inverse trigonometric functions, as in [9]:

$$\arcsin z = \arctan\frac{z}{\sqrt{1-z^2}} + \pi\mathcal{K}(-\ln(1+z)) - \pi\mathcal{K}(-\ln(1-z)). \tag{20}$$

No really new issues arise when looking at the other inverse trigonometric functions, or at the inverse hyperbolic functions.

## 5 The case for ATAN2

It is common to say, or at least believe, that, for real $x$ and $y$,

$$\arg(x+iy) = \arctan\left(\frac{y}{x}\right), \tag{21}$$

but a moment's consideration of ranges (a tool that we have found very valuable in this area) shows that it cannot be so: the left-hand side has a range of $(-\pi, \pi]$ with the standard branch cuts, and certainly has a range of size $2\pi$, whereas the right-hand side has a range of size $\pi$.

The fundamental problem is, of course, that considering $\frac{y}{x}$ immediately confuses $1+i$ with $-1-i$. This fact was well-known to the early designers of FORTRAN, who defined a two-argument function `ATAN2`, such that

$$\texttt{ATAN2}(y,x) = \arctan\left(\frac{y}{x}\right) \stackrel{?}{\pm}\pi. \tag{22}$$

More precisely, the correction factor is 0 when $x > 0$, $+\pi$ when $x < 0$ and $y \geq 0$, and $-\pi$ when $x, y < 0$. For completeness, one should also define what happens when $x = 0$, when the answer is $+\pi/2$ when $y > 0$ and $-\pi/2$ when $y < 0$.

[8] The situation with addition of arcsin is complicated: see the discussion around equation (37).

This has been added to OpenMath, as the symbol `arctan` in the `transc2` Content Dictionary. Use of this enables us to rescue the incorrect equation [2, 6.1.24] $\arg \Gamma(z+1) = \arg \Gamma(z) + \arctan \frac{y}{x}$ (where $x$ and $y$ are the real and imaginary parts of $z$) as

$$\arg \Gamma(z+1) \equiv \arg \Gamma(z) + \arctan(y,x) \pmod{2\pi}. \tag{23}$$

We should note the necessity to think in terms of congruences.

## 6 Multivalued functions

Mathematical texts often urge us (and we have found this idea useful in [6, 5]) to treat these functions as multivalued (which we will interpret as set-valued), defining, say, $\mathrm{Ln}(z) = \{y \mid \exp y = z\} = \{\mathrm{Ln}\, z + 2n\pi i \mid n \in \mathbf{Z}\}$ (therefore $\mathrm{Sqrt}(z) = \pm\sqrt{z}$) and $\mathrm{Arctan}(z) = \{y \mid \tan y = z\} = \{\arctan(z) + n\pi \mid n \in \mathbf{Z}\}$ (the notational convention of using capital letters for these set-valued functions seems helpful). It should be noted that Ln and Arctan are deceptively simple in this respect, and the true rules for the inverse trigonometric functions are [2, (4.4.10–12)]

$$\mathrm{Arcsin}(z) = \{(-1)^k \arcsin(z) + k\pi \mid k \in \mathbf{Z}\} \tag{24}$$
$$\mathrm{Arccos}(z) = \{\pm \arccos(z) + 2k\pi \mid k \in \mathbf{Z}\} \tag{25}$$
$$\mathrm{Arctan}(z) = \{\arctan(z) + k\pi \mid k \in \mathbf{Z}\} \tag{26}$$
$$\mathrm{Arccot}(z) = \{\mathrm{arccot}(z) + k\pi \mid k \in \mathbf{Z}\} \tag{27}$$
$$\mathrm{Arcsec}(z) = \{\pm \mathrm{arcsec}(z) + 2k\pi \mid k \in \mathbf{Z}\} \tag{28}$$
$$\mathrm{Arccsc}(z) = \{(-1)^k \mathrm{arccsc}(z) + k\pi \mid k \in \mathbf{Z}\} \tag{29}$$

where we have changed to our set-theoretic notation, and added the last three equations, which are clearly implied by the first three.

[2, (4.4.26–31)] give equivalent multivalued expressions in terms of Ln, as in table 1 (we have preserved their notation). To get the correct indeterminacy

**Table 1.** Multivalued functions in terms of Ln

| | | |
|---|---|---|
| (4.4.26) | $\mathrm{Arcsin}\ x = -i\,\mathrm{Ln}\left[(1-x^2)^{\frac{1}{2}} + ix\right]$ | $x^2 \le 1$ |
| (4.4.27) | $\mathrm{Arccos}\ x = -i\,\mathrm{Ln}\left[x + i(1-x^2)^{\frac{1}{2}}\right]$ | $x^2 \le 1$ |
| (4.4.28) | $\mathrm{Arctan}\ x = \frac{i}{2}\,\mathrm{Ln}\,\frac{1-ix}{1+ix} = \frac{i}{2}\,\mathrm{Ln}\,\frac{i+x}{i-x}$ | $x$ real |
| (4.4.29) | $\mathrm{Arccsc}\ x = -i\,\mathrm{Ln}\left[\frac{(x^2-1)^{\frac{1}{2}}+i}{x}\right]$ | $x^2 \ge 1$ |
| (4.4.30) | $\mathrm{Arcsec}\ x = -i\,\mathrm{Ln}\left[1 + i\frac{(x^2-1)^{\frac{1}{2}}}{x}\right]$ | $x^2 \ge 1$ |
| (4.4.31) | $\mathrm{Arccot}\ x = \frac{i}{2}\,\mathrm{Ln}\,\frac{ix+1}{ix-2} = \frac{i}{2}\,\mathrm{Ln}\,\frac{x-i}{x+1}$ | $x$ real |

from equation (24), it is in fact necessary to interpret $z^{\frac{1}{2}}$ as $\mathrm{Sqrt}(z)$ throughout

this table. The range restrictions are in fact unnecessary (as proved in [12]), and it is possible (and consistent with the decisions in the univariate case) to accept these as definitions.

One might think that the move to multivalued functions was a simplification. Indeed many statements that needed caveats (unwinding numbers, exceptional cases) before are now unconditionally true: we give a few examples below, where, for example, $\mathrm{Ln}(z_1) + \mathrm{Ln}(z_2)$ is to be interpreted as $\{x + y \mid x \in \mathrm{Ln}(z_1) \wedge y \in \mathrm{Ln}(z_2)\}$.

$$\begin{aligned} \mathrm{Sqrt}(z_1)\,\mathrm{Sqrt}(z_2) &= \mathrm{Sqrt}(z_1 z_2) \\ \mathrm{Ln}(z_1) + \mathrm{Ln}(z_2) &= \mathrm{Ln}(z_1 z_2) \\ \mathrm{Ln}(\overline{z}) &= \overline{\mathrm{Ln}\, z} \\ \mathrm{Arcsin}(\overline{z}) &= \overline{\mathrm{Arcsin}\, z}. \end{aligned}$$

However, all is not perfect. Equation (20), which needed caveats (but only on the branch cuts), now becomes the strict containment

$$\mathrm{Arcsin}\, z \subset \mathrm{Arctan}\, \frac{z}{\mathrm{Sqrt}(1 - z^2)}, \tag{30}$$

and the true identity is

$$\mathrm{Arcsin}\, z \cup \mathrm{Arcsin}(-z) = \mathrm{Arctan}\, \frac{z}{\mathrm{Sqrt}(1 - z^2)}. \tag{31}$$

Note that it is not true that $\mathrm{Arcsin}\, z = \mathrm{Arctan}\, \frac{z}{\sqrt{1-z^2}}$: the right=hand side has values alternately in $\mathrm{Arcsin}\, z$ and $\mathrm{Arcsin}(-z)$, and misses half the values in each.

### 6.1 Addition laws

[2] quotes several addition laws for the multivalued inverse trigonometric functions. We give below (4.4.32–4).

$$\mathrm{Arcsin}(z_1) \pm \mathrm{Arcsin}(z_2) = \mathrm{Arcsin}\left(z_1\sqrt{1 - z_2^2} \pm z_2\sqrt{1 - z_1^2}\right). \tag{32}$$

$$\mathrm{Arccos}(z_1) \pm \mathrm{Arccos}(z_2) = \mathrm{Arccos}\left(z_1 z_2 \mp \sqrt{(1 - z_1^2)(1 - z_2^2)}\right). \tag{33}$$

$$\mathrm{Arctan}(z_1) \pm \mathrm{Arctan}(z_2) = \mathrm{Arctan}\left(\frac{z_1 \pm z_2}{1 \mp z_1 z_2}\right). \tag{34}$$

Equation (34) is, as the layout suggests, shorthand for the two equations

$$\mathrm{Arctan}(z_1) + \mathrm{Arctan}(z_2) = \mathrm{Arctan}\left(\frac{z_1 + z_2}{1 - z_1 z_2}\right) \tag{35}$$

and

$$\operatorname{Arctan}(z_1) - \operatorname{Arctan}(z_2) = \operatorname{Arctan}\left(\frac{z_1 - z_2}{1 + z_1 z_2}\right). \tag{36}$$

It would be tempting to think the same of equation (33), but in fact $\operatorname{Arccos}(x) = -\operatorname{Arccos}(x)$, so the $\pm$ on the left-hand side is spurious. Modulo $2\pi$, each of $\operatorname{Arccos}(z_1)$ and $\operatorname{Arccos}(z_2)$ has two values, so the left-hand side has, generically, four values modulo $2\pi$. Therefore we seem to need (see the proof in [12]) both values of $\mp$, and this is indeed true. The equation could also be written as

$$\operatorname{Arccos}(z_1) + \operatorname{Arccos}(z_2) = \operatorname{Arccos}\left(z_1 z_2 + \operatorname{Sqrt}\left((1 - z_1^2)(1 - z_2^2)\right)\right).$$

When it comes to equation (32), the situation is more complicated, but in fact it is possible to prove (see [12]) that any containment of the form $\operatorname{Arcsin}(z_1) + \operatorname{Arcsin}(z_2) \subset \operatorname{Arcsin}(A)$ must also have the property that $\operatorname{Arcsin}(z_1) - \operatorname{Arcsin}(z_2) \subset \operatorname{Arcsin}(A)$. So the equation should be read as

$$\operatorname{Arcsin}(z_1) \pm \operatorname{Arcsin}(z_2) = \operatorname{Arcsin}\left(z_1 \operatorname{Sqrt}(1 - z_2^2) + z_2 \operatorname{Sqrt}(1 - z_1^2)\right), \tag{37}$$

with each side taking on eight values modulo $2\pi$ (counting special cases like $\operatorname{Arcsin}(1)$ as a "double root").

It is unfortunate that the desire to save space led the compilers of [2] to compress equations (35) and (36) into equation (34), since the $\pm$ notation here actually has a completely different meaning from its use in the adjacent equations (32) and (33). For completeness, let us say that in [2, (4.4.35)] —

$$\begin{aligned}\operatorname{Arcsin} z_1 \pm \operatorname{Arccos} z_2 &= \operatorname{Arcsin}\left(z_1 z_2 \pm \sqrt{(1 - z_1^2)(1 - z_2^2)}\right) \\ &= \operatorname{Arccos}\left(z_2\sqrt{1 - z_1^2} \mp z_1\sqrt{1 - z_2^2}\right)\end{aligned}$$

the convention is as in (4.4.32), i.e. the equation cannot be split and $\sqrt{w}$ means $\operatorname{Sqrt}(w)$, whereas in [2, (4.4.36)] —

$$\begin{aligned}\operatorname{Arctan} z_1 \pm \operatorname{Arccot} z_2 &= \operatorname{Arctan}\left(\frac{z_1 z_2 \pm 1}{z_2 \mp z_1}\right) \\ &= \operatorname{Arccot}\left(\frac{z_2 \mp z_1}{z_1 z_2 \pm 1}\right)\end{aligned}$$

the convention is as in (4.4.34), i.e. the equation can be split.

## 7 Couthness

[9] introducted this concept. If $h$ is any hyperbolic function, and $t$ the corresponding trigonometric function, we have a relation

$$t(z) = ch(iz) \text{ where } c = \begin{cases} 1 & \cos, \sec \\ i & \cot, \operatorname{cosec} \\ -i & \sin, \tan \end{cases}. \tag{38}$$

From this it follows *formally* that

$$h^{-1}\left(\frac{1}{c}z'\right) = it^{-1}(z'). \tag{39}$$

**Definition 1.** *A choice of branch cuts for $h^{-1}$ and $t^{-1}$ is said to be a* couth *pair of choices if equation (39) holds except possibly at finitely many points.*

[9] show that, with their definitions (the definitions of [2] with the values on the branch cuts prescribed) all pairs were couth except for:

**arccos/arccosh** Here equation (39) only holds on the upper half-plane (including the real axis for $\Re z \leq 1$);

**arcsec/arcsech** Here equation (39) only holds on the lower half-plane (including the real axis for $\Re z > 1$).

However, [2, (4.4.20–25)] show that all pairs are couth in the multivalued case (where equation (39) is interpreted as equality of sets).

## 8 Conclusion

This paper has, as is perhaps inevitable at this stage of Mathematical Knowledge Management, posed more questions than it answers. For convenience, we recapitulate them here.

1. Should we codify a branch cut, e.g. for ln as a direct subset of $\mathbf{C}$, or via a specification such as equation (5).
2. If the former, what formal semantics can we attach to the phrase "branch cut"? Can one do this in a way independent of the specification of $\mathbf{C}$ and $\mathbf{C} \to \mathbf{C}$?
3. What should be the correct encoding of false equations such as equation (1): unwinding numbers, conditional or multivalued? How does one cope with equations such as (7) and (8) in the conditional formalism — aren't we just rewriting the unwinding number formalism? Conversely, equation (16) is distinctly simpler than equation (15), and equation (19) currently has no unwinding equivalent. Should a Mathematical Knowledge Management system (in this area) have to support more than one such encoding?
4. How do we support the restriction of these functions to (partial) functions $\mathbf{R} \to \mathbf{R}$? In this case *most* of the unwinding number terms or conditions drop out. It is harder to see how the multivalued formalism supports this restriction.
   The obvious case where some caveat is still necessary is $\sqrt{z^2} \stackrel{?}{=} z$, where the formalisms might be:
   $$z \geq 0 \Rightarrow \sqrt{z^2} = z;$$
   $$\sqrt{z^2} = (-1)^{\mathcal{K}(2\ln z)} z.$$
   The second has the disadvantage of still introducing complex numbers, via $\ln z$ when $z < 0$, though it could clearly be massaged into $\sqrt{z^2} = (\operatorname{sign} z)z$.

5. It appears that, contrary to popular belief, the multivalued semantics are not simply a tidier version of the univalued (branch cut) semantics: contrast equation (20) and its conditional equivalent

$$(z \notin \mathbf{R}) \vee (z^2 \leq 1) \Rightarrow \arcsin z = \arctan \frac{z}{\sqrt{1-z^2}}$$

   with equation (31). Does this mean that we need two separate Mathematical Knowledge Repositories for the two cases?
6. Can a Mathematical Knowledge Repository for these facts (either case, or both cases) be simply a database, or must it be much more intelligent, possibly incorporating ideas along the lines outlined in [5].

We also deduce the following differences between the "Abramowitz & Stegun" (A+S) era and the MKM era.

- In the A+S era, it was not necessary to specify the values of the functions on branch cuts: numerical analysts for the most part did not care (but see [15]) since the branch cuts were of measure zero, and the intelligent reader could choose the adherence most suitable to the problem. In the MKM era, both computer algebra systems and theorem provers need to know correctly what the answer is. For interoperability, they must agree on what the answers is — see the examples in [9].
- In the A+S era, it was acceptable (maybe only just) to use the $\pm$ notation to mean two different things: in the MKM era it is not, and the notation should only be used (if at all) with $A \pm B$ being shorthand for $\{a+b, a-b\}$ (or, in the set-valued case $(A+B) \cup (A-B)$).
- A+S was ambivalent about whether it was talking about $\mathbf{C} \to \mathbf{C}$ or (partial) $\mathbf{R} \to \mathbf{R}$. Many of the formulae are stated with (unnecessary) restrictions to the $\mathbf{R}$ case — see equation (13) and [2, 4.4.28] relating Arctan to Ln, which restricts $z$ to be real.
- In the A+S era "everyone knew" what a branch cut was. To the best of the author's knowledge, no computer algebra system or theorem prover does.

It is hoped that these thoughts, limited as they are to one chapter of one book, will stimulate debate about the difficulties of managing this sort of mathematical knowledge.

## References


1. Abbott,J.A., Díaz,A. & Sutor,R.S, OpenMath: A Protocol for the Exchange of Mathematical Information. SIGSAM Bulletin **30** (1996) 1 pp. 21–24.
2. Abramowitz,M. & Stegun,I., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. US Government Printing Office, 1964. 10th Printing December 1972.
3. Aslaksen,H., Can your computer do complex analysis?. In: *Computer Algebra Systems: A Practical Guide* (M. Wester ed.), John Wiley, 1999. `http://www.math.nus.edu.sg/aslaksen/helmerpub.shtml`.

4. Beeson,M. & Wiedijk,F., The Meaning of Infinity in Calculus and Computer Algebra Systems. Artificial Intelligence, Automated Reasoning, and Symbolic Computation (ed. J. Calmet *et al.*), Springer Lecture Notes in Artificial Intelligence 2385, Springer-Verlag, 2002, pp. 246–258.
5. Bradford,R.J. & Davenport,J.H., Towards Better Simplification of Elementary Functions. Proc. ISSAC 2002 (ed. T. Mora), ACM Press, New York, 2002, pp. 15–22.
6. Bradford,R.J., Corless,R.M., Davenport,J.H., Jeffrey,D.J. & Watt,S.M., Reasoning about the Elementary Functions of Complex Analysis. *Annals of Mathematics and Artificial Intelligence* **36** (2002) pp. 303–318.
7. de Bruijn,N., The Mathematical Vernacular, a language for mathematics with type sets. Proc. Workshop on Programming Logic, Chalmers U., May 1987.
8. 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.
9. 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.
10. Corless,R.M. & Jeffrey,D.J., The Unwinding Number. SIGSAM Bulletin **30** (1996) 2, pp. 28–35.
11. Davenport,J.H., Table Errata — Abramowitz & Stegun. To appear in *Math. Comp.*
12. Davenport,J.H., "According to Abramowitz and Stegun" II. OpenMath Thematic Network Deliverable , 2002. `http://www.monet.nag.co.uk/cocoon/openmath/documents/AS2.pdf`
13. Dewar,M.C., OpenMath: An Overview. *ACM SIGSAM Bulletin* **34** (2000) 2 pp. 2-5.
14. IEEE Standard 754 for Binary Floating-Point Arithmetic. IEEE Inc., 1985.
15. 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.
16. Rich,A.D. and Jeffrey,D.J., Function evaluation on branch cuts. SIGSAM Bulletin 116(1996).
17. Stoutemyer,D., Crimes and Misdemeanors in the Computer Algebra Trade. *Notices AMS* **38** (1991) pp. 779–785.