Complex Analysis

Harold V. McIntosh
Departamento de Aplicación de Microcomputadoras,
Instituto de Ciencias, Universidad Autónoma de Puebla,
Apartado postal 461, 72000 Puebla, Puebla, México.

April 5, 2001

Abstract

When the School of Computation was established at the University of Puebla, it inherited a course on complex variables from an earlier curriculum. Teaching the course was always passed off to the Mathematics Department, but in recent years even they have been reluctant to accept the responsibility. In the meantime, a requirement has arisen for the inclusion of complex analysis in a course on Mathematical Methods related to solid state physics (band gaps, Bloch's theorem, ...). Both of these circumstances have provided the opportunity to review materials last seen in graduate school. There are so many books on complex variable theory in existence that there hardly seems room for still another; nevertheless written material is needed for the entertainment of the students. Consequently these notes cover some of the why's and wherefore's of complex variables; ranging from the role of the cross ratio and the Schwartz derivative to topics such as the Mandelbrot Set, Elliptic Curves, and spectral densities.

Contents

1 Complex number arithmetic

There is a certain historical sequence of difficulties and their solution which arose as the use of numbers was extended to ever more complicated calculations. Fractions were needed to divide entities up into parts and rearrange the pieces. Negative numbers as the equivalent of debts represent another level of abstraction, requiring some explanations about the negatives of negatives, and expecially about how the product of two negative numbers ought to be positive. That was decided upon long before concepts like associative laws or distributive laws were formalized, but the ideas behind them were appreciated and lent a certain order to rules which could have been hard to understand.

The solution of equations containing products now and then leads to having to multiply two equal numbers together to obtain a negative result The product of positive numbers can't be negative, but trying to square negative numbers to get a negative result goes against the conclusion that such a product of negative numbers should also be positive. Sometimes this contradiction can be avoided by reformulating the problem, or the terms of its solution.

Still, the basic contradiction persists. In order not to remain a permanent obstacle, some adjustment in the idea of what constitutes a number is needed. One approach is to postulate a quantity whose square is minus one, afer having decided that it can be scaled to get square roots of other negative quantities. But there aren't any such "numbers," which seems to be why they got to be called *imaginary.* Even so, it is not clear that such an invention solves all problems; one of the first reactions of persons who hear about i and understand (but maybe not too deeply) the reasons for its use is to ask: "What about the square root of $-i$?"

A recent book of Paul J. Nahin [23] considers the historical evolution of the concept of imaginary numbers and their symbolism at some length. In turn, the book has been reviewed by Brian E. Blank [3], who finds that mathematicians and electrical engineers think somewhat differently.

1.1 complex numbers in the real plane via the Argand diagram

Although the symbol manipulation surrounding the use of a symbol i which fulfils the arithmetical rule that $i^2 = -1$ is extremely convenient and widely used, complex numbers can be given a much more mundane explanation by making up special rules for performing arithmetic operations on *pairs* of numbers.

There are antecedents for such circumlocutions; for example it is possible to work with fractions without giving up integers. Simply keep the numerators and denominators separate, notice that to multiply fractions it suffices to multiply the numerators together, and then the denominators. Other rules of convenience, such as to omit common factors in numerator and denominator, keep the process from getting out of hand. The rule for adding fractions involves common denominators; to get reciprocals, exchange numerator and denominator. The point is that all the arithmetic operations for fractions can be covered, yet nothing ever actually gets broken into pieces.

Even the negative numbers succumb to this treatment: simply keep two accounts, one for what is on hand and one for what is owed, making opportune transfers from one to the other as occasion demands.

Apparently William Rowan Hamilton, who later invented quaternions, was the first to propose that $x + iy$ be written instead as the pair (x, y) with the arithmetical rules for pairs

$$\begin{aligned}(a, b) + (c, d) &= (a + c, b + d) && (1)\\ (a, b) \times (c, d) &= (ac - bd, ad + bc). && (2)\end{aligned}$$

From this definition follows

$$\begin{aligned}(a,b)+(0,0) &= (a,b) \quad (3)\\ (a,b)\times(1,0) &= (a,b) \quad (4)\\ (0,1)\times(0,1) &= (-1,0), \quad (5)\end{aligned}$$

as do all the other things that ought to be checked, such as the associative, commutative, and distributive laws. The importance of doing things this way seems to be that all the manipulation is done with concrete, readily visible objects; there is no such thing as an "imaginary" number with some speculative properties.

1.2 polar and cartesian coordinates

Since the time that coordinates were taken up as a way to give an algebraic treatment to geometric ideas, there has been an association between pairs of numbers and points in the plane. In the traditional representation, the (real, imaginary) pair (x, y) has been seen as a point with x-coordinate x and y coordinate y; in fact the correspondence is so familiar that it seems redundant to try to describe it.

Besides the rectangular cartesian coordinates, there are many others which can be used to locate points in the plane; one of the most convenient is polar coordinates, wherein

$$\begin{aligned} x &= r\cos\theta & \qquad r &= \surd(x^2+y^2)\\ y &= r\sin\theta & \qquad \theta &= \arctan y/x\end{aligned}$$

Familiarity with the power series expansions respectively of the exponential, sine, and cosine functions,

$$\begin{aligned} e^x &= 1+x+\frac{1}{2}x^2+\frac{1}{3!}x^3+\frac{1}{4!}x^4+\cdots \quad (6)\\ \cos x &= 1-\frac{1}{2}x^2+\frac{1}{4!}x^4+\cdots \quad (7)\\ \sin x &= x-\frac{1}{3!}x^3+\frac{1}{5!}x^5+\cdots, \quad (8)\end{aligned}$$

suggests combining the two trigonometric series series to get Euler's formula

$$e^{ix} = \cos x + i\sin x \quad (9)$$

or the more general polar form

$$x+iy = re^{i\theta} \quad (10)$$

for a complex number.

1.3 absolute value, phase, modulus

The radius of a complex number, written in polar form, is zero only when both the real part and the imaginary part of the number are zero, and conversely. It therefore serves as an absolute value, or norm, symbolized in the usual manner as $|z|$.

$$|x+iy| = \surd(x^2+y^2). \quad (11)$$

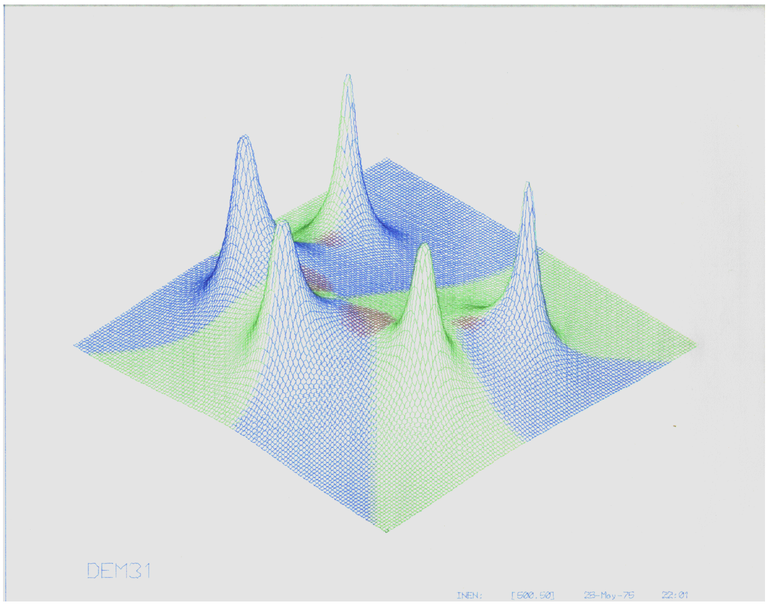


Figure 1: Graph of the absolute value of $(z-1)/(z^6-1)$.

The factorization

$$x^2 + y^2 \quad = (x + iy)(x - iy). \tag{12}$$

prompts the introduction of the *complex conjugate* $\bar{z}$ of a complex number z, by setting

$$\overline{x + iy} \quad = \quad x - iy. \tag{13}$$

It thereby follows that $|z|^2 = z\bar{z}$.

The angle θ in the polar decomposition of a complex number is called its *phase*, or sometimes its *amplitude*, but beware that amplitude does not mean magnitude, unless magnitude refers to the angular distance of the point representing the complex number from the real axis. On the other hand, modulus does mean magnitude.

The absolute value is multiplicative, which can be checked by comparing the algebraic expressions involved; in other words $|wz| = |w||z|$. It also obeys the triangle inequality, $|w+z| \leq |w| + |z|$. Amongst other things, those two properties are sufficient to check the convergence of the sine, cosine, and exponential series for complex numbers, to verify Euler's formula. It would still be premature to interpret the coefficients in the Taylor's series as derivatives, however.

1.4 stereographic projection

Stereographic projection is often used to reduce the size of the complex plane while preserving some of the characteristics of the plane, such as the angles between intersecting curves. In practice various lines are drawn from a point in such a way that they intersect both a sphere and a plane,

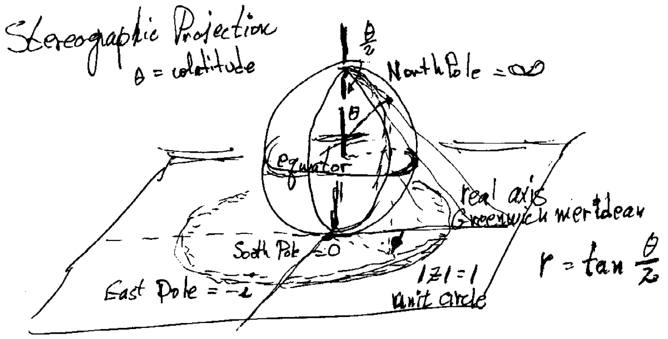


Figure 2: Represention of complex numbers via stereographic projection of a sphere of unit diameter onto a tangent plane.

although it is easy to imagine two spheres or two planes. The points of intersection common to a given line as it intersects the two surfaces establish the correspondence between the plane and the sphere. In order to use up all the sphere as well as all the plane in the mapping, the plane is usually set up tangent to the sphere, and the source point for the projection is either the opposite pole, resulting in a one-to-one mapping, or else the center of the sphere, by which two diametrically points map to the plane, resulting in a two-to-one mapping.

In the case of the polar *stereographic* projection, lines tangent to the pole do not intersect the plane, to which they are parallel. That leaves the pole without an image, but it can be called "infinity" and treated as though it were part of the complex plane. In this representation, the unit circle corresponds to the equator of the sphere, whose axis lies perpendicular to the plane. Zero is the point of tangency which could be supposed to be the south pole, infinity is accordingly the north pole, and the real axis might be regarded as the Greenwich meridian. Great circles map into straight lines, other circles into lesser circles.

In this mapping, inversion (mapping z into $1/z$) turns the sphere over, allowing infinity to be treated as the reciprocal of zero, and conversely. There is only one infinity; by including it the complex plane can be made compact. Numbers of sufficiently large modulus constitute neighborhoods of infinity, but in working with the sphere rather than the plane, the use of large numbers is avoided.

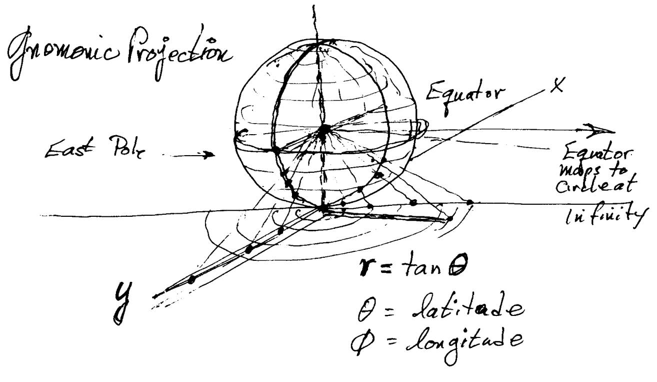


Figure 3: Represention of complex numbers via gnomonic projection of a sphere of unit radius onto a tangent plane.

The central, or *gnomonic*, projection maps the equator into infinity, which then consists of a "line," and not a point. Consequently there would be directions at infinity; plus infinity would be distinct from minus infinity, and both from the two purely imaginary infinities. The arrangement can still be compact by taking the sphere with opposite points identified as the representative of the complex plain, but all those different directions still have to be respected.

1.5 Smith Chart

The stereographic representations of the complex plane are obtained by real constructions. Instead, the complex analogue of the representation of the trigonometric functions could be used, in which the tangent of an angle is the point of intersection of the radius of the unit circle prolonged to intersect the vertical tangent at $x = 1$. The basic representation holds that the x-projection of the radius is $\cos\theta$, and that the y-projection is $\sin\theta$.

The complex equivalent of this mapping needs a complex angle to work with, and ought to correspond to the polar stereographic projection rather than the central stereographic projection. In the former case, the modulus of the projection is $\tan\theta/2$ rather than $\tan\theta$, so the suggested mapping is $w = \tan\theta/2$. But then,

$$\begin{aligned}\tan\frac{\theta}{2} &= \frac{\sin\frac{\theta}{2}}{\cos\frac{\theta}{2}} && (14)\\ &= -i\,\frac{e^{i\theta/2}-e^{-i\theta/2}}{e^{i\theta/2}+e^{-i\theta/2}} && (15)\\ &= -i\,\frac{e^{i\theta}-1}{e^{i\theta}+1} && (16)\\ &= -i\,\frac{z-1}{z+1} && (17)\end{aligned}$$

after introducing the abbreviation $e^{i\theta} = z$.

Contour plots for this mapping constitute nomograms which, after having been labelled and drawn artistically, are known as *Smith Charts*. They are of considerable use in transmission line theory, and are used without the factor i.

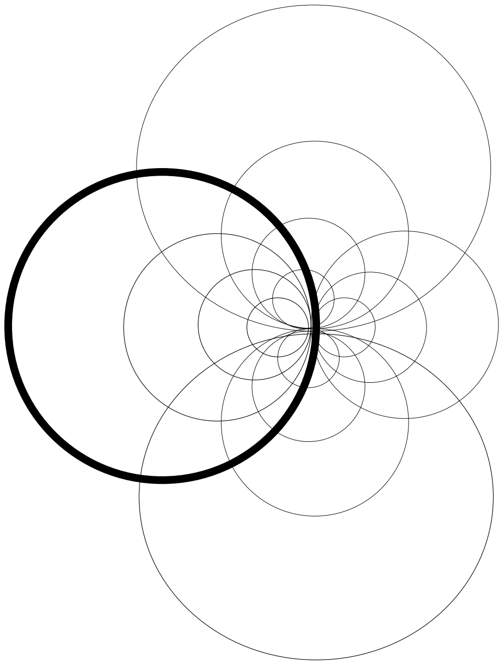


Figure 4: contours of $\tan\theta/2$ generate a useful nomogram, the Smith Chart.

One great advantage of this representation is that the whole right half-plane, the one whose numbers have positive real parts, is mapped into the unit circle, the imaginary axis taking up residence on its circumference. The real axis maps into the real axis, but given that infinity maps into 1, the whole coordinate grid of lines parallel to the real and to the imaginary axis ends up as as two families of mutually orthogonal circels, all passing through 1.

1.6 graphical representation of complex functions

Although the use of either cartesian or polar coordinates in a plane gives a way to illustrate complex numbers, it is still not so easy to work with all the properties of complex numbers. While complex addition is nothing but vector addition, complex multiplication has two different aspects. A real factor changes scale, just as it does for the real line and for vectors in general. But multiplication by i rotates by 90° counterclockwise about the origin, with a further change of scale if i has a multiplier of its own. It is an interesting historical remark, to recall that Hamilton is remembered as having invented quaternions by trying to do the same thing in three dimensions.

Still, thinking of a complex plane brings a certain amount of order to working with complex numbers. But what is to be done with complex functions, even of a single variable? Real tradition graphs the function in the real plane as a cartesian product of the two number sets, the range and the domain, paired by the function. To do the same with complex numbers would require four dimensions – the cartesian product of two planes – which lies beyond anyone's ordinary experience.

A compromise is to graph part of the function in three dimensions, or in three dimensions as seen in perspective. The usual choice is to graph the absolute value, a positive real number, as a function of the real and imaginary parts of the variable. Variations on this presentation consist of graphing only contours while confining the whole presentation to a plane, or in coloring the three dimensional surface with additional information, such as the phase of the function value.

Figure 1 shows how the function $(z-1)/(z^6-1)$ may be described in these terms. The absolute value is shown via its traces along lines parallel to the real and imaginary axes, yet colored according to the signs of the real and imaginary parts of its value.

As a further item of interest, there are some patches drawn in red, where the derivative of the function is small. This anticipates a result to be proven later on, that the zeroes of the derivative of an analytic function lie within the convex hull of the function's own zeroes. Looking arounf for saddle points gives an idea of why this theorem should be true.

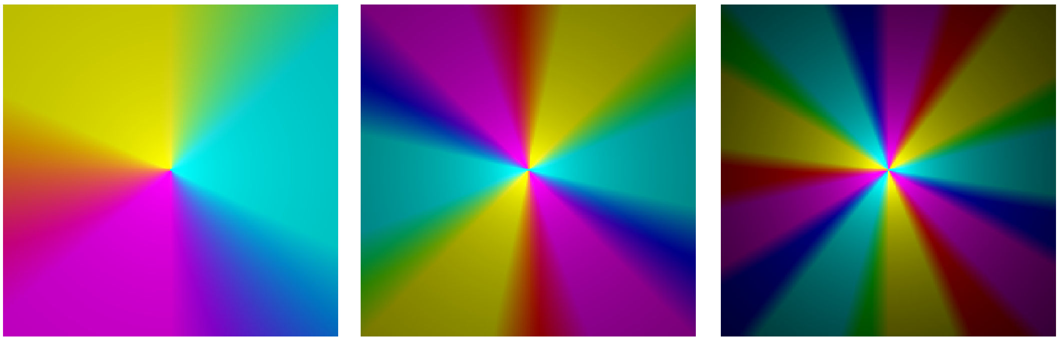


Figure 5: The representation of complex functions by color coded contours. Left: the identity mapping, which shows the natural coloring of the complex plane under this scheme. Notice an ever so slight radial darkening. Middle: the function $w = z^2$ showing its values according to the same scheme. The function is lower than its argument within the unit circle, but then it increases rapidly outside, a difference reflected in the brightness of the colors. Right: the function z^3 continues the tendency established by the first two powers, of running n times around the color wheel on account of the n^{th} power, Meanwhile the interior and exterior of the unit circle is more sharply differentiated, but the bottom is flattish even though the function is odd (with respect to negating its argument). Of course, that is a consequence of taking the absolute value which flips everything so that it can always be positive.

When graphs were drawn by hand or with line plotters, it was quite an art to choose appropriate lines, perspectives, and contours to get a good artistic rendition. With color television style monitors, there is a temptation to work more with areas filled with little colored dots, and to forget all the lines. From such a view point, the absolute value of a function could be represented in shades of grey, leaving the color wheel for the phase of the function.

Although all these color images are very beautiful, they still have to be chosen with care to give them the greatest information content. Using them to make stereopairs is a possibility, although few people seem to be able to visualize raw stereopairs without some form of optical enhancement.

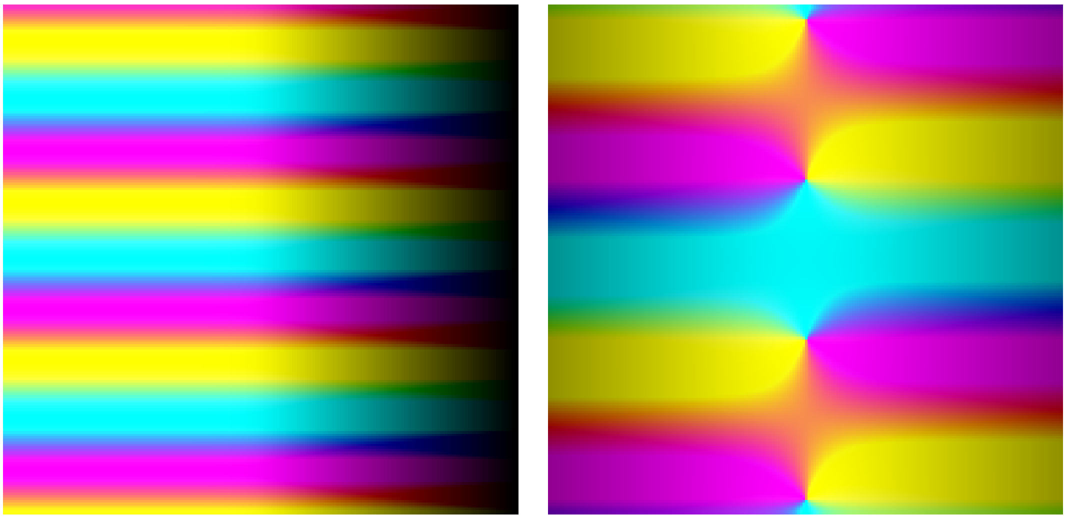


Figure 6: The complex exponential (left) rises slowly from zero at the far left to infinity at the far right. The hyperbolic cosine adds this to its rotated (because of the differing signs of the exponents) image.

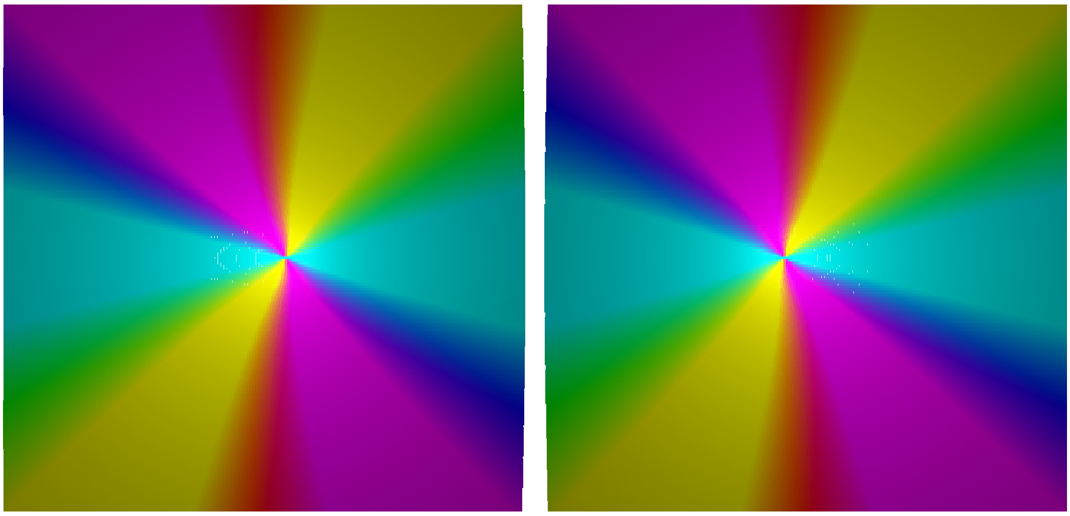


Figure 7: A stereopair constructed from slightly shifted renditions of the colorized version of the contours of the square function $w = z^2$.

2 Functions of a complex variable

The usual way of generating functions apply to complex variables as well as to real variables. By combining sums and products one arrives at polynomials, which are the simplest functions to define. By taking limits, polynomials can be extended to infinite series. Trigonometric functions, logarithms and exponentials arise from the process without much difficulty. In another direction, the inclusion of division among the constructing operations leads to rational functions and eventually to quotients of series and series of quotients.

2.1 one-to-one and invertible

Before looking at the complications caused by limiting processes, it is convenient to begin with the simpler functions, and to look for functions with simpler properties. Functions can be defined explicitly as well as implicitly, for example by equating polynomials in two variables to zero. Bearing in mind that polynomials have as many roots as their degree, the polynomials should be limited to the first degree to avoid multiple valuedness in the inverse function. Therefore the relation

$$dwz + cw + bz + a \quad = \quad 0 \tag{18}$$

is reccommended. It can be rendered explicit in either direction,

$$w \quad = \quad -\frac{bz+a}{dz+c} \tag{19}$$

$$z \quad = \quad -\frac{cw+a}{dw+b} \tag{20}$$

Inspection, and particularly familiarity with projective transformations, suggests a matrix representation of these fractional linear transformations. To get such a representation, the complex variables should be represented as quotients, such as $w = s/t, z = u/v$. Then the numerator and denominator of the fractional linear transformation are subject to linear transformations for which the matrix notation is appropriate.

$$\begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} b & a \\ d & c \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

Just because the complex variable is represented as a quotient (and observe that the members of the numerator-denominator pair are *not* the real and imaginary parts of the variable, but are themselves complex numbers) ambiguity exists both in the representation of the variable and in the matrix describing the transformation.

Since the matrix should correspond to an invertible transformation, its determinant should not vanish. Accordingly the matrix could be multiplied by a factor making the determinant unity. Such a choice which will later on prove to be consistent because the matrices of composite transformations multiply and so do the determinants.

2.2 Möbius transformations represented as cross ratios

One of the things which can be done with the Möbius transformation

$$dwz + cw + bz + a \quad = \quad 0 \tag{21}$$

is to find relationships amongst the variables which do not depend on the coefficients of the transformation. One procedure is to try it out for four points, which is the number of coefficients, to get simultaneous equations.

$$\begin{aligned} dw_1z_1 + cw_1 + bz_1 + a &= 0 \qquad (22)\\ dw_2z_2 + cw_2 + bz_2 + a &= 0 \qquad (23)\\ dw_3z_3 + cw_3 + bz_3 + a &= 0 \qquad (24)\\ dw_4z_4 + cw_4 + bz_4 + a &= 0 \qquad (25)\end{aligned}$$

The corresponding matrix equation,

$$\begin{bmatrix} w_1z_1 & w_1 & z_1 & 1 \\ w_2z_2 & w_2 & z_2 & 1 \\ w_3z_3 & w_3 & z_3 & 1 \\ w_4z_4 & w_4 & z_4 & 1 \end{bmatrix} \begin{bmatrix} d \\ c \\ b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

can only hold when the determinant of the square matrix vanishes. That relationship invites a series of transformations which can ultimately produce the result

$$\frac{w_4 - w_1}{w_4 - w_2}\frac{w_3 - w_1}{w_3 - w_2} = \frac{z_4 - z_1}{z_4 - z_2}\frac{z_3 - z_1}{z_3 - z_2} \qquad (26)$$

This quantity equates the cross ratios of the two sets of numbers, $\{w_1, w_2, w_3, w_4\}$ and $\{z_1, z_2, z_3, z_4\}$, it being that the cross ratio is an invariant of the Möbius transformation.

The equivalence of this result to the vanishing determinant is not immediately obvious, although it can be verified by careful attention to algebraic rearrangement. A line of reasoning which suggests a result which then can subsequently be confirmed begins with the fractional linear form of the Möbius transformation,

$$w = \frac{z - z_1}{z - z_3} \qquad (27)$$

which clearly maps z_1 to zero and z_3 to infinity, as long as they are distinct. As for counterimages, zero maps into z_1/z_3, and infinity to unity. However, this is not the only formula which could map those pairs of points; any multiple of the function will do the same; a third pair of points could be introduced to resolve the ambiguity. A convenient set of points is $\{0, 1, \infty\}$, which suggests selecting a point z_2 to map into unity and removing the ambiguity. Thus consider

$$w = \frac{z - z_1}{z - z_3}\frac{z_2 - z_3}{z_2 - z_1}, \qquad (28)$$

which is one of the cross ratios in the invariant expression Eq 26. In fact, by setting up matching expressions in w and z, any three distinct points $\{z_1, z_2, z_3\}$ can be mapped into any other three distinct points $\{w_1, w_2, w_3\}$ by going through the virtual intermediary of $\{0, 1, \infty\}$.

2.3 Möbius transformations representable as 2x2 matrices

Although there may be some question of whether the implicit or the explicit form of one-to-one and invertible mappings should be called *Möbius transformations*, that is the name given to this class of transformations. Mappings often form a group. Since the associative law for their composition is a given, the items to be checked are closure, invertibility and the existence of the identity.

To check closure, consider the composite of two maps. First, suppose we have

$$t = \frac{Aw+B}{Cw+D} \qquad (29)$$

$$w = \frac{az+b}{cz+d} \qquad (30)$$

which gives, on substituting,

$$t = \frac{A\frac{az+b}{cz+d}+B}{C\frac{az+b}{cz+d}+D}, \qquad (31)$$

and, on simplifying,

$$t = \frac{(Aa+Bc)z+(Ab+Bd)}{(Ca+Dc)z+(Cb+Dd)}, \qquad (32)$$

which mimics the matrix product

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} Aa+Bc & Ab+Bd \\ Ca+Dc & Cb+Dd \end{bmatrix}$$

From this it is apparent that the 2×2 unimodular matrices form a representation of the group of Möbius transformations; in particular the identity is represented by the unit matrix,

$$z = \frac{1z+0}{0z+1}, \qquad (33)$$

and the inverse transformation by the inverse matrix.

2.4 eigenvalues and eigenvectors of a Möbius transformation

The fixed points of a Möbius transformation satisfy the requirement

$$z = \frac{az+b}{cz+d} \qquad (34)$$

which translates into the quadratic equation

$$cz^2+(d-a)z-b = 0 \qquad (35)$$

with roots

$$\frac{1}{c}\left\{\frac{a-d}{2} \pm \sqrt{\left[\left(\frac{a+d}{2}\right)^2-1\right]}\right\}.$$

In case $c = 0$ the quadratic term is absent and the single fixed point would be $b/(d-a)$ unless $(d-a)$ were zero. That alternative would leave z without constraint and require b to vanish as well as c. But then all points would be fixed and the transformation would be the identity. And if b did not vanish, the equation would read $z = z + b$ which could not be satisfied unless z were ∞, which might as well be considered admissible.

Another case of a single root occurs when the radicand is zero leaving $(a-d)/2c$ as the only fixed point. Otherewise there are always two distinct fixed points

2.5 hyperbolic, parabolic, elliptic transformations

If the eigenvectors of the matrix representation of a Möbius transformation are its fixed points, there remains the question of interpreting the eigenvalues. A good way to find this out is to use a coordinate system in which the representation is diagonal (or failing that, in the Jordan normal form). It also helps to recall that the representation matrix might just as well be unimodular, with reciprocal eigenvalues. In canonical form, the fixed points are zero and infinity, or else just infinity.

In the former case, represent the complex number z by u/v and consider the eigenvalue equation

$$\left[\begin{array}{cc} \lambda & 0 \\ 0 & \lambda^{-1} \end{array}\right] \left[\begin{array}{c} u \\ v \end{array}\right] = \left[\begin{array}{c} \lambda u \\ \lambda^{-1} v \end{array}\right], \tag{36}$$

which transforms u/v into $\lambda u/\lambda^{-1}v$, or in other words z becomes $\lambda^2 z$, and the eigenvalue squared is a multiplier.

There are essentially three cases to consider. If the multiplier is real, the transformation moves points radially – outward if the factor is greater than 1, inward if it is less than 1. In the first case, ∞ is a stable attractor and 0 is an unstable repeller; in the second case their roles are reversed. Transformations in this class are usually called *hyperbolic* mappings.

If the multiplier is complex, but with absolute value 1, both 0 and ∞ are neutral fixed points, the general movement of points according to the transformation being a rotation. Such transformations are usually called *elliptic* mappings.

In both these restricted cases it is convenient to think of the circles fixed with respect to the transformation. In the first case, they are radii which are arcs of constant angle passing through zero and ∞; in the second case they are concentric circles surrounding the two fixed points.

The third case combines the previous two, contemplating a general complex multplier. It results in a composite of the other types, given the polar representation of a complex number. These are called *loxodromic* mappings.

The confluent case, where the two fixed points have coalesced, gives the Jordan normal form of the Möbius transformation, which would read

$$\left[\begin{array}{cc} 1 & a \\ 0 & 1 \end{array}\right] \left[\begin{array}{c} u \\ v \end{array}\right] = \left[\begin{array}{c} u + av \\ v \end{array}\right], \tag{37}$$

and transforms z into $z + a$, which is a translation. The invariant circles in this case are straight lines in the direction of the translation, which are circles through ∞. The confluent mappings are called *parabolic* mappings, to complete the analogy with the classification of conic sections.

There is a degenerate case not so far mentioned, in which z maps into $1/z$, and it to be seen as inversion in a circle – the unit circle.

When the matrix of the Möbius transformation is not diagonal, the ability to map any three points into any other three points should be used to place the fixed points at 0 and ∞, the diagonal form of the transformation executed, and the inverse mapping be used to restore the fixed points. Of course, all other points will have moved in the process but knowing that circles map into circles, the mapping can be followed along their arcs.

Hyperbolic mappings will move points away from the repelling fixed point along circular arcs toward the attracting point. Elliptic transformations will shepherd points towards the line joining the fixed points or draw them away into the vast space remote from the fixed points. Parabolic transformations behave similarly, without there being any space between the coalesced fixed points. The Smith Chart is based on a parabolic transformation.

3 The derivative of a function of a complex variable

To begin with, the usual calculus definition of a derivative can be applied to complex functions just as well as to real functions. It is just that the norm $|\Delta z|$ of an increment must be sent to zero, rather than the absolute value of a real difference, $|\Delta x|$.

$$\frac{d}{dz}f(z) \quad = \quad \lim_{|\Delta z|\rightarrow 0} \frac{f(z+\Delta z)-f(z)}{\Delta z}, \tag{38}$$

supposing that the limit exists.

From this, the usual rules governing the derivatives of sums, products, differences and quotients can be deduced. Derivatives of powers can be obtained from the binomial theorem

$$(z+\Delta z)^n \quad = \quad \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} z^{n-i}\Delta z^i \tag{39}$$

in the usual way with the result

$$\frac{d}{dz}z^n \quad = \quad nz^{n-1}. \tag{40}$$

The chain rule for composite functions is applicable, so that the derivatives of inverse functions can be obtained. About all that remains from introductory calculus courses is to obtain the derivative of an exponential and of the trigonometric functions. At least for the exponenial, the derivative of products and the definition holding that

$$e^z \quad = \quad \lim_{n\rightarrow\infty}(1+\frac{z}{n})^n \tag{41}$$

will take the same course as in real analysis if it is sure that the order of the limits can be exchanged.

There is one detail which can easily be overlooked in applying the limit in the definition of the derivative, which is that the result not depend on the phase of Δz, just as in the real case, there would be a difference in left hand or right hand limits which could cause an ambiguity defeating the definition.

3.1 Cauchy-Riemann equations

To check that the derivative is well-defined, separate the complex function into the sum of two real functions, just as z can be written as the sum of a real and an imaginary part:

$$z \quad = \quad x+iy \tag{42}$$

$$f \quad = \quad u(x,y)+iv(x,y). \tag{43}$$

By first taking Δz real,

$$\frac{df}{dz} \quad = \quad \lim_{\Delta z\rightarrow 0}\frac{f(z+\Delta x)}{\Delta x} \tag{44}$$

$$\quad = \quad \lim_{\Delta x\rightarrow 0}\frac{u(x+\Delta x)+iv(x+\Delta x)-(u(x,y)+iv(x,y))}{\Delta x} \tag{45}$$

$$\quad = \quad \frac{\partial u}{\partial x}+i\frac{\partial v}{\partial x}, \tag{46}$$

and then imaginary,

$$\begin{aligned}
\frac{df}{dz} &= \lim_{\Delta z\to 0}\frac{f(z+i\Delta y)}{i\Delta y} && (47)\\
&= \lim_{\Delta y\to 0}\frac{u(x,y+\Delta y)+iv(x,y+\Delta y)-(u(x,y)+iv(x,y))}{i\Delta y} && (48)\\
&= -i\frac{\partial u}{\partial y}+\frac{\partial v}{\partial y}. && (49)
\end{aligned}$$

Supposing that the derivative is independent of phase, at least in the context taken here, comparison shows

$$\begin{aligned}
\frac{\partial u}{\partial x} &= \frac{\partial v}{\partial y} && (50)\\
\frac{\partial v}{\partial x} &= -\frac{\partial u}{\partial y} && (51)
\end{aligned}$$

The members of this pair of equations are known as the Cauchy - Riemann equations, relating the real and imaginary parts of a complex function which is supposed to have a derivative. Ignoring the relationship is equivalent to failing to check whether or not a supposed derivative is well-defined. Strictly the relation should be checked for all other phases, but in the end all of them depend on just these two.

Consider the Jacobian matrix for the real functions $u(x,y)$ and $v(x,y)$ of the real variables x and y,

$$\begin{aligned}
\frac{\partial(u,v)}{\partial(x,y)} &= \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{bmatrix} && (52)\\
&= \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ -\frac{\partial v}{\partial x} & \frac{\partial u}{\partial y} \end{bmatrix} && (53)\\
&= \frac{\partial u}{\partial x}\mathbf{1}+\frac{\partial v}{\partial x}\mathbf{i} && (54)
\end{aligned}$$

The impact of the Cauchy-Riemann equations is to give the Jacobian matrix the form of a complex number in quaternion disguise; none other will suffice. Writing the matrix as an exponential shows how the derivative is a complex number with absolute value and a phase.

The conclusion is not only that not any old pair of functions could be joined to get an *analytic function* (as differentiable functions of a complex variable are called), but that by knowing one of them, one effectively knows the other. That is,

$$\begin{aligned}
v(x,y)-v(a,y) &= \int_a^x \frac{\partial v(\sigma,y)}{\partial\sigma}d\sigma && (55)\\
&= -\int_a^x \frac{\partial u(\sigma,y)}{\partial y}d\sigma && (56)
\end{aligned}$$

Knowing u, take its derivative and integrate to get v. For example, according to Euler's formula,

$$e^z = e^x(\cos y+i\sin y). \qquad (57)$$

Then

$$\begin{aligned} u &= e^x \cos y & (58)\\ \frac{\partial u}{\partial y} &= -e^x \sin y & (59)\\ -\int_a^b e^\sigma \sin y d\sigma &= e^\sigma \sin y|_a^x & (60)\\ v(x,y) - v(a,y) &= e^x \sin y - e^a \sin a & (61)\end{aligned}$$

which is consistent with

$$v(x,y) = e^x \sin y \qquad (62)$$

So trying to think of something like $e^x(\cos y + i \tan y)$ as an analytic function just wouldn't work.

The reason there aren't analytic quaternion functions is twofold. First, they anticommute (complicating division by Δq), and besides, there is getting their Jacobian matrices to act like numbers.

Consider anticommutativity and the derivative of q^2. Having written

$$(q + \Delta q)^2 - q^2 = q^2 + q\ \Delta q + \Delta q\ q + \Delta q^2 - q^2, \qquad (63)$$

should we divide by Δq on the left, on the right, take half the sum of the foregoing, or divide by $\surd(\Delta q)$ on both sides. No matter what, it is questionable whether the result would be $2q$, or anything else free of Δq

Historically, some progress has been made by requiring quaternion functions to satisfy linear partial differential equations similar to the Cauchy-Riemann equations. However, examining the possibilities in more detail would be a distraction from our concern with complex functions,

Remaining with functions of a complex numbers, it seems that all the manipulations which work for real variables seem to work for complex variables. That is because they are confined to polynomials and perhaps their limits, where there is always a term free of Δz along with others having Δz as a factor which can vanish in the limit.

But those are not the only functions of two real variables which can take complex values. Consider the complex conjugate,

$$\overline{x + iy} = x - iy. \qquad (64)$$

Its Jacobian matrix is

$$J(x,y) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad (65)$$

which certainly doesn't satisfy the Cauchy-Riemann equation. Neither does the squared norm

$$z\bar{z} = x^2 + y^2, \qquad (66)$$

whose Jacobian matrix is

$$J(x,y) = \begin{bmatrix} 2x & 0 \\ 2y & 0 \end{bmatrix}. \qquad (67)$$

Therefore any polynomial which mixes z and $\bar{z}$, or depends on $\bar{z}$ alone, will fail to be analytic unless it is purely zero.

3.2 harmonic functions of real variables

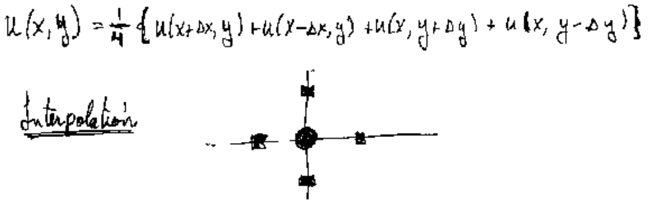

u(x,y) = 1/4 {u(x+Δx,y) + u(x−Δx,y) + u(x,y+Δy) + u(x,y−Δy)}
Interpolation

Figure 8: A harmonic function averages the values of its neighbors.

If an additional derivative is taken in the Cauchy - Riemann equations, we have

$$\begin{array}{rclcrcl} \frac{\partial^2 u}{\partial x^2} & = & \frac{\partial^2 v}{\partial x \partial y} & \qquad & \frac{\partial^2 u}{\partial y \partial x} & = & \frac{\partial^2 v}{\partial y^2} \\ \frac{\partial^2 v}{\partial x^2} & = & -\frac{\partial^2 u}{\partial x \partial y} & \qquad & \frac{\partial^2 v}{\partial x \partial y} & = & -\frac{\partial^2 u}{\partial y^2} \end{array}$$

From the equality of mixed second partial derivatives we get

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{68}$$

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \tag{69}$$

which means that u and v satisfy the two dimensional Laplace equation and are harmonic. An important practical consequence is that such functions never have maxima nor minima (unless they are confined by a boundary). This is most easily seen by writing the second derivatives in finite difference form

$$\frac{d^2 f}{dx^2} = \frac{f(x - \Delta x) - 2f(x) + f(x + \Delta x)}{\Delta x^2} \tag{70}$$

so that

$$u(x,y) = \frac{1}{4}\left\{u(x + \Delta x, y) + u(x - \Delta x, y) + u(x, y + \Delta y) + u(x, y - \Delta y)\right\} \tag{71}$$

$$v(x,y) = \frac{1}{4}\left\{v(x + \Delta x, y) + v(x - \Delta x, y) + v(x, y + \Delta y) + v(x, y - \Delta y)\right\}, \tag{72}$$

or, in other words, each central value is the average of its neighbors. No average can exceed all the values it is averaging, nor fall below all of them, either. This observation is a precursor of the maximum modulus principle, which holds that the critical points of an analytic function are saddle points, and that extreme values of such a function can only occur on the boundary of a region over which it is examined.

3.3 the Schwartz derivative

Consider the basic linear fractional transformation

$$w = \frac{az + b}{cz + d} \tag{73}$$

which is equivalent to the vanishing of the implicit expression

$$cwz + dw - az - b \quad = \quad 0. \tag{74}$$

To get a relation between w and z which does not depend on a, b, c, and d, take three derivatives with respect to a still unspecified parameter:

$$\begin{aligned} c(wz)' + dw' - az' &= \quad 0 \qquad (75)\\ c(wz)'' + dw'' - az'' &= \quad 0 \qquad (76)\\ c(wz)''' + dw''' - az''' &= \quad 0. \qquad (77) \end{aligned}$$

In equivalent matrix form,

$$\begin{bmatrix} (wz)' & w' & -z' \\ (wz)'' & w'' & -z'' \\ (wz)''' & w''' & -z''' \end{bmatrix} \begin{bmatrix} c \\ d \\ a \end{bmatrix} \quad = \quad \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{78}$$

The determinant of this matrix, which is a Wronskian, must vanish to get a nontrivial solution. But

$$\begin{aligned}
\begin{vmatrix} (wz)' & w' & -z' \\ (wz)'' & w'' & -z'' \\ (wz)''' & w''' & -z''' \end{vmatrix} &= \begin{vmatrix} w'z + wz' & w' & -z' \\ w''z + 2w'z' + wz'' & w'' & -z'' \\ w'''z + 3w''z' + 3w'z'' + wz''' & w''' & -z''' \end{vmatrix} \\
&= \begin{vmatrix} 0 & w' & -z' \\ 2w'z' & w'' & -z'' \\ 3w''z' + 3w'z'' & w''' & -z''' \end{vmatrix} \\
&= -w'z' \begin{vmatrix} 0 & w' & -z' \\ 2 & w'' & -z'' \\ 3\frac{w''}{w'} + 3\frac{z''}{z'z} & w''' & -z''' \end{vmatrix} \\
&= -w'z' \left\{ -w' \begin{vmatrix} 2 & z'' \\ 3(\frac{w''}{w'} + \frac{z''}{z'}) & z''' \end{vmatrix} - z' \begin{vmatrix} 2 & w'' \\ 3(\frac{w''}{w'} + \frac{z''}{z'}) & w''' \end{vmatrix} \right\} \\
&= -w'z' \left\{ -w'(2z''' - 3z''(\frac{w''}{w'} + \frac{z''}{z'})) + z'(2w''' - 3w''(\frac{w''}{w'} + \frac{z''}{z'})) \right\} \\
&= -w'z' \left\{ -2w'z''' + 3w'z''(\frac{w''}{w'} + \frac{z''}{z'}) + 2z'w''' - 3w''(\frac{w''}{w'} + \frac{z''}{z'}) \right\} \\
&= -(w'z')^2 \left\{ -2\frac{z'''}{z'} + 3\frac{z''}{z'}(\frac{w''}{w'} + \frac{z''}{z'}) + 2\frac{w'''}{w'} - 3w''(\frac{w''}{w'} + \frac{z''}{z'}) \right\}
\end{aligned}$$

Since this expression vanishes, the factor $(w'z')^2$ can be discarded (unless it happens to vanish everywhere) to leave

$$2\frac{w'''}{w'} - 3\left(\frac{w''}{w'}\right)^2 \quad = \quad 2\frac{z'''}{z'} - 3\left(\frac{z''}{z'}\right)^2, \tag{79}$$

which is a relationship which is expected to hold after applying a linear fractional transformation to a variable, no matter what the linear fractional transformation. Besides, if z is the variable, the right hand is zero, which leaves the left hand side as a constraint on any function that purports invariance. Some rearrangement can make the invariant read

$$\frac{2w'w''' - 3(w'')^2}{(w')^2},$$

half of which is called a *Schwartz derivative.* The symbol for the Schwartz derivative of w with respect to z is $\{w, z\}$, some variants of whose definition are:

$$\begin{aligned} \{w, z\} &= \frac{w'''}{w'} - \frac{3}{2}\left(\frac{w''}{w'}\right)^2, && (80) \\ &= \left\{\frac{w''}{w'}\right\}' - \frac{1}{2}\left(\frac{w''}{w'}\right)^2, && (81) \\ &= \frac{d^2}{dz^2}\left[\ln\frac{dw}{dz}\right] - \frac{1}{2}\left\{\frac{d}{dz}\left[\ln\frac{dw}{dz}\right]\right\}^2. && (82) \end{aligned}$$

The Schwartz derivative is not really a derivative (it is a differential invariant) because it doesn't follow the product rule when applied to products, nor does it even work linearly on sums, and it is furthermore immune to non-zero multiplicative constants. As the derivative was derived, w and z depended on a parameter with respect to which two differential expressions were identical. When z itself was the parameter, one result was obviously zero, forcing the other to assume the same value; which of the two was chosen creates an asymmetry by which $\{w, z\} \neq \{z, w\}$.

Nevertheless, the invariant can be calculated for various w's all depending on the same z, for which the following rules and particular results can be verified through calculation:

$$\begin{aligned} \left\{\frac{aw+b}{cw+d}, z\right\} &= \{w, z\}, && (83) \\ \left\{w, \frac{az+b}{cz+d}\right\}\frac{(ad-bc)^2}{(cz+d)^4} &= \{w, z\}, && (84) \\ \{w, z\} &= \{w, u\}\left(\frac{du}{dz}\right)^2 + \{u, z\}, && (85) \\ \{w, z\} &= -\{z, w\}\left(\frac{dw}{dz}\right)^2 && (86) \\ \{w, z\} = 0 &\Rightarrow \quad w = \frac{az+b}{cz+d}, && (87) \\ \{z^n, z\} &= \frac{1-n^2}{2z^2}, && (88) \\ \{e^{\lambda z}, z\} &= -\frac{1}{2}\lambda^2. && (89) \end{aligned}$$

The last three lines suggest the task of inverting the Schwartz derivative; that is, of finding $w(z)$ for which $\{w, z\} = Q(z)$, for some prescribed function Q. The answer is that one should solve the differential equation

$$y''(z) + \frac{1}{2}Q(z)y(z) = 0, \tag{90}$$

for which one can expect to obtain two linearly independent solutions $y_1(z)$ and $y_2(z)$. Their quotient

$$w(z) = \frac{y_2(z)}{y_1(z)} \tag{91}$$

will then solve the Schwartzian equation. This is a subject better postponed until differential equations, second order linear differential equations, and Ricatti equations have all been introduced. However, the connections can be foreseen in the logarithmic derivative version of the Schwartz derivative, and a few comments can already be made.

For example, the conclusion in equation (87) follows from observing that a solution to

$$y''(z) = 0$$

could be $az + b$ for two constants a and b, that a second solution could be $cz + d$ with $ad - bc \neq 0$, and the fractional linear quotient follows.

Given the privleged status of exponentials with respect to differentiation, we have already seen in equation (89) that $\{\exp(\lambda z), z\} = -\lambda^2/2$, whereby complex exponentials generate constant Schwartz derivatives. How does this reconcile with the case where $\lambda = 0$ where this might lead us to think that just constants have a zero Schwartz derivative?

Again we need to consider that the most general form of the solution of the associated Ricatti equation is a fractional linear combination of $\exp(\lambda z)$ and $\exp(-\lambda z)$. But $\lambda = 0$ represents the confluent case where the two solutions are $\exp(\lambda z)$ and its derivative with respect to λ, namely $z \exp(\lambda z)$. The limit gives a fractional linear combination of 1 and z, the Möbius transformation already seen.

A somewhat similar discussion could be based on equation (88) where both $n = 1$ and $n = -1$ give zero Schwartz derivatives, but it suffices to say that these two exceptions are already fractional linear transformations. Aside from that, the denominator of $\{z^n, z\}$ tends to zero with large z, so it is again possible to wonder about Schwartz derivatives which are small, but not necessarily zero.

In turn it could be asked, what it means to have a mapping which is almost, but not quite, single valued? The example of the exponential shows that it could introduce a very long periodicity, with an anomaly which concentrates around infinity. Powers put a definitive multiple valuedness right at the origin, which of course extends right out to infinity. We could also ask, "Can singlevaluedness be assured, at least in a finite region?" It is a question which can be better answered later on.

That singlevaluedness is a delicate and unstable concept is already apparent in the implicit polynomial which was used to obtain its properties. At the level of polynomials, any addition whatsoever will create additional roots which may possibly be located far away, but whose vestiges could nevertheless be far reaching. Consider $\varepsilon z^2 + z + 1 = 0$, for example. An extremely flat parabola has been added to a straight line which, although not changing the existing zero by much, introduces a new root near infinity.

4 Iterated functions

The use of zeroes to obtain a factored form for a polynomial is well known, as is the expansion of the coefficients as symmetric functions of the roots. If

$$\begin{aligned} P(z) &= (z - z_1)(z - z_2)\dots(z - z_n) && (92) \\ &= z^n + a_1 z^{n-1} + \cdots + a_n && (93) \end{aligned}$$

then

$$\begin{aligned} a_1 &= z_1 + z_2 + \cdots + z_n && (94)\\ a_2 &= z_1z_2 + z_1z_3 + \cdots + z_{n-1}z_n && (95)\\ a_3 &= z_1z_2z_3 + z_1z_2z_4 + \cdots && (96)\\ \cdots & & & \\ a_n &= z_1z_2z_3\cdots z_n && (97)\end{aligned}$$

In particular, if any $z_i = 0$, then $a_n = 0$ and so on for additional vanishing roots.

Of course, the reverse process, determining the zeroes given the coefficients, is more difficult and is the subject of much numerical analysis.

4.1 fixed points and their stability

Another characteristic of polynomials is their fixed points. They are points for which

$$P(z) = z \qquad (98)$$

By defining an auxiliary polynomial $Q(z) = P(z) - z$, the problem of finding fixed points is turned into finding zeroes, to which the previous analysis applies. Although zeroes and fixed points are expected to be different, their polynomials both have the same symmetric functions, with the exception of c_{n-1}, which is the coefficient of z and which differs by 1 between the two polynomials.

Zeroes play an important role in the multiplication of polynomials whereas fixed points dominate the iteration of polynomials. Besides identifying points where nothing changes on account of a polynomial mapping, the derivative at the fixed point describes the behavior of nearby points. That is because a Taylor's series expansion about the fixed point calls for a contraction or an expansion according to the absolute value of the derivative, extending over a region whose size depends on the remaining terms in the series.

According to the size of the derivative, fixed points are classified as

unstable	$\|f'\| > 1$
neutral	$\|f'\| = 1$
stable	$\|f'\| < 1$
superstable	$\|f'\| = 0$

The principal characteristics of iteration can be summarized as:

1. Fixed points persist,
2. Critical points proliferate,
3. Stability becomes ever more pronounced,
4. Critical points respond to fixed point stability.

For the first item it suffices to note that if $P(x) = z$, then $P(P(z)) = P(z) = z$ and so on for any further application of P.

For the second item, the chain rule for derivatives is pertinent. Set $G(z) = P(P(z))$ and observe that

$$\frac{d}{dz}P(P(z)) = P'(P(z))\ P'(z). \qquad (99)$$

Therefore, if z is a point for which $P'(z) = 0$, its effect as the right hand factor in equation (99) is to make $G'(z) = 0$, indicating a critical point for the composite function. So it is that critical points persist, just as fixed points do.

But suppose that z maps into a point $P(z) = y$ for which $P'(y) = 0$. Then $G'(z) = 0$ because of the left hand factor in equation (99). So all the counterimages of any of the function's critical points turn up as critical points for the iterated function, which accounts for the claimed proliferation.

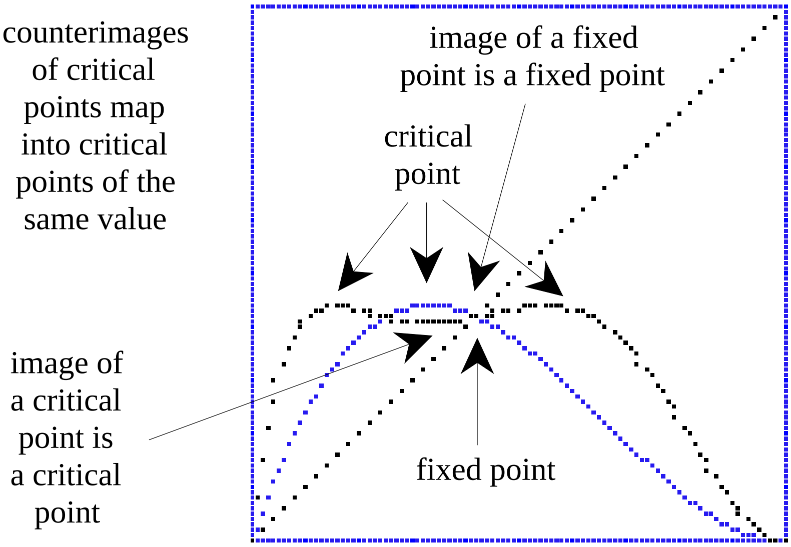


Figure 9: critical points and fixed points of a map and its composite.

The chain rule also describes the behavior of the derivatives considered as multipliers at fixed points. Namely, their values at each point in a trajectory $z, P(z), P(P(z)), \ldots$ multiply. If the point is fixed this product is a power and so a stable fixed point becomes ever more stable, neutral points remain neutral, and unstable points become increasingly unstable. This accounts for the intensification of the stability parameters of as fixed point, but just looking at a formula does not always convey the severity of its consequences. A factor as small as 2 becomes 1,024 after just ten iterations.

The extreme flattening of a function around its stable fixed points and accompanying steepness in the vicinity of unstable fixed points reflects itself in attracting critical points towards the stable points and repelling them from unstable points.

Imaging and counterimaging define trajectories in the complex plane. There is no reason for a series of images to close, but if it does, then there is a sequence of values which must repeat forever after.

The part which repeats is a *cycle*, the sequence leading into the cycle is a *transient*. The cycle is stable or not according to the magnitude of the product of all the derivatives taken around the cycle, the same whatever the reference point as long as the complete cycle is consulted.

There is likewise no guarantee that the set of counterimages, even of a fixed point or of a cycle, closes. Generally it will not, especially in view of the number of roots of the polynomial establishing the counterimage. In fact, the collection of counterimages will resemble a rooted tree, which offers the opportunity of introducing an invariant polar coordinate system for the tree. The radius is the number of iterations, the angle divides the previous arc by the number of roots and puts them down in some order – say by their phases.

The counterimage tree, carried to extreme, is infinite. However it doesn't fill up the plane, but more nearly an image of the unit circle.

For an unstable fixed point this collection of counterimages is called a *Julia set.* The classical studies of iteration were performed around the year 1920 by Gaston Julia and the somewhat older Pierre Fatou, and were related to a prize offered by the French Academy for an earlier problem of Cayley. In the forties, Carl Ludwig Siegel made an important contribution to the theory of neutral fixed points, but still more recently computer processing and computer graphics have permitted experimental work which has stimulated the theory even further, beginning with the work of Benoit Mandelbrot during the seventies.

4.2 Mandelbrot set for second degree polynomials

Linear mappings and fractional linear mappings have the structure which we have already seen. The next most complicated, in terms of polynomial degree and hence in number of fixed points and their stability, are quadratic mappings:

$$P(z) \quad = \quad az^2 + bz + c. \tag{100}$$

However, the use of a scale factor and shift of origin can transform any quadratic polynomial to the form

$$w \quad = \quad z^2 + c. \tag{101}$$

which depends on one single complex parameter. The fixed points are

$$\begin{aligned} z &= z^2 + c & (102)\\ z &= \frac{1 \pm \surd(1-4c)}{2}. & (103)\end{aligned}$$

at which the derivatives are $2z$, or

$$w' \quad = 1 \pm \surd(1-4c). \tag{104}$$

Unless $c = \frac{1}{4}$, where the fixed points are both neutral, one is stable, the other unstable (or maybe both are unstable).

The critical points are where

$$w' = 2z = 0.$$

so there is just one, at the origin.

Counterimages satisfy

$$\begin{aligned} w &= z^2 + c & (105)\\ z &= \pm\surd(w-c) & (106)\end{aligned}$$

and so are negatives of one another (as is evident from the defining equation).

Since everything depends on this one number, c, consider values for which the fixed points can be neutral, the dividing line between stability and instability. There set $w' = e^{i\phi}$:

$$\begin{aligned} e^{i\phi} &= 1 \pm \surd(1-4c) && (107)\\ (e^{i\phi}-1)^2 &= 1-4c && (108)\\ c &= \frac{1}{4}\left(1-(e^{i\phi}-1)^2\right) && (109)\\ &= \frac{1}{4}(1-e^{2i\phi}+2e^{i\phi}-1) && (110)\\ &= \frac{1}{4}(2e^{i\phi}-e^{2i\phi}) && (111)\end{aligned}$$

This is the sum of two contrarotating vectors, one double the length of the other and the shorter spinning at twice the angular velocity.

The figure is a cardioid, the interior of which contains values of c for which there is a stable fixed point.

Passing to the iterated function,

$$\begin{aligned} w &= (z^2+c)^2+c && (112)\\ &= z^4+2cz^2+c^2+c, && (113)\end{aligned}$$

and using long division,

$$z^4+2cz^2+c^2+c = (z^2-z+c)(z^2+z+(c+1)), \qquad (114)$$

so the iterate still has the fixed points of the primitive function, and two more besides, which are:

$$z_f = \frac{-1 \pm \surd(1-4(c+1)}{2} \qquad (115)$$

To get the stability of the iterate, we need its derivative

$$w' = 4z(z^2+c) \qquad (116)$$

at the new fixed points, where

$$\begin{aligned} z^2+c &= -(z+1) && (117)\\ z^2+z &= -(c+1). && (118)\end{aligned}$$

That results, in sequence, in

$$\begin{aligned} w'_f &= -4z_f(z_f+1) && (119)\\ &= 4(c+1). && (120)\end{aligned}$$

To get the contour of neutrality this derivative must have modulus 1, leaving

$$c = -1+\frac{1}{4}e^{i\phi} \qquad (121)$$

which is a circle at center -1, radius $\frac{1}{4}$, which has to be joined with the previous cardioid. It extends leftward to $-\frac{3}{4}$, so they are tangent.

All this calculation can be repeated for longer and longer cycles, although the algebra becomes ever more intricate and quickly lies beyond solution by formula.

The locus of stability can be located by more direct numerical processes, and in its totality forms the boundary of the Mandelbrot set. The higher loci will still resemble cardioids or circles, although they interfere with one another sufficiently that none of them will have that exact form although approximating it recognizably.

The Mandelbrot set is a map describing the stability of the fixed points and cycles of $w = z^2 + c$, in terms of the single complex parameter c. Because the results can be shown visually on a single sheet of paper, this function has been a popular one to study. Of course, other functions could be studied by the same procedure, without a corresponding graphical visualization, although the term "Mandelbrot set" could still be used.

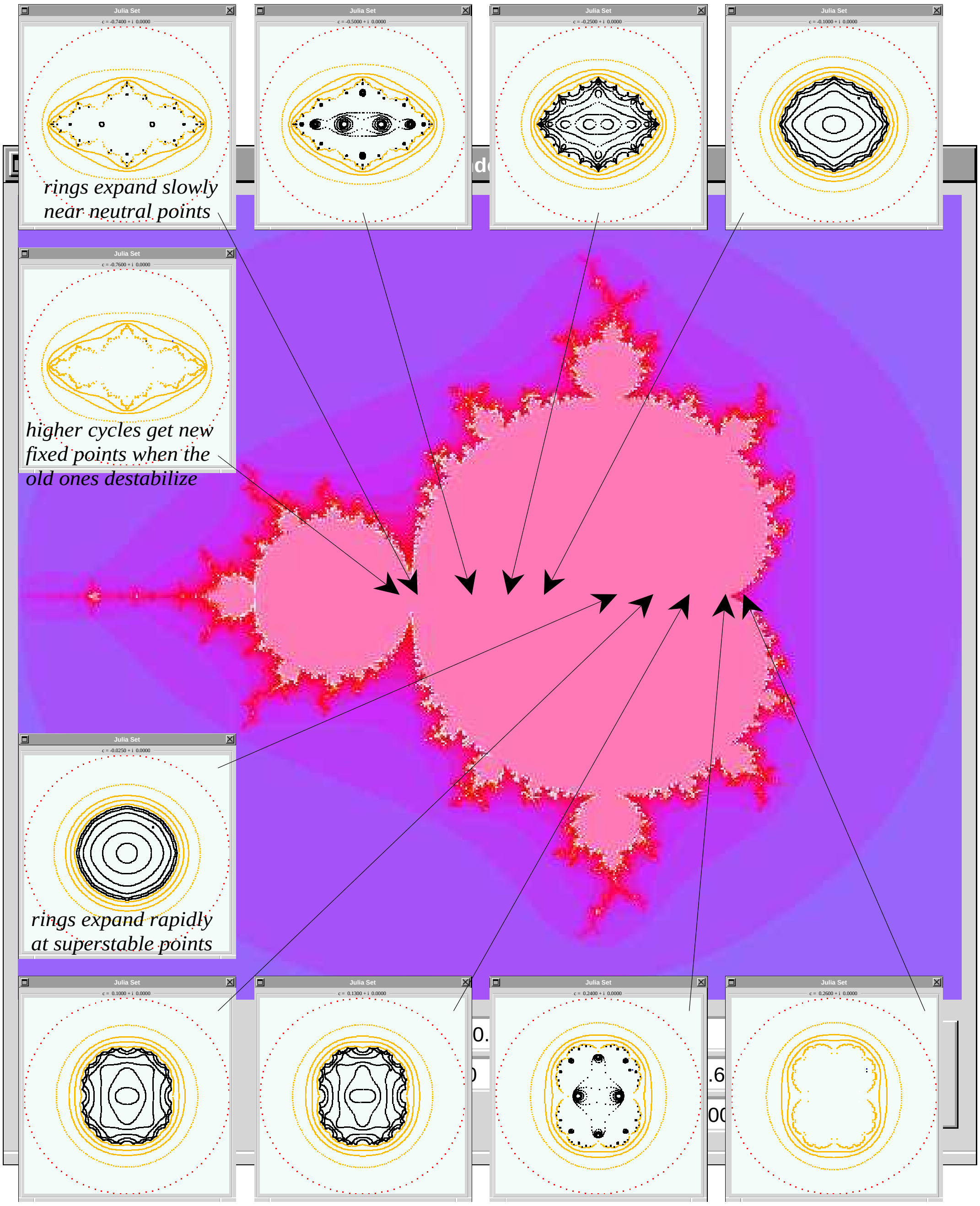
Also bear in mind that *each point* in the Mandlebrot plane represents a different complex function, whose values and the values of whose iterates can also be graphed on a sheet of paper. The contours for high iterates of $z^2 + c$ show a step function which jumps from one value to another at points which are members of the Julia set. The number of steps depends on the location of c within the Mandelbrot set, but more precisely in which of the little bubbles it lies.

Figure 10 contains an image of the Mandelbrot set, surrounded by a sequence of insets showing the behavior of the iterates of $w = z^2 + c$ for some values of c lying along the real axis. Of coures, when $c = 0$, which sits at the center of the stable cardioid, the figure is a parabola, flattening more and more inside the unit circle, while rising more and more rapidly outside.

The vignettes show the effect of iteration backwards, since the contours have been produced by computing counterimages of a small circle surrounding the origin. Correspondingly, a "small" circle surrounding infinity, $|z| = 2$ has been counterimaged to produce a series of contours approaching the Julia set from the outside.

It might have been better if the circle surrounding 0 had surrounded the stable fixed point instead, but this version reveals some interesting aspects of the mapping process. The counterimage depends on a square root and degenerates when the argument of the square root vanishes, which is when the radius of the circle reaches the absolute value of c. Up to that point, the circle deforms into ellipses and figures with additional Fourier components, but then a critical point has been reached where the counterimage lacks definition. Beyond that value the circle splits into two components, and then it may split further as additional critical points are encountered.

In any event, the expanding counterimages approach the Julia set from the inside, which has to be confined between the two sequences of counterimages, approaching from the inside and from the outside. Of course, there may not be an "inside," as would happen outside the Mandelbrot set, with both finite fixed points unstable.



Julia sets running from c = -0.76 to c = 0.26

November 8, 1996

Figure 10: A survey of points in the Mandelbrot set, showing their corresponding Julia sets.

Creating a polynomial with certain fixed points is just the same as creating a polynomial with particular zeroes: write

$$P(z) \quad = \quad (z - f_1)(z - f_2) \ldots (z - f_n) + z. \tag{122}$$

When that is done, the derivatives have all been assigned, because

$$P'(z) \quad = \quad 1 + \sum_j \prod_{\text{omit } j} (z - f_j) \tag{123}$$

$$P'(f_i) \quad = \quad 1 + \prod_{\text{omit } i} (f_i - f_j). \tag{124}$$

Some ingenuity could be exercised in setting aside one of the fixed points and moving it around to get a desired derivative at some other fixed point, but a better way to make a certain fixed point have a desired modulus of stability would be to use Hermite interpolation. For polynomials of even degree, any derivatives at all can be given to half of the data, the other half being the actual locations of the fixed points.

There is a precaution to be observed when passing between zeroes and fixed points in describing a polynomial. Infinity is not a zero of a polynomial because the dominant term will always increase towards infinity, even though the analysis of an essential singularity may lead to such a zero; think of $\exp(-\infty)$. On the other hand, infinity is a fixed point for polynomials, at least on the Riemann sphere where there is only one infinity.

The precaution consists in counting fixed points accurately; a polynomial of degree d has exactly d zeroes, and apparently d fixed points. But the number is actually $d + 1$ because of having to include infinity.

5 Contour Integrals

The basic differential relations for complex variables are the Cauchy - Riemann equations. For a function $f(z) = u(x, y) + iv(x, y)$ of the complex variable $z = x + iy$,

$$\frac{\partial u}{\partial x} \quad = \quad \frac{\partial v}{\partial y} \tag{125}$$

$$\frac{\partial v}{\partial x} \quad = \quad -\frac{\partial u}{\partial y}, \tag{126}$$

whose extension to the requirement that u and v be harmonic has already been discussed in terms of finite differences. But the Cauchy - Riemann equations can be approximated directly in finite differences

$$\frac{u(x + \Delta x, y) - u(x, y)}{\Delta x} \quad = \quad \frac{v(x, y + \Delta y) - v(x, y)}{\Delta y} \tag{127}$$

$$\frac{v(x + \Delta x, y) - v(x, y)}{\Delta x} \quad = \quad \frac{u(x, y) - v(x, y + \Delta y)}{\Delta y} \tag{128}$$

Suppose that f has been defined along the real axis and that we want to know about its behavior in the upper half plane, $y > 0$. Then

$$u(x, y + \Delta y) \quad = \quad u(x, y) - \frac{\Delta y}{\Delta x} \{v(x + \Delta x, y) - v(x, y)\} \tag{129}$$

which means that to get u on a line just above the real axis, it should be decremented everywhere by horizontal differences in v, weighted by the ratio of Δy to Δx, which could be 1. The new u depends only on the old u and the v differences, both taken along the real axis. But similar results would apply to any contour.

Meanwhile v is changing according to

$$v(x, y + \Delta y) \quad = \quad v(x, y) + \frac{\Delta y}{\Delta x} \left\{ u(x + \Delta x, y) - u(x, y) \right\}, \tag{130}$$

to which similar comments apply.

Taken together, these two equations tell how to work up the positive y-axis strip by strip. Somehow that says that an analytic function is defined everywhere else by knowing its two halves on the real axis. Actually some kind of line drawn anywhere else ought to work, but this approach avoids worrying about corners and the like.

This is a kind of initial value problem, where both u and v are specified and are leapfrogged to get values somewhere else. But numerical values could be given for u and symbolic values for v. In the next strip, the v's would get numerical increments, the u's symbolic increments, but of first degree in v. Repeating, first degree expressions keep recurring until at last, in some final strip some exact values could be proposed. Then some linear equations woud have to be solved to find out what the values of the v's should have been to get the required values. The matrix just has $+1$'s or -1's because of the sums and differences, and has to be raised to a power to get the number of parallel lines used in climbing.

Rather than flipping between u's and v's, the fact that each is harmonic can be used to work up the ladder of strips using averages. Then a row depends on two predecessors, so initial values and derivatives could be used (derivatives of u's gotten from values of v by using one of the Cauchy - Riemann equations). Also, instead of strips, think of any area, such as a rectangle. Points outside the rectangle aren't part of the differential equation but define averages, so one could try to accomodate the interior to a prescribed boundary.

Reasoning with finite differences means getting the matrices in detail and getting a good symbolic description. Usually ordinary calculus is applied to finding values at one place in terms of another.

First, line integrals.

In real analysis, an integral (at least a Riemann integral) is defined

$$\int_a^b f(x)dx \quad = \quad \lim_{\Delta x \to 0} \sum_{i=0}^{n} f(x_i) \Delta x_i \tag{131}$$

where an interval is subdivided and the function averaged over the points of subdivision. Much anguish can be expended over where to evaluate the function, how large to take the intervals, and so on. To do the same thing with an analytic function one could write

$$\int_a^b f(z)dz \quad = \quad \lim_{\Delta z \to 0} \sum_{i=0}^{n} f(z_i) \Delta z_i \tag{132}$$

with the proviso that $\sum \Delta z_i = b - a$ and subject to the ambiguity that many sequences of Δz_i's lead from a to b. If that mattered, the integral would depend explicitly on the path. Thus the first order of business is to decide whether it does or it does not.

The dependence can be checked out beginning with a $\Delta x \Delta y$ rectangle. From the definition, we need

$$\begin{aligned} f(z)\Delta z &= (u+iv)(\Delta x + i\Delta y) && (133)\\ &= (u\Delta x - v\Delta y) + i(u\Delta y + v\Delta x) && (134) \end{aligned}$$

With Δz decomposed into its real and imaginary increments, consider the alternative orders in which they can be taken. In keeping with the spirit of Riemann integration, the function should be evaluated at the beginning of each interval of subdivision and not be allowed to retain the same value from one step to the next.

Using Δx, then Δy we get

$$f(z)\Delta x + if(z+\Delta x)\Delta y \quad = \quad u\Delta x + iv\Delta x + i(u + \frac{\partial u}{\partial x}\Delta x)\Delta y - (v + \frac{\partial v}{\partial x}\Delta x)\Delta y. \qquad (135)$$

The other order, Δy, then Δx, leads to

$$if(z)\Delta y + if(z+\Delta y)\Delta x \quad = \quad iu\Delta y - v\Delta y + (u + \frac{\partial u}{\partial y}\Delta y)\Delta x + i(v + \frac{\partial v}{\partial y}\Delta y)\Delta x. \qquad (136)$$

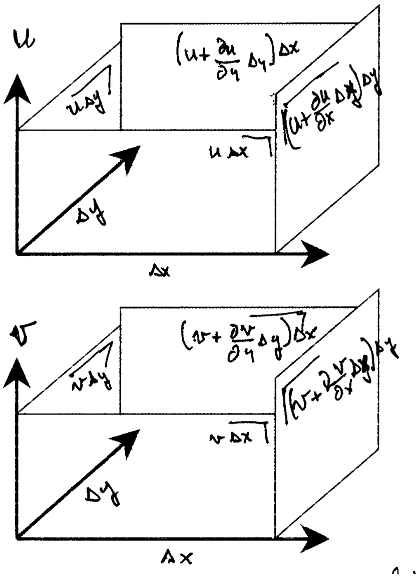

u
$(u+\frac{\partial u}{\partial y}\Delta y)\Delta x$
$u\Delta y$
$(u+\frac{\partial u}{\partial x}\Delta x)\Delta y$
$u\Delta x$
Δy
Δx
v
$(v+\frac{\partial v}{\partial y}\Delta y)\Delta x$
$v\Delta y$
$(v+\frac{\partial v}{\partial x}\Delta x)\Delta y$
$v\Delta x$
Δy
Δx

Figure 11: An integral can be shifted across a rectangle without changinging its value.

It is now a matter of comparing

$$\begin{aligned} \frac{\partial u}{\partial x} &\quad \text{with} \quad \frac{\partial v}{\partial y} && (137)\\ -\frac{\partial v}{\partial x} &\quad \text{with} \quad \frac{\partial u}{\partial y} && (138) \end{aligned}$$

and the Cauchy - Riemann equations say that they are the same. Of course, this is based on one definition of a Riemann integral, in which the function at the left endpoint is multiplied by the length of the interval, and must be corrected at the midpoint if the interval is split in two. Otherwise terms proportional to the arc length will be neglected and the limiting sum will not be correct. In general, function values anywhere in the interval can be used in the definition, but first order effects will accrue if values from outside the interval creep in.

In general, the contour can be shifted across any rectangle for which a linear approximation to the function is valid, which would exclude any singularities or branch points – points where the function is not invertible and so making its definition suspect.

A consequence of path-shifting is that when the endpoint matches the initial point, the value of the integral is truly zero. Again the requirement is that the contour never has to cross over a singularity.

A useful side effect of the way Riemann integrals are defined is that the function can be defined in different ways in different places – just so long as the definitions overlap sufficiently that the Cauchy - Riemann equations hold for both definitions.

A common source of multiple definitions is to have two different Taylor's series with finite radii of convergence, centered on different points; the definitions can be compared where their disks of convergence overlap.

The vanishing of a closed contour integral is usually verified by using Green's formula, which is a useful technique in its own right. Note that the harmonicity condition makes the average over an area vanish, so that a double integral over a plane region should be zero except for boundary values which are nc

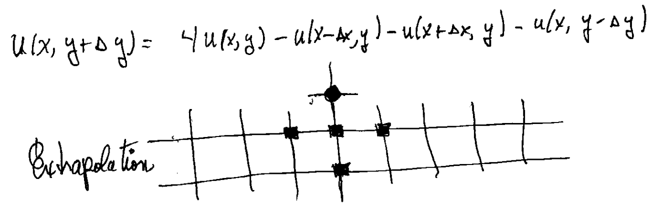


Figure 12: A harmonic function can be extrapolated beyond a neighborhood as long as there is enough information to approximate a derivative.

So add a rim to any region and fill it with values to get the zero average. To make this work, the sum over the interior – an area integral – must be the negative of the sum of the boundary values – a contour integral. In the form needed for complex variable theory, suppose that $u(x, y)$ and $v(x, y)$ are two real valued functions such as the real and imaginary parts of an analytic function. Further suppose that it is desired to calculate

$$\int\int_R \left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\right) dxdy$$

and that is possible (although not explicitly shown in the formula) to parameterize x and y as functions of t. To avoid confusion, write functions $x(\alpha)$ and $y(\beta)$ using separate variables, and rewrite the integral

$$\int\int_R \left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\right) \frac{dx}{d\alpha}\frac{dy}{d\beta} d\alpha d\beta$$

and then integrate. Note

$$\begin{aligned}\int_{\text{bottom}}^{\text{top}} \frac{\partial u}{\partial y}\frac{dy}{d\beta}d\beta &= u(\text{top}) - u(\text{bottom}) && (139)\\ \int_{\text{left}}^{\text{right}} \frac{\partial v}{\partial x}\frac{dx}{d\alpha}d\alpha &= u(\text{right}) - u(\text{left}). && (140)\end{aligned}$$

It remains to evaluate

$$\int [u(\text{top}) - u(\text{bottom})]\, dx$$

and to match it with

$$\int [v(\text{right}) - v(\text{left})]\, dy$$

to get

$$\int\int_R \left(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\right) dxdy = -\oint_{\partial R} (udx + vdy) \qquad (141)$$

Then, by using the Cauchy - Riemann equations in the area integral to make it vanish, the vanishing of the contour integral follows.

5.1 evaluation of integrals by using the residues at poles

The way to deal with singularities inside a contour is to wrap the contour around them so as not to include the singularities and then cancel parts of the contour which overlap and run in opposite directions. For example, consider a singularity at the origin and surround it with two circles, one large and the other small. By breaking the circles and connecting them by two close radii, an overall singularity-free contour can be created.

By taking very close radii, their integrals can be made to cancel, leaving he difference (because of running in opposite directions whilst counting counterclockwise contours as positive) of integrals over the circles. The outer circle can be almost anything else unless it is considered to surround infinity, whereas the circular form of the inner circle is important. Surrounding its singularfity symmetrically, it can be approximated by transferring to polar coordinates, whereupon it becomes

$$i\int_{\varepsilon}^{2\pi-\varepsilon} rf(z)e^{i\phi}d\phi.$$

Here it is a question of what the singularity is like. If $f(z)$ behaves like a power, or a sum of powers, we get terms like

$$i\int_0^{2\pi} r(re^{i\phi})^n e^{i\phi}d\phi = ir^{n+1}\int_0^{2\pi} e^{i(n+1)\phi}d\phi \qquad (142)$$

which is an integral of functions averaging to zero unless $n = -1$. For that term, the integral turns into $2\pi i$. Therefore isolated power-like singularities don't change the value of a contour integral enclosing them (just the power -1). Square roots and such things are another matter; the difference of the integrals along the radial arcs also has to be taken into account.

The *residue* of a function $f(z)$ at a point of singularity z_0 is defined by

$$r = \lim_{z\to z_0} (z - z_0) f(z - z_0), \tag{143}$$

which looks very much like the definition of a derivative. However, since z_0 is a point of singularity, there is no derivative there, and besides, the factor $(z - z_0)$ sits in the numerator, where its role is to cancel out a similar divisor in the makeup of the singularity rather than in the denominator where the definition of a derivative would place it.

If $f(z)$ had a power series expansion beginning with negative powers (a Laurent series) the residue would be the coefficient a_{-1} in that series. Therefore suppose

$$f(z) = \sum_{-n}^{\infty} a_i z^i \tag{144}$$

which implies that an integral around a contour enclosing some or all of its singularities would be

$$\oint f(z)dz = 2\pi i \sum_{\text{interior poles}} r_j. \tag{145}$$

The use of the residues of a complex function gives a way to evaluate many definite integrals, including what seem to be real integrals. The way to get a real definite integral is to close the half-plane above the real axis with a huge semicircle, and hope that the function vanishes sufficently rapidly as one rises in the plane. The integral over the semicircle then approaches zero as a limit as its radius increases, leaving the poles in the upper half plane to contribute their residues. There exist extensive integral tables which have been constructed in this way.

5.2 existence of derivatives of all orders

Another application is to deliberately introduce a singularity and then to evaluate the new integral. Consider

$$f(z_0) = \frac{1}{2\pi i} \oint \frac{f(z)}{z - z_0} dz$$

which has a simple pole at z_0 and residue $f(z_0)$.

This is *Cauchy's integral representation* of an analytic function, which is nothing more than an integral version of the observation that an analytic function averages its surrounding values. Here each value is given the weight $dz/(z - z_0) = d \ln(z - z_0)$. If $z - z_0$ is written in polar form, the weight is a combination of the reciprocal of the distance of an element of arc and its angular aperture. That normalizes all arcs to an equivalent distance (that's the factor r), and performs the average.

Exploiting the fact that the derivative of an integral is the integral of the derivative of its integrand,

$$\begin{aligned} f'(z_0) &= \frac{1}{2\pi i} \frac{d}{dz_0} \oint \frac{f(z)}{z - z_0} dz \\ &= \frac{1}{2\pi i} \oint \frac{f(z)}{(z - z_0)^2} dz \end{aligned}$$

which surely ought to converge as well as the original integral. Note that the minus sign arising from the negative power is cancelled by the sign of $-z_0$ in the denominator, which is the variable for this differentiation. Repeating the process turns up a whole infinite series of derivatives, which stands in contrast to what may happen for a real variable, that derivatives may eventually suffer discontinuities and other faults. In particular, there is no analytic function mimicking the peaks in the sawtooth function.

For convenience of reference, we could just collect the formulas for the first few derivatives:

$$\begin{aligned} f(z_0) &= \frac{1}{2\pi i}\oint \frac{f(z)}{z-z_0}dz & (146)\\ f'(z_0) &= \frac{1}{2\pi i}\oint \frac{f(z)}{(z-z_0)^2}dz & (147)\\ f''(z_0) &= \frac{2}{2\pi i}\oint \frac{f(z)}{(z-z_0)^3}dz & (148)\\ &\cdots \\ f^{(n)}(z_0) &= \frac{n!}{2\pi i}\oint \frac{f(z)}{(z-z_0)^{n+1}}dz & (149)\end{aligned}$$

Still another useful induced singularity makes use of the logarithmic derivative, supposing that there is a region containing only a finite number of zeroes and poles of $f(z)$. Then.

$$\frac{1}{2\pi i}\oint \frac{f'(z)}{f(z)}dz = \text{(number of zeroes)} - \text{(number of poles)}, \qquad (150)$$

both being counted with their respective multiplicities. Such a function could be the quotient of two polynomials. In fact, if f were a polynomial and the contour were a sufficiently large circle, the highest power would dominate the logarithmic derivative, provoking the conclusion that there were just as many zeroes inside, as the degree of the polynomial.

This result, being true for any polynomial with complex coefficients, lays to rest the question of whether anything new can be gotten from taking the square root of $-i$ (bearing in mind the humble origin of i itself). Neither x^2+i nor any other combination of powers and complex numbers needs anything beside complex numbers as roots. An interesting point here is the way that the contour integral transforms the result that a power has n roots into the fact that any polynomial of that degree also has roots, although they are not necessarily (real multiples of) roots of unity.

5.3 Liouville's theorem: a bounded analytic function is constant

Suppose that $|f(z)| < M$ and that Cauchy's integral formula is used to compare the values of $f(z)$ at two points. Then

$$f(z_2) - f(z_1) = \frac{1}{2\pi i}\oint \left\{\frac{f(z)}{z-z_1} - \frac{f(z)}{z-z_2}\right\}dz. \qquad (151)$$

A little rearrangement produces

$$\frac{f(z_2)-f(z_1)}{z_2-z_1} = \frac{1}{2\pi i}\oint \frac{f(z)dz}{(z-z_0)^2}\left\{\frac{z-z_0}{z-z_1}\right\}\left\{\frac{z-z_0}{z-z_2}\right\}, \qquad (152)$$

which could very easily be used in an alternative derivation of the Cauchy formula for a derivative. Now, choose a huge circular contour centered on the auxiliary point z_0, and note that $f(z)$ is still bounded however large the region. In Cauchy's formula the factor $1/(z - z_0) \approx 1/r$ gets the better of the situation and just compensates the $2\pi r$ in the arc length of the surrounding circle, but that just repeats what we already know, that the function is bounded.

But in the difference we are calculating, the common denominator introduces an additional power of r, leaving little doubt that the combination $(M \times r)/r^2$ approaches zero. Very well, take the limit, and arrive at the conclusion a bounded analytic function can only be a constant. That is the content of *Liouville's theorem.*

5.4 The maximum modulus principle

To get some actual inequalities to work with, Cauchy's integral formula ought to be subjected to absolute values (depending on all four combinations arising from positive maxima and negative minima as found in the real as well as the imaginary part of the complex function), This is most conveniently done in polar coordinates. Let the contour be a circle of radius ρ centered at the origin. Then, supposing that

$$M(\rho) = \max_{0 \leq \theta \leq 2\pi} f(\rho, \theta) \tag{153}$$

we get for the function

$$\begin{aligned} |f(0)| &\leq \frac{1}{2\pi} \int_0^{2\pi} |f(\rho, \theta)| d\theta && (154) \\ &\leq M(\rho), && (155) \end{aligned}$$

for its derivative,

$$\begin{aligned} |f'(0)| &\leq \frac{1}{2\pi} \int_0^{2\pi} \frac{|f(\rho, \theta)|}{\rho} d\theta && (156) \\ &\leq \frac{M(\rho)}{\rho}, && (157) \end{aligned}$$

and in general

$$\begin{aligned} |f^{(n)}(0)| &\leq \frac{n!}{2\pi} \int_0^{2\pi} \frac{|f(\rho, \theta)|}{\rho^n} d\theta && (158) \\ &\leq \frac{n! M(\rho)}{\rho^n}. && (159) \end{aligned}$$

This series of inequalities relates the value of the function (as well as its derivatives) at the center of a circle of analyticity to the maximum value on the circumference of the circle. Any bound would establish an inequality, but the bound used in this series of equations is optimal, given that it actually occurs somewhere on the perimeter of the bounding circle.

But the inequality can be read in both directions, in the sense that there must be a place on the circumference where the absolute value of the function is greater than (or at least equal to) the absolute value of the function at the center. That is because the integral expresses an average of certain data, yet an average can never be strictly greater than all its data. But the more inclusive "greater than or equal to" could actually hold, supposing that all the data were equal.

Saying "place" rather than "point" is intended to invoke enough continuity to avoid sets of measure zero (like individual points) while evaluating the integral. Presumably the analyticity of $f(z)$ provides this assurance.

Since the extreme case of equality in equation (155) reduces $|f(\rho,\theta)|$ to the real constant $M(\rho)$, equations (157) through (159) and onwards experience the same reduction, establishing equality along the line.

Supposing a Maclaurin series for $f(z)$

$$f(z) \quad = \quad \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n, \tag{160}$$

taking absolute values would lead to an inequality

$$|f(z)| \quad \leq \quad \sum_{n=0}^{\infty} \frac{|f^{(n)}(0)|}{n!} |z|^n. \tag{161}$$

Using the inequalities just established would weaken this last result, without saying exactly by how much. Substituting the extreme case equalities gives an exact replacement, but the series can be summed whether extreme or not:

$$|f(z)| \quad \leq \quad \sum_{n=0}^{\infty} M(\rho) \left(\frac{|z|}{\rho}\right)^n, \tag{162}$$

$$\leq \quad M(\rho) \frac{\rho}{\rho - |z|}. \tag{163}$$

which is consistent with what we have already learned at $z = 0$, but but thoroughly ineffective when $|z| = \rho$.

There are corresponding minimum inequalities, for functions which do not vanish within the circle; it is only necessary to apply the foregoing reasoning to the then finite-valued $g(z) = 1/f(z)$ in place of $f(z)$.

For a single circle, these results are of limited utility, but it is easy enough to extend them to more general domains, as long as they accomodate the intermediate circles needed in the following argument. The main concern is with taking the bounding modulus from anywhere within the domain, rather than just from the boundary.

There are two possibilities. Suppose the greatest modulus sits at an interior point rather than on the boundary. Enclosing it within a circle small enough to still lie in the domain, we observe that there must be a point on the circumference of that little circle with a still greater modulus, which would be a contradiction, or all the values on that little circle would equal the purportedly greater modulus. So enlarge the circle until it makes contact with the boundary. If contact is made through the equality alternative, the contradiction consists in having a greater modulus than the one taken from the boundary (even though it isn't necessarily a circle any more).

The final result of all this reasoning is the *maximum modulus principle*, that the maximum absolute value of an analytic function lies on the boundary of any domain of analyticity.

5.5 Schwartz's lemma

Just as there are variants to Cauchy's integral formula yielding useful results, applying the bounds already obtained to more specific situations can give equally specific results. One idea is to examine the stability of a fixed point, which we know is influenced by its derivative. By translating and scaling, most functions can be made to vanish at zero with unit derivative; just introduce the new function

$$g(z) = \frac{f(z)-f(0)}{f'(0)}, \tag{164}$$

and if necessary, forget the scaling until it becomes convenient to use it.

This preparation yields an analytic function vanishing at the origin, precluding the singularity of $f(z)/z$ which would otherwise be found there. Again, relative to a circle of radius ρ,

$$\left|\frac{f(z)}{z}\right| \leq \frac{M(\rho)}{\rho} \tag{165}$$

$$|f(z)| \leq \frac{M(\rho)}{\rho}|z| \tag{166}$$

$$\leq M(\rho)\left|\frac{z}{\rho}\right|, \tag{167}$$

or even,

$$\frac{|f(z)|}{M(\rho)} \leq \frac{|z|}{R}, \tag{168}$$

which is a sort of concavity result for an analytic function. which can never increase faster than by the first power while runnning from zero towards a known bound, although constant multiples can just barely meet the challenge (if two complex numbers have the same modulus, their quotient is a complex number of absolute value 1). This result is known as *Schwartz's lemma.*

The idea behind Schwartz's lemma can be followed in two directions. If zero is a superstable fixed point the maximum modulus inequality could be applied to $f(z)/z^2$ get

$$\left|\frac{f(z)}{z^2}\right| \leq \frac{M(\rho)}{\rho^2} \tag{169}$$

$$\frac{|f(z)|}{M(\rho)} \leq \left(\frac{|z|}{\rho}\right)^2, \tag{170}$$

so a superstable point is superattractive relative to the pertinent geometry.

The other direction would be to check for derivatives, which for an ordinary f with a fixed point at 0 would read

$$\left|\left(\frac{f(z)}{z}\right)'\right| \leq \frac{M(\rho)}{\rho^2} \tag{171}$$

$$\frac{|zf'(z)-f(z)|}{M(\rho)} \leq \left(\frac{|z|}{\rho}\right)^2 \tag{172}$$

5.6 residues and the stability of fixed points

Residues can show up in defining fixed points for rational functions. Suppose that the function r,

$$r(z) = \frac{P(z)}{Q(z)}, \tag{173}$$

is a quotient of two polynomials. If Q has lesser degree than P, division may be used to write

$$r(z) = g(z) + \frac{p(z)}{q(z)} \tag{174}$$

wherein p and q have the desired property.

Consider a "resolvent" $R(z)$ defined by

$$R(z) = \frac{1}{r(z) - z} \tag{175}$$

which will have poles wherever $r(z)$ has fixed points. As previously discussed, $r(z) - z$ can be factored, allowing the writing of

$$\begin{aligned} R(z) &= \frac{1}{g(z) + \frac{p(z)}{q(z)} - z} && (176)\\ &= \frac{q(z)}{g(z)q(z) + p(z) - zq(z)}. && (177)\end{aligned}$$

This denominator has greater degree than the numerator, so it can be factored and R can be written as a sum of partial fractions (due allowance should be made for repeated factors)

$$R(z) = \sum \frac{a_j}{z - z_j} \tag{178}$$

for fixed points of r, namely z_j. The a_j's are residues of R, but we are interested in the stability of $r(z)$.

$$\begin{aligned} r(z) &= \frac{1}{R(z)} + z && (179)\\ r'(z_f) &= \lim_{z \to z_f} \frac{1}{z - z_f} \left(\frac{1}{R(z)} + z \right) && (180)\\ &= 1 + \frac{1}{a_j}. && (181)\end{aligned}$$

So stability depends on negative (but not too negative) a_j's.

5.7 representation of a function by a power series

Evidently polynomials and rational fractions – quotients of polynomials – are analytic functions, moreover defined in the whole complex plane if "infinity" is accepted as a number. There is not much trouble in extending a polynomial to an infinite series provided that its convergence is checked; a power series can be expected to have a radius of convergence. In fact, we have used he complex exponential from the beginning to get such things as Euler's formula or the solution of systems of differential equations with constant coefficients.

But the concept of a function is quite general, since the only thing required is a certain kind of set of pairs of values. That creates the problem of representing the function in a manageable form, which a table of numbers is not. Some help is given by the Cauchy integral formula,

$$f(z_0) = \frac{1}{2\pi i} \oint \frac{f(z)}{z - z_0} dz \tag{182}$$

to the extent that the value of the function at one place can be compared to its value at another. Write

$$
\begin{aligned}
f(z_1) &= \frac{1}{2\pi i}\oint \frac{f(z)}{z+(z_0-z_1)-z_0}dz && (183)\\
&= \frac{1}{2\pi i}\oint \frac{f(z)dz}{(z-z_0)-(z_1-z_0)} && (184)\\
&= \frac{1}{2\pi i}\oint \frac{f(z)dz}{(z-z_0)\left(1-\frac{z_1-z_0}{z-z_0}\right)} && (185)\\
&= \frac{1}{2\pi i}\oint \frac{f(z)dz}{z-z_0}\left(1+\frac{z_1-z_0}{z-z_0}+(\frac{z_1-z_0}{z-z_0})^2+(\frac{z_1-z_0}{z-z_0})^3+\cdots\right) && (186)\\
&= \frac{1}{2\pi i}\oint \frac{f(z)dz}{z-z_0}+(z_1-z_0)\frac{1}{2\pi i}\oint \frac{f(z)dz}{(z-z_0)^2}+(z_1-z_0)^2\frac{1}{2\pi i}\oint \frac{f(z)dz}{(z-z_0)^3}+\cdots && (187)\\
&= f(z_0)+(z_1-z_0)f'(z_0)+\frac{(z_1-z_0)^2}{2}f''(z_0)+\cdots && (188)
\end{aligned}
$$

which is the Taylor's series for $f(z_1)$ relative to the point z_0. The requirement for the convergence of the geometric series is that $|z-z_0| > |z_1-z_0|$, which means that z_1 is closer to z_0 than z_0 is to the boundary. Thus the boundary can be taken as far away as $f(z)$ is still analytic – up to a pole or a branch point or whatever.

Is it safe to conclude that a function is zero if its Taylor's series has all zero coefficients? For example, the function $e^{-1/x}$ has derivative $(1/x^2)e^{-1/x}$ in which the zero due to $e^{-1/x}$ overwhelms the pole due to $1/x^2$ – for positive x, that is. Consider e^{-1/x^2} which has zero derivatives from both sides at zero. But along the imaginary axis, that fails. Such functions cannot be analytic.

But these examples depend upon a failure of analyticity, so it seems that a zero series for an analytic function really represents the zero function and nothing else.

If not all coefficients are zero, but some of the leading coefficients vanish, then there is a zero at the expansion point due to a factor $(z-z_0)^k$. Then by continuity there is a small disk surrounding z_0 where the function can't be zero either. Hence the zeroes of an analytic function may be multiple, but their degree is finite unless the function is actually zero.

Stated in another form, zeroes may occur in finite clusters, which must be isolated from one another, and the function can still be analytic. This excludes analyticity at a limit point of zeroes, or at zeroes of infinite multiplicity.

The same conclusions hold when the zeroes arise from the difference of two analytic functions (Can the difference of two functions be analytic when the functions themselves are not? Consider $\frac{1}{2}(f+\bar{z})$ and $\frac{1}{2}(f-\bar{z})$.) Therefore if the functions coincide over a set which has a limit point, they are identical in any domain including the limit point. (Can infinity be a limit point? In other words if two functions agree for integer values, are they identical: Try adding $\sin z/2\pi$ to one of them.)

5.8 the monodromy principle

Although Taylor's series have radii of convergence, the circle in which they are defined can sometimes be shifted, which means that sometimes it is possible to work one's way around singularities. One way is to write $z-z_0=(z-z_1)+(z_1-z_0)$, which is best seen by turning a Maclaurin series into a Taylor's series. If

$$\begin{aligned} f(z - z_0) &= \sum c_n (z - z_0)^n && (189) \\ &= \sum \sum c_n \frac{n!}{i!(n-i)!} (-1)^i z^{n-i} z^i && (190) \end{aligned}$$

The transformation has a matrix representation

$$\begin{bmatrix} 1 \\ z - z_0 \\ (z - z_0)^2 \\ \cdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ -z_0 & 1 & 0 & 0 & 0 & \cdots \\ z_0^2 & -2z_0 & 1 & 0 & 0 & \cdots \\ \cdots \end{bmatrix} \begin{bmatrix} 1 \\ z \\ z^2 \\ \cdots \end{bmatrix}$$

which, as a triangular matrix is readily inverted by shifting by $-z_0$ rather than z_0.

The expectation is that shifting and reexpansion can be used to circumvent a singularity. It should not be thought that bypassing a singularity by two routes should always give the same result; but two routes which don't surround a singularity must. This is the *monodromy theorem*; the process is called *analytic continuation.*

Sometimes an attempt at continuation will strike a barrier, no matter what route is taken. That would lead to a region densely bounded by singularities.

If all the coefficients of a Taylor's series vanish at a point of analyticity, and the radius of its disk of convergence is non-zero, then the function not only vanishes throughout that disk, but everywhere else that can be reached by analytic continuation. Conversely, if a function vanishes on any convenient point set large enough to take derivatives, or to apply Cauchy's integral formula, or to perform a continuation, the Taylor's series, must be zero.

By taking differences, two functions which coincide in a similar fashion must turn out to be identical.

One way of making a continuation that sometimes works is to observe that $\bar{f}(\bar{z})$ is an analytic function whenever f is, in spite of the complex conjugations. The two conjugates cancel one another, so to speak. In terms of a Taylor's series, it is a new function of the same variable z, whose coefficients are the complex conjugates of those of the old function.

Then, if a region straddles the real axis, and the function is real for real arguments, we have two functions with an arc of common values. Thus values from the upper half plane can be reflected into the lower half plane and taken as an analytic continuation of the original function.

6 Periodic functions

Given a general familiarity with the periodicity of trigonometric functions it is narural to look for periodic functions of a complex variable. However, the trigonometric functions themselves won't do because they turn into hyperbolic functions when given an imaginary argument; for example $\cos(i\theta) = \cosh(\theta)$. On the other hand, trying something like $\sin(x)\sin(y)$ readily exhibits two-dimensional periodicity, but trying to combine x and y into the complex variable $z = x + iy$ runs afoul of the Cauchy-Riemann equations, which reveal that such a product is not analytic.

Those two examples are only hasty attempts at guessing functions that could be periodic in the complex plane, whereas a more successful approach would be to start with a periodic configuration like a square grid, followed by taking averages over the grid.

Actually any two vectors will suffice to establish a grid, as long as they are not parallel. By tradition they are called 2ω and $2\omega'$ so that the lattice consists of the assemblage $\Omega = \{m\omega + n\omega'\}$ wherein m and n range over all the signed integers. As a matter of notation, it is convenient to use Ω' when the origin $(0, 0)$ has been dropped from Ω.

6.1 Eisenstein series

Next, there is a question of what should be averaged, to get a convergent double sum containing the order of n^2 summands, for a large integer n. Inverse powers of the distance, z^{-k} come to mind, with k large enough for the size of n^{-k} to compensate the number of terms yet remaining as one approaches infinity.

An *Eisenstein series*,

$$S_k(z) \quad = \quad \sum_{\Omega} \frac{1}{(z - z_{ij})^k}, \tag{191}$$

is an uncomplicated way to get a function for which $S_k(z + m\omega + n\omega') = S_k(z)$. It has poles when $z\varepsilon\Omega$.

When $k \geq 3$ the size of far terms falls off fast enough that there is no problem with the convergence of the double sum, whereas the series for S_1 can reasonably be expected to diverge. The case $k = 2$ sits on the borderline, but the series can be rescued by noticing that the partial sums for one value of z differ by little from those for another value of z.

But it is important to work with the difference of the two series from the outset, and not try to take the difference of their limits. Therefore what might be considered as $S_2(Z) - S_2(0)$ could be written as

$$\wp(x) \quad = \quad \frac{1}{z^2} + \sum_{\Omega'} \left\{ \frac{1}{(z - z_{ij})^2} - \frac{1}{z_{ij}^2} \right\}, \tag{192}$$

where the somewhat asymmetrical leading term and the use of Ω' result from avoiding the formal presence of $1/(0^2)$ in $S_2(0)$.

6.2 Weierstrass $\wp$ function

This function is Weierstrass's $\wp$, traditionally written with a gothic p, on which he founded the theory of elliptic functions. However, $\wp'(z) = -2S_3(z)$ and from there on, $S_k'(z) = -kS_{k+1}(z)$. Looking in the other direction, integrals of $\wp$ exist, although they are no longer represented by Eisenstein series, and not necessarily periodic nor single valued because of the different paths that could be taken by contour integrals defining them.

Nevertheless, there are traditional names and symbols associated with the first two integrals, according to which

$$\wp(z) \quad = \quad -\zeta'(z) \tag{193}$$

$$\zeta(z) \quad = \quad \frac{\sigma'(z)}{\sigma(z)} \tag{194}$$

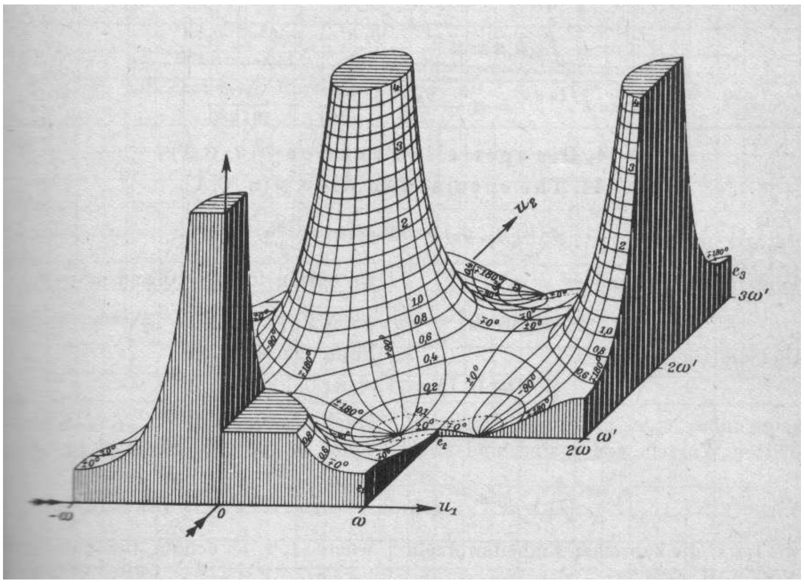


Figure 13: Jahnke and Emde's drawing of the Weierstrass $\wp$ function [16, page 99, figure 55]. The unique double pole is repeated four times because it sits on the corner of a unit cell in the periodic lattice. By uniform multiplicity there are two zeroes, distinct unless forced to coincide by symmetry. Two of the three critical points sit between copies of the double pole, while the third straddles the two zeroes.

6.3 restraints on periodic functions

The fact that there are many $\wp$ functions is somewhat hidden from view, especially when the two periods ω and ω' are not given explicitly. There could easily be still more doubly periodic functions, even for the same lattice, gotten by just averaging something else besides the inverse powers, just as periodic real functions can be built up from sines and cosines of different frequencies.

Cauchy's integral formula and the various consequences derived from it, such as the the constancy of bounded functions or the use of the logarithm to count zeroes and poles, imposes corresponding constraints on periodic functions.

To start, should a periodic function be bounded within its unit cell, repetition bounds it throughout the whole plane, making it constant by Liouville's theorem. Constancy is trivially periodic for whatever period, but any other alternative requires some singularity within the unit cell. The simplest possibility would be a finite number of poles, but even these are subject to some further restrictions.

For example, counting the difference between poles and zeroes by integrating the logarithm finds that integrals along opposite sides of the unit cell cancel because they traverse the same values in opposite directions. So there are just as many poles (counting multiplicity) as zeroes (again counting multiplicity).

If there were just a simple pole with residue, that residue would have to be zero because the boundary integral evaluating it would would always vanish by symmetry. So single isolated poles in periodic functions can't be found. In fact, the residues of whatever collection of poles would have to sum to zero; in particular, a pair of poles would have to have residues of opposite sign.

The Weierstrass function has a pole of order 2 and residue 0 at the origin (and indeed, at every lattice point), requiring every unit cell to have exactly two zeroes. In fact, the number of occurrences of any particular value within a unit cell must be the same as the number of zeroes. The reason is that subtracting a constant from a function does not change the location of its poles, but of course shifts the values and locations of its zeroes. So including infinity, there is a uniform multiplicity of values within each unit cell of a periodic function.

Moreover, all possible values must appear, although this is an observation which ought to be discussed in its own right. Isolated intervals are excluded because otherwise a Móbius transformation could be used to put the center of the interval at infinity while excluding a neighborhood of infinity. Then the composite function would have to be bounded, hence constant by Liouville's theorem, with only trivial periodicity. Once it is known that there are no missing intervals, continuity precludes excluding discrete values.

6.4 the differential equation for $\wp$

Liouville's theorem can be used to infer a differential equation for functions, given that they depend on poles for their existence. Consider two devices for creating new functions of the same polarity. One is squaring, or in general raising to powers, which will only change the multiplicity of whatever zeroes or poles are present, but not their location. Division, of course, will interchange zeroes and poles. Besides that, there exists the possibility of taking derivatives, which will probably change the number and location of zeroes, but only increases the order of poles without changing their location.

In the case of $\wp$, there is a double pole at the origin, implying that $\wp'$ has a triple pole there. Consequently both the cube of the function and the square of the derivative have poles of sixth order which can be cancelled by dividing them. But zeroes in the denominator can create new poles, so they must be removed; by forming linear combinations, if that is possible. Expecting to see three zeroes in $\wp'$ corresponding to the number of its poles, the attempt could be made to match them with a linear combination of $\wp^2$, $\wp$, and 1.

The algebra involved is fairly messy, but there does exist a quotient which, lacking poles, is constant by Liouville's theorem, leading to the differential equation

$$\wp'(z)^2 \quad = \quad 4\wp(z)^3 - g_2\wp(z) - g_3 \tag{195}$$

(the use of the symbols g_2 and g_3 being traditional). There is a corresponding integral,

$$\int \sqrt{4z^3 - g_2 z - g_3}\; dz, \tag{196}$$

which is related to the historical exercise of calculating the length or an arc on an ellipse. Consequently affiliated integrals are called *elliptic integrals* and their inverses *elliptic functions.*

Interestingly, because of the technique of balancing poles in quotients by which equation (195) was derived, the whole theory of elliptic functions can be based on $\wp$ and $\wp'$, notwithstanding that there are whole alternative families of variants which are more useful for some purposes or in other situations. It is probably not an exaggeration to say that the advanced mathematics of the late nineteenth century consisted in the study of elliptic functions in their manifold variations.

6.5 the Jacobi functions sn, cn, and dn

Akthough the Weierstrass function and its derivative constitute a logical basis for the theory of elliptic functions, there is a second tradition which is both of historical and practical importance.

It can be related to the Weierstrass form by writing the right hand side of the basic differential equation (195) in factored form (using traditional letters for the roots),

$$\wp'(z)^2 \quad = \quad 4(\wp(z) - e_1)(\wp(z) - e_2)(\wp(z) - e_3). \tag{197}$$

Given that the derivative vanishes at the roots, they are not only critical points, but good places to take square roots of what locally is nearly a square, in the expectation that the two branches will be distinguishable. Although that was not the historical reasoning, it is a good pretext to introduce the three *root functions* [6, §29]

$$f_i^2(z) \quad = \quad \wp(z) - e_i. \tag{198}$$

There is a very high symmetry inherent in these definitions; in fact the substitution of $\wp = f_i^2 + e_i$ for any choice of i into equation (197) results in

$$(2f_i f_i')^2 \quad = \quad 4f_1^2 f_2^2 f_3^2. \tag{199}$$

Subject to a sign ambiguity which works out to $f_i' = f_j f_k$ for i, j, k in cyclic order, the derivative of each of these functions is the product of the other two. With a little manipulation, various sums of squares can be shown to have vanishing derivative, and thus to be constant.

Such a system of equations as this arises in many applications, such as the equation of motion for a heavy top about its principal axes and so was one of the historical inspirations for the study of its solutions in whatever form they took.

There is an elaborate naming system for quotients of the root functions, plus schemes for standardizing scale factors and getting convenient combinations of the roots. To this end, define

$$\begin{aligned} u &= \sqrt{}(e_1 - e_3)\ z & (200) \\ k^2 &= \frac{e_2 - e_3}{e_1 - e_3}. & (201) \end{aligned}$$

and observe (by checking a fair amount of algebra) that the three most commonly encountered combinations are

$$\begin{aligned} \mathrm{sn}(u) &= \sqrt{}\left(\frac{e_1 - e_3}{\wp(z) - e_3}\right) & (202) \\ \mathrm{cn}(u) &= \sqrt{}\left(\frac{\wp(z) - e_1}{\wp(z) - e_3}\right) & (203) \\ \mathrm{dn}(u) &= \sqrt{}\left(\frac{\wp(z) - e_2}{\wp(z) - e_3}\right) & (204) \end{aligned}$$

Figure 14 shows a perspective view of the absolute value of the Jacobi function sn(z), which can be contrasted with the analogous view in Figure 13 of the Weierstrass function $\wp(z)$. They look much the same, due to the occurrence of poles and zeroes in each unit cell. However, the Jacobi functions have two simple poles per cell whereas the Weierstrass function has just one double pole. Examination of the figures reveals many symmetries, not only the expected translational symmetry but also by reflection and combinations of translation and reflection. Beyond that, if the period lattice is square or rhombic, the functions follow suit.

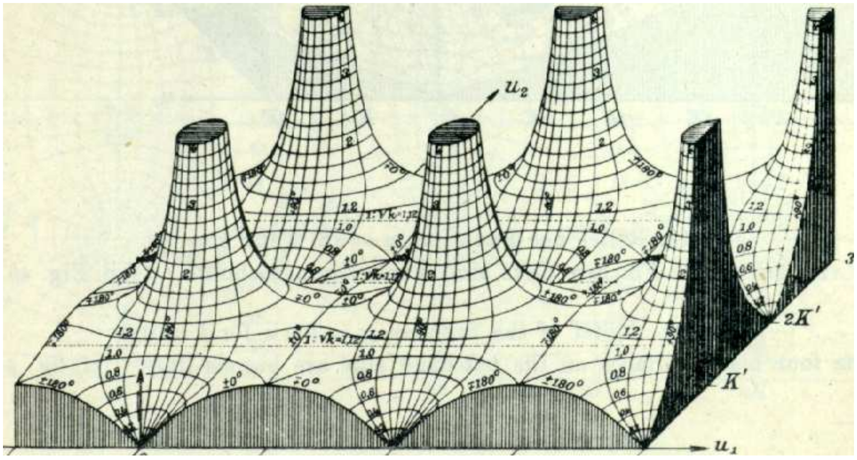


Figure 14: Jahnke and Emde's drawing of the Jacobi function sn(z,0.8) [16, page 92, figure 48].

The root form (197) of the Weierstrass differential equation lends itself to calculating a Schwartz derivative. Since we already have

$$\wp'^2 \quad = \quad 4(\wp - e_1)(\wp - e_2)(\wp - e_3), \tag{205}$$

the derivative,

$$2\wp'\wp'' \quad = \quad 4\left\{(\wp - e_2)(\wp - e_3) + (\wp - e_1)(\wp - e_3) + (\wp - e_1)(\wp - e_2)\right\}\ \wp', \tag{206}$$

transforms into

$$2\frac{\wp''}{\wp'^2} \quad = \quad \frac{1}{\wp - e_1} + \frac{1}{\wp - e_2} + \frac{1}{\wp - e_3}. \tag{207}$$

Another derivative,

$$-2\frac{\wp'''}{\wp'^3} + 4\frac{\wp''^2}{\wp'^4} \quad = \quad \frac{1}{(\wp - e_1)^2} + \frac{1}{(\wp - e_2)^2} + \frac{1}{(\wp - e_3)^2}, \tag{208}$$

completes the ingredients required for the Schwartz derivative, which subtracts half of $\wp''/\wp'$, squared, from its derivative. Then,

$$\frac{3}{4}\frac{\wp''^2}{\wp'^4} - \frac{1}{2}\frac{\wp'''}{\wp'^3} \quad = \quad \frac{3}{16}\left\{\frac{1}{(\wp - e_1)^2} + \frac{1}{(\wp - e_2)^2} + \frac{1}{(\wp - e_3)^2}\right\} - \frac{3}{8}\frac{\wp}{(\wp - e_1)(\wp - e_2)(\wp - e_3)},$$

for which there is an inconvenient $\wp'^2$ in the denominator of the left hand terms. But, recalling equation (86), this factor can be seen as having arisen from having taken the inverse Schwartz derivative. Therefore, finally,

$$\{z, \wp(z)\} \quad = \quad \frac{3}{16}\left\{\frac{1}{(\wp - e_1)^2} + \frac{1}{(\wp - e_2)^2} + \frac{1}{(\wp - e_3)^2}\right\} - \frac{3}{8}\frac{\wp}{(\wp - e_1)(\wp - e_2)(\wp - e_3)}, \tag{209}$$

This result, set forth as an exercise in Whittaker and Watson's *A Course of Modern Analysis* [28, chap XX, p. 439], does not seem to be used thereafter, but it invites some speculation. A periodic lattice would be expected to be invariant under affine transformations — composites of translations with matrix products — but Möbius transformations also include inversions, which would produce a different network entirely.

Therefore the Weierstrass equation might well not be invariant; what is interesting is that the inverse transformation is well behaved, although in the end, evaluating the Schwartz derivative is itself moderately complicated. Naturally, the Schwartz derivative of any function whatsoever can be calculated, which does not thereby automatically guarantee a useful interpretation.

In fact, the Ricatti style resolution of Schwartz derivative equations shows that the inverse Weierstrass function is a fractional linear combination of two linearly independent solutions of the second order linear differential equation

$$\frac{d^2u}{dt^2} = \left\{\frac{3}{16}\left\{\frac{1}{(t-e_1)^2}+\frac{1}{(t-e_2)^2}+\frac{1}{(t-e_3)^2}\right\}-\frac{3}{8}\frac{t}{(t-e_1)(t-e_2)(t-e_3)}\right\} u, \quad (210)$$

This observation prompts two lines of thought. One is to repeat the derivation of the Schwartz derivative for functions which could be double valued functions of a cubic polynomial, to obtain a differential invariant for this more general class of functions. The other would be to look for automorphic functions — those which retain their values with respect to some discrete subgroup of the unitary unimodular group of Möbius transformations. This latter alternative has been extensively studied, for example in Ford's *Automorphic Functions* [8].

7 Mapping theorems

The question often comes up, of designing a function taking given values at specified places. Naturally the answer depends on knowing what kinds of values, and at what places. For example, assigning values which violated the maximum modulus principle or Schwartz's lemma (without introducing singularities) would be futile.

Three levels at which the task may be considered are: i) mapping points, ii) mapping lines, and iii) mapping regions. The first is solved by interpolation, the second by the Schwartz-Christoffel process, and the third by the Riemann mapping theorem.

7.1 interpolation

An already existing linear relationship, such as the expansion of a polynomial of n^{th} degree in terms of powers x^i of degrees zero through n,

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad (211)$$

can be written as the inner product of two vectors, say a row of coefficients $\{a_0, a_1, \ldots, a_n\}$ and a column of powers, $\{1, x, \ldots, x^n\}$. Any change of basis, such as by replacing powers x^i with Newton polynomials $n^{(i)}(x) = x(x-1)\cdots(x-i+1)$, gotten by linear transformation, would be represented by a matrix whose inverse transforms the coeffients, keeping the representation intact.

If the expansion is not available, but it is desired to have a certain basis together with a family of linear functionals from which the coefficients are obtained, there is a determinant which can be used. For example, the combination of powers $\{x^i\}$ and values at a set of points $(x_0, x_1, \ldots, x_n\}$ will produce the Lagrange interpolation polynomials via the Vandermonde determinant. Write

$$\begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_1^n & x_2^n & x_3^n & \cdots & x^n \\ y_1 & y_2 & y_3 & \cdots & y \end{vmatrix} = 0. \tag{212}$$

The rows correspond to basis elements, the columns to linear functionals; the vanishing of the determinant results from duplicating a column whenever the value of the functional at the additional point matches the interpolation. Assigning suitable names to the pieces resulting from partitioning the determinant according to its last row and column, and the corresponding Laplace development, the equation reads

$$y \quad = \quad Y^T M^{-1} X. \tag{213}$$

This time there is a metric matrix which has to be calculated by inverting the matrix of mutual interactions, which will be possible if the two sets of constituents are actually linearly independent and form bases. The resulting representation formula will always work, but the question of its convergence as the number of reference points increases is usually fairly complicated. With power or polynomial bases, the degree of the polynomials is correspondingly high leaving the interior of the convex hull of the points as the only region in which a useful approximation should be expected. Whether that expectation can be realized, and many other points of interest, are discussed in Walsh's Colloquium Publication [27].

7.2 mapping the real line to a polygon

If integrals are used instead of interpolating polynomials, increments in the integrand can be used to guide the curve followed by an indefinite integral. A readily visualized process is to use increments of constant direction to create a straight line, varying the direction from time to time to outline a polygon. Doing the same by interpolation would mean assigning vertices to the data points, without the assurance of their being connected adequately by intermediate values.

Increments of constant direction are best gotten by taking real increments and multiplying them by an overall phase factor. The real axis could contribute the increments, relying upon roots to contribute the phase factors; the minus in a square root for a reversal of direction, a cube root for a 120° shift, a fourth root to make a square corner, and so on. The roots should be negative fractional powers to give finite integrals, leaving the question of their placement along the real axis as one to be resolved. It is desirable that the terms vanish strongly enough at infinity that an excursion along the whole real axis produces a closed polygon.

Consider the derivative defined by

$$\frac{dw}{dz} \quad = \quad (z - z_1)^{-\mu_1}(z - z_2)^{-\mu_2} \cdots (z - z_n)^{-\mu_n} \tag{214}$$

as it appears on the real line. Writing each singularity in the form

$$(z - z_p)^{-\mu_p} \quad = \quad |z - z_p|^{\mu_p} e^{i\mu_p \arg(z - z_p)}, \tag{215}$$

we can examine the effect of moving z from left to right along the real axis. The only effect will be to change the argument of $z - z_p$ from π to 0, and thus change the argument of the derivative by $\pi\mu_p$, which sends the integral off in a new direction. The condition for closure of the polygon is

$$\mu_1 + \mu_2 + \cdots + \mu_n \quad = \quad 2, \tag{216}$$

restoring the increments to their original direction. Putting angles in the exponents is easy enough, but deciding where to place the z_p to get desired edge lengths, and thus the w_p, requires more careful planning.

This whole process is called a *Schwartz-Christoffel transformation* To map one polygon to another, consider inverting the procedure, sending the first polygon to the real axis. Composing the inverse with the mapping of the real axis to the second polygon makes the transition directly.

7.3 functions with boundary values

A third viewpoint is to consider an analytic function as a solution to the Cauchy-Riemann partial differential equation with a boundary condition. The function to be defined has a boundary, at which it is supposed to acquire a certain set of values. It might be questioned whether such a thing is possible, but deciding whether or not is part of getting the solution.

Of course, it is possible to iterate the Cauchy-Riemann equations to get Laplace equations expressing the harmonicity of the real and imaginary parts of the complex function separately. It then becomes a question of solving Laplace's equation with boundary conditions, which leads to the Dirichlet integral and Riemann's mapping theorem.

8 Complex functions solving differential equations

Given derivatives and integrals, it is natural to look for relationships connecting them, and to develop a theory of differential equations. Of all these, systems of linear differential equations, and expecially single equations of second order, or pairs of equations of first order, are relatively easy to analyze. They are of especial importance because of the range of applications, and the great frequency with which they occur in all those applications.

8.1 kinds of equations

A differential equation is simply a relationship between a function and its derivatives, usually expressed through the vanishing of such an expression. An equation is considered to be *linear* if any linear combinations of its solutions is also a solution. The *degree* of the equation refers to the highest power of a term containing a derivative, whilst its *order* refers to the highest derivative present.

Thus a second order differential equation contains second derivatives, but there exists a trick whereby it may be replaced by a pair of first order equations. Suppose that the equation reads

$$a(s)\frac{d^2f(s)}{dt^2} + b(s)\frac{df(s)}{dt} + c(s)f(s) \quad = \quad 0. \tag{217}$$

Then the introduction of an auxiliary function $g(s) = df(s)/dt$ opens the possibility of studying the system

$$a(s)\frac{dg(s)}{dt} + b(s)g(s) + c(s)f(s) \quad = \quad 0 \tag{218}$$

$$\frac{df(s)}{dt} = g(s), \tag{219}$$

and even of writing it in matrix form

$$\frac{d}{dt}\left[\begin{array}{c} g(s) \\ f(s) \end{array}\right] = \left[\begin{array}{cc} -\frac{b(s)}{a(s)} & -\frac{c(s)}{a(s)} \\ 1 & 0 \end{array}\right]\left[\begin{array}{c} g(s) \\ f(s) \end{array}\right], \tag{220}$$

Given that higher derivatives may be replaced by a whole series of lower derivatives by using a companion matrix of coefficients, and that the same strategy serves as well for systems of equations, we are free to concentrate our attention on first order linear matrix differential equations. It is moreover desirable to foresee the existence of several solutions according to the selection of an initial vector by subsuming them all into a one matrix rather than several vectors. The linearity of the equation justifies working with square matrices throughout.

8.2 the differential calculus of matrices

There is really no such thing as the derivative of one matrix with respect to another, so the differential calculus of matrices refers to matrices whose elements are functions of a complex variable, and the results of taking derivatives with respect to that variable. It is not unreasonable to work with functions of several complex variables and to take partial derivatives, but the present discussion foresees but one independent variable.

Given a matrix with functional elements, its derivative would be a matrix filled with the derivatives of the individual elements, all in their corresponding locations. Rules for the derivatives of sums, differences, and products are readily obtained from the definitions of the respective operations:

- $\frac{d(A(t)+B(t))}{dt} = \frac{dA(t)}{dt} + \frac{dB(t)}{dt}$
- $\frac{d(A(t)-B(t))}{dt} = \frac{dA(t)}{dt} - \frac{dB(t)}{dt}$
- $\frac{d(A(t)B(t))}{dt} = \frac{dA(t)}{dt}B(t) + A(t)\frac{dB(t)}{dt}$,

although in the case of the derivative of a matrix product, the order of the factors in the derivative must be carefully respected given the noncommutativity of matrix multiplication.

Scalar factors get the same treatment as matrix products, although advantage can be taken of commutativity to simplify results, if necessary. The derivative of a constant matrix is the zero matrix, a result which can be used to calculate the derivative of an inverse matrix. Since $\frac{d}{dt}MM^{-1} = \mathbf{0}$ we get

$$\frac{dM}{dt}M^{-1} + M\frac{dM^{-1}}{dt} = \mathbf{0} \tag{221}$$

$$\frac{dM^{-1}}{dt} = -M^{-1}\frac{dM}{dt}M^{-1}, \tag{222}$$

which can be compared with the traditional commutative formula.

In general, the computation of the derivative of a function defined as an infinite series, or even as a polynomial, will not give the expected result unless the matrix commutes with itself for all values of the parameter, Thus even such a simple result as the derivative of a square becomes

$$\frac{dM^2}{dt} = \frac{dM}{dt}M + M\frac{dM}{dt} \tag{223}$$

and it is only possible to expect

$$\frac{de^{M(t)}}{dt} = Me^{M(t)} \tag{224}$$

under conditions of full commutativity. Indeed, since the solutions of systems of linear differential equations are exponential in nature, the major portion of their discussion consists in examining the appropriate replacement of the exponential.

8.3 linear matrix differential equations

Even agreeing to write differential equations entirely in terms of matrices, without using vectors, there are two different first order equations, depending on whether the coefficient matrix is a left factor or a left factor. Either of the two forms

$$\frac{dZ}{dt} = MZ \tag{225}$$

$$\frac{dW}{dt} = WM \tag{226}$$

is possible, not to mention combinations with coefficient matrices on both sides of the unknown matrix and even more complicated combinations. However, the one-sided equations can be interchanged by taking transposes, so it is reasonable to emphasize one of them in preference to the other. Even so, if the coefficient matrix is sufficiently asymmetical, there may be practical advantages to working with the coefficient on one side rather than the other. In any event, we have the rule

$$\frac{dZ^t}{dt} = \left\{\frac{dZ}{dt}\right\}^t, \tag{227}$$

so

$$\frac{dZ}{dt} = MZ \tag{228}$$

becomes

$$\frac{dZ^t}{dt} = Z^tM^t. \tag{229}$$

Another way to put the coefficient matrix on the other side of the variable matrix is to use the inverse matrix. So, if $dZ/dt = MZ$,

$$\frac{dZ^{-1}}{dt} = -Z^{-1}\frac{dZ}{dt}Z^{-1} \tag{230}$$

$$= -Z^{-1}MZZ^{-1} \tag{231}$$

$$= -Z^{-1}M. \tag{232}$$

Note the interesting combination which results from combining the transpose with the inverse, on the one hand, and using an antisymmetric coefficient matrix on the other. That the derivative of the combination ${Z^{-1}}^T$ is zero makes it constant; once the unit matrix, always the unit matrix, and Z is an orthogonal matrix.

Writing a differential equation for an inverse matrix supposes that there is such a thing, which could be settled by looking at the determinant of the solution. Given that a determinant is made up from products of matrix elements, there should be a formula for the derivative of a determinant as a sum of determinants, in which each column of the original determinant is replaced one at a time by its derivative. But the derivative of the column is just its multiple by the coefficient matrix, so we have

$$\begin{aligned} \frac{d}{dt}|Z_1, Z_2, \ldots, Z_n| &= |\frac{dZ_1}{dt}, Z_2, \ldots| + |Z_1, \frac{dZ_2}{dt}, \ldots| + \cdots && (233) \\ &= |MZ_1, Z_2, \ldots| + |Z_1, MZ_2, \ldots| + \cdots && (234) \\ &= \text{Trace } (M) \ |Z_1, Z_2, \ldots, Z_n| && (235) \end{aligned}$$

This is an ordinary differential equation with an exponential solution:

$$|Z(t)| = |Z(0)| e^{\int_0^t \text{Trace } (M(\sigma)) d\sigma}, \qquad (236)$$

and the characteristic common to exponentials of never vanishing unless it always vanishes. Consequently, the solution matrix of a linear differential equation starting out from a basis of linearly independent initial contions is forever nonsingular.

8.4 the matrizant

Generally speaking, the solution of a system of linear differential equations is an exponential, subject to the complication that matrix multiplication is noncommutative. The ordinary exponential can usually be approached either as a power series, or as the product of numerous factors nearly equal to unity. In the matrix context, the first alternative is realized by Picard's method of iterative refinement, while the second corresponds more closely to solving the equation by Euler's method: distance = velocity x time, applied repetitively over small intervals.

8.4.1 Picard's method

One way to obtain a representation for the solution of a system of equations is to integrate the equation $dZ/dt = MZ$, transforming it into an integral equation,

$$Z(t) - Z(0) = \int_0^t M(\sigma) Z(\sigma) d\sigma, \qquad (237)$$

and then iterate it by successive substitutions to obtain

$$\begin{aligned} Z(t) &= Z(0) + \int_0^t M(\sigma) Z(\sigma) d\sigma, && (238) \\ &= Z(0) + \int_0^t M(\sigma) \left[Z(0) + \int_0^\sigma M(\sigma_1) Z(\sigma_1) d\sigma_1 \right] d\sigma, && (239) \\ &= Z(0) + \int_0^t M(\sigma) \left[Z(0) + \int_0^\sigma M(\sigma_1) \left[Z(0) + \int_0^{\sigma_1} M(\sigma_2) Z(\sigma_2) d\sigma_2 \right] d\sigma_1 \right] d\sigma, && (240) \end{aligned}$$

$$
\begin{aligned}
&= \left[I + \int_0^t M(\sigma)Z(\sigma)d\sigma + \int_0^t \int_0^\sigma M(\sigma)M(\sigma_1)d\sigma_1 d\sigma + \ldots \right] Z(0) && (241)\\
&= \Omega(M,t,0)Z(0) && (242)
\end{aligned}
$$

This final matrix series, called a *matrizant*, possesses properties very similar to those of the matrix exponential. For example, by bounding the individual matrix elements and using the comparison test, it can be seen to converge as well as the exponential series.

If the coefficient matrix commutes for all values of its argument, for example when it is a matrix of constants, the ordering implicit in the nesting of the variables of integration can be relaxed by extending each integral to the full range $\{0, t\}$. Then, dividing the resulting n-fold integral by the exponentially increasing $n!$ to compensate the ensuing repetition by permutation of the basic domain of integration, the series simplifies to

$$
\begin{aligned}
Z(t) &= \left[I + \int_0^t M(\sigma)Z(\sigma)d\sigma + \frac{1}{2}\int_0^t \int_0^t M(\sigma)M(\sigma_1)d\sigma_1 d\sigma + \ldots \right] Z(0) && (243)\\
&= \left[I + \int_0^t M(\sigma)Z(\sigma)d\sigma + \frac{1}{2}\left(\int_0^t M(\sigma)d\sigma\right)^2 + \ldots \right] Z(0) && (244)\\
&= e^{\int_0^t M(\sigma)d\sigma} Z(0) && (245)
\end{aligned}
$$

The iterative process just descrined is known as *Picard's method* for solving differential equations. The gradual disappearance of the residual term from the result depends upon bounding the elements of the coefficient matrix as well as the rapidly diminishing volume of integration, which only encompasses the one ordered subset of the unit cube, whose volume is $1/n!$. Still, the error term must sometimes be retained. In general, linear systems should be solved in regions free of singularities in the coefficients; one way to take advantage of the complex plane is to integrate along trajectories which avoid the singularity.

The actual singularity can sometimes be confronted by taking limits, as can the problems at infinity arising from bounded coefficients which nevertheless increase as infinity is approached.

8.4.2 Euler's method

Another scheme of solution is to discretize the differential equation, approximating $dZ/dt = MZ$ by

$$
\begin{aligned}
Z(t+\Delta t) &= Z(t) + \Delta t M(t) Z(t) && (246)\\
&= (I + \Delta t M(t))Z(t) && (247)\\
Z(t+2\Delta t) &= (I + \Delta t M(t+\Delta t)))(I + \Delta t M(t))Z(t) && (248)\\
Z(t+n\Delta t) &= \left(\prod_{i=0}^{n-1}(I + \Delta t M(t_i))\right) Z(t) && (249)\\
&= \left[I + \Delta t \sum_{i=0}^{n-1} M(t_i) + \Delta t^2 \sum_{i=0}^{n-1}\sum_{j<i} M(t_i)M(t_j) + \cdots \right] Z(t) && (250)\\
&= \Omega(M,t,0)Z(0) && (251)
\end{aligned}
$$

These sums can be recognized as approximations to the integrals in Picard's method. However, this approximation is most useful in the product form, where the matrizant is sometimes called a *product integral.* For example, by observing the limits in the product, it is easy to conclude the rules

$$\begin{aligned}\Omega(M,t,t) &= I && (252)\\ \Omega(M,s,t)\Omega(M,t,u) &= \Omega(M,s,u) && (253)\\ \Omega(M,s,t)^{-1} &= \Omega(M,t,s) && (254)\end{aligned}$$

Two further rules, which follow from the definitions, are

$$\begin{aligned}\frac{\partial\Omega(M,s,t)}{\partial s} &= M(s)\ \Omega(M,s,t) && (255)\\ \frac{\partial\Omega(M,s,t)}{\partial t} &= \Omega(M,s,t)\ M(t). && (256)\end{aligned}$$

8.4.3 sum of coefficients

The sum rule for exponents requires modification for matrizants, due as usual to problems of noncommutativity. Suppose that the coefficient matrix M is the sum of two others, A and B, and that it is desired to solve he system $dZ/dt = (A+B)Z$. Suppose furthermore that Z is factored into $Z = UV$, and that the equation $dU/dt = AU$ has already been solved subject to the initial condition $U(0) = I$. Then we have, in succession,

$$\begin{aligned}\frac{dZ}{dt} &= (A+B)Z && (257)\\ \frac{dU}{dt}V + U\frac{dV}{dt} &= (A+B)UV && (258)\\ AUV + U\frac{dV}{dt} &= AUV + BUV && (259)\\ \frac{dV}{dt} &= U^{-1}BU\ V && (260)\end{aligned}$$

Given the presumption that the differential equation for U with the coefficient matrix A has already been solved, the new coefficient modifying B can now be taken as a known quantity. In matrizant notation:

$$\begin{aligned}\Omega(A+B,s,t) &= \Omega(A,s,t)\Omega(B^*,s,t) && (261)\\ B^* &= \Omega(A,t,s)B\Omega(A,s,t). && (262)\end{aligned}$$

There is no requirement that the solution matrix be a product of solutions to linear equations, even though the overall equation is linear. Suppose that $Z = PQ$ for some inveertible P and Q, but that the equation is still $dZ/dt = MZ$. Then

$$\begin{aligned}\frac{dZ}{dt} &= MZ && (263)\\ P^{-1}\frac{dZ}{dt} &= P^{-1}MZ && (264)\end{aligned}$$

$$P^{-1}\frac{dP}{dt}Q + \frac{dQ}{dt} = P^{-1}MPQ \tag{265}$$

$$\frac{dQ}{dt} = \left(P^{-1}MP - P^{-1}\frac{dP}{dt}\right)Q \tag{266}$$

The coefficient matrix has been split into a sum to which the previously described UV factorizatioon can be applied.

The *WKB method* uses this transformation to diagonalize the coefficient matrix M, and then treat the new term which is introduced as a correction to which the procedure for a sum of exponents can be applied. Actually there are two parts to the WKB method, of which the more delicate is the treatment of the singularity which inevitably arises because the place where U^{-1} becomes singular is usually the most interesting point in the differential equation. Although it can be sidestepped by moving around in the complex plane, *Stokes' phenomonon* refers to the discrepancy in analytic continuations pass by one side or the other of such singularities.

In using the WKB decomposition (266) it is advantageous to begin with the largest term, treating the other as a perturbation. Normally the large term would be the diagonal matrix $P^{-1}MP = \Lambda$, solved as a system of one dimensional equations with the individual solutions

$$q_i(t) = q_i(0)e^{\int_0^t \lambda_i(\sigma)d\sigma}. \tag{267}$$

These solutions would then be mixed by the smaller term. A single iteration by Picard's method, often called the *first Born approximation*, might suffice to achieve reasonable accuracy.

Where P is rapidly changing, or changing slowly in a context of a nearly singular P, the second term should get priority. The classical turning point in quantum mechanical problems meets this specification, as would any other equation where P had the Jordan normal form and thus confluent eigenvalues.

8.5 uniqueness and periodicity

Consider the derivative of a quotient, $U^{-1}V$, and suppose that U and V satisfy the same linear differential equation $dZ/dt = MZ$:

$$\frac{d}{dt}U^{-1}V = -U^{-1}\frac{dU}{dt}U^{-1}V + U^{-1}\frac{dV}{dt} \tag{268}$$

$$= -U^{-1}MV + U^{-1}MV \tag{269}$$

$$= O. \tag{270}$$

Only the derivative of a constant matrix, C, can be the zero matrix, showing that $U^{-1}V = C$, or $V = CU$. In other words, two different solutions of the same left handed linear differential equation can only differ by multiplication on the left by a constant matrix. Evidently this requires that if their initial conditions are the same, the solutions must be exactly identical.

It should also be borne in mind that the equations could be solved starting from singular initial conditions, which would interfere with forming quotients. The conclusion that two solutions of the same equation are related by a factor supposes that neither is singular, in which case the proportionality constant C would be nonsingular too.

There is an interesting application of this result when the matrix of coefficients is periodic; that is, when $M(t+\tau) = M(t)$ for a constant period τ. We have

$$\frac{dZ(t+\tau)}{d(t+\tau)} = M(t+\tau)Z(t+\tau) \tag{271}$$

$$\begin{aligned} \frac{dZ(t+\tau)}{dt}\frac{dt}{d(t+\tau)} &= M(t)Z(t+\tau) && (272)\\ \frac{dZ(t+\tau)}{dt} &= M(t)Z(t+\tau). && (273)\end{aligned}$$

Accordingly, $Z(t)$ and $Z(t+\tau)$, satisfying the same differential equation, are multiples,

$$Z(t+\tau) = CZ(t). \qquad (274)$$

Again it is implicit that the solution was nonsingular to begin with.

Mathematicians call this the *Floquet theorem*, whereas it is known to physicists as the *Bloch theorem.*

Symmetries other than translational are possible, for example reflection sending t into $-t$. The conclusions are similar: although the solutions do not need to have reflective symmetry, they can be built up from basic solutions which are themselves either even or odd.

8.6 the coefficient matrix as a tensor sum

The matrix of coefficients for a linear differential equation can be decomposed into the trace, a traceless symmetric tensor, and an antisymmetric tensor:

$$\begin{aligned} M &= \text{Trace(M)}\ I + M^o + M^a && (275)\\ M^o &= \frac{1}{2}(M+M^t) - \text{Trace(M)}\ I && (276)\\ M^a &= \frac{1}{2}(M-M^t) && (277)\end{aligned}$$

Since the trace part commutes with the rest, it can be separated at once, integrated to get a scalar exponential, and multiplied by the remainder of the solution. Note that this factor corresponds to the determinant of the full solution, so the remainder must be unimodular. In reality, both parts of the remaining decomposition are unimodular, because both of their traces are zero.

Solving for the antisymmetric part produces a rotation, because the differential equation for the inverse of the solution matrix has the same negative factor as the differential equation for the transpose. We have seen that two solutions of the same linear differential equation coincide whenever their initial values coincide.

When the system is two dimensional, there is only one antisymmetric matrix apart from scalar multiples, so that the solution is directly a matrix exponential, similar to what happens when the trace generates an exponential scale factor.

In any event, once the trace and antisymmetric part have been attended to, the final equation which remains has to deal with a rotating traceless symmetric tensor; this may or may not resemble an actual simplification.

One way to get a traceless coefficient matrix is to change the independent variable. In that case,

$$\frac{dZ(s(t))}{dt} = \frac{dZ(s)}{ds}\frac{ds(t)}{dt}, \qquad (278)$$

so choosing $\frac{ds}{dt}\ I$ as the factor P in equation (266) would have

$$\begin{aligned} -\frac{1}{s}\frac{ds}{dt} &= \text{Trace } M(t) && (279)\\ s(t) &= s(0) + e^{-\int_0^t \text{Trace}M(\sigma)d\sigma}, && (280)\end{aligned}$$

It would seem that the expansion inherent in the exponential-like solution to a system of linear differential equations introduces a natural scale for the independent variable, namely the duration of one half-life. In general, changing the independent variable seems to be called a *Liouville transformation*, and can be used to generate various scale modifications.

9 Second order differential equations

Second order linear differential equations for functions of a real variable ore of common occurrence in physical and engineering applications. Some equations are of second order as a consequence of Newton's laws, wherein accelerations rather than velocities play the predominant role. Others result from separating variables in Laplace's equation or Poisson's equation, where sexond derivatives are a consequence of curvature in a variational principle.

Complex numbers inevitably figure in the solution of such equations because they unify the occurrence of sines and cosines, a situation which can be traced in turn to working with eigenvalues and eigenvectors of antisymmetric matrices. They require complex numbers for the same reason that complex numbers were introduced into the solution of algebraic equations in the first place.

Since a single second order equation begs the introduction of two first order equations, the solutions of those equations are conveniently represented by points in the real plane. There is no particular reason to consider those points as complex numbers, although that is sometimes done. However it is not a question of working with the real and imaginary parts of a single analytic function.

In contrast, it is often useful to work with second order differential equations in a single *complex* variable, even though its coefficients make it look like the same real equation which would have been supplanted by two equations in real variables. This time the supplementary variable is complex, requiring four dimensions to give everything the same treatment as before. Obviously whatever analytic gains are achieved in the process, they are compensated by much more restricted graphical and visual opportunities.

Much insight into complex equations can be gained by using the graphical insight obtained from real equations. Conversely, puzzling aspects of the solutions of real differential equations are often clarified by their behavior in the complex plane, especially via the influence of complex singularities on real behavior. That may take the form of an unsuspected singularity, or the existence of an unanticipated branch point.

Whatever their origin, consider a pair of linear first order differential equations whose constituents could be either real or complex:

$$\frac{dy(s)}{ds} = a(s)y(s) + b(s)x(s) \tag{281}$$

$$\frac{dx(s)}{ds} = c(s)y(s) + d(s)x(s), \tag{282}$$

which could be written in matrix form as

$$\frac{d}{ds}\left[\begin{array}{c} y(s) \\ x(s) \end{array}\right] = \left[\begin{array}{cc} a(s) & b(s) \\ c(s) & d(s) \end{array}\right]\left[\begin{array}{c} y(s) \\ x(s) \end{array}\right].$$

To foresee some applications, the coefficient matrix for the Schrödinger equation takes the form

$$\left[\begin{array}{cc} 0 & V(x)-E \\ 1 & 0 \end{array}\right] \tag{283}$$

for energy E, potential $V(x)$, and position x as the independent variable. The dependent variables are $\Psi(x)$, the wave function, and $d\Psi(x)/dx$. On ther other hand, either a one dimensionad Dirac equation, or the radial part of a three-dimensional Dirac equation, employ a rather more symmetric version of the same matrix:

$$\begin{bmatrix} 0 & m - E + V(x) \\ m + E - V(x) & 0 \end{bmatrix} \tag{284}$$

The components are the positive and negative energy components – electron and positron wave functions. The matrix has diagonal components when spin has to be included as well.

The plane whose cartesian coordinates are x and y is called the *phase plane*, on account of that's being its name in classical mechanics where x is position and y is momentum.

9.1 polar coordinates and Prüfer's transformation

One of the most insightful visual techniques is to introduce polar coordinates into the phase plane; it is called *Prüfer's transformation*

$$\begin{array}{rclcrcl} \rho & = & \surd(x^2+y^2) & \quad & x & = & \rho\cos\theta \\ \theta & = & \arctan(y/x), & & y & = & \rho\sin\theta. \end{array}$$

with variables which satisfy the differential equations,

$$\begin{aligned} \frac{d\rho}{ds} &= \frac{1}{2}\frac{1}{\surd(x^2+y^2)}(2x\frac{dx}{ds}+2y\frac{dy}{ds}) && (285) \\ &= \frac{x(cy+dx)+y(ay+bx)}{\surd(x^2+y^2)} && (286) \\ &= \rho(a\sin^2\theta+(b+c)\sin\theta\cos\theta+d\cos^2\theta) && (287) \end{aligned}$$

for ρ, and

$$\begin{aligned} \frac{d\theta}{ds} &= \left(\frac{1}{1+\frac{y^2}{x^2}}\right)(y\frac{dx^{-1}}{ds}+\frac{dy}{ds}x^{-1}), && (288) \\ &= \left(\frac{1}{1+\frac{y^2}{x^2}}\right)\frac{(x\frac{dy}{ds}-y\frac{dx}{ds})}{x^2} && (289) \\ &= x(ay+bx)-y(cy+dx) && (290) \\ &= -c\sin^2\theta+(a-d)\sin\theta\cos\theta+b\cos^2\theta && (291) \end{aligned}$$

for θ. In summary,

$$\begin{aligned} \frac{1}{\rho}\frac{d\rho}{ds} &= d\cos^2\theta+(b+c)\sin\theta\cos\theta+a\sin^2\theta && (292) \\ \frac{d\theta}{ds} &= b\cos^2\theta+(a-d)\sin\theta\cos\theta-c\sin^2\theta, && (293) \end{aligned}$$

the form of which suggests using the double angle formulas:

$$\begin{aligned} \frac{1}{\rho}\frac{d\rho}{ds} &= \frac{a+d}{2}+\frac{a-d}{2}\cos 2\theta+\frac{b+c}{2}\sin 2\theta && (294) \\ \frac{d\theta}{ds} &= \frac{b-c}{2}+\frac{a-d}{2}\sin 2\theta+\frac{b+c}{2}\cos 2\theta. && (295) \end{aligned}$$

An interesting comparison and check results from setting $V = 0$ in the matrix of the Schrödinger equation, Eq. (283). Then $a = d = 0$, $b = -E$ and $c = 1$. The Prüfer equations are

$$\begin{aligned} \frac{1}{\rho}\frac{d\rho}{ds} &= (1-E)\sin\theta\cos\theta & (296) \\ \frac{d\theta}{ds} &= -E\cos^2\theta - \sin^2\theta, & (297) \end{aligned}$$

whose solution for θ can be put in one or the other ot two forms:

$$\begin{aligned} \theta(s) - \theta(0) &= (1-E)\ \text{arctan}\ s - s & (298) \\ &= (1-E)\ \text{arccot}\ s - Es, & (299) \end{aligned}$$

leaving the corresponding quadrature for ρ still to be performed.

Often the Prüfer equations are not pursued beyond the θ-equation because the sign and magnitude of the derivative establish the oscillatory nature of the solutions. If the derivative is of constant sign and bounded away from zero, the solutions will necessarily oscillate indefinitely with a period implied by the bound. Moreover, given two equations — say the same equation with different values of the eigenvalue parameter E — if the derivative of one of them is always greater than that of the other, conclusions about the interlacing of nodes can be drawn.

Even in the simple example shown, there is an unexpected jitter in the angle θ, when the result is compared to the more familiar solution consisting of sines and cosines derived from uniform motion around the circumference of a circle, at an angular velocity which is the square root of E, not E itself. However, the result shown is not incorrect, and can be used to illustrate the advantages of preparing a system of differential equations for the optimal interpretation of results.

In this case, the coefficient of the differential equation is not a multiple of the quaternion **i** whose exponential produces a simple rotation, but contains and admixture of the quaternion **j** which would generate hyperbolic motion. The combination leads to an ellipse in the phase plane, rather than a circle. Furthermore, the expected $\sqrt{E}$ is the quaternion norm of the coefficient, whilst terms such as $(E \pm 1)/2$ are the coefficients of the individual unit quaternions.

An appropriate preparation for the Prüfer transformation, actually one close to the self-adjoint form which is the version originally proposed by Prüfer, is to revise the matrtix equation by writing

$$\begin{bmatrix} 1/\sqrt{E} & 0 \\ 0 & 1 \end{bmatrix} \frac{d}{ds} \begin{bmatrix} y(s) \\ x(s) \end{bmatrix} = \begin{bmatrix} 1/\sqrt{E} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -E \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{E} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{E} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y(s) \\ x(s) \end{bmatrix}.$$

which, due to the constancy of E, takes the form

$$\frac{d}{ds} \begin{bmatrix} y(s)/\sqrt{E} \\ x(s) \end{bmatrix} = \begin{bmatrix} 0 & -\sqrt{E} \\ \sqrt{E} & 0 \end{bmatrix} \begin{bmatrix} y(s)/\sqrt{E} \\ x(s) \end{bmatrix}.$$

for which the appropriate Prüfer transformation would be

$$\begin{aligned} \rho &= \sqrt{Ex^2 + y^2} & \qquad x &= \rho\cos\theta \\ \theta &= \text{arctan}\left(\frac{y}{\sqrt{E}x}\right), & \qquad y &= -\sqrt{E}\ \rho\sin\theta. \end{aligned}$$

It then follows that

$$\begin{aligned} \frac{d\rho}{ds} &= 0 & (300) \\ \frac{d\theta}{ds} &= -1 & (301) \end{aligned}$$

which is the expected result of high symmetry and great simplicity.

9.2 projective coordinates and Ricatti's equation

Another useful conversion of a second order differential equation into a nonlinear first order equation, which works just as well for a pair of first order equations, is to introduce a projective transformation:

$$\begin{array}{rclcrcl} u &=& xy & \qquad & y^2 &=& uv \\ v &=& y/x & & x^2 &=& u/v \end{array}$$

whose variables satisfy

$$\begin{aligned} \frac{du}{ds} &= \frac{dx}{ds}y + x\frac{dy}{ds} && (302)\\ &= (cy+dx)y + x(ay+bx) && (303)\\ &= cuv + (a+d)u + buv^{-1} && (304)\\ u^{-1}\frac{du}{ds} &= (a+d) + cv + bv^{-1} && (305) \end{aligned}$$

$$\begin{aligned} \frac{dv}{ds} &= y\frac{dx^{-1}}{ds} + \frac{dy}{ds}x^{-1} && (306)\\ &= -yx^{-2}(cy+dx) + (ay+bx)x^{-1} && (307)\\ &= -cv^2 + (a-d)v + b && (308) \end{aligned}$$

which gives the interesting result that both the logarithmic derivative of u and the derivative of v itself, depend only on v. The equation for v is nonlinear, but quadratic, and of the first order. An equation with this format is called a *Ricatti equation.*

Again summarizing,

$$\begin{aligned} u^{-1}\frac{du}{ds} &= (a+d) + cv + bv^{-1} && (309)\\ \frac{dv}{ds} &= -cv^2 + (a-d)v + b. && (310) \end{aligned}$$

Of course, the objective is to obtain x and y through the intermediary of u and v.

If the same example used for the Prüfer transformation, namely the zero potential Schrödinger equation, is used as a test case for the Ricatti equation, the result is:

$$\begin{aligned} u^{-1}\frac{du}{ds} &= v - Ev^{-1} && (311)\\ \frac{dv}{ds} &= -v^2 - E && (312) \end{aligned}$$

whereupon

$$v(s) = v(0) - \sqrt{E}\ \tan s, \qquad (313)$$

which is an agreeable result. Just as the self-adjoint form of the Prüfer transformation inproved the appearance of its results, similar adjustments can be made in the Ricatti equation. Incorporating a $\sqrt{E}$ in place of E would improve the format of the equation for u, but both Ricatti and Prüfer transformations would benefit from a change of independent variable that would absorb $\sqrt{E}$.

9.3 solving a Schwartz derivative

A computation related to the link between a second order linear differential equation and a Ricatti equation gives a solution to a Schwartz differential equation

$$\{w, z\} \quad = \quad 2Q(z) \tag{314}$$

where

$$\{w, z\} \quad = \quad \left\{\frac{w''}{w'}\right\}' - \frac{1}{2}\left(\frac{w''}{w'}\right)^2 . \tag{315}$$

Suppose that y_1 and y_2 are two solutions of the second order equation

$$y'' + 2Q(z) \quad = \quad 0 \tag{316}$$

whose nonvanishing and linear indepencence is established by their having a unit Wronskian,

$$\left| \begin{array}{cc} y_1(z) & y_2(z) \\ y_1'(z) & y_2'(z) \end{array} \right| \quad = \quad 1. \tag{317}$$

Starting with the definition

$$w(z) \quad = \quad \frac{y_1(z)}{y_2(z)}, \tag{318}$$

there follow

$$w'(z) \quad = \quad [y_2(z)]^{-2}, \tag{319}$$

$$\frac{w''}{w'} \quad = \quad -2\frac{y_2'(z)}{y_2(z)}, \tag{320}$$

$$\left(\frac{w''(z)}{w'(z)}\right)' \quad = \quad -2\frac{y_2(z)'}{y_2(z)} + 2\left(\frac{y_2'(z)}{y_2(z)}\right)^2, \tag{321}$$

$$= \quad 2Q(z) + \frac{1}{2}\left(\frac{w''(z)}{w'(z)}\right)^2, \tag{322}$$

according to which such a quotient solves the equation. The involvement of the Ricatti equation arises from concluding that the only other solutions of the Schwartz equation result from a change of basis from the two solutions $y_1(z)$ and $y_2(z)$, which would enter w in the form of a fractional linear transformation. But we already know that invariance under fractional linear transformation and invariance of the Schwartz derivative are two sides of the same coin.

10 Functions of mathematical physics

Besides an extensive study of elliptic functions and integrals, classical complex analysis was heavily occupied with solving Laplace's equation in one form or another. Several configurations were sufficiently symmetrical that it was possible to solve the equation by separation of variables, leading to a variety of second order real differential equations. The involvement of complex analysis was a usual one, given that real functions inherit many of their properties from the complex plane, particularly through the location of singularities therein.

With the advent of quantum mechanics, Schrödinger's equation took the form of Laplace's equation with a potential term, so that most of the same functions were still useful, even if they appeared in a somewhat different form or with different parameters.

Each of the special cases has an extensive literature which it is unnecessary to repeat. However it is worth listing the principal functions, with an in indication of where they arose and why they are interesting. Some illustrtate general points and so are worth disscussing in more detail. Having originated from separating variables in a partial differential equation, they all present themselves as eigenvalue equations. Analytic continuation of eigenvalues, parameters, or both, is often useful, aside from the already informative consequence of complexifying real functions.

10.1 survey of equations

The following table lists some frequently encountered differential equations.

name	**equation**	**comment**
Plane wave	$-\frac{d^2}{dx^2}\Psi = k^2\Psi$	free particle
Bessel function	$\frac{d^2}{dr^2}W + \frac{1}{r}\frac{d}{dr}W + \frac{(1-\nu^2)}{r^2}W = 0$	radial function in spherical coordinates
hydrogen atom	$\frac{d^2}{dr^2}R + \frac{2}{r}\frac{d}{dr}R + (-\frac{1}{4} + \frac{n}{r} - \frac{\ell(\ell+1)}{r^2})R = 0$	Coulomb potential in Schrödinger equation
Legendre polynomial	$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) - \frac{m^2}{\sin^2\theta}\Theta + \ell(\ell+1)\Theta = 0$	colatitude function for spherical coordinates
Mathieu function	$\frac{d^2}{dx^2}Y + (a - 2q\cos(2x))Y = 0$	separated elliptic coordinates, periodic potential
Lamé equation	$\frac{d^2}{dz^2}\Lambda + (b + n(n+1)\wp(z))\Lambda = 0$	ellipsoidal harmonics; doubly periodic potential
Schrödinger equation	$-\frac{d^2}{dx^2}\Psi = (V(x) - E)\Psi$	potential V, energy E
Dirac equation	$\frac{d}{dt}\phi = (m - E + V)\psi$ $\frac{d}{dt}\psi = (m + E - V)\phi$	mass m, potential V, energy eigenvalue E

Some of their interesting features are discussed in the following subsections.

10.2 Bessel functions

Bessel functions made their first appearance by relating the angular position of a planet moving along a Keplerian ellipse to elapsed time. However the integral and power series shows up in other places, generally concerning the radial variable after separating Laplace's equarion in polar or spherical polar coordinates. However, the radial functions in the Schrödinger equation are Laguerre polynomials, and the one dimensional Schrödinger equation for a constant force are Airy functions which can be transformed into Bessel functions of order 1/3.

Bessel functions obey the differential equation

$$\frac{d^2W}{dr^2} + \frac{1}{r}\frac{dW}{dr} + \frac{(1-\nu^2)}{r^2}W \quad = \quad 0 \tag{323}$$

whose distinguishing characteristic is the inverse square singularity at the origin. Consistent with its relationship to polar coordinartes, solutions would be sought in the open interval $(0, \infty)$, with particular attention being paid to their limiting behaviour at both ends of the interval. That does not mean that efforts have not been made to continue such solutions onto the negative real axis, nor that Bessel functions of imaginary argument are not useful. They bear the same relation to Bessel functions that the hyperbolic cosine bears to the trigonometric cosine.

Consider Sommerfeld's contour integral

$$J_\nu(r) \quad = \quad \frac{1}{2\pi}\int_C e^{ir\cos(z)}e^{i\nu(z-\pi/2)}dz. \tag{324}$$

which can be verified as solving the Bessel equation. Although it is a question of finding the right exponential, the technique can be applied to a variety of differential equations.

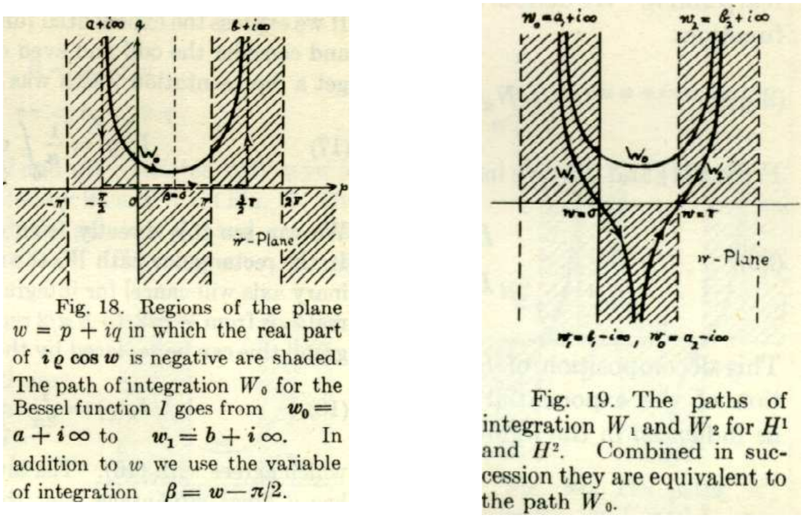


Fig. 18. Regions of the plane $w = p + iq$ in which the real part of $i\varrho \cos w$ is negative are shaded. The path of integration W_0 for the Bessel function I goes from $w_0 = a + i\infty$ to $w_1 = b + i\infty$. In addition to w we use the variable of integration $\beta = w - \pi/2$.

Fig. 19. The paths of integration W_1 and W_2 for H^1 and H^2. Combined in succession they are equivalent to the path W_0.

Figure 15: Two figures taken from Arnold Sommerfeld's treatise on partial differential equations [26, Figures 18, 19]. He was outstandingly versatile in bringing complex analysis into almost any physical problem.

Changing the contour of integration gives different varieties of Bessel functions. The common functions resemble sines and cosines. Kelvin's functions are more akin to hyperbolic sines and cosines, for being defined along the imaginary rather than the real axis. Finally, Hankel functions are the analogues of complex exponentials at the unit circle.

It would hardly be possible not to include a sample of Jahnke and Emde's drawings in a discussion of Bessel Functions, such as the one which can be seen in Figure 16.

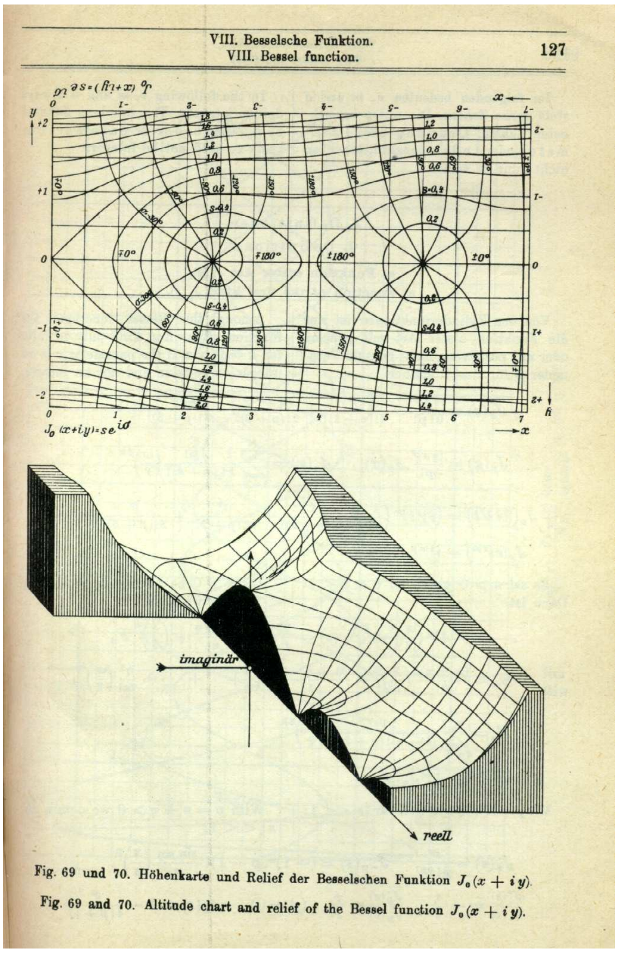

VIII. Besselsche Funktion.
VIII. Bessel function. 127

Fig. 69 und 70. Höhenkarte und Relief der Besselschen Funktion $J_0(x+iy)$.

Fig. 69 and 70. Altitude chart and relief of the Bessel function $J_0(x+iy)$.

Figure 16: A perspective view of the modulus of a Bessel function in the complex plane. Typically of such graphs, the function oscillates along the real axis while rising exponentially along the imaginary axis. The figure is a replica of the page containing Jahnke and Emde's drawing of the Bessel function $J_0(x+iy)$ [16, page 127, figures 69 and 70].

10.3 Legendre polynomials

Associated Legendre polynomials are the colatitudinal part of the spherical harmonics which are common to all separations of Laplace's equation in spherical polar coordinates. The radial part of the solution varies from one potential to another, but the harmonics are always the same and are a consequence of spherical symmetry. Associated polynomials have to be used when the solutions have an azimuthal component $e^{2m\pi i}$, for which reason the term dependent on m appears in the differential equation:

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) - \frac{m^2}{\sin^2\theta}\Theta + \ell(\ell+1)\Theta \quad = 0. \tag{325}$$

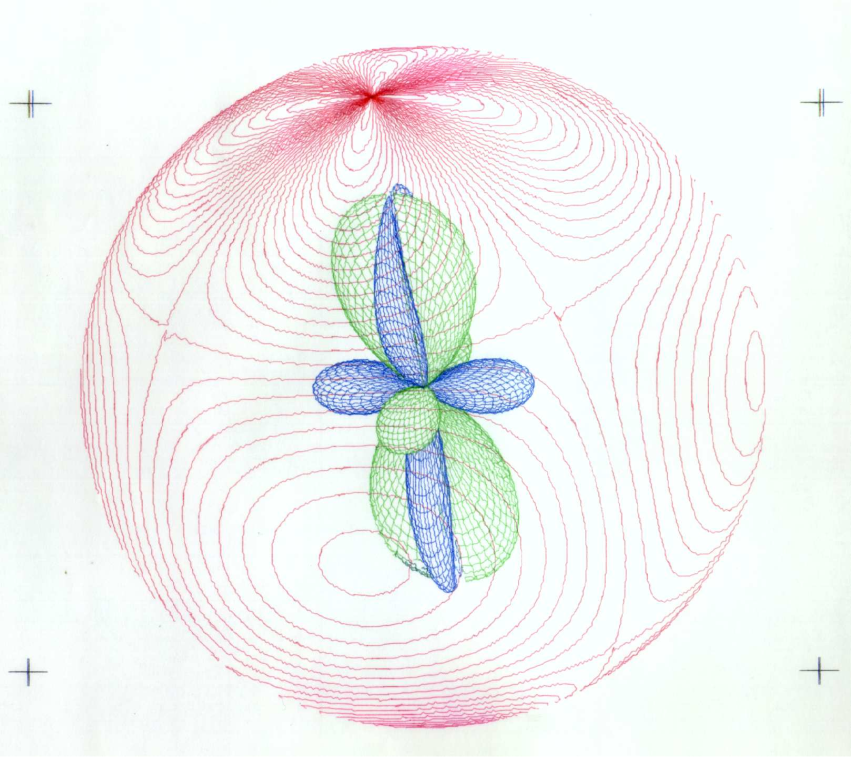


Figure 17: A spherical harmonic embedded in a contour map.

Figure 17 shows a visual representation of the spherical harmonic with four azimuthal nodes and two colatitudinal nodes. Each spherical harmonic subdivides the surface of a sphere into squares analogous to the action of $\sin(x)\sin(y)$ in the plane, except that they have to fit onto the surface of a sphere and respect the averaging principle which makes functions harmonic. Quite general functions defined over the sphere can be constructed as linear combinations, end even infinite series, of spherical harmonics.

Generally speaking, the solutions of equation (325) are only polynomials for integer values of ℓ and m, and even then it may be only one of the two linearly independent solutions which is a polynomial, but the full implications of this situation will only become fully apparent when the spectral density is discussed. In any event, the coefficients in the differential equation become singular at the poles, where the colatitude is either 0 or π, so especial attention must be given to the solutions at these points.

Even though the associated Legendre functions satisfy a real differential equation, the possibility esists of treating either ℓ, or m, or both, as complex variables, with respect to which continuation into the complex plane is possible.

10.4 Mathieu functions

One of the easiest ways to get a differential equation with periodic coefficients is to insert a sine or cosine in place of the frequency in the equation for simple harmonic motion; with stylized symbols, the result is Mathieu's equation:

$$\frac{d^2}{dx^2}Y + (a - 2q\cos(2x))Y \quad = \quad 0 \tag{326}$$

The outstanding new feature in such an equation is the possibility that there will be periodic solutions to match the periodic coefficient, but we already know that that is not to be expected. Rather, although some solutions may be periodic, the result of a translation through a single period will be to form a specific linear combination of solutions which will be reapplied for any further translations.

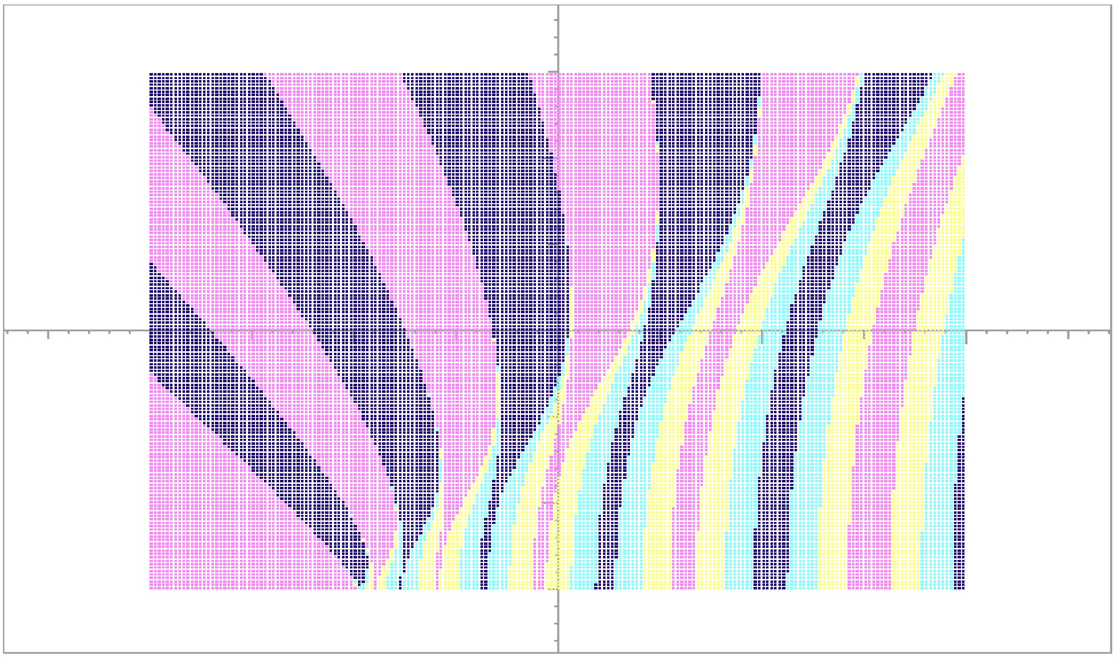


Figure 18: stability chart for solutions to the Mathieu equation.

The quantities of interest are then the eigenvalues and eigenvectors of the period matrix. Unless an eigenvalue is exactly equal to 1, there will be no strictly periodic solution, but the failure reduces to a multiplying factor, which is the eigenvalue. The eigenvalue -1 is exceptional, for implying subperiodic solutions — those which repeat having skipped a unit cell rather than immediately. The main feature of qualitative interest is the absolute value of the eigenvalues, which describes the rate of growth of solutions with respect to the translational lattice.

There are two parameters in the Mathieu equation, a multiplier a which would be an energy level in a quantum mechanical equation, and an intensity q. The main item of interest in the equation is often its *stability chart*, rather than the exact solutions themselves. The stability chart is a contour plot of the absolute value of the eigenvectors of the period matrix as a function of the two parameters a and q. Given that the solution matrix of the equation is unimodular, half its trace is the cosine of the logarithm of the eigenvalue. Therefore plotting the trace gives the required information, relative to the values 1 and -1. Contours are more informative than absolute values because all the quantities in the discussion are real and signs can be retained.

Figure 18 is just such a contour plot, although it bears some explanation. For $q = 0$ the equation reduces to the harmonic oscillator, a being the square of the frequency. When a is positive, the solutions are trigonometric, when a is negative, they are hyperbolic. The place where $a = 0$ is slightly off center, to the left. Half the trace of the period matrix is $\cos(\sqrt{a}\ s)$, whose own period is $2\pi/\sqrt{a}$ and therefore whose contour values will reach ± 1 at ever lengthening intervals. The thickness of the contour bands along the axis at the lower margin is an artifact caused by the horizontal tangents of the cosine at those points.

Along the line where $a = 2q$, the amplitude of the cosine matches the harmonic oscillator parameter; quantum mechanically that would be a classical turning point, with the solution matrix taking the Jordan normal form; classically the same anomaly just gets a different interpretation. Visually, the line separates the region where the hyperbolic solutions represent a narrow intrusion into a region of essentially periodic solutions, and the reverse. In the outer region, hyperbolic solutions are the norm, oscillatory solutions requiring narrowly defined parameter ranges for their existence.

A much more detailled analysis of the Mathieu equation and periodic potentials in general can be found in course notes [22].

10.5 Airy functions for a linear potential

The Schrödinger equation for a linear potential has Airy functions as its solution. One of its interesting properties is that a translation of the origin can be compensated by a shift in energy level, so that in principle the solutions for any one energy suffice for all. The equation is

$$\frac{d^2\Psi}{dx^2} \quad = \quad (\alpha x - E)\Psi, \tag{327}$$

whose matrix coefficient, with respect to a pair of first order equations, would be

$$\left[\begin{array}{cc} 0 & \alpha x - E \\ 1 & 0 \end{array}\right]. \tag{328}$$

This matrix has a negative pair of eigenvalues, one of which is $\lambda = \sqrt{(\alpha x - E)}$; real or imaginary according to the relative sizes of αx and E. There are also eigenvector matrices, P for columns and $P^{-1}$for rows:

$$P \quad = \quad \tfrac{1}{\sqrt{2}}\left[\begin{array}{cc} \lambda & -\lambda \\ 1 & 1 \end{array}\right], \qquad P^{-1} \quad = \quad \tfrac{1}{\sqrt{2}}\left[\begin{array}{cc} 1/\lambda & 1 \\ -1/\lambda & 1 \end{array}\right]. \tag{329}$$

With this preparation, the WKB factorization embodied in equation (266) can be applied. We need the integrated eigenvalue to produces an angle

$$\phi \quad = \quad \int_0^x \sqrt{\alpha\sigma - E}\; d\sigma \tag{330}$$

$$\quad = \quad \frac{2\sqrt{\alpha}}{3}(\alpha x - E)^{\frac{3}{2}}, \tag{331}$$

whose rate of increase, a $\frac{3}{2}$ power, lies between linear and quadratic.

At first appearance, the solution of the Airy equation would be

$$Z(x) \quad = \quad \frac{1}{\lambda^2}\left[\begin{array}{cc} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{array}\right], \tag{332}$$

which combines an inverse first power drop in amplitude with the square-root like increase in frequence with increasing distance. That is an asymptotic behavior near infinity; near the classical turning point, there is an important correction to the diagonal matrix $P^{-1}MP$, namely

$$P^{-1}\frac{dP}{dx} \quad = \quad \frac{\alpha}{4\lambda^2}\left[\begin{array}{cc} 1 & -1 \\ -1 & 1 \end{array}\right], \tag{333}$$

which requires its own treatment. As a scalar multiple of a constant matrix, the solution is an exponential of that matrix, this time with the multiplier

$$\theta \quad = \quad \int_0^x \frac{\alpha d\sigma}{4(\alpha\sigma - E)} \tag{334}$$

$$\quad = \quad \frac{1}{4}\ln(\alpha x - E). \tag{335}$$

Figure 19 shows the result of a numerical integration of the Airy equation, choosing the two basic solutions as the one similar to a sine (initial value 0, derivative 1) together with the one similar

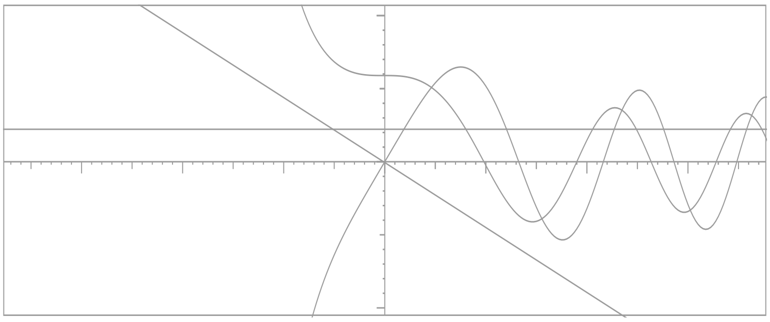


Figure 19: wave functions for a linear potential, which is solved with Airy functions. Generally, they act like $\sin(r^{3/2})/r$ and $\cos(r^{3/2})/r$.

to a cosine (initial value 1, derivative 0). Of course, calculating the exponential of a logarithm gives the argument back, which is a way to avoid the singularity implied by the logarithm. It would be more in keeping with the spirit of complex variable theory to give the energy a small imaginary component, replacing E by $E + i\varepsilon$, and narrowly avoiding the singularity. The main preoccupation is to get the trigonometrric functions on one side of the singularity to connect smoothly with the hyperbolic functions on the other side.

Splitting the coefficent matrix for the Schrödinger equation has mainly historical interest, given the ease of performing numerical integrations nowadays. Nevertheless, it is often useful to have an approximation in terms of more ordinary functions, either for symbolic calculations or as a starting point for mixed numerical computations.

Figure 20 shows the same result in the phase plane, which is the natural domain of the Prüfer transformation.

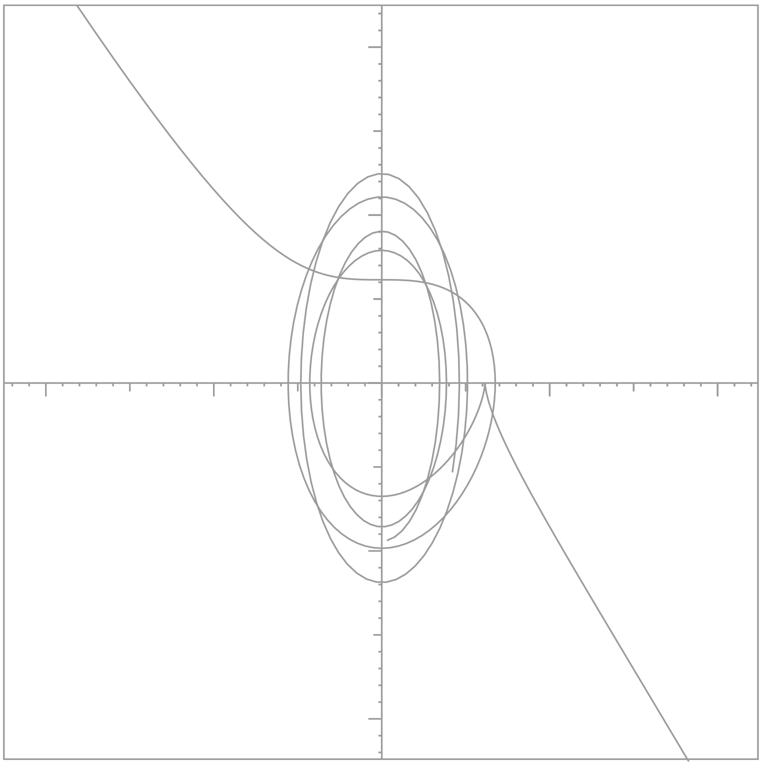


Figure 20: wave functions for a linear potential in the phase plane.

10.6 Dirac harmonic oscillator

As an example of a one-dimensional Dirac equation, consider the harmonic oscillator with potential $\frac{1}{2}x^2$ and mass m

$$\frac{d}{dx}\begin{bmatrix}\phi(x)\\ \psi(x)\end{bmatrix} = \begin{bmatrix}0 & m-E+\frac{1}{2}x^2\\ m+E-\frac{1}{2}x^2 & 0\end{bmatrix}\begin{bmatrix}\phi(x)\\ \psi(x)\end{bmatrix} \tag{336}$$

In normal physical problems with units compatible with atomic dimensions the mass is of the order 137, which completely dominates the equation. For mathematical purposes it is more convenient for the mass to be comparable to 1, although both extremes lead to illustrative partitions of the coefficient matrix. If the mass dominates, the solutions are hyperbolic functions, whereas they are completely trigonometric in nature when the mass is zero.

For moderate mass, the region “inside” the potential well gives oscillatory solutions, similar to the Hermite polynomials fount in the Schroedinger equation. In the “mass shell” the solutions are exponential, as expected. However there is a new feature of the Dirac equation, in the “outside” region the solutions are again trigonometric, although their phase advance with distance is opposite to the “inside” behavior. For particles in the actual physical world, the mass shell is very thick, becoming infinitely so as the nonrelativistic limit is approached.

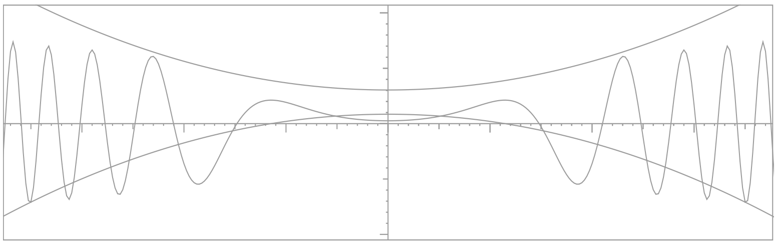


Figure 21: wave functions for a Dirac harmonic oscillator. The potential varies with z^2 making the wave function resemble $\sin(z^3)$ and $\cos(z^3)$. An even function is shown, the two parabolas are matrix elements of the coefficient matrix, and delimit the mass shell.

Figure 21 shows a solution for an energy value that gives a very slight oscillation within the well, but oscillations with rapidly increasing frequency outside the well.

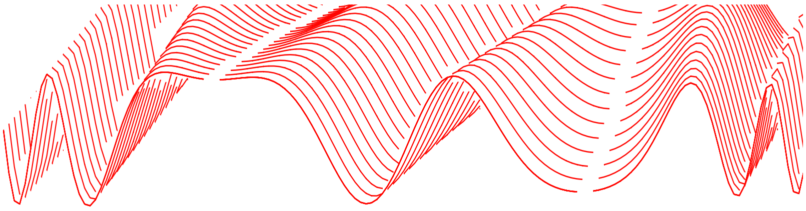


Figure 22: wave functions for a Dirac harmonic oscillator. Wave functions for a series of energies are shown in perspective view. The solutions are odd; gaps locate the classical turning points.

Figure 22 shows a perspective view, with hidden line suppression, of a series of solutions over an energy range. The movement of the classical turning points can be seen, as well as the variation of frequency with energy. There is some jaggedness on acount of the relative large increment in the horizon function which suppresses hidden lines.

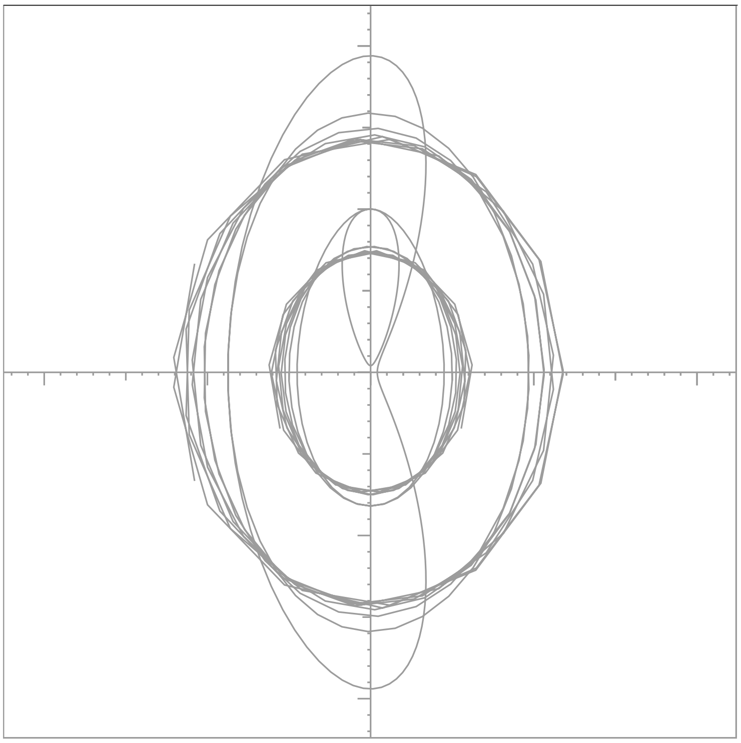


Figure 23: wave functions for a Dirac harmonic oscillator in the phase plane.

Figure 23 shows the solution of the Dirac harmonic oscillator in the phase plane. The solution consists of three sets of circles, connected by hyperbolic arcs. Two of them are superposed, by symmetry, representing the solution at large positive and large negative distances, respectively. As the energy varies, the ratio of the raduis of the inner circles to the outer circles will change; this is the quantum mechanical phenomonon of resonance and antiresonance. At resonance, the inner amplitude is a maximum relative to the asymptotic amplitude.

All one-dimensional Dirac equations afford an oportunity to split the mass term from the energy term, analogously to the WKB splitting already used for the Airy function. The kinetic energy term multiplies a unit antisymmetric matrix, so the solution is an Euler-formula exponential with an angle gotten by integrating the kinetic energy. For the Dirac harmonic oscillator, the angle is

$$\phi = \int_0^x (E - \frac{1}{2}\sigma^2) d\sigma \tag{337}$$

$$= Ex + \frac{1}{6}x^3 \tag{338}$$

Accordingly the matrix U is $\exp(\phi\mathbf{i})$, the usual matrtix representing a rotation. But then the equation for the second factor, V has the coefficient

$$e^{-\phi \mathbf{i}}(m\mathbf{j})e^{\phi \mathbf{i}} \quad = \quad m \left[\begin{array}{cc} \cos(2\phi) & \sin(2\phi) \\ \sin(2\phi) & -\cos(2\phi) \end{array} \right] \tag{339}$$

corresponding to a rapidly spinning mirror.

The solution for this coefficient is shown in Figure 24, whose most notable feature is the small asymptotic amplitude of oscillation of the solution. In the asymptotic region the mass, however large, is eventually dominated by the increasing potential energy, so the existence of mass or not is inconsequential. Near the origin, of course, the effects are more tangible. It is precisely in that region that the effects of resonance will be manifest.

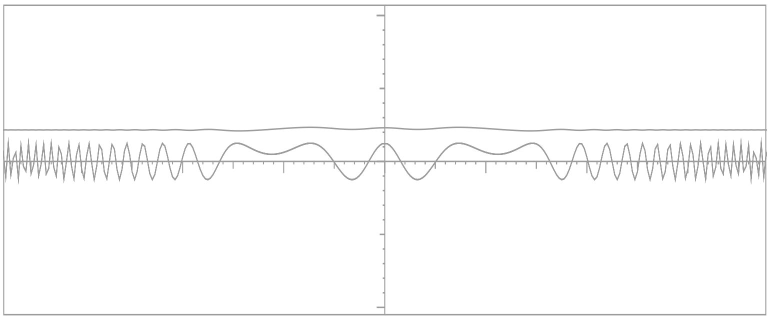


Figure 24: Wave functions for a factorized Dirac harmonic oscillator. The massless solution is a rotation with argument proportional to z^3. The factored equation adjusts for the mass.

Figure 25 shows these results more dramatically than the configuration space plots, and give some idea of the practical utility of the factorization. A much more detailled analysis of the Dirac Harmonic Oscillator can be found in course notes [21].

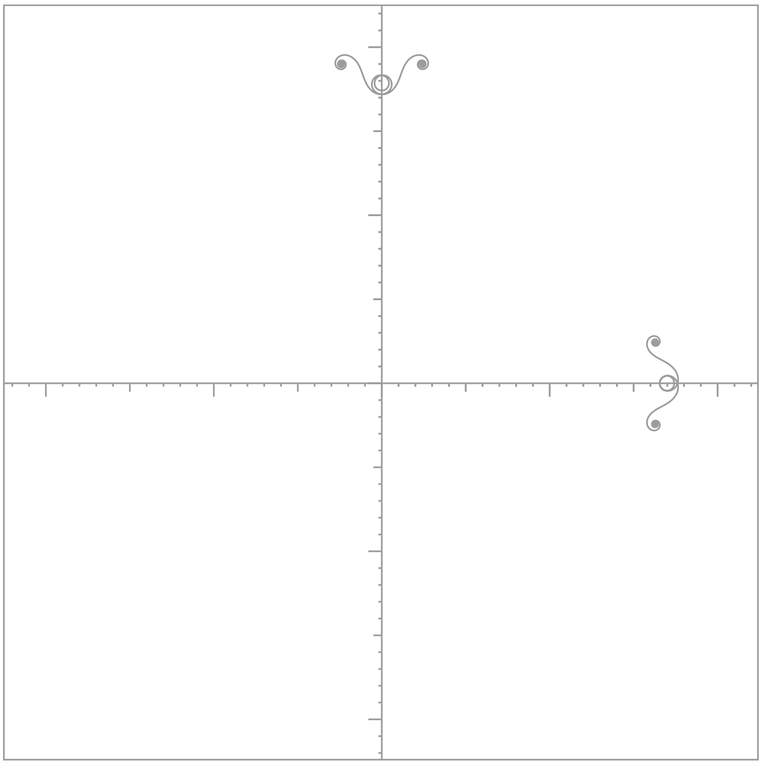


Figure 25: wave functions for a factorized Dirac harmonic oscillator in the phase plane.

11 Sturm-Liouville boundary conditions

The solution of a linear matrix differential equation with prescribed initial conditions is a relatively straightforward process, although the result can be complicated by the noncommutativity of the matrix of coefficients for different values of the independent variable. The solutions generally behave as exponentials, whose law of exponents is further distorted by noncommutativity. Nevertheless, the distinction between real and imaginary exponents holds, so that some solutions behave like hyperbolic functions and others behave like trigonometric functions. Those of trigonometric type oscillate, have zeroes, and remain bounded when the independent variable is real. Of course, when the solutions are taken as functions of a complex variable, all this structure rotates into the complex plane and may have to contend with branch points and other singularities.

A variant on the problem of initial values is the problem of final values. Taken over a finite interval, the nonsingularity of the solution matrix permits its inversion with an immediate solution to the problem. Asymptotically the situation is complicated by the fact that although the solution matrix is technically invertible, the difference between growing and diminishing eigenvalues is so great that only certain kinds of limiting information can be obtained.

Straddling these two extremes, the so-called *Sturm-Liouville* boundary conditions consist in a partial specification of initial values together with additional final values, sufficient to get as unique a solution as possible. Taken at face value, such a mixture is not likely to have a solution, but in practice, the differential equation often depends upon a parameter, usually the consequence of separating variables to solve some partial differential equation. The constant of separation would then be adjusted until the split boundary conditions could be satisfied.

That such a procedure has hope of success is encouraged when it is observed that varying the separation parameter, for given initial conditions, runs the terminal conditions through a range of values, one of which could be just the one sought after. In fact it is separation of variables which typically leads to an eigenvalue equation, because in the separated form it is the quotient of a differential operator depending on a single variable applied to its function by that separated function itself which is equal to a constant. So algebraic rearrangement sets the operator up as producing a multiple when it acts on its function:

$$\mathcal{H}\Psi \quad = \quad E\Psi, \tag{340}$$

anticipating the notation of quantum mechanics. If $\mathcal{H}$ is a second order differential operator, it can be replaced by a matrix combination of a pair of first order operators.

11.1 self adjoint equations and the symplectic metric

The form in which the initial value differential equation for a pair of first order equations is written can be called its *standard form*,

$$\frac{dZ(z)}{dz} \quad = \quad M(z)Z(z), \tag{341}$$

but for Sturn-Liouville applications, it is preferable to write it in *canonical form*, which is

$$\alpha(z)\frac{dZ(z)}{dz} + \{\beta(z) + \frac{1}{2}\frac{d\alpha(z)}{dz}\}Z(z) \quad = \quad \lambda\gamma(z)Z(z). \tag{342}$$

Here β takes the place of M from the standard equation, but without any eigenvalue which M might have contained. The factor α, only half of whose derivative is written explicitly (the other half being combined with M in β). allows for the possibility of attaching multipliers to the derivatives, which is sometimes either necessary or convenient. Finally, γ positions the eigenvalue where it belongs in the matrix multiplier of Z.

If α is invertible, it is easy to interconvert the standard and canonical forms; if it is not, the order of the system of differential equations ought to be reduced. By inspection,

$$M(z) = \alpha^{-1}(z)\{\lambda\gamma(z) - \beta(z) - \frac{1}{2}\frac{d\alpha(z)}{dz}\}. \tag{343}$$

So the only outstanding question concerns the motivation for the strange way of writing the canonical equation.

There is a further innovation, which consists in introducing the *adjoint* of the canonical equation,

$$-\alpha^T(z)\frac{dW(z)}{dz} + \{\beta^T(z) - \frac{1}{2}\frac{d\alpha^T(z)}{dz}\}W(z) = \lambda\gamma^T(z)W(z). \tag{344}$$

If the combination of signs and transposes seems mystifying, think of the adjoint W as the transpose of the inverse of Z. More exactly, direct substitution and invoking the uniqueness of solutions relative to initial conditions,

$$Z^A = (\alpha Z)^{-1\ T} \tag{345}$$

By regarding the left hand side of the canonical equation as the application of an operator $\mathcal{L}$ to Z, and likewise the left hand side of the adjoint equation as the application of $\mathcal{M}$ to W, the two can be written in the abbreviated form

$$\mathcal{L}(Z) = \lambda\gamma Z \tag{346}$$
$$\mathcal{M}(W) = \lambda\gamma^T W. \tag{347}$$

There are conditions under which an operator is just the same as its adjoint; by comparison they are

$$\alpha = -\alpha^T \tag{348}$$
$$\beta = \beta^T \tag{349}$$
$$\gamma = \gamma^T. \tag{350}$$

Once the definitions have all been established, *Green's formula*, which asserts that

$$\int_a^b \{\phi^T\mathcal{L}(\psi) - \mathcal{M}(\phi)^T\psi\}dz = \phi^T\alpha\psi|_b - \phi^T\alpha\psi|_a, \tag{351}$$

can be derived. In general terms, it is a consequence of $\mathcal{L}$ and $\mathcal{M}$ acting like derivatives applied to a product which evaluates into the product evaluated at its endpoints. The α sandwiched between the vectors follows from the detailed structure of these differential operators.

If the vectors ψ and ϕ are eigenvectors,

$$\mathcal{L}(\psi) = \lambda\gamma\psi \tag{352}$$
$$\mathcal{M}(\phi) = \mu\gamma^T\phi \tag{353}$$

the left hand simplifies to give the *Christoffel-Darboux formula*,

$$(\lambda - \mu)\int_a^b \phi^T \gamma \psi dz \quad = \quad \phi^T \alpha \psi|_a^b. \tag{354}$$

Among other things, it justifies eigenfunction expansions with respect to the solutions of differential equations.

The matrix form of these results is more complicated than the version which is usually seen in textbooks, but it has the advantage of broad applicability. For example, for Schrödinger's equation,

$$\gamma \quad = \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{355}$$

Although γ is singular, rather than disrupting the results, it makes the norm of functions depend on their values alone, and not at all on their derivatives. Spaces in which functions have this more complicated norm are called *Sobolev spaces* rather than *Banach spaces.* For the Dirac equation, $\gamma = \mathbf{1}$, and both positive and negative energy components figure in the calculation of norms.

The selfadjoint α is antisymmetric – a multiple of $\mathbf{i}$ – turning the inner product $\phi^T\alpha\psi$ into a determinant; concretely, the Wronskian of the solutions ψ and ϕ, in the case of the Schrödinger equation.

The Christoffel-Darboux formula establishes the orthogonality of solutions belonging to different eigenvalues, but cannot establish a norm because of the zero factor resulting from equal eigenvalues. The norm might be gotten as a confluent case, taking limits as two eigenvalues approach, or the undefined norm might just be left in place in expansion formulas.

By using Hermitean conjugates instead of transposes in the relevant formulas, and treating the eigenvalue as a complex variable, the factor turns into $(\lambda^* - \mu)$ which leaves the imaginary part of the eigenvalue when the two eigenfunctions are the same. The real result can be gotten by taking the limit as the imaginary part vanishes.

Supposing that

$$f(z) \quad = \quad \sum_{i=0}^{\infty} c_i \psi^i(z), \tag{356}$$

and setting aside questions of convergence,

$$\int_a^b \psi_i^T(z) f(z) dz \quad = \quad c_i \int_a^b \psi_i^T(z)\psi_i(z) dz, \tag{357}$$

so that

$$c_i \quad = \quad \frac{\int_a^b \psi_i^T(\sigma) f(\sigma) d\sigma}{\int_a^b \psi_i^T(\sigma)\psi_i(\sigma) d\sigma}, \tag{358}$$

The Stieltjes integral comes into play when we find that the eigenfunctions are quite numerous and the actual separation between eigenvalues is very small but nevertheless the eigenvalues are packed irregularly. It is also desirable to separate the parts of the coefficient which are due to integrating to function to be represented, and the small factor occasioned by the large norm.

Therefore, keep the integral of the function as a coefficient

$$\xi_i \quad = \quad \int_a^b \psi_i^T(\sigma) f(\sigma) d\sigma, \tag{359}$$

and introduce the *distribution function* $\rho(\lambda)$ which is a step function vanishing at $-\infty$, with increments at the eigenvalues λ_i in the amount of the reciprocals of those denominators, which are actually the square of the norm of the eigenfunction.

Altogether,

$$f(z) \quad = \quad \sum_{i=0}^{\infty} \xi_i \psi_i(z)(\rho_{i+1} - \rho_i), \tag{360}$$

$$\quad = \quad \int_{-\infty}^{\infty} \xi(\lambda)\psi(\lambda, z) d\rho(\lambda) \tag{361}$$

There still remains the selection of an appropriate right hand boundary condition, and the choice of a uniform normalization for all the eigenfunctions. Help in making the selections can be had from looking at *Parseval's equality*,

$$\int_a^b |f(x)|^2 dx \quad = \quad \sum_{i=0}^{\infty} |c_i|^2, \tag{362}$$

which is a generalization of the Pythagorean theorem to function space, and which can be written in the more symbolic form

$$(f, f) \quad = \quad \sum_{i=0}^{\infty} |(\psi_i, f)|^2. \tag{363}$$

By the Christoffel-Darboux formula, and using the complex version of Green's formulas rather than the real version,

$$\frac{[f, f](b) - [f, f](a)}{\lambda - \lambda^*} \quad = \quad \sum_{i=0}^{\infty} \frac{|[f, \psi_i(b)] - [f, \psi_i(a)]|^2}{|\lambda - \lambda_i|^2} \tag{364}$$

Having turned integrals in function space into boundary sums (an interval has just two boundary points, a and b), it would be convenient to eliminate the dependence on b. One mechanism is to look at a solution $f = \phi + m\psi$ which is a combination of the two standard solutions, and to suppose that f satisfies a *real* boundary condition at b, resulting in $[f, f](b) = 0$.

Whatever that boundary condition, it should be used for the ψ's as well; in other words,

$$[f, \psi_i](b) \quad = \quad 0 \tag{365}$$

Having removed the influence of the right boundary point by working with real boundary values in the complex domain, there remains the left boundary to assign some standard form. Using real values there too, and recalling the two linearly independent solutions ψ and ϕ, altogether,

$$[\psi, \psi_i](a) \quad = \quad 0 \tag{366}$$

with

$$[f, \psi_i](a) = [\phi, \psi_i](a) = r_i.$$

This quantity r_i, which is the increment in the Stieltjes integral, is the initial amplitude of a real, normalized solution of the differential equation over the finite interval $a,\ b$. That is another way to get the step in the spectral distribution function, because Parseval's equality now reads

$$\frac{m - m^*}{\lambda - \lambda^*} = \sum_{i=0}^{\infty} \frac{r_i^2}{|\lambda - \lambda_i|^2} \tag{367}$$

$$= \int_{-\infty}^{\infty} \frac{d\rho(\mu)}{|\lambda - \mu|^2} \tag{368}$$

$$= \int_{-\infty}^{\infty} \frac{\rho'(\mu) d\mu}{|\lambda - \mu|^2}. \tag{369}$$

The last line is admissable for points in the continuous spectrum of the differential operator, but the Stieltjes form must be retained for the point spectrum. If ρ' exists, it is called the spectral density.

The description of the spectral density calls for some care in its presentation. In works on solid state theory especially, the spectral density is often considered to be "the number of eigenvalues per unit of frequency interval," which is only a part of the story. This works well for plane waves, or functions which are more or less homogeneous throughout their extent, but a much more important consideration is the relationship between near amplitude and far amplitude. That determines the weight any given function requires to build up a given wave packet, and it is that weight which is properly the spectral density.

A good way to appreciate the difference is to return to the Dirac Harmonic Oscillator previously discussed. The spectrum is continuous, and there is no particular reason to think of the number of eigenvalues per unit interval because they are pretty much uniformly distributed. But most of them dissipate the presence of their particle by their large relative amplitude at infinity. Only eigenvalues in small, selected, intervals contribute to the presence of a particle near at hand, and it is this emphasis which assigns them a high spectral density.

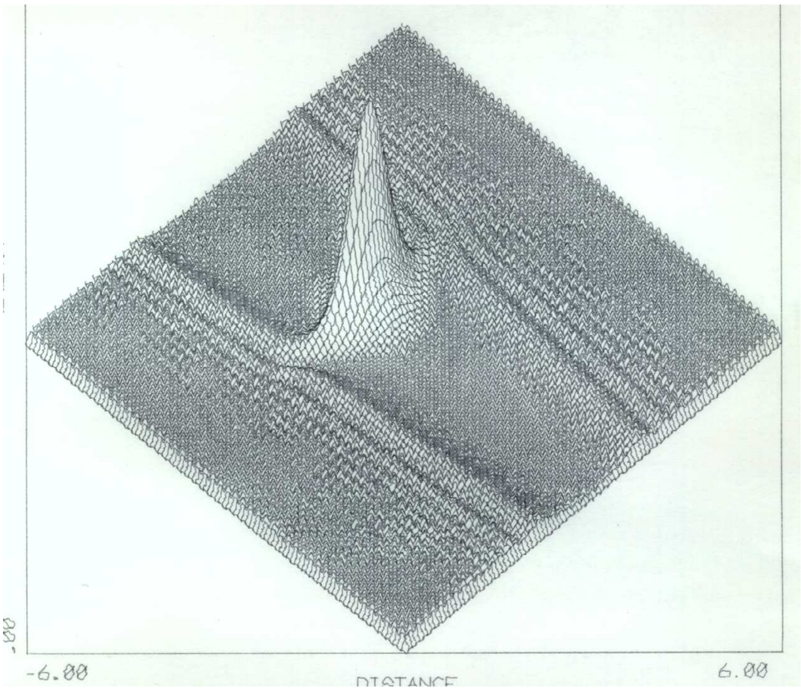


Figure 26: The spectral density function for the Dirac Harmonic Oscillator is the section of this surface bisecting the figure. Actually, there is a spectral density *matrix* [5], of which this section only reveals the even, or (0,0), element of that matrix.

Figure 26, although it graphs probability density as a function of both energy and distance, has been normalized to unit amplitude at infinity, so the values over zero distance portray the Titchmarsh-Weyl m-function, or in other words, the spectral density.

Whereas it might be overly ambitious to write

$$m(\lambda) \quad = \quad \int_{-\infty}^{\infty} \frac{\rho'(\mu)d\mu}{\lambda - \mu} \tag{370}$$

and compare it to Cauchy's integral formula, the relation certainly holds for the imaginary part of the equation, and invites considering ρ' as the boundary value of an analytic function which could be extrapolated throughout a half-plane, at least [19].

References

[1] Lars V. Ahlfors, *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable*, McGraw-Hill Book Company 1990 (ISBN 0070006571).

[2] Umberto Botazzini, *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass*, Springer Verlag, New York, 1986 (ISBN 0-387-96302-2).

[3] Brian E. Blank, "Book Review: An Imaginary Tale: The Story of the Square Root of Minus One," *Notices of hte American Mathematical Society* **46** 1233-1236 (1999).

[4] Lennart Carleson and Theodore W. Gamelin, *Complex Dynamics*, Springer Verlag, New York, 1993 (ISBN 0-387-97942-5).

[5] Earl A. Coddington and Norman Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill Book Company, New York, 1955.

[6] Patrick Du Val, *Elliptic Functions and Elliptic Curves*, Cambridge University Press, Cambridge, 1973 (ISBN 0-521-20036-9).

[7] A. Erdèlyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions, Volume II*, McGraw-Hill Book Company, New York, 1953.

[8] Lester R. Ford, *Automorphic Functions*, McGraw-Hill Book Company, New York, 1929.

[9] Nanny Fröman and Per-Olof Fröman. *Phase-Integral Method,* Springer Verlag, New York, 1995 (ISBN 0-387-94520-2).

[10] E. A. Guillemin, *The Mathematics of Circuit Analysis*, The M.I.T. Press, Cambridge, 1949.

[11] Liang-shin Hahn, *Complex Numbers and Geometry*, Mathematical Association of America, 1994 (ISBN 0-88385-510-0).

[12] Philip Hartman, *Ordinary Differential Equations*, John Wiley and Sons, New York, 1964.

[13] Einar Hille, *Lectures on Ordinary Differential Equations*, Addison-Wesley Publishing Company, Reading, Massachussetts, 1969.

[14] Adolf Hurwitz and Richard Courant, *Vorlesungen über Allgemeine Funktionentheorie und Elliptische Funktionen*, Verlag von Julius Springer, Berlin, 1925.

[15] E. L. Ince, *Ordinary Differential Equations*, Dover Publications, Inc., New York, 1956 [Original: Longmans, Green, and Co. 1926].

[16] Eugene Jahnke and Fritz Emde, *Tables of Functions with Formulae and Curves*, Dover Publications, New York, 1945.

[17] Felix Klein, *Lectures on the Icosahedron*, Dover Publications, New York, 1956.

[18] Morris Marden, *The Geometry of the Zeroes,* American Mathematical Society, Mathematical Surveys, Number 3. New York, 1949.

[19] Harold V. McIntosh, "Quantization as an Eigenvalue Problem," In *Group Theory and Its Applications*, Vol. 3, (Ernest M. Loebl Editor), Academic Press, New York, 1975. pp. 333-368

[20] Harold V. McIntosh, "Quantization and Green's Function for Systems of Linear Ordinary Differential Equations," In *Quantum Science: Methods and Structure*, Edited by J. L. Calais, O. Goscinsky. J. Linderberg and Y. Ohrn, Plenum Press, New York, 1976. pp. 227-294

[21] Harold V. McIntosh, *Resonances in the Dirac Harmonic Oscillator*, University of Puebla course notes, (16 pages) November, 2000.

[22] Harold V. McIntosh, *Periodic Potentials in One Dimension*, University of Puebla course notes, (19 pages) December, 2000.

[23] Paul J. Nahin, *An Imaginary Tale: The Story of the Square Root of Minus One*, Princeton University Press, Princeton, 1998 (ISBN: 0691027951).

[24] H.-O. Peitgen and P. H. Richter, *The Beauty of Fractals*, Springer Verlag, New York, 1986 (ISBN 0-387-15851-0).

[25] William T. Reid, *Ricatti Differential Equations*, Academic Press, New York, 1972 (ISBN 0-12-586250-4).

[26] Arnold Sommerfeld, *Partial Differential Equations in Physics,* Academic Press, Inc., New York, 1949.

[27] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, American Mathematical Society, Colloquium Publications, Volume XX. Providence, Rhode Islasnd, 1969.

[28] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge, at the University Press, 1963.

August 19, 2005