Why Computer Algebra Systems Can't Solve Simple Equations

Richard J. Fateman

January 3, 1996

1 Introduction

Among the basic equations one might wish a computer to solve symbolically is the inverse of the power function, solving $y = z^w$ for z. While many special cases, easily solved, abound, the general question is fraught with implications: if this is so hard, how can we expect success in other ventures? Having solved this, we can naturally use it in a "composition" of solution methods for expressions of the form $y = f(z)^w$.

Can't we already do this? Is it not the case that the solution of $y = z^{a+bi}$ is trivially $z = y^{1/(a+bi))}$?

Not so. if this were the case, then a plot of the function $t(y) := y - (y^{1/(1+i))})^{1+i}$ would be indistinguishable from $t(y) \equiv 0$. For many values, $t(y)$ is (allowing for round-off error), zero. But if your computer system correctly computes with values in the complex plane, then, (to pick two complex points from a region described later), $t(-10000 + 4000i)$ is not zero, but about $-9981 + 3993i$ and $t(-0.01 + 0.002i)$ is about $5.34 - 1.06i$. These strange numbers are not the consequence of round-off error or some other numerical phenomena. The alleged solution is just not mathematically correct.

2 How to solve $y = z^w$ for z

We assume that y and w are complex-valued variables in general, and the solution sought for z is permitted to be complex as well. If we know specific values for y and w, we can simplify the question and answer it. For example, $9 = z^2$ has the solution set $\{-3,\ 3\}$. The equation $-9 = z^2$ has the solution set $\{-3i,\ 3i\}$. The equation $3 = z^{1/2}$ has the solution set $\{9\}$.

But $-3 = z^{1/2}$ has no solution for real or complex numbers, at least given the conventional meaning for this power function: if $z := r \cdot \exp(i\theta)$ then $z^{1/2}$ must be $\sqrt{r} \cdot \exp(i\theta/2)$ where the positive $\sqrt{r}$ of the positive value r is taken. There is no value for r and for $\pi < \theta \leq \pi/2$ to satisfy this equation. If you feel like arguing this point, read the footnote[1].

To some extent we can try to limit the scope of the answer even though we may not have specific values for w and y. One way of furthering this exploration is to inquire about the real and imaginary parts (or perhaps the argument and magnitude) of w and y. We've already seen situations above where the solution set has zero, one, or more distinct solutions, giving us some hint as to what to expect.

3 A systematic attempt

There are any number of ways of approaching this problem from a naive complex-variables direction. We've tried quite a few, and believe this is about as simple as it gets.

Let use define $y := s \cdot \exp(i\rho)$, $z := r \cdot \exp(i\theta)$ and $w := a + bi$. That's right, even though we have three complex variables, we don't use the same representation for w as for y and z. This is a matter of convenience; any alternative representation can be changed to this.

Note: s and r are nonnegative real, a and b are real, ρ and θ are in the half-open real interval $(-\pi, \pi]$. These are all conventional restrictions to make the representations of complex values canonical, and do not limit the "values" they can assume[2].

Then

$$z^w = \exp((a + bi) \cdot (\log r + i\theta))$$

[1] Even if you wish to specify that $()^{1/2}$ means a set of two values, then the equation still has no solution. If you think that a solution is $z = 9$, observe that $\{3\} \neq \{3,\ -3\}$. If you uniformly choose (somewhat perversely) the negative square root, then $3 = z^{1/2}$ has no solution. It appears that a solution would entail being able to magically distinguish the number 9 whose square-root is 3 from the number 9 whose square-root is -3.

[2] If $z = 0$ we will say that $r = 0$ and $\theta = 0$ for definiteness.

$$= \exp(a \log r - b\theta + i(a\theta + b \log r))$$

So

$$z^w = \exp(a \log r - b\theta) \cdot \exp(i\phi) \tag{1}$$

where $\phi = b \log r + a\theta$. We do not assume that ϕ is in $(-\pi, \pi]$. Note however, that the first factor in equation (1) is necessarily real because r is non-negative and a, b and θ are real.

Our objective is for z^w, so expressed, to be equal to y:

$$y = s \cdot \exp(i\rho). \tag{2}$$

The magnitude and the argument (modulo 2π) of the two expressions must be equal, and so we are provided with 2 equations:

$$s = \exp(a \log r - b\theta) \tag{3}$$

and

$$\rho = b \log r + a\theta + 2n\pi \tag{4}$$

(for some integer value n) which must be solved simultaneously for r and θ. Solving for r, a real value, in (3) yields:

$$r = \exp((\log(s) + b\theta)/a). \tag{5}$$

Equation (5) should alert us to a possible problem at $a = 0$. Proceeding nevertheless, we substitute for $\log r$ (note, r is non-negative) in (4) and get

$$\rho = b/a \cdot (\log s + b\theta) + a\theta + 2n\pi$$

The solutions to the latter are

$$\theta = (-b \log s + a\rho + 2an\pi)/(a^2 + b^2). \tag{6}$$

Now what remains is for n to be chosen appropriately. Given a set of values for a, b, ρ, and s in equation (6) we can find some set of integer values for n, namely when

$$\frac{-\pi\ (a^2 + b^2) + b\ \log s - a\rho}{2a\pi} < n \leq \frac{\pi\ (a^2 + b^2) + b\ \log s - a\rho}{2a\pi}$$

which then imposes the condition that θ is in $(-\pi,\ \pi]$. We then use those values to get corresponding values for r from equation (5).

It would be nice if this were the end of it. Unfortunately, it is not so simple.

4 Branch Cuts

The solutions of the previous section fall apart in various ways because of singularities and the necessity of defining a branch cut in the logarithm function. The branch cut is normally along the negative real axis, and the values along the cut are pasted to the "top" part. In more detail, let us consider the situation.

1. For $a = b = 0$, we are solving at a singular point, and the equation degenerates to $y = z^0$. The only solution is when $y = 1$, and then z is arbitrary.

2. If $b = 0,\ a \neq 0$ (the real exponent case), then the solution exists for the simpler equations

$$r = \exp((\log s)/a)$$

and

$$\theta = \rho/a$$

Since $-\pi < \theta \leq \pi$, a solution can exist only when

$$-a\pi < \rho \leq a\pi. \tag{7}$$

Thus there is no solution for $y = z^a$ unless ρ (which is $\arg y$) abides by condition (7). Two examples: if $a = 1/2$ then y must be in the right half plane with $\rho \in (\pi/2,\ \pi/2]$. If $a = 1/3$, then y must be in a wedge in the half-open interval with $\rho \in (\pi/3,\ \pi/3]$.

3. If $a = 0$ (but $b \neq 0$) we must avoid the division in equation (5) and go back to equations (3) and (4) giving us

$$\theta = -\log(s)/b$$

$$r = \exp(\rho/b)$$

The restrictions: since $-\pi < \theta \leq \pi$,

$$-b\pi < -\log(s) \leq b\pi$$

or if b is negative,

$$b\pi \leq -\log(s) < -b\pi.$$

Since exp is monotonic, we impose one of

$$\exp(-b\pi) \leq s < \exp(b\pi). \tag{8}$$

or

$$\exp(b\pi) < s \leq \exp(-b\pi). \tag{8'}$$

Thus there is no solution for $y = z^{ib}$ unless s (which is $|y|$), abides by condition (8) or (8'). For example, $\exp(\pi) = z^i$ has no solution, but $\exp(-\pi) = z^i$ has one. Geometrically, *the acceptable values of y appear in an annulus: the region between two concentric circles about the origin in the complex plane of radii $p = \exp(-|b|\pi)$ and $1/p$.* We would draw a figure, except this is not very hard to visualize.

4. If $a \neq 0$ and $b \neq 0$, consider, once again from (6) that $-\pi < \theta \leq \pi$ and therefore

$$-\pi(a^2 + b^2) < -b \ \log s + a\rho + 2an\pi \leq \pi(a^2 + b^2). \tag{9}$$

Consider the border curve of this region as determined by this equation:

$$-b \log s + a\rho = \pm\pi(a^2 + b^2 - 2an\pi).$$

The curve, in polar form is

$$s = K \exp((a/b) \cdot \rho) \tag{10}$$

for the constants

$$K = \exp(\pm\pi(a^2 + b^2 - 2an)/b).$$

Equation (10) defines a spiral starting on the real axis ($\rho = -\pi$) and ending on the real axis ($\rho = \pi$) after one revolution. The interior of the acceptable region includes one of these spirals (the one with the $-\pi$) but not the other. It is as though the annulus of the previous case were cut along the negative real axis and distorted. The two curves are joined in two places by segments of the negative real axis.

5 An example of the complex case

If we look at a particular instance of our equation, namely

$$y = z^{i+1}$$

we can easily but rather cavalierly solve it as

$$z = y^{1/(i+1)} = y^{1/2-i/2}.$$

For this case, the two spirals are governed by $K = exp(\pm\pi(2-2n))$. For $n = 0$, the values of K are $\{535.492,\ 0.00186744\}$. The outer logarithmic spiral hits the negative real axis at about -12,392 and the inner spiral hits the negative real axis at about -0.0432. The acceptable values for y are between the spirals and the line connecting them on the negative real axis (this line is joined upward to the figure).

Is $n = 0$ acceptable? By equation (6), using $a = 1$ and $b = 1$, we require $\theta = (-\log s + \rho)/2$ to be in $(-\pi,\ \pi]$ when ρ is in $(-\pi,\ \pi]$. This is no problem, since for every value of ρ there is a satisfactorily corresponding s. There may be other n possible, in which case there is a pair of spirals such as illustrated below. The solution exists between them. We give two figures to show the general shape; first, two spirals, and then a close-up of them near the origin.

6 Conclusion

In the case of solving equations, one would like to have as much of the following information as possible: a description of the inverse function's domain; the number of distinct solutions; a formula to numerically evaluate each inverse.

If particular cases simplify, and the result can be expressed as lists of equations, this seems most plausible to be useful.

If one cannot make determinations as to whether $w = 0$, or if its real part is 0, or whether various conditions hold on y, then some alternatives must be considered.

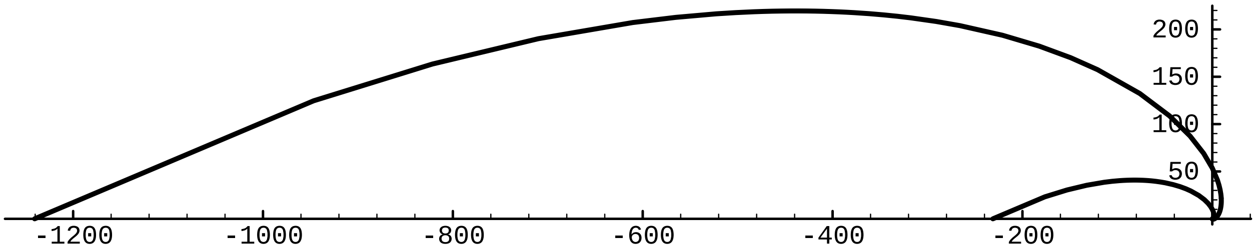

Figure 1: Two spirals.

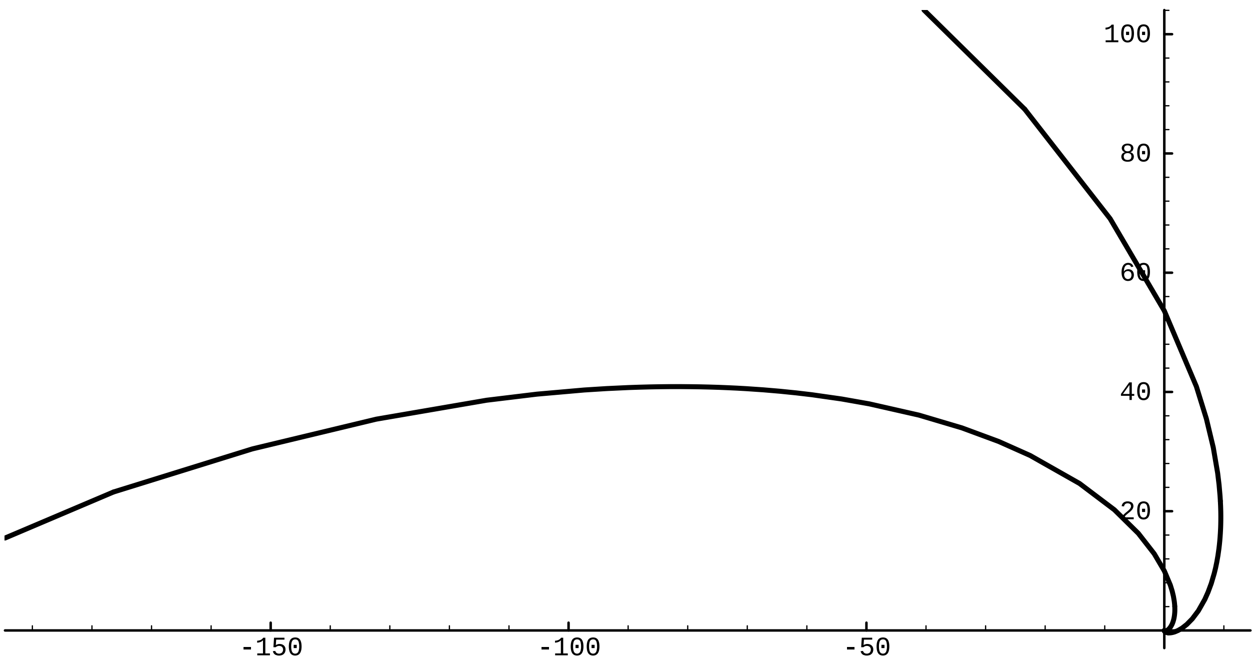

Figure 2: A closer view spirals.

1. We could ask the user, or inquire of some assumption "knowledge base" in the system to try to determine appropriate information. Some CAS designers (e.g. Maple) are opposed to halting the computation to ask the user. Others (e.g. Macsyma) are not so shy. A reference to a set of "assumptions" is possible in either of these systems.

2. We could give a symbolic `If [ ...]` answer along the lines of the construction above. This is rarely useful unless (a) it collapses substantially as a consequence of arithmetic and/or logical simplification, and (b) the CAS is able to continue computation with "conditional" expressions, including, for example, adding and multiplying them, inverting them (!), etc. This is a challenge, but in some sense inevitable if the answer to the questions of domain are not and cannot yet be "known". Although we've written this out, it does not seem to warrant repetition.

3. We could assume that if one cannot prove (say) $a = 0$, then it is definitely the case that $a \neq 0$. Although some CAS use this "closed world assumption" (that all true things are known, and anything that is not known or provable is false), the consequences are potentially dreadful.

4. We could defer execution of the program that evaluates the test conditions until the time that it is in fact needed to proceed with a computation. At that time our computer program would insist that any information that is indeed needed is provided. This so-called lazy evaluation is rarely used in computer algebra systems with the exception of some series computations in which terms are computed "as needed".

5. One could refuse to solve the equation as given.

We recommend some version of 2, as being almost inevitable in cases where "symbolic" parameters must be used, although in some circumstances, approach 4 is workable. Most existing computer algebra systems seem to use some version of 1 or 3.

7 Acknowledgements

Thanks to D. Halpern, T. Tokuyasu, T. Einwohner and P. Liao for discussions of this problem. Further detailed work, including programs and representations for algebraic and geometric objects are subjects for continuing work.This work was supported in part by NSF Grant number CCR-9214963 and by NSF Infrastructure Grant number CDA-8722788.