Complex Analysis, 2nd Ed. **Lars V. Ahlfors** PhL 8.12.09

As I was reading Stakgold Vol 2 today, I ran into something about "the famous Cauchy-Riemann equations" and drew a complete blank. So I thought I might review this old book at bit and try to reload a bit.

Chapter 1. The Basics of Complex Numbers **2**

Chapter 2. Complex Functions (21) **3**

1. Introduction to the concept of "analytic function". 3
   1.1. Limits and Continuity. 4
   1.2 Analytic Functions. 4
   1.2 Polynomials (28). 8
   1.3 Rational Functions (30). 9
2. Elementary theory of power series (33) 11
   2.1 Sequences 11
   2.2 Series 11
   2.3 Uniform Convergence 12
   2.4 Power Series (38) 12
   2.5 Abel's Limit Theorem (42) 15
3. The Exponential and Trigonometric Functions (43) 16
   3.1 The Exponential. 16
   3.2 The trig functions (44) 17
   3.3. The Periodicity (45) 17
   3.4 The Logarithm (46) 17

Chapter 3. Analytic Functions as Mappings (49) **18**

1. Elementary Point Set Topology. 18
   1.1 Sets and Elements (50). 19
   1.2 Metric spaces(51). 19
   1.3 Connectedness (54). 19
   1.4 Compactness (59). 20
   1.5 Continuous Functions (64). 26
   1.6 Topological spaces (67). 28
2. Conformality (68) 29
   2.1 Arcs and closed curves (68). 29
   2.2 Analytic functions in regions (69). 30
   2.3 Conformal Mapping (73). 31
3. Linear Transformations (76) 32
   3.1 The linear group (76). 32
   Huge Digression on the Subject of Projective Transformations and Homogeneous Coords 33
   Digression on Affine Transformations 38
   Review of Projective Transformations 38
   My own notes added: about "going complex" with the projective line 39
   Question: What does this "complex projective line" have to do with the Riemann Sphere? 40
   We return now to Ahlfors page 76 in section 3.1 there. 41
   3.2 The cross ratio (78) 41

Theorem 12: the cross ratio is invariant under linear transformations. 42
Question: How would you find a linear transform that takes z_1, z_2, z_3 to w_1, w_2, w_3 ? 43
Theorem 13. Claim that (z_1,z_2,z_3,z_4) is real $\Leftrightarrow$ the four points lie on a circle or line. 45
Exercise added: Prove theorem 13. 45
Theorem 14: Linear transformation maps Circles into Circles. 49
3.3 Symmetry (80) 49
3.4 Oriented Circles (83) 57
3.4 Families of Circles (84) 59
4. Elementary Conformal Mapping (89) 67
4.1 The Use of Level Curves (89) 67
4.2 A survey of elementary mappings (93) 70
4.3 Elementary Riemann Surfaces (97) 73

History. I have not looked much at this book for a very long time. Looking at my class history, I see that this book was used in my Berkeley Spring 1972 course ,

Math 185 (PN)	Complex Variables	Behrens	Ahlfors

So it has been a mere 37 years since I have looked at this stuff. The only inscription in the book is my name and a phone number 841-0677. I looked through my UCB old phone book and found several people with an 841 phone number. This must have been the Haste St apartment number,

1971-72 2725 Haste St #301 (Victor Chow, he got married, then Dixon) 8/71- 8/72

I have now found some hand-written notes in a blue bound folder with lots of other classes. These notes say that in this course we did only Chapters 2 and 4 of the Ahlfors book, skipping the topology chapter 2. My notes are fairly minimal, but I do have a long typed list of main ideas we were supposed to have learned. I will just start over I think here with these notes.

Chapter 1. The Basics of Complex Numbers

1.1 Basic arithmetic. Adding is easy for a+ib and c+id, multiplication not too hard, division takes more work. Reciprocal is special case of division (p 2).

1.2 Square roots. There is a way to write the square root of a + ib. He likes using $\alpha + i\beta$.

1.3 Justification. (4) Yes, the reals **R** are a field with field properties and an order relation and a certain completeness condition (p 5). He wanders off a bit into axiomatic talk. We end up with **C** being the field of complex numbers.

1.4 Complex Conjugates and Absolute Value (6). All the usual stuff here, lots of simple "rules" that I use every day.

1.5 Inequalities (9). C does not have an order relation, so cannot say $z_1 < z_2$, you can only have inequalities with the real or imag parts of complex numbers. We get the famous "triangle inequality". Notation $c > 0$ means two things: c is real, and $c > 0$. He then quotes and derives the Cauchy-Schwarz Inequality which roughly says $A \bullet B \leq AB$.

1.6 Geometric representation (12). Now we have the complex plane and all the usual rules. He notes that you don't derive things from this view, it is just a view. The triangle and other famous inequality turn out here to be just saying the length of one side of a triangle cannot exceed the sum of the other two side lengths. Then we have polar notation, and on p 13 A we get a rule for multiplication. Notice that he has not introduced the notion of $z = r e^{i\theta}$ yet, he is holding off on that. That of course makes it harder to prove various things, but you can still prove them.

2.2 The Binomial Equation (15). By this title phrase he means this:

$(\cos\varphi + i \sin\varphi)^n = \cos(n\varphi) + i\sin(n\varphi)$ de Moivre's formula

It is after all a binomial expansion. If we could use expo, it is trivial of course: $(e^{i\varphi})^n = e^{i(n\varphi)}$. Again, this is something you can prove without expo by brute force. He claims at this point that a complex number has n n^{th} roots, meaning if you try to solve $z^n = a$, there are n solutions for z. He shows this by messing with the angles. In expo we say $r^n e^{in\varphi} = ae^{i\theta}$ so $r^n = a$ and $n\varphi = \theta + 2\pi N$ for any integer N, so I guess you get that $\varphi = \theta/n + 2\pi N/n$ and this gives n distinct values for φ, answer is p 16. He of course uses the little de M formula shown above in his derivation of this fact.

2.3 Analytic geometry (17). Here he just points out that sometimes complex notation gives you an easy way to describe a geometric shape, such as $|z-a| = r$ for a circle. A straight line is $z = z_1 + t\, z_2$. An ellipse with foci at a and b would be $|z-a| + |z-b| =$ constant.

2.4 The Riemann Sphere (19). We get a picture and we get the math which relates a point z in the complex plane to a point Z on the sphere. This causes ∞ to be just like any other sphere point. Also, the two points z and Z are on a line which goes to the north pole. The claims are made that (1) z plane lines become sphere circles. (2) a circle on the sphere maps into either a circle or line in the s-plane. He grinds a little algebra to show all these things are true. Due to the projection from the north pole, this whole method is called a **stereographic projection**. I doubt he will ever use it, but we shall see.

So OK, this was a nice 20 page voyage through the most basic concepts of "complex analysis". Each section has some exercises that you are supposed to do. "regard as part of the text".

Chapter 2. Complex Functions (21)

1. Introduction to the concept of "analytic function".

Holomorphic and **analytic** are regard as the same by A. He will use z,w for complex, but x,y can be either, and t is always real. A function is by definition single-valued.

1.1. Limits and Continuity (22). The limit notion f(x)→A involves abs values, so can easily be extended to the notion of f(z)→A all complex. Basic limit rules like f(z)→A = Im[f(z)] → ImA, etc. The formal definition of a "derivative" is written in (4) exactly as in "real" calculus. BUT, you can now at least think about having the differential Δx being in either the real or in the imaginary direction, a bit new feature. These two cases are shown in p 23 B and C. If "the derivative exists", it must be the same in whichever direction you compute it. If f:C→R as in B and C there, answer B is real and answer C is imaginary, so for such an f, the derivative must either not exist or be 0 ! Totally new to me. How would you write such a function? $f(z) = x^2 + 2$, say, where x = Re(z). Real direction gives normal f'(x), real. The imaginary direction gives 0/h = 0 as the limit since this f(z) does not depend on y. So I guess in this example the derivative does not exist since the two directions give unequal results. This would NOT be an analytic function.

1.2 Analytic Functions (24). Analytic function f(z) on R is one for which f '(z) exists on R (R is a region of the C plane). But f '(z) exists only if you get the same result in both real and imaginary directions (and hence in any direction). You might say $f'(z) = \hat{\mathbf{n}} \bullet \nabla_2 f(x,y)$ where $\hat{\mathbf{n}}$ can be in any direction. Ie, you get the same result for any $\hat{\mathbf{n}}$. This does seem a highly restrictive condition on an analytic function!

In the previous section, f(z) is "continuous at a" if f(z)→f(a). In real calculus, we think of this as being either from the left or from the right. If the limit is finite (I think) and the SAME in both directions, then f(z) is continuous at a. For complex f(z) , continuous means get same limit when you approach from any direction in the C plane, but really that means from the real or the imaginary directions I think.

Suppose f '(z) exists, meaning same in all directions. Then on page 24, A implies B which says yes, f(z) is continuous if f '(z) exists. And we note that f '(z) must be finite in the limit as I "thought" above.

Now, what are the consequences of derivative f '(z) being the same in "both directions" ? This is the main act, so pay attention! If you compute in the real direction [z = x+iy], you get p 24C for f '(z). And if you compute in the imaginary direction, you get p 25 A for f '(z). Here we have f = u + iv . If we insist that these two computations of f '(z) be the same [as they must be if f(z) is "analytic"] then equating our two forms gives us two equations which are: The **Cauchy Riemann Equations** (6) ! [~ 1815]

Note added 1.29.2010. I guess I did not pay enough attention, so let's do this with micro baby steps.

One confusion is that it seems we can write f(z) as f(x+iy) or as f(x,y). But the first is a function of one complex variable, the second is a function of two real variables, these cannot be the same in terms of "functional form". You could say f(z) = f(x+iy) = F(x,y). You could then talk about $\partial F/\partial x$ and $\partial F/\partial y$ but f(z) is a function of only a single variable, so you cannot talk about $\partial f(z)/\partial x$ as a partial derivative. These little details are swept under the rug by Ahlfors; my "book" would put issues like this front and center. My green book page 63 makes it clear how this is resolved. We don't bother with F(x,y), but we say this:

$f(z) = u(x,y) + i\, v(x,y)$

where now u and v are functions of two real variables, and thus you can talk about $\partial_{\mathbf{x}} v$, for instance.

Now, the derivative of f is defined this way:

$df/dz = [f(z+dz) - f(z)]/(dz)$

$f(z) = u(x,y) + iv(x,y)$

$\partial_{\mathbf{x}} f(z) = \partial_{\mathbf{x}} u(x,y) + i\partial_{\mathbf{x}} v(x,y)$

$$\partial_y f(z) = \partial_y u(x,y) + i\partial_y v(x,y)$$

If we set dz = idy, what does this say? I set dz = idy in all thee places it appears:

$$df/dz = df/d(idy) = [f(z+idy) - f(z)]/(idy)$$

$$= [\, u(x,y+dy) + iv(x,y+dy) - u(x,y) - iv(x,y)]/(idy)$$

$$= [\, u(x,y+dy) - u(x,y)]\,/(idy) + i\,[v(x,y+dy) - v(x,y)]/(idy)$$

$$= (-i)\,[\, u(x,y+dy) - u(x,y)]\,/(dy) + [v(x,y+dy) - v(x,y)]/(dy)$$

$$= (-i)\,\partial_y u(x,y) + \partial_y v(x,y) \qquad = -i\,\partial_y u + \partial_y v \qquad \text{// p 25 A}$$

So this is our result if we examine df/dz in the "imaginary direction". Now repeat all the above in the real direction: Paste and edit:

If we set dz = dx, what does this say? I set dz = dx in all thee places it appears:

$$df/dz = df/d(dx) = [f(z+dx) - f(z)]/(dx)$$

$$= [\, u(x+dx,y) + iv(x+dx,y) - u(x,y) - iv(x,y)]/(dx)$$

$$= [u(x+dx,y) - u(x,y)]\,/(\,dx) + i\,[v(x+dx,y) - v(x,y)]/(\,dx)$$

$$= \partial_x u(x,y) + i\,\partial_x v(x,y) \qquad = \partial_x u + i\,\partial_x v \qquad \text{// p 24 C}$$

So now we have an precise understanding of where these two equations 25A and 24C are coming from. Now, if we want our derivative df/dz to be exactly the same in both real and imaginary directions, we have to equate the two results obtained above. If we equate the real and imaginary parts, we get

$$\partial_x u = \partial_y v \quad \text{and} \quad \partial_x v = -\partial_y u \qquad \text{CRE} \quad \text{p 25 (6)}$$

Now just for fun, consider $g(z) \equiv f(z^*) = u(x,y) - iv(x,y)$. Since we have basically replaced v with -v, we will end up with WRONG CRE, and we conclude that g(z) is not an analytic function of z.

Also just for fun: suppose analytic f(z) were entirely real, so $v(x,y) \equiv 0$. Then CR says that both $\partial_x u = 0$ and $\partial_y u = 0$ at every point in space. This means u = C, some constant. Thus, the only analytic function which is real everywhere is f(z) = a real constant, could think of as $C\,z^0$.

Here now are a few theorems that Stak either assumes without detailed proof, or has some proof.

Theorem 1: If f(z) is an analytic function of z, then f(z*) is NOT an analytic function of z.

Theorem 2: If f(z) is analytic and purely real, then f(z) = constant.

Theorem 3: if f(z) and g(z) are analytic, then h(z) = f(z)g(z) and f(z) + g(x) are analytic . [Ahlfors claims this in 3rd sentence of section 1.3 on page 28, it would be easy to show.]

Theorem 4: If we define h(z) = f(z) g(z*) where f and g are analytic in z, then h(z) is NOT analytic.

Corollary 4: The function h(z) = f(z)f(z*) = $|f(z)|^2$ is not analytic in z. This follows from Theorem 4, and is consistent with Theorem 2. This would be true if f(z) were a constant, which is not our intention.

Definition: f(z) is **real analytic**. I see this is not a popular term these days. Here is someone who says what I would like to be true:

"If a real function f(x) can be extended to an analytic function f(z), f is called real analytic."

For example, if f(x) = $\sqrt{x}$ or ln(x), then this is a "real function" on the positive real axis. That is to say, for the argument on the real axis, f(x) is real. (You would not say this for g(x) = i$\sqrt{x}$) . Neither of these functions is real on the *entire* real axis, just an interval of it, a key point. On that interval, if we assume we can expand in a convergent Taylor series at any point, those coefficients will be real. We know how to extend both of these functions to the complex plane as f(z) = $\sqrt{z}$ or ln(z). So both these functions of a complex variable would be said to be "real analytic" since on their real intervals we can expand in real Taylor series.

Definition wiki: This omits the key idea of "extension" noted above.

Formally, a function f is *real analytic* on an open set D in the real line if for any x_0 in D one can write

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots$$

in which the coefficients a_0, a_1, ... are real numbers and the series is convergent to f for x in a neighborhood of x_0.

Definition Phil: If an analytic function f(z) is real on the real axis, it is real analytic.

Theorem 5: if f(z) is real analytic and has a branch cut on the real axis, then the discontinuity across that branch cut is twice Im f(z) above the cut. That is to say, disc[f(z)] = 2 Im f(z+).

Proof for a power branch point (cuts pulled off to the left here, and α = real)

f(z) = $(z-a)^{\alpha}$ $\quad$ $f(z)_{\text{above cut}} = e^{+i\pi\alpha}\ |z-a|^{\alpha}$ $\quad$ $f(z)_{\text{below cut}} = e^{-i\pi\alpha}\ |z-a|^{\alpha}$

disc[f(z)] = $f(z)_{\text{above cut}} - f(z)_{\text{below cut}} = 2\ \sin\pi\alpha\ |z-a|^{\alpha}$

Im $f(z)_{\text{above cut}} = \sin\pi\alpha\ |z-a|^{\alpha}$ $\quad$ QED

For a log branch point: $\quad$ $z-a = |z-a|e^{i\pi}$ above cut

f(z) = ln(z-a) $\quad$ $f(z)_{\text{above cut}} = \ln|z-a| + i\pi$ $\quad$ $f(z)_{\text{below cut}} = \ln|z-a| + i\pi$

disc[f(z)] = $f(z)_{\text{above cut}} - f(z)_{\text{below cut}} = 2i\pi$

Im $f(z)_{\text{above cut}} = i\pi$ QED

Fact: Basically, all elementary and all special functions are real analytic. We pull branch cuts to the left, so the "interval" on which the function is real is usually some (a,∞), but we can pull cuts differently and have finite interval, such as (-1,1) for Legendre functions. In this case, however, Theorem 5 might not be true for a cut not pulled to the left. Lets try it:

Proof for a power branch point (cuts pulled off to the right here, and α = real. Define phase 0 on the negative real axis where function is real. Define CW phase up to +π, CCW phase to -π.)

$f(z) = (a-z)^{\alpha}$ $f(z)_{\text{above cut}} = e^{+i\pi\alpha}\ |z-a|^{\alpha}$ $f(z)_{\text{below cut}} = e^{-i\pi\alpha}\ |z-a|^{\alpha}$

disc[f(z)] = $f(z)_{\text{above cut}} - f(z)_{\text{below cut}} = 2\ \sin\pi\alpha\ |z-a|^{\alpha}$

Im $f(z)_{\text{above cut}} = \sin\pi\alpha\ |z-a|^{\alpha}$ QED

Theorem 6: if f(z) is real analytic, then f(z*) = [f(z)]* . This follows from $f(z) = \Sigma\ A_n\ (z-a)^n$ knowing that A_n are real coefficients and a is also real. Ie, we have this somewhere on the real axis. Then we get that $f(z)^* = \Sigma\ A_n^*(z-a)^{n\,*} = \Sigma\ A_n(z^*-a)^n = f(z^*)$.

These tell us that when f = u + iv and f is analytic, the two real functions are "correlated" with each other according to these equations. $\partial_x u = \partial_y v$ and $\partial_y u = -\partial_x v$. Their partial derivatives must be "cross related" in this manner! So certainly u and v cannot be arbitrary real functions! When they are correlated in this way, they are said to be "**harmonic conjugates**". Well, v is conjugate of u, and -u is conjugate of v, since we have a minus sign.

There is more! First, in p 25 B we see that there are 3 different ways to write $|\ f'(z)\ |^2$. Basically we start with the middle way which just comes from A, then the other two ways come from using the C-R equations. Second, and shocking to me, is result C. If you apply ∇^2 (he calls this Δ) to either u or v where f = u+iv, you can use the C-R to show that $\nabla^2\mathbf{u = 0}$ **and** $\nabla^2\mathbf{v = 0.}$ There it is, as Stakgold claimed on page 164 of Vol II. In order for a function to be "analytic", its derivative must exist (be finite and same from real and imag directions), and that in turn implies that the real and imag parts of f must be harmonic! But then of course f(z) itself must be harmonic!

I need an example fast! Consider $f(z) = z^n = (x+iy)^n = r^n e^{in\varphi} = r^n\ [\ \cos(n\varphi) + i\ \sin(n\varphi)]$. But yes, we know well from Stakgold that $r^n\ \cos(n\varphi)$ is harmonic, just look at p 93 6.8. And of course same for the imaginary part. You just separate Laplace in polars and radial solution is r^n in unit disk and $e^{in\varphi}$ is solution of azimuthal, we throw out r^{-n} and log(r) for n = 0 (we want a disk around z = 0). So, I have just shown that any power z^n is analytic because the real and imaginary parts are harmonic. But this at once tells me that any polynomial must be analytic as well. Consider $f(z) = iz^n$. The coefficient being non-real

makes no difference as to whether solves Laplace. If $f(z) = A + Bz^3$, each term in the poly is harmonic, and constant can be complex.

[p 26] Here he expends the top half of the page **proving** that: if u and v are harmonic conjugate real functions, then f = u+iv is an analytic function. At the end of his proof he ends up with the derivative definition being true with approach in an arbitrary direction, ie, $\Delta z = h+ik$. These are his favorite letters by the way for a differential, h and k.

Question: if you are **given harmonic u, how might you compute harmonic v**? One way is to integrate the C-R equations. He shows that bottom page 26 in example where $u = x^2 - y^2$ and then $v = 2xy+C$. This example is precisely $f(z) = z^2$ with C = 0.

[p 27] Now we have a slightly strange page, and I found it strange in my original reading of this book in 1972. He makes two points here.

(1) He shows in p 27 A that if f(z) is really an analytic function, then $\partial f/\partial \bar{z} = 0$, where you think of z and $\bar{z}$ as independent variables, and therefore f(z) is "a function only of z".("we are tempted to say this"). If we CC everything on line A, we conclude then that $\partial \bar{f}/\partial \bar{z} = 0$, so this $\bar{f}$ is a function only of $\bar{z}$ and he wants to write this as $\bar{f}(\bar{z})$. But this thing must be $\overline{f(z)}$. Think polynomial: $\bar{f}$ conjugates the constants, and $\bar{z}$ conjugates the powers, and $\overline{f(z)}$ conjugates both. Then of course $Re(f) = 1/2\ [f(z) + \overline{f(z)}\ ]$ and therefore $Re(f) = 1/2\ [f(z) + \bar{f}(\bar{z})]$ and that, finally, is where result p 27 A is coming from (it took me a while).

(2) If we assume result A stays true for complex x and y, we can insert the special values shown for x and y and get result B. He then argues that $\bar{f}(0)$ can be taken real (first term in poly series, can make real if you want), and thus identify this with u(0) and you end up then with result C.]

Result C says this: if you know a harmonic function u(x,y) , you can find the full f(z) whose real part is that u! Let's try it. Suppose $u(x,y) = x^2-y^2$. Then $f(z) = 2[(z/2)^2 - (z/2i)^2] = z^2$, so it works! His point here is that you don't have to integrate anything here, you just plug things in. He then says this method is viable only for u(x,y) = rational functions which I think means poly / poly, but with more work you can see that it works "in the general case" as well.

I am not so sure that result A is connected with this "function only of z" business. The page remains "confused" I think.

1.2 Polynomials (28). He quickly shows that 1 and z are analytic. Since sums and products of analytics are analytic, we see that a polynomial must be analytic.

The "fundy theorem of algebra" says a poly has at least one root, to be proven later. If you keep factoring out one root at a time, you get result (8) which is a complete factorization of your $P_n(z)$. Some roots might be the same, I would call that algebraic multiplicity, he calls it the order of the zero. First order zero (mult = 1) is a simple zero.

He then on p 29 proves **Theorem 1** "Lucas's Theorem": the zeros of P'(z) are in the same half-plane as the zeros of P(z). Now if I just take the factored form in (8), I get

P'(z) = sum of terms where each term is missing one of the (z-a) factors.

but this does not tell me where the zeros of P'(z) are located, the theorem is non-trivial and he proves it, but I skip the proof (bottom page 29). He says there is a stronger form of the theorem which says the zeros of P'(z) are in the same convex polygon as those of P(z). If we need this fact later on, I will come back here and prove it or verify his proof.

1.3 Rational Functions (30).

R = P/Q, and the zeros of Q are the poles of R and can be of multiple order. Easy then to show that R' has the same poles but each is of doubled order.

The big issue here is poles and zeros at ∞, they are a little harder to see. Write out P/Q as in p 30 A where we don't know which power is higher, m or n.

It seems clear to me that if n > m, then as $z \to \infty$ you are getting a blow-up z^{n-m} and you would call this "a pole at ∞ of order n-m". A pole is a place where an expression "blows up". He likes the idea of writing out $R_1(z) = R(1/z)$ and then looking for poles at 0 in R_1 and these are obvious from the leading factor. These poles at z = 0 of R_1 are of course the poles at z=∞ of R.

What if m > n? For me, this is harder to see looking at R. Looking at R_1, I can see that in this case there is a zero at z=0 of order m-n, so there must then be a zero at z=∞ of this order in R. $R(z) = R_1(1/z)$. How can you see this looking directly at R(z) ? For large z you can see that $R = (1/z)^{m-n}$ and yes, you can regard this as a zero at z = ∞, because you then get $(0)^{m-n}$.

Recall that "**the extended plane**" includes the point at ∞.

Question: how many poles and zeros does $R = P_n/Q_m$ have? If we factor up and down, we now for starters that there are n finite zeros and m finite poles (counting a order h zero as h zeros, and including poles and zeros at z = 0).

Case 1: n > m, there are n-m poles at z=∞, so total poles = m + (n-m) = n. In this same case, the number of zeros is n just from P_n. So # poles = # zeros = n.

Case 2: n < m, there m-n zeros at z = ∞, so total zeros = n + (m-n) = m. In this same case, the number of poles is m. So #poles = # zeros = m.

We can combine these results into this **simple rule**:

For a rational polynomial $R = P_n/Q_m$, we have

#poles = # zeros = max(m,n) = "the order of R" which he likes to call p.

This includes poles and zeros at z= ∞ as well as the finite poles and zeros. Notice that zeros and poles at z = 0 are counted in the normal consideration of the two polys.

Now consider R(z) - a. We can see that the # finite poles is the same. The large z blowup shows that the number of poles at ∞ is the same. Therefore, R and R-a must have the same order, call it p. This means that R-a must have p zeros, so the equation R-a=0 has p roots if R=0 has p roots. (top p 31). Of course some "roots" (read zeros) might be at z = ∞.

The most general p = 1 R is shown in p 31A (one finite pole and one finite zero in this case, he calls it S instead of R), and this is called **"a linear transformation"**, though I don't know where this name comes from, wait till Chap 3 he says. He goes on to solve R = a (ie, S = w) which we know also has one zero and one pole. See p 31 B.

Comment: I think a linear transformation is going to mean we *change variable* from z to z' = (αx+β)/(γz+δ). If you have some f(z) with some poles and zeros in z, then F(z') = f(z(z')) will have the poles and zeros MOVED in terms of z'. A simple example is f(z) = z which has a zero at z=0. Then F = 1/z' has the zero at z' = ∞. I think you can arrange to move the one pole and one zero anywhere you want.

The partial fraction story. Start with an example:

$(z^3+3z+5)/(z-2) = z^2 + 2z + 7 + 19/(z-2)$ // I did this by "long division"

$= z^2 + 2z + 7 + 19/(z-2)$

$R(z) = H(z) + G(z)$

We are just taking out the "improper" part H. I agree that H has a pole at z=∞ which is the same order as the pole in R -- we have taken it out, yes. G is "finite at z = ∞". He calls H(z) "the singular part of R at ∞" because it has all the pole action of R at this point.

Now he does a very strange thing. Let β_i be the finite poles of R. Construct $R(\beta_i + 1/\xi)$ where we have picked out one of the pole values. As $\xi \to \infty$, I agree that this R thing will have a pole due to β_i. So let's "take out the singular part at β_i " and write

$R(\beta_i + 1/\xi) = H_i(\xi) + G_i(\xi)$ // which is p 31 C

Then let $z = \beta_i + 1/\xi$ and rewrite the above, $1/\xi = z-\beta_i$, and we have p 32 A. Note that $G_i([z-\beta_i]^{-1})$ contains the singular part of R at β_i.

In expression (13) he has "taken out" all the singular parts, finite and infinite. Thus, the object I called T in p 32 (13) has no poles at all! It must therefore be order 0 which means it must be a constant. Absorb that into G and get (14), QED. Remember that G_j is a function, and it blows up to some order at $z = \beta_i$.

Here is a wiki web example where we start out with a proper fraction,

$$\frac{10x^2 - 63x + 29}{x^3 - 11x^2 + 40x - 48} = \frac{-70}{x-3} + \frac{80}{x-4} + \frac{-63}{(x-4)^2}.$$

Here the last TWO terms together would be $G_i(1/(x-4))$ for $\beta_i = 4$.

The purpose of the partial fraction expansion is to make integration easier, you just do the terms separately.

Rational functions of order p = 2. Imagine we are allowed to do linear transformations for free, meaning we can move poles and zeros around at will (but cannot change their number). For R with p = 2, we can have a double pole or two single poles. He likes to write these in a sort of canonical form:

double pole $R = az^2 + bz + c$ double pole is "thrown to ∞"

two single poles $R = Az + B + C/z$ one pole thrown to 0, the other to ∞

I am not sure what his point is here.

2. Elementary theory of power series (33)

2.1 Sequences

The big claim here is that

(a sequence converges) $\Leftrightarrow$ (sequence is a Cauchy sequence) " Cauchy's Condition"

the latter something I am familiar with. You could restate this as

$d(x_n,x) \rightarrow 0 \quad \Leftrightarrow \quad d(x_n,x_m) \rightarrow 0$

Aside: When A $\Leftrightarrow$ B, the part A $\Rightarrow$ B is called the "necessity": A must be true in order for B to be true -- A is necessary for B. As we know, A $\Leftarrow$ B is the same as notA $\Rightarrow$ notB, If A is "sufficient" for B to get true, then nothing else is needed. Then if A is not true, B won't be true. I never like these two words, but authors use them.

So if a sequence converges, easy to show Cauchy by triangle inequality, he does that, "necessity".

He then spends the next 1.5 pages proving "sufficiency", meaning

(a sequence converges) $\Leftarrow$ (sequence is a Cauchy sequence)

His presentation, although precise, is just not clear. The main points are not emphasized so the reader just sees a long sequence of details.

In particular, he talks about "**limes superior**" and inferior, but does not define these terms. I had to go off and write a document "limes inferior and superior.doc" to get these things understood. There I call these two things M and L. The only possibilities for a sequence are M < L and M=L. If M<L, your sequence has more than one accumulation point and the sequence does not converge regularly or Cauchy. If M=L, then series does converge regularly and Cauchy. So Ahlfors method seems to be to show that M = L if you have a Cauchy sequence and therefore you must have regular convergence as well.

Stakgold says the two sides of the above claim are the same as long as vector space is complete and is not missing the point x, so for Stak this is just a lot of hot air since the reals are a complete vector space.

2.2 Series

(1) Consider $s_n = a_1 + a_2 + \ldots . a_n$

I would call s_n a "partial sum". It is also a sequence of real numbers. Cauchy convergence says

$d(s_n, s_m) \rightarrow 0$

But the difference of these two partial sums is a sort of "tail sum" between n and m. So Cauchy just tells us that all tail sums must $\rightarrow 0$. The tail sum of length one is just a_n and so $a_n \rightarrow 0$. Thus, for any convergence partial sum, we know that $a_n \rightarrow 0$.

(2) If $d(b_m, b_n) \leq d(a_m, a_n)$, then surely b Cauchy converges if a does. He calls b_n a **contraction** of a_n which is an odd word to pick, and he admits it is "non standard".

(3) if $a_k \leq |a_k|$, then by 2 above we know that if $|a_k|$ converges (absolute convergence) so does a_k

2.3 Uniform Convergence

The idea of uniform convergence is defined in the usual way and compared to pointwise convergence. Then we get a **theorem** concerning continuity:

Theorem 1: continuity of limit function

If the sequence $f_n(x)$ converges uniformly to f(x) for x in E, then if the $f_n(x)$ are continuous, so is f(x).

A proves this on the bottom of page 36, fine, I skip these details for now. On page 37 we find this:

Theorem 2: Cauchy uniform convergence = regular uniform convergence

$f_n(x) \rightarrow f(x)$ uniformly for x in E $\Leftrightarrow$ $d(f_n(x), f_m(x)) \rightarrow 0$ as usual but for all x in E

Really this just says uniform regular convergence $\Leftrightarrow$ uniform Cauchy convergence. Then comes

Theorem 3: contraction against sequence of constants test

If $f_n(x)$ is contraction of a_n for all x in E, and if $a_n \rightarrow a$, then $f_n(x) \rightarrow$ some f(x).

Theorem 4: **Weierstrass M Test**

Consider $s_n(x)$ as the partial sum of some $f_k(x)$ for x in E. Suppose we can form another sequence of positive elements a_k such that $|f_k(x)| \leq M\, a_k$ for some constant M. Suppose the partial sums S_n of this new sequence a_k converges. Then if the partial sum sequence S_n converges (sum is finite), so does the partial sum sequence $s_n(x)$ for all x in E. In fancy lingo, if the **majorant** converges, then the **minorant** converges uniformly on E.

2.4 Power Series (38)

We first look at the basic power series where all coefficients are 1. He shows that it converges for $|z| < 1$, and diverges for $z \geq 1$. This is not hard to show, and the sum itself for convergence is shown p 38 A. The divergence at $z = 1$ appears as a pole. For $n = \infty$, the numerator is just 1 when it converges.

Now what do we know about a power series with constant complex coefficients a_n ?

He quotes **Abel's Theorem on Series Convergence** which has several parts:

There exists a radius of convergence R for $|z|$. R could be 0 or finite or ∞. In other words, if there is a region of convergence, that region is a disk, not some region of some other shape. Many things are true:

(1a) for $|z| < R$, the series converges.
(1b) it also converges absolutely since we can get another series with the same radius (as we shall see) by just changing the sign of all negative coefficients.
(1c) Convergence is uniform for $|z| \leq \rho < R$

(2a) If $|z| > R$, the series diverges.
(2b) The terms of the series keep growing so that $|a_n z^n| >$ any C for n large enough. Terms are unbounded. There is no bound C such that $|a_n z^n| \leq C$ for all n.

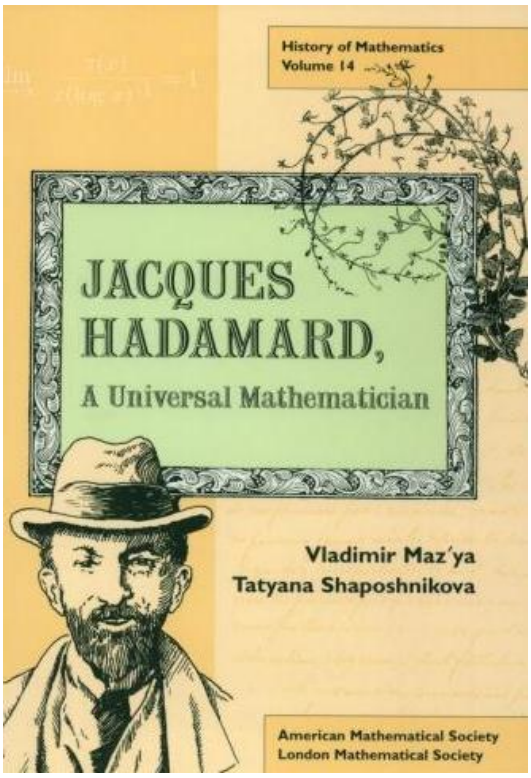
(3a) For $|z| < R$, the sum of the series not only converges, but is in fact an *analytic function*,
(3b) You can compute the derivative of the series termwise.
(3c) this derivative series has the same radius of convergence R.

(4) Nothing is claimed about convergence or divergence for a value of z such that $|z| = R$.

(5) There is a formula for R: $1/R = \limsup |a_n|^{1/n}$ = limes superior of sequence $|a_n|^{1/n}$ (Hadamard)

Notice that even if the sequence of *coefficients* cycles through multiple accumulation points, there will still be some very-large-n least upper bound on the accumulation points which is the limes superior. Obviously if $a_n = 1$, this formula gives $R = 1$.

What comes next is detailed proofs of most of these claims! But first some history: here is a nice little book on Hadamard,



Jacques Salomon Hadamard (1865-1963)

He presented his radius formula in 1892 at age 27 in his PhD thesis which this book talks about. It points out

$$R = (\limsup |c_n|^{1/n})^{-1}. \qquad (1.2)$$

In fact, this formula has been discovered much earlier by Cauchy, but Hadamard was unaware of this and he was the first who gave an accurate proof.

So nowadays this thing is called the **Cauchy-Hadamard Formula**.

Meanwhile we have this guy **Niels Henrik Abel** (1802-1829, lived 28 years!) I think his work on series came out perhaps 1827 or so, and I think it was for real, not complex series, but the main idea was there. A PDF on line talks a lot about his series work, but complex numbers don't appear. Notice that Abel died 35 years before Hadamard was born.

Now onto the technical matters of this section.

Proof of (5) and (1a) and (1c)

Assume that 1/R is given by the Hadamard's formula (19). Then he is saying that my limes superior L = 1/R = lim sup $|a_n|^{1/n}$. So for some ε, we know that $|a_n|^{1/n} < 1/R + \varepsilon$ for large enough n because we are coming down from above. Then think of $1/R + \varepsilon = 1/\rho$, compatible with $1/\rho > 1/R$. Then instead of choosing an ε, we choose a ρ. Then yes, we have $|a_n|^{1/n} < 1/\rho$ for $n \geq$ some n_o which is a function of this ρ or this ε. And then $|a_n| < 1/\rho^n$. And then $|a_n z^n| < (|z|/\rho)^n$. This inequality relates two series, let's look hard at corresponding terms:

$|a_n z^n| < |z|/\rho^n$

The LHS is a term in our "series of interest" where z in E which is a circle of radius R > ρ. I guess you could say this series was a minorant of the series on the RHS, though that does not fit Ahlfors' definition that you should be comparing to a series of *constants*, but OK. What do we know about this RHS?

$\Sigma_{n=0}^{\infty} (|z|/\rho)^n = 1/[1 - (|z|/\rho)]$ according to p 38 A

This series converges to this sum when $(|z|/\rho) < 1$. Since this is a majorant of our "series of interest", **we know that our series of interest also converges**. This is sort of a *variation* of the Weierstrass M Test we just saw on page 37, but the majorant is not a set of constants. A has been hazy.

So far then he has shown that for |z| < R of Hadamard, our series of interest converges any |z| inside our disk ρ. Now he picks ρ' such that ρ < ρ' < R and garbles up another proof, but the idea is that he now gets our series to have a constant majorant, and this then says our series **converges uniformly for |z| < R**.

He then shows that for |z| > R the series diverges. This A guy does not think the way I do.

Proof of (2a) and (2b). This is done in text near "B" on page 39.

Proof of (3c). You show here basically that $|n\, a_n|^{1/n} = |a_n|^{1/n}$ in the large n limit because $n^{1/n} \to 1$. So that is why the radius of convergence of the term by term diff series is the same as the original series.

Proof of (3b). This starts bottom of page 39 and consumes all of page 40. You need to show that even though the series f(z) has an infinite number of terms, you can find the correct f'(z) by doing term by term differentiation. I am skipping this detail! He of course starts out with the official formal definition of f(z) as shown LHS of (20) and he ends up with result A which says $f'(z_0) \to f_1(z_0)$ where f_1 is the term by term thing, and f ' (z) is the "official" thing.

"Proof" of Taylor-Maclaurin Series. We now know that we can differentiate a convergent infinite series all we want and we keep getting the same radius of convergence and term by term is always OK. Writing out all these derivatives on top of page 41, we get p 41 A that "suggests" you can perhaps write a power series expansion for *any* function f(z) which has all derivatives. Note that the Maclaurin Series is about z=0, while Taylor is about general z = a. We know of course that it will turn out that you can only do this if f(z) if analytic at z = 0 or at z = a for Taylor. See my chart, Taylor seems to have come first.

"Though Taylor series were known before Newton and Gregory (Grabiner 1997), and in special cases by Madhava of Sangamagrama in fourteenth century India, Maclaurin attributed Taylor in his work on approximating functions by series. [5]. At the time, Maclaurin was unaware and published his work in in Methodus incrementorum directa et inversa. Maclaurin series, which are Taylor series expanded around 0, and are not attributed to Maclaurin due to the past discoveries, but he still receives credit because of his use of them. In particular, he used these series to characterize maxima, minima, and points of inflection for infinitely differentiable functions."

2.5 Abel's Limit Theorem (42)

Just getting this thing stated takes a little work. Imagine a circle of convergence for some series with a_n coefficients, assume radius 1. Suppose also that it happens that $\Sigma a_n = B$, some finite number. Our series is

f(z) =$\Sigma a_n z^n$. We would like to say that $\lim_{z \to 1}$ f(z) = "f(1)" = B. The theorem says that this is true, as long as you DON'T approach z = 1 along the circle boundary! You have to approach on a line segment ending at the point z = 1 such that the segment has a slope that is finite -- that is not tangent to the circle. Doing this, you avoid the pole 1/(1-|z|) during your approach, so that (1-z)/(/(1-|z|) is bounded during the approach. Of course you want your approach segment to lie inside the circle so you have convergence AS you approach the limit point. So you approach at some angle in a 180 degree range. This type of approach angle is called a **Stolz angle** (1875), and here is why: [a slightly generalized version of Ahlfors]

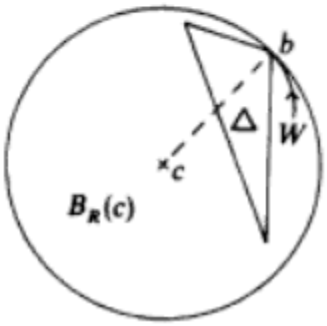
The limit theorem can also be phrased this way: for a boundary point at which the series converges, the function defined by the series has a limit under *radial* approach to this point and it is the sum of the series at this point. In 1875 O. STOLZ proved the following generalization (*Zeitschrift Math. Phys.* **20**, 369-376):

Let $\sum a_\nu (z-c)^\nu$ *be a power series with radius of convergence* R*; suppose that it converges at the point* $b \in \partial B_R(c)$*. Then the series converges uniformly in every closed triangle* Δ *which has one vertex at* b *and its other two vertices in* $B_R(c)$ (cf. the figure). *In particular, the function*

$$f : \Delta \to \mathbb{C}\ ,\ \zeta \mapsto \sum a_\nu (\zeta - c)^\nu$$

is continuous throughout Δ *and, consequently, satisfies*

$$\lim_{\substack{z \to b \\ z \in \Delta}} f(z) = \sum a_\nu (b-c)^\nu .$$



On page 42 Ahlfors proves his version of Abel's Limit Theorem, and I skip his details.

3. The Exponential and Trigonometric Functions (43)

3.1 The Exponential.

This is very strange. We DEFINE expo(x) as the solution to f' =f with f(0) = 1 as the BC. But we write it as e^x just as a notation. This is not some number "e" raised to a power, it is just a notation. The first order of business is to find the power series that solves the ODE + BC and that is (22) for e(x). We compute Hadamard's convergence radius and it is R = ∞, so converges everywhere. (those big n! denominators do it for us).

Now he proves an addition theorem which says e(a+b) = e(a)e(b). He does this in a sort of tricky way. This proof is NOT based on any rule like $a^x a^y = a^{x+y}$ for powers.

Since the series has only real coefficients, we know that $e(\overline{z}) = \overline{e(z)}$ so for example $|e(iy)|^2 =$ e(iy)e(-iy) = e(0) = 1, using the addition theorem. Thus |e(iy)| = 1 and |e(x+iy)| = e(x), etc.

3.2 The trig functions (44)

Sine and cosine are defined in terms of e(±iz) in the usual way. We then obtain the series for each directly, and Euler's formula, and in fact we can derive ALL the formulas of trig in this way. So everything is defined and we have never talked about "geometry" or "angles". This is a completely non-geometric view of trig, that is why it is so strange.

3.3. The Periodicity (45)

If we write e(z+c) = e(z) and if this is true for some c, the smallest such c would be called the period. But tradition makes us do this like so: e(iy+ω) = e(iy). In a very unusual fashion A proves there is some period, done bottom page 45. He calls this period $\omega = 4y_0$. He shows it is the smallest period and renames it to be ω_0. Easy then to show that period in general is $N\omega_0$. He defines the number 2π as being the size of this period, $\omega_0 = 2\pi$. For the first time on page 46 he refers to the number e. Well, that number would be e(0) from the power series, so we have a relation between e(0) and 2π.

Finally he allows as e(iy) describes a unit circle as y increases.

Again, no geometry! No drawings of triangles.

3.4 The Logarithm (46)

Defined as the solution for z of the equation w = e(z), and written z = log(w) in this book. We get the usual log(ab) rule and then of course we know $\log(Re^{i\theta}) = \log(R) + i\theta$ as top page 47, but θ is just called "the argument", not called an angle. Page 47 A says log(w) = log|w|, log|w| + i2π, log|w| + i4π and so on, so the mapping is that a single complex value w has many logarithms, a 1 to many mapping. So this would not then be a "function", but he does not bring that out yet.

His last gasp is to show how you define things like $\cos^{-1}w$ all in terms of what we have.

Now let's look at **how a transcendental function is defined**, versus an algebraic function:

In mathematics, an **algebraic function** is informally a function which satisfies a polynomial equation whose coefficients are themselves polynomials. For example, an algebraic function in one variable x is a solution y for an equation

$$a_n(x)y^n + a_{n-1}(x)y^{n-1} + \cdots + a_0(x) = 0$$

where the coefficients $a_i(x)$ are polynomial functions of x. A function which is not algebraic is called a transcendental function.

Here then are some facts about **algebraic functions**:

First, note that any polynomial is an algebraic function, since polynomials are simply the solutions for *y* of the equation

$$y - p(x) = 0.$$

More generally, any rational function is algebraic, being the solution of

$$q(x)y - p(x) = 0 \implies y = \frac{p(x)}{q(x)}.$$

Moreover, the *n*th root of any polynomial is an algebraic function, solving the equation

$$y^n - p(x) = 0 \implies y = \sqrt[n]{p(x)}.$$

Surprisingly, the inverse function of an algebraic function is an algebraic function. For supposing that *y* is a solution of

$$a_n(x)y^n + \cdots + a_0(x),$$

for each value of *x*, then *x* is also a solution of this equation for each value of *y*. Indeed, interchanging the roles of *x* and *y* and gathering terms,

$$b_m(y)x^m + b_{m-1}(y)x^{m-1} + \cdots + b_0(y) = 0.$$

Writing *x* as a function of *y* gives the inverse function, also an algebraic function.

You can loosely regard algebraic functions as functions which can be formed by the usual algebraic operations: addition, multiplication, division, and taking an nth root

So a transcendental function is any function that is not algebraic. The expo, log and trig functions are all in this class. Ahlfor's point is that we can think of there being **only ONE basic transcendental function**, and it is our e(x), and all other transcendental functions we know about can be written in terms of this one.

Notice that in this section there is no mention of what the value of e(0) might be, nor of 2π which is the period of e(x).

Ahlfors credits Gauss with really first understanding the Euler formula, but says no more.

I imagine Ahlfors's students might have had a rough time with a section like this one. It is a bit "radical" in claiming that trig exists without regard to geometry. No angles, no ratios of triangle sides, etc.

Ahlfors has that same quality as Stakgold of wanting to jam in a densepack proof of every little detail. Some day those proofs will help the aged student no doubt.

Chapter 3. Analytic Functions as Mappings (49)

1. Elementary Point Set Topology.

Topology includes many subfields. The most basic and traditional division within topology is **point-set topology**, which establishes the foundational aspects of topology and investigates concepts inherent to topological spaces - basic examples being compactness and connectedness; **algebraic topology**, which generally tries to measure degrees of connectivity using algebraic constructs such as homotopy groups and homology; and **geometric topology**, which primarily studies manifolds and their embeddings (placements) in other manifolds. Some of the most active areas, such as low dimensional topology and graph theory, do not fit neatly in this division.

1.1 Sets and Elements (50).

Here we have a quick review of all the famous buzzwords of "set theory". All notations familiar to me, ~X is the complement of set X. The famous union and intersection. Notice that this subject exists on its own outside the world of a metric space, there need be no distance concept. Nor vector concept.

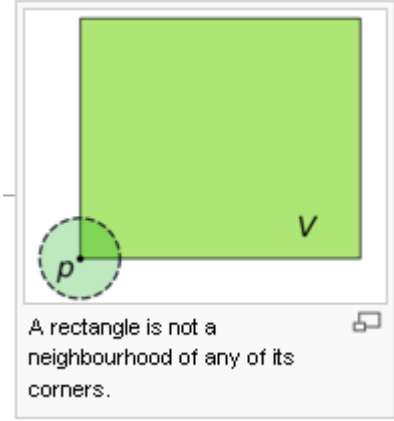
1.2 Metric spaces(51).

The metric d(x,y) has its usual 3 properties listed (if they are met, then d is a metric). Now we have a notion of DISTANCE between points in a set or in a space. The first concept is the **δ-neighborhood** of a point: the set of all points which are less than δ in *distance* from the point. So you can only have such thing as a δ-neighborhood in a set which has a metric, ie, in a metric space. This then is true from all the things that follow on.

A has been a little unclear about his definition here in the following sense. Consider these two pictures: In the right picture, what is the δ-neighborhood of the rectangle corner with respect to the rectangle? According to his definition, it would be the quarter disk in V. Then of course V contains that quarter disk so V would then be a neighborhood by the definition to come below. But this is NOT what is meant. The δ-neighborhood is meant to include all of the small disk. It really is the open ball concept. So in the case on the right, you cannot put an open ball around the corner which ball is in V, so the rectangle is not a neighborhood for the point p (nor would it be for any boundary point!)



A set *V* in the plane is a neighbourhood of a point *p* if a small disk around *p* is contained in *V*.



A rectangle is not a neighbourhood of any of its corners.

A **neighborhood** of a point is any set which contains at least one δ-neighborhood of that point. You could not do this for a point on a boundary. A set is **open** if you can jam an open ball around each of its elements. It is **closed** if its complement is open. A goes on to produce a list of little claims concerning open and closed sets, then on page 53 there are more definitions like **interior** and **exterior** and **boundary** and **closure**, each with its little symbol (Int X, Ext X, ∂X, X^-). If a point has only itself in small neighborhoods, it is an **isolated point**, otherwise it is an **accumulation point**.

I have merely read this section and don't have things nailed with complete precision. The notion of openness to me means some set that does not include any part of its boundary.

1.3 Connectedness (54).

First we have a great confusion about the word "**relative**". He says you can take the interval (closed in R) [0,1] and call it S. Relative to S, consider [0,1] or [0,1). I think both these are not open relative to R, but both are open relative to S, because S is all there is. Then we have this idea applied to a relative topology. A has avoided defining the word "topology", so here it is:

Let **X** be any set and let ***T*** be a family of subsets of **X**. Then ***T*** is a **topology** on **X** if

1. Both the empty set and **X** are elements of ***T***.
2. Any union of arbitrarily many elements of ***T*** is an element of ***T***.
3. Any intersection of finitely many elements of ***T*** is an element of ***T***.

It is something involving a set and its subsets. It involves only set concepts like union and intersection, nothing about distance / metric.

A's main interest in this section is the idea of a subset of a space being connected. The rough idea is that if you can partition such a subset E of space S into 2 disjoint pieces (where neither piece is empty!), then E is not connected.

The connected subsets of the real line are intervals (open or closed or half). He expends two pages proving this seemingly obvious fact, p 55-56. In doing so, he introduces the notion of g.l.b. = inf, and l.u.b = sup. He mentions the idea of a set being **bounded**, but only in a passing manner.

What are the non-empty connected open sets (= **regions**) of the plane? In **Theorem 3** these are those such that if you pick any two points in the set, you can connected them by a polygonal line in the set. I think that is why he means by the word "polygon" here. There follows a long proof. These proofs give the reader a laboratory to practice using the various tools of the trade here, but I am skipping them for now.

Next he talks about expressing a disconnected set in terms of its connected **components**. This partitioning is unique, given that component implies as large as it can be. This is Theorem 4, and again lots of practice text concerning the proof.

Theorem 5 says the components of an open set in R^n are themselves open.

A space is **locally connected** if every point in it has a connected neighborhood. So I guess if that space were two open non-overlapping disks in R^2, the set is disconnected but is locally connected.

Number of components is always **countable**. Notion of E being **dense** in S. If space S contains at last one subset E which is dense in S, then S is **separable**. These are all Stakgold words.

1.4 Compactness (59).

Here is the topic that always confuses me, we shall see what A has to say. Well, again A has provided a sandbox in which I can some day practice on compactness if I want to -- just that in itself is quite useful.

He first defines that a metric space is **complete** if every Cauchy sequence converges. I know that this could mean the space is missing some of its Cauchy limit points, and I know you can add them if they are missing. But I guess it could also mean the sequence just bounces around forever without converging. Maybe it has multiple accumulation points.

Now jump to p 60.

definition of compact: A set X within metric space S is **compact** if X has the **Heine-Borel Property**: every open covering of X has a finite subcovering. **(Def 6)**

Theorem X: If X is compact, then it is complete.

The proof is given in paragraphs A and B on page 60, but every phrase needs full decoding. I will use this one theorem to attempt to "practice" the subject.

You at first ask: what on earth is the connection here between a finite subcover and Cauchy convergence? They seem to be completely unrelated topics.

Paragraph A. So let's now attempt to "read" the paragraph marked A on page 60. x_n is some Cauchy sequence. In order to posit that y is a point that is NOT the limit of this sequence, I guess we have to assume that it has a limit. But OK, if y is not the limit, then as we go to the Cauchy limit (call it y_0), our shrinking distance ball excludes y more and more. So sure, after some n we can say that $d(x_n, y) > 2\varepsilon$. You give me ε, I will given n_1 beyond which this is true. But just call this value of n_1 simply "n". But all he cares about is that "there are an infinite number of n values for which $d(x_n, y) > 2\varepsilon$". Sure, they are all those with $n > n_1$ for example.

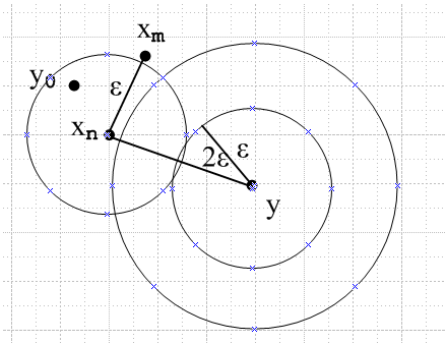
Next, there will exist some n_0 for both indices such that $d(x_n, x_m) < \varepsilon$. But let's make sure that n is large enough so we still have $d(x_n, y) > 2\varepsilon$. Now here is a triangle rule for the metric:

$$d(x_n, y) \leq d(x_n, x_m) + d(x_m, y)$$

which we rearrange to say

$$d(x_m, y) \geq d(x_n, y) - d(x_m, x_n) \geq 2\varepsilon - \varepsilon = \varepsilon$$

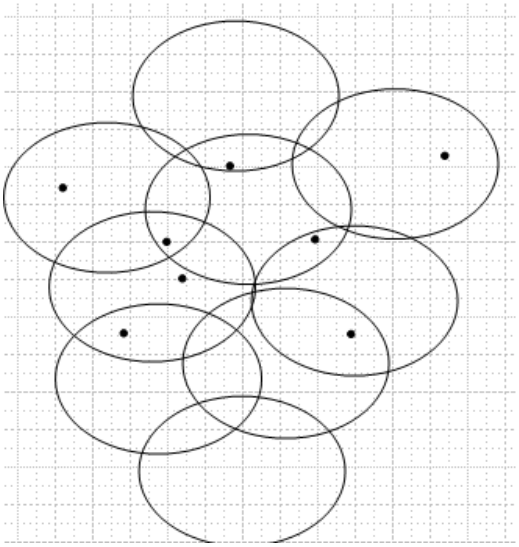
Now I have to draw a picture (an Ahlfors weakness, attributable to the cost of pictures in books at that time, and to the need to keep the page count small).



The two Cauchy guys x_n and x_m are converging to some y_0. By the time we are looking (for n and m large enough), we have shown that the distance from y to the x_m is > ε, so the m tail of the Cauchy sequence lies outside our y ε disk, which means our y ε disk only contains a finite number of the x_n

points! **That is to say, $N_\varepsilon(y)$ contains x_n for *finitely* many n.** So far so good. The real test is to see if I can find where this fact we just worked hard to prove gets used in the next paragraph somewhere.

Paragraph B: We now start paragraph B. Imagine the x_n all splattered around and we cover the entire space X with some open sets each of which captures only a finite number of the x_n. Here is my picture of this:



I suppose if x_n converges, one of my sets would have an infinite number of x_n in it. So in order to keep a finite number of points in each set, we assume x_n is not convergent and we have our open covering. BUT, we are supposed to be assuming (condition of the theorem we are now proving) that X is compact, so by definition of compact this covering must have a finite subcovering, label the sets U_1 through U_N. But if we add up all the points in all these sets, we get a finite number. But that is a contradiction because a sequence does not have a finite number of points. Thus, our assuming that x_n did not converge was wrong, so x_n must converge. Thus we have proven that compact => complete.

But sadly I don't see where the result of paragraph A got used Maybe it just shows that you can create a set which has as finite number of elements of x_n. That set was $N_\varepsilon(y)$.

OK, end of practice session. Somehow the proof involved that fact that a sequence must have an infinite number of points, and we constructed a covering which has only a finite number of points and that made a contradiction.

Let's now just quote theorems so at least we can know WHAT is being claimed, even if we cannot prove anything.

1. A Heine-Borel compact space is complete, meaning all Cauchy sequences converge. (Theorem X)

2. A compact set is bounded.

3. If you can cover a set X with a finite number of $N_\varepsilon(y)$ neighborhoods, set is "totally bounded". **(Def 7)**

(skip numbers 4,5 so can use 6 label to match theorem number)

6. (set is compact) $\Leftrightarrow$ (set is complete and totally bounded) **[Theorem 6]**

Break this down:

6a. (set is compact) $\Rightarrow$ (set is complete) // same as Theorem X
6b. (set is compact) $\Rightarrow$ (set is totally bounded)

Pause: I guess if compact, it has that finite covering capability, and those sets could be little $N_\varepsilon(y)$ neighborhoods, so can do with finite number of them, so totally bounded by definition.

6c (set is complete and totally bounded) $\Rightarrow$(set is compact)

He proves this piece starting page 61 A, but I am not going to be suckered into trying to read the proof because I know it will be confused and will take me several hours, as was the case with 1 above.

7. (A subset of R or C is compact) $\Leftrightarrow$ (it is closed and bounded)

This is the traditional big result that is what applies to us here.

Now getting to page 62, he claims to have three distinct "characterizations" of compactness.

1. The Heine-Borel definition of compactness : any open cover has a finite subcover

2. The Theorem 6 idea that set is compact if it is complete and totally bounded.

3. The idea that every subsequence must have a convergent subsequence (limit point). [Theorem 7].
This is the **Bolzano-Weierstrass** property if you make this be your definition of compactness, or it is the BW Theorem if you use the Heine-Borel definition.

Conclusion: if you really want to learn about this stuff, you have to find a book that is written for the beginning student of the subject. Such books are hard to find. Of course my sources so far are always trying to jam the entire subject into a few pages. What would the name of this subject even be? I think the right word is "topology" and in particular "point set topology", the title of our current section.

Sample book:



Introduction to Topology:
Third Edition (Paperback)
by Bert Mendelson

Preface *ix*

1 Theory of Sets 1

1 Introduction 1
2 Sets and subsets 2
3 Set operations: union, intersection, and complement 4
4 Indexed families of sets 7
5 Products of sets 9
6 Functions 11
7 Relations 15
8 Composition of functions and diagrams 17
9 Inverse functions, extensions, and restrictions 21
10 Arbitrary products 25

2 Metric Spaces 29

1 Introduction 29

2 *Metric spaces* 30
3 *Continuity* 35
4 *Open balls and neighborhoods* 40
5 *Limits* 47
6 *Open sets and closed sets* 52
7 *Subspaces and equivalence of metric spaces* 58
8 *An infinite dimensional Euclidean space* 66

3 Topological Spaces 70

1 *Introduction* 70
2 *Topological spaces* 71
3 *Neighborhoods and neighborhood spaces* 75
4 *Closure, interior, boundary* 81
5 *Functions, continuity, homeomorphism* 87
6 *Subspaces* 92
7 *Products* 97
8 *Identification topologies* 101
9 *Categories and functors* 107

4 Connectedness 112

1 *Introduction* 112
2 *Connectedness* 113
3 *Connectedness on the real line* 119
4 *Some applications of connectedness* 122
5 *Components and local connectedness* 130
6 *Path-connected topological spaces* 133
7 *Homotopic paths and the fundamental group* 139
8 *Simple connectedness* 151

5 Compactness 157

1 *Introduction* 157
2 *Compact topological spaces* 158
3 *Compact subsets of the real line* 164
4 *Products of compact spaces* 168
5 *Compact metric spaces* 172
6 *Compactness and the Bolzano-Weierstrass property* 179
7 *Surfaces by identification* 186

Bibliography 201

Index 203

So this book is doing exactly what Ahlfors is doing in this chapter. Heine Borel comes up p 167 which I cannot look at. This is a 1990 Dover 8.50 new. Amazon shows LOTS of books like this one, so I could some day pursue this I suppose.

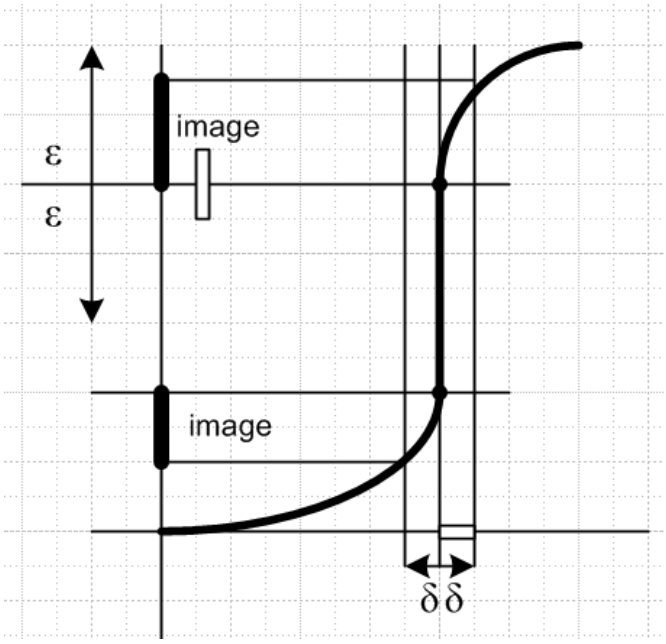
1.5 Continuous Functions (64).

Comment before starting to read this section: In the previous sections, we worried about spaces and properties of sets in these spaces, such as a set being bounded, or being compact, or being connected, or being complete, or being dense in some other space. The space has a metric so we can do all these things (eg, determine if bounded). He did not use the phrases vector space or linear space, we were dealing with points in a metric space, we only used distance ideas, not vectors and norms and such.

Now in this section I am guessing we are going to look at functions which map such spaces into other such spaces, and we are maybe going to be interested particularly in functions which are continuous -- in general one is usually interested in such

So, he starts off with the usual ideas, writes f:S→S'. Talks about an inverse mapping f^{-1}. The metric in S is called d, and in S' is called d'. We are talking metric spaces, there is no mention of vectors, just points in a metric space! A function is continuous if, as you approach x' in the image, you approach x in the domain. So for any ε of closeness in the range, you can find domain closeness δ that achieves that ε.

Let's remind ourselves of why a discontinuous function f:R→R violates this ε δ rule.



Here is a discontinuous function. I have shown an ε. (I use 2ε to make it look as it might in C). There is no δ we can pick that gets the image into the ε ball shown! We always have that lower image piece due to the left δ that we cannot get rid of no matter how small δ gets. The problem is that we have an infinite slope. If this slope were finite, we would not have the problem.

This same picture shows another fact that we are about to state. The vertical white rectangle shows an open set in the range. The inverse map of this thing is the horizontal rectangle shown on the right δ side. This rectangle (really a line segment) obviously includes its left end, and includes it quite solidly! Every

lower part of the image rectangle maps into this left end. Another way to say this is that as we come down in the vertical rectangle, when we get to the center, we get to the center of the horizontal rectangle. This point is not on the boundary of the vertical rectangle, but it *is* on the boundary of the horizontal rectangle. So the point is that the horizontal segment is NOT an open set, whereas the image one was.

The claim (p64 bot) is that if the function were continuous, then every range open set would map back into a domain open set, and here we have seen how that fails to happen.

Note: Ahlfors does not use the word "range", he uses the word "image".

Definition: f is **continuous** at x_1 means this: for any ε, I can find a δ small enough such that such that $|x-x_1| < \delta$ causes $|f(x)-f(x_1)| < \varepsilon$. He states this as $d(x,x_1) < \delta$ causes $d'(f(x),f(x_1) < \varepsilon$. The d' reminds us that the metric in the image space could in theory be different from the metric in the domain. My first notation does not make that quite clear, since I just use an unlabelled norm.

Theorem Y: (f continuous) ⇔ (the *inverse* image of any open set is an open set)

You can replace the word open with closed in both places and get another theorem. However, I am unable to find an image closed set which maps back into a domain *non*-closed set using my picture, but it must be possible somehow. I cannot find any position for my vertical rectangle with both ends included such that I get a non-closed set in the domain. Obviously it would have to somehow get onto the discontinuity region of the curve. If I put the vertical box in the middle and include both ends so it is closed, the inverse map is a single point, or maybe an empty set. Maybe somehow this is not closed in the domain.

[p 65]

Theorem 8: If f is continuous, it maps any compact set into a compact set.

Time to practice again, we are at paragraph A on page 65. Turn on your decoder ring! f(X) is the image of our compact set X. Cover this image with some U_i (open sets). There is a corresponding covering of X in the domain you get from the open sets $f^{-1}(U_i)$. But X is compact, let's select a finite subcover from our set of $f^{-1}(U_i)$. The corresponding selected U_i will then be a finite subcover of f(X) in the image. Thus, the original possibly infinite cover $\{U_i\}$ of the image has a finite subcover, so the image is compact! One key thing is that the inverse maps $f^{-1}(U_i)$ are open sets because f is continuous, by our last theorem! A cover has to be *open* sets.

Theorem 9: If f is continuous, it maps any connected set into a connected set.

If f is **one-to-one** and uses up all the range (is **onto**), then you can invert any point in the range (image) and therefore $\mathbf{f^{-1}}$ **exists** and everything is nice. IF in addition both f and f^{-1} are continuous, then such a mapping is called a **topological mapping** (= **homeomorphism**) -- "the isomorphism of the topological world". Such a mapping preserves certain topological properties. Two such properties that are preserved are compactness and connectedness, according to our theorems 8 and 9. Openness is not such a property.

[p 66]

Definition: **uniform continuity**. Let x' = f(x), etc.

For any ε, can find δ such that d(x,y)< δ causes d'(x',y') ≤ ε where δ does not depend on the points

Examples: Consider $f(z) = z^2$. Then d(x,y) = |x-y| and $d'(x',y') = |x^2-y^2| = |x-y|\,|x+y|$. Now suppose

$|x-y| < \delta$

Then we know that

$|x-y|\,|x+y| < \delta\,|x+y|$

Here δ is just a free parameter, so let's define another parameter $\varepsilon = \delta\,|z+y|$ and $\delta = \varepsilon/|z+y|$. Then we have

$|x-y| < \varepsilon/|x+y| \quad \Rightarrow \quad |x-y|\,x+y| < \varepsilon$

So the largest δ we can use is $\delta = \varepsilon/|x+y|$. But we want δ NOT to depend on x,y and to just depend on ε. We want a δ that will work for ALL x and y -- that is the idea of "uniform". But if S = the full plane, you would need δ = 0 to make it go. So there is no δ>0 that works! Thus, $f = z^2$ is *not* uniformly convergent *on the z plane*. It would be on some finite piece of the z plane.

In contrast, f(x) = z is uniformly continuous on the entire z plane, you just pick δ = ε and you have it.

I am used to uniform convergence where a parameter is in a set, but this idea is new to me. You need a δ that works for every pair in your little domain ball.

Theorem 10: If continuous f acts on a compact set, then f is uniformly continuous on that set.

This sort of matches my idea above that if $f = z^2$ is defined on a bounded set, then you would get uniform continuity.

1.6 Topological spaces (67).

This is very strange again. He wants us to get rid of the notion of "distance" which we have used as our main tool! He wants us to replace this tool with the tool of "open sets". He defines the topo space in the standard manner in Def 8, I already quoted this above. It is not a metric space with a metric (distance). It is a topological space with a "topology", namely, with a set of open sets which have the usual properties.

Example: how would you talk about a **neighborhood** without talking about distance? Here it is:

N is a neighborhood of x if we can find open set U such that $x \in U \subset N$.

Distance was never mentioned. This approach is discussed here:

http://en.wikipedia.org/wiki/Neighbourhood_(mathematics)

So this is the "thing that happened" in math since the early days! There has been a conversion from metric space thinking (with its distance between points) to topological space thinking (with its open sets). I wonder WHEN this happened. [My guess: old way 1860, new way 1920]

To the extent our earlier ideas and theorems involved only open sets (such as the Heine Borel definition of compactness), they are true for topological spaces as well as for metric spaces. But theorems that involve convergence, with its strong distance requirement, have problems. These are fixed somehow if we add the magic **Hausdorff Property** (1914) to our topo space:

Hausdorff Property: you can surround any two distinct points in topo space with disjoint open sets

This reminds me of developing geometry "axiomatically" and see what is the minimum you have to assume, your minimal axiom set. This Hausdorff thing, he claims, prevents a sequence from converging to two different points. This is pretty technical. What is the meaning of "convergence" in this topo space world in the first place?

> $x_n \rightarrow x$ in topo space means every "neighborhood" (defined above) of x contains all x_n except a finite number (which would be the head of the sequence).

So no matter how you place your open set U around x, U contains a tail of the sequence. Suppose there were two tail convergence points x and y. Well, I need help here, but I sort of see the idea. He says it is "obvious" that the Hausdorff condition makes the tail converge to a unique point. It is not obvious to the beginning student reader who has not been in the field for 35 years! Suppose there are two limit points. Suppose you can separate them with two open sets that are disjoint. Suppose the sequence just jumps back and forth between x and y, so each open set has an infinite number of points. Well I guess the entire tail has to fit in one open set. A good text on this subject would have lots of examples, but here he is just mentioning this extra "axiom" so we will have heard about it once in our lives.

A topo space with this property is a Hausdorff Space and that is the only kind of space we care about.

Complaint about Ahlfors' book: <u>There is not a single reference anywhere</u>! If you want to read more on some subject, he has zero advice for you! I wonder why he chose that path? Goldstein, Jackson, Stakgold all have references at the end of each chapter. BD have footnote references.

2. Conformality (68)

Now we are descending from the clouds, feet back on the ground with R and C, and we are going to talk about the fact that analytic mappings I think "preserve angles" and this is what conformality refers to. Probably this has lots of significance. He says this section is going to be "descriptive".

2.1 Arcs and closed curves (68).

An arc γ in C is closed and bounded hence compact. It is also connected. Describe it with a real parameter so that z(t) is your arc as t ranges (α,β) say. You could change the parameter t using some $t' = \varphi(t)$. If φ is monotonic, you "run the arc" in the same direction I guess, the speed varies, so to speak. This **change of parameter** would be **reversible** in that $\varphi^{-1}(t)$ exists if φ is monotonic.

The **tangent** to an arc z(t) would be z'(t). That is, [z(t+dt)-z(t)]/dt. Think of this as a vector in C space pointing between these two points, so this vector is tangent to the curve at z(t). The angle of this tangent vector would bet arg z'(t). If z'(t) is never 0, the arc is **regular**. If z'(t) exists and is continuous, the arc is **differentiable**, just as for a normal function defined on R.

Now let w = f(z) be a differentiable mapping. Then some arc z(t) maps into an arc w(t) = f(z(t)) in the image space. Then the tangent vector to the image arc at point w is w'(t) = f '(z) z '(t) by the chain rule.

On page 69 we get lots of little definitions. If the arc does not intersect itself (one z does not have two values of t), the arc is **simple** or a **Jordan arc**. If at the two ends of the parameter interval z(t) is the same, the arc is a **closed curve**. You could change parameters to **shift** the point of closing to a different location on the curve. A closed curve could be simple if it does not intersect itself, in which case it is called a **Jordan curve**. It could be differentiable. It could be regular if it avoids z'(t) = 0. [But how can a closed curve avoid having a zero slope at two places?] I think the **opposite arc** is what you get reversing your parameter, just run the arc the other way. If the entire arc z(t) is point A as t varies, that is called a **point curve**.

2.2 Analytic functions in regions (69).

Recall that a region is a non-empty connected set in C (see p 57).

And recall that an analytic function f(z) is one which has a derivative f '(z) on some region of interest. In order to be able to examine the derivative from all directions around z, it really makes sense to only work with regions which are open.

Comment: on page 72 it is shown that the existence of f '(z) means that f(z) is continuous, so the phrase "continuous analytic function" is a bit redundant. All analytic functions are continuous.

Def 10 states our definition of f(z) being analytic on a region Ω . I thought he was going to say here that the region had to be open, but he does not.

Def 11 says f(z) is analytic on a point set A if it is analytic in a region containing A. This seems a little forced, we get the point. You have to have some space around each z to be able to compute the derivative in all directions.

At this point (p 70 A pencil line), he wanders off and talks about how you need to define branches of analytic functions which have branch cuts, because we must have functions which are single-valued. First he does $w(z) = z^{1/2}$ in the usual manner if your open region in C is the complement of the closed negative real axis. Ie, this open region is the entire C plane minus the neg real axis including the origin. The principle branch here is where this w(z) has a positive real part. He shows that this w(z) is continuous in our selected region. He then shows how you get a cleanly defined w'(z) as shown p 71 A (though his comments there seem strange). The same branch of the square root is used here as in w(z).

Next he has the same discussion for w = ln(z) with the exact same open region. The principle branch has arg(w) in range $-\pi,\pi$ and of course this arg is 0 on the positive real axis (he has not mentioned real analytic functions, by the way).

And next is does $w = \cos^{-1}(z)$. I have ignored all the detail chatter in these sections, but will read it if I need to later.

[72] We are reminded again that you have to specify your branch clearly. Then we are reminded that analytic is the same as having C-R equations as on p 72 A.

Theorem 11: If $f'(z) = 0$ for an analytic function $f(z)$ everywhere in a region Ω, then $f(z)$ = constant.

Theorem 11A: If arg(f(z)) = constant everywhere in Ω, then f(z) = constant.
Theorem 11B: If Re((z)) = constant everywhere in Ω, then f(z) = constant.
Theorem 11C: If Im((z)) = constant everywhere in Ω, then f(z) = constant.
Theorem 11D: If |f(z)| = constant everywhere in Ω, then f(z) = constant.

Fact: if $f'(z) = 0$, then all the C-R equation derivatives are 0. I sort of skip now text below Theorem 11 and the end of this section, regarding these C-R equations. He is proving the Fact I just stated.

2.3 Conformal Mapping (73).

Finally we arrive where I want to be in this book, since this is going to relate to using conformal mapping in relation to Laplace's equation in Stakgold Vol II p 164. This served as a good excuse to reread Ahlfors at least to this point.

As noted above, we know that the tangent to an image arc is $w'(t) = f'(z(t))\, z'(t)$ and we are going to stay away from places where any of these derivatives is 0 (both domain and image arcs are "regular"). The above equation implies at once that

$\arg [w'(t_0)] = \arg [f'(z_0)] + \arg [z'(t_0)]$ $\qquad z_0 = z(t_0)$

$\arg [w'(t_0)] - \arg [z'(t_0)]$ = change in angle of the tangent going domain → image = $\arg [f'(z_0)]$

Suppose two domain arcs pass through z_0 in the domain, and have an angle of 32 degrees between them at the point z_0. Since the quantities $f'(z_0)$ and hence $\arg [f'(z_0)]$ don't know about these arcs, we find that these two arcs will be mapped into arcs in the image which have a relative angle of 32 degrees at the image of the point z_0 (ie, at $f(z_0)$. This is a big major result:

Any analytic mapping f(z) preserves angles of intersecting arcs -- the "conformal property"

I don't see why we had to exclude $f'(z) = 0$ in this discussion (tangent =horizontal). The arg addition equation has no problem with this. I suppose though that if z = 0, then $z = Re^{i\theta}$ does not have much meaning for θ.

Next, starting at p 74 A, he shows that "scale" is also preserved under analytic mapping. Consider the equation shown there which is obviously true. Imagine now a short line segment through z_0 which maps into a short line segment through $f(z_0)$. The relative lengths of these line segments is given by $|f'(z_0)|$ which again does not know anything about our line segments. Thus, the scaling of a short line segment is the same regardless of the angle of the line segment. This really means that you can think of a "relative scale factor" between a point in the domain and the mapped point in the image, which applies to segments of all angles, so it applies to the entire 2D surface. It is of course a local scale factor. Very good.

Starting at p 74 B, he proves two "converse facts" :

if angles are preserved as described above, a mapping must be analytic
if scale is preserved as described above, a mapping must be analytic

In the second case, we find that either f(z) must be analytic, or g(z) ≡ $\overline{f(z)}$ is analytic. Both f(z) and g(z) scale distances the same way, but g(z) of course flips the signs of angles. This g(z) is said to be **indirectly conformal**.

Comment: f(z) analytic does imply that g(z) = $\overline{f(z)}$ is analytic since in both cases derivative exists.

We are now at p 75 A. Recall that mapping f(z) is topological means that the inverse exists and is also analytic. Here we DO have to assume that f'(z) ≠ 0 in our region of interest, because if f'(z) = 0, the inverse derivative is ∞. Why is that?

w = f(z) z = $f^{-1}(w) \equiv g(w)$ $1 = \partial_z z = \partial_z g(w(z)) = g'(w)\,\partial_z w = g'(w)\,\partial_z(f) = g'(w)f'(z)$.
=> g'(w) = 1/ f '(z) so if f '(z) = 0, g'(w) there does not exist, not analytic

He then claims that f'(z) ≠ 0 makes your 1-1 inverse sense safe at least in a local region to a point z, something you show from the Jacobian. But he then brings up the famous picture on page 75. Even if f'(z) ≠ 0 everywhere, we see some points in the image which map into two different points in the domain, so f(z) is not single valued and we are not 1-1 and such a mapping is then not "topological". A comments on thinking of this image picture as a "transparent film" and says this idea is used in our notion of Riemann Surfaces, (sheets).

So to summarize: *(verbatim from raw notes)* ANY analytic function f(z) provides a mapping from domain to image plane which preserves both angles and scale and is thus a "conformal map". This f(z) does not have to be a rational function or anything particular, just as long as it is analytic. This seems to mean that for a tiny piece of the domain near z, we map into a tiny piece of image where things are merely rotated and scaled by some amount. There is no distortion! If you put a tiny pixel image in this domain region, the image region would show the image merely rotated and scaled, but a circle would remain a circle. Of course for *large* pixel image, this would not be the case, because conformality is a local property. I would add that this means if you define a local x,y orthogonal coordinate system in the domain, and a similar local one in the range at the corresponding point, then the image x',y' coordinate system is merely rotated and scaled relative to x,y. It seems to me this fact ought to be useful.

3. Linear Transformations (76)

A says this stuff is very useful, so pay attention.

3.1 The linear group (76).

We now restrict our interest to analytic functions which are rational Q = P/Q where Q and P are first order, we get plenty of mileage out of this simple form, called a "**linear transformation**". Notice that you can scale a,b,c,d all up or down together without changing the function f(z), so as long as ad-bc ≠ 0, you can normalize things so that ad-bc = 1. Fine.

Comment: How many degrees of freedom or free parameters does a LT have? There are four complex numbers, so 8 real degrees. But scaling all four complex numbers by complex α gives the same, so really only 6 degrees. You can remove these 2 scaling degrees by requiring ad-bc = 1. The isomorphic group SL(2,C) is of course itself isomorphic to SO(3,1), the Lorentz Group, which we know has 6 free parameters. So I would just think of a LT has having 6 true degrees of freedom.

Now comes the strange act. We are supposed to represent the one complex number z as the ratio of two complex numbers z_1/z_2. Obviously you could scale these together and not change $z = z_1/z_2$. If we do this for w and z in w = f(z), our order 1 rational polynomial can be written as a 2x2 linear matrix equation shown p 76 A. I am a little surprised at this fact, and am confused by it. We have a new 2D space of some sort in which space our f(z) appears as a linear matrix transformation! You would not call the ratio of two linear polynomials a "linear" function.

I agree that with normalization, our set of matrices form a group that is exactly SL(2,C). Thus, we can say this: "the set of linear transformations p 76 (5) form a group which is isomorphic to SL(2,C) ".

****

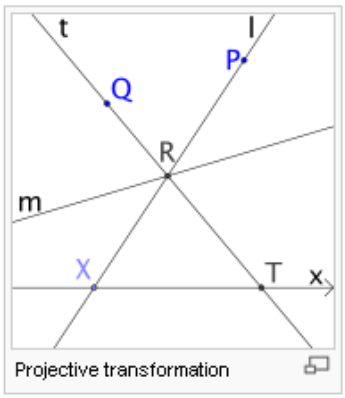
Huge Digression on the Subject of Projective Transformations and Homogeneous Coords

****

This is something I have never even heard of (age 61). A good discussion is here,

http://en.wikipedia.org/wiki/Projective_transformation

(1) First, we will do **"transformations on the projective line".** The picture is this:



The idea is that P and Q are observers looking at real object R which is sitting on some real (objective) line "m" (which we later say also has slope m). We imagine however some sort of "viewing screen" (my term) which is the x axis (subjective line). P sees R as being at X, and Q sees R as being at T. Each viewer has a different "**projection**" of the point R on the screen. The **projective transformation** is then define as T = f(X). Now if you give all the points the obvious Cartesian coordinates, like X = (x,0) and T = (t,0), you can do the geometry and here is what you find:

$$t(x) = \frac{\alpha x + \beta}{\gamma x + \delta}$$

where

$$\alpha = mQ_xP_y - Q_yP_y + bQ_y,$$
$$\beta = bQ_xP_y - bQ_yP_x,$$
$$\gamma = m(P_y - Q_y),$$
$$\delta = mP_xQ_y + P_y(b - Q_y).$$

and the line "m" was given by y = mx + b. Notice that everything is real here. We then identify this projective transformation with a first order rational function. This thing has several famous names. One is that it is a "**bilinear fractional transformation**" and it is a "**Mobius transformation**".

You can of course find the inverse transformation x = x(t), we know how to do that just from the fractional form.

The screen (x axis) is called the **subjective line**, and the actual line (line "m") is called the **objective line**.

Now, what happens if you chain two of these things in a row like this:

$$t_1(x) = \frac{\alpha_1 x + \beta_1}{\gamma_1 x + \delta_1},$$
$$t_2(x) = \frac{\alpha_2 x + \beta_2}{\gamma_2 x + \delta_2}.$$

$$\begin{aligned} t_2(t_1(x)) &= \frac{\alpha_2 \left(\frac{\alpha_1 x+\beta_1}{\gamma_1 x+\delta_1}\right) + \beta_2}{\gamma_2 \left(\frac{\alpha_1 x+\beta_1}{\gamma_1 x+\delta_1}\right) + \delta_2} \\ &= \frac{\alpha_2\alpha_1 x + \alpha_2\beta_1 + \beta_2\gamma_1 x + \beta_2\delta_1}{\gamma_2\alpha_1 x + \gamma_2\beta_1 + \delta_2\gamma_1 x + \delta_2\delta_1} \\ &= \frac{(\alpha_2\alpha_1 + \beta_2\gamma_1)x + (\alpha_2\beta_1 + \beta_2\delta_1)}{(\gamma_2\alpha_1 + \delta_2\gamma_1)x + (\gamma_2\beta_1 + \delta_2\delta_1)}. \end{aligned}$$

You see that the composition of two transformations is again a linear fractional transformation.

Now "it happens" (more later) that the new a,b,c,d coefficients can be obtained in this way:
(it is trivial to show this is true, but later we will try to learn more about it)

$$\begin{bmatrix} \alpha_2 & \beta_2 \\ \gamma_2 & \delta_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & \beta_1 \\ \gamma_1 & \delta_1 \end{bmatrix} = \begin{bmatrix} \alpha_2\alpha_1 + \beta_2\gamma_1 & \alpha_2\beta_1 + \beta_2\delta_1 \\ \gamma_2\alpha_1 + \delta_2\gamma_1 & \gamma_2\beta_1 + \delta_2\delta_1 \end{bmatrix}$$

Therefore, we can make an **isomorphism** between "linear transformations" and "2x2 matrices". The matrices would be elements of the group representation SL(2,R) where S if we normalize a,b,c,d. So we can say that the "**transformations of the projective line**" in fact form a group which is isomorphic to the group known as SL(2,R). It happens that if we negate a,b,c,d we get a new element of SL(2,R), but we get the same transformation, so there is a 2:1 issue here that is just fine by me.

The web page discusses the **cross ratio** at this point, which I skip for now. Well here is from another site that gives it in a nutshell:

Cross-ratio preservation [edit]

Cross-ratios are invariant under Möbius transformations. That is, if a Möbius transformation maps four distinct points z_1,z_2,z_3,z_4 to four distinct points w_1,w_2,w_3,w_4 respectively, then

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

If one of the points z_1,z_2,z_3,z_4 is the point at infinity, then the cross-ratio has to be defined by taking the appropriate limit; e.g. the cross-ratio of z_1, z_2, z_3, ∞ is

$$\frac{(z_1 - z_3)}{(z_2 - z_3)}.$$

(2) Second, we will do "**transformations on the projective plane**".

We redo all this line stuff in an obvious manner moving from 2D space to 3D space. Then the screen is the subjective plane (z = 0), the line "m" becomes a "objective plane " of the form z(x,y) = mx +ny + b. Write this as - mx - ny + z = b and it describes a plane with normal $\hat{\mathbf{n}}$ = (-m,-n,1) and if x=y=0, z = b. We can then ask how X = (x,y) is related to T = ($T_{\mathbf{x}}$, $T_{\mathbf{y}}$). The answer is this:

$$T_x = \frac{x(-mQ_xP_z - nQ_zP_y + Q_z(P_z - b)) + (ny + b)(Q_zP_x - Q_xP_z)}{(mx + ny)(Q_z - P_z) - (mP_x + nP_y)Q_z + (Q_z - b)P_z},$$
$$T_y = \frac{y(-nQ_yP_z - mQ_zP_x + Q_z(P_z - b)) + (mx + b)(Q_zP_y - Q_yP_z)}{(ny + mx)(Q_z - P_z) - (nP_y + mP_x)Q_z + (Q_z - b)P_z}.$$

where are observers are still at P and Q. You can put this into the following form,

$$T_x = \frac{\alpha x + \beta y + \gamma}{\delta x + \epsilon y + \zeta}. \qquad T_y = \frac{\eta x + \theta y + \kappa}{\delta x + \epsilon y + \zeta}.$$

$$\alpha = -mQ_xP_z - nQ_zP_y + Q_z(P_z - b),$$
$$\beta = n(Q_zP_x - Q_xP_z),$$
$$\gamma = b(Q_zP_x - Q_xP_z),$$
$$\delta = m(Q_z - P_z),$$
$$\varepsilon = n(Q_z - P_z),$$
$$\zeta = -(mP_x + nP_y)Q_z + (Q_z - b)P_z,$$

$$\eta = m(Q_z P_y - Q_y P_z),$$
$$\theta = -mQ_z P_x - nQ_y P_z + Q_z(P_z - b),$$
$$\kappa = b(Q_z P_y - Q_y P_z),$$

and these babies are called **"*trilinear* fractional transformations"**. We have 9 parameters, the denominators each have 3 of the parameters and they are the same.

As in the previous case, you can show that chaining two of these things gives a third trilinear transformation, and the group is isomorphic to SL(3,R) and we have this rule:

$$\begin{bmatrix} \alpha_2 & \beta_2 & \gamma_2 \\ \eta_2 & \theta_2 & \kappa_2 \\ \delta_2 & \epsilon_2 & \zeta_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \eta_1 & \theta_1 & \kappa_1 \\ \delta_1 & \epsilon_1 & \zeta_1 \end{bmatrix} = \begin{bmatrix} \alpha_3 & \beta_3 & \gamma_3 \\ \eta_3 & \theta_3 & \kappa_3 \\ \delta_3 & \epsilon_3 & \zeta_3 \end{bmatrix}.$$

Now, we have said that the **"viewing screen" or subjective plane is the x-y plane**. Suppose we consider not just a point on this plane, but a **straight line on this plane**. Take the line y = mx+b on this viewing screen. This is the line seen by viewer P as the "objective point R" moves along some unknown line in the objective plane. The other viewer Q sees some other line on the viewing screen which we will call the transformed line. It will be $T_y = nT_x + c$, where this n is different from our previous n. In fact,

$$n = \frac{m(\epsilon\kappa - \zeta\theta) + b(\delta\theta - \epsilon\eta) + (\delta\kappa - \zeta\eta)}{m(\epsilon\gamma - \zeta\beta) + b(\delta\beta - \epsilon\alpha) + (\delta\gamma - \zeta\alpha)}$$

and

$$c = \frac{m(\beta\kappa - \gamma\theta) + b(\alpha\theta - \beta\eta) + (\alpha\kappa - \gamma\eta)}{m(\beta\zeta - \gamma\epsilon) + b(\alpha\epsilon - \beta\delta) + (\alpha\zeta - \gamma\delta)}.$$

So we might call this the "transformation of a line under a planar projective transformation" . There is no analogous thing for the "linear projective transformation" since the only line on the screen is the x axis.

Here then is an interesting claim":

Planar transformations of conic sections

A trilinear transformation such as *T* given by equations (14) and (15) will convert a conic section

$$Ax^2 + By^2 + Cx + Dy + Exy + F = 0 \qquad (17)$$

into another conic section

$$A'T_x^2 + B'T_y^2 + C'T_x + D'T_y + E'T_xT_y + F' = 0. \qquad (18)$$

That is to say, for the input side instead of looking at a line that P sees, suppose he sees a conic section. Then viewer Q ALSO sees a conic section. The input side might be a circle, and the output side an ellipse, for example (this example is in fact given in this long web page).

(3) We can continue this idea up to higher dimensions all we want. For example, if our viewers could exist in a 4D space, we could have the subjective and objective items be 3D "planes" and then we end up with SL(4,R) and this situation:

$$T_x = \frac{\alpha x + \beta y + \gamma z + \delta}{\epsilon x + \zeta y + \eta z + \theta},$$
$$T_y = \frac{\iota x + \kappa y + \lambda z + \mu}{\epsilon x + \zeta y + \eta z + \theta},$$
$$T_z = \frac{\nu x + \xi y + o z + \rho}{\epsilon x + \zeta y + \eta z + \theta}.$$

$$M_T = \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \iota & \kappa & \lambda & \mu \\ \nu & \xi & o & \rho \\ \epsilon & \zeta & \eta & \theta \end{bmatrix}$$

and now we have "quadrilinear fractional transformations".

The idea of **homogeneous coordinates** is this: We can achieve the above quadrilinear transformations if we use a 4-vector (x,y,z,1) to represent our domain point in the subjective space. Then if we apply the above 4x4 matrix to this vector, we get a new 4-vector (A_x,A_y,A_z,A_t) whose first three components are the *numerators* for $T_{x,y,z}$ and whose fourth component T is common *denominator* the T_i all have! We can then construct the image point of this transformation from this transformed 4-vector as follows:

$T_x = A_x/A_t$	x = x/1
$T_y = A_y/A_t$	y = y/1
$T_z = A_z/A_t$	z = z/1

We can think of the input coordinate in this way as well, shown on the right above. So when we "represent" a point T = (T_x, T_y, T_z) by A = (A_x,A_y,A_z,A_t), the T coordinates are the *actual* coordinates, while the A coordinates are the "*homogeneous* coordinates". Our web author has a strange way to write this:

$$T : [x : y : z : 1] \rightarrow [\alpha x+\beta y+\gamma z+\delta : \iota x+\kappa y+\lambda z+\mu : \nu x+\xi y+oz+\rho : \epsilon x+\zeta y+\eta z+\theta].$$

where he is just saying our transformation maps (x,y,z,1) into (A_x,A_y,A_z,A_t).

One feature of the homogeneous coordinates is that if you scale all four components by some constant K, the new 4-vector (KA_x,KA_y,KA_z,KA_t) corresponds to the same (T_x, T_y, T_z). So there is a many to 1 relation here in the sense of this scaling. I have seen this represented by a group notation maybe P(V).

It seems that the *tradition* is to write homogeneous coordinates in this way: (A_x:A_y:A_z:A_t)

In our isomorphism between the group of actual transformations and the group of linear 4x4 matrix transformations, the points like (x,y,z) correspond to 4-vectors (Kx : Ky : Kz : K) for any K. This somehow reminds me of Pauli spin matrix stuff somehow, but I forget the connection. The 4x4 stuff is a "matrix representation" of the group SL(4,R), but the quadrilinear transformations "form a group". In similar fashion, we say that rotations form the rotation group, but of course we know lots of matrix

representations of the rotation group. I guess you would call our current stuff "the projective group" in N dimensions, name then P(N,R) , and we started above with N = 1, and we are on the reals R.

In our original projective line transformation above, our domain points were just thinks like x, while the homo coordinate representation would be (Kx : K). If we write this as (A_x:A_y), then we get'

$Kx = A_x$ $K = A_y$ $x = A_x / A_y$

So this is what Ahlfors is talking about when he says to represent $z = z_1/z_2$.

Digression on Affine Transformations

In geometry, an **affine transformation** or **affine map** or an **affinity** (from the Latin, *affinis*, "connected with") between two vector spaces (strictly speaking, two affine spaces) consists of a linear transformation followed by a translation:

$$x \mapsto Ax + b.$$

If A were a rotation (det A = 1), this would be a Galilean transformation, but any matrix A is allowed, including "shear" and "scaling" matrices and such. The claim is that distance between points and collinearity of points is preserved by such an affine transformation. It is of course a map between a vector space and itself.

From a computer graphics computational point of view, you can compute the affine transformation using matrix methods if you are willing to "add a dimension" in this way: (matrix is still square)

$$\begin{bmatrix} \vec{y} \\ 1 \end{bmatrix} = \begin{bmatrix} A & \vec{b} \\ 0, \ldots, 0 & 1 \end{bmatrix} \begin{bmatrix} \vec{x} \\ 1 \end{bmatrix}$$

is equivalent to the following

$$\vec{y} = A\vec{x} + \vec{b}.$$

So you could think of this as a homogeneous coordinates situation as we just described above in the sense that we have added a fourth coordinate, but here that fourth coordinate never seems to change from 1. But I suppose in the above projective stuff, you could always scale K to make the 4th coordinate 1, then the first coordinates would be your true "answer", and that then looks just like this affine transformation.

Obviously you can use matrix concatenation to compute the chaining of these things. We probably did that in the FGS-4000 days, but I cannot remember clearly.

Review of Projective Transformations

Now that we have the basics understood, we can comprehend the start of our original wiki page:

A **projective transformation** is a transformation used in projective geometry: it is the composition of a pair of perspective projections. It describes what happens to the perceived positions of observed objects when the point of view of the observer changes. Projective transformations do not preserve sizes or angles but do preserve incidence and cross-ratio: two properties which are important in projective geometry. A projective transformation can also be called a **projectivity**. Projectivities form a group.[1]

As important special cases, a projective transformation can be in the (real) one-dimensional projective line RP^1, the two-dimensional projective plane RP^2, and the three-dimensional projective 3-space RP^3.

Notice the first appearance of the word "perspective" and "projectivity". Everything makes sense. The comment about sizes and angles not being preserved is good. That only happens when we talk about sizes and angles between arcs in a complex version of these transformations, and that subject has not come up yet, despite the fact that we do have a linear transformation here. This is a smart author. When we later do things complex, those names will become CP^1 and CP^2 .

My own notes added: about "going complex" with the projective line .

Now let's go back to the "projective transformation on a line" section above. Recall that X = (x,0) and T = (t,0) where everything was real. Suppose we now allow x and t to be complex numbers. Then X is no longer a point in R^2 , it is a point in C^2. All the math of course goes through exactly the same and we get our isomorphism now between the "projective linear transformation on a complex line" (y = mx + b where everything is now complex) and the group SL(2,C) if we normalize. This then explains Ahlfors mysterious reference on page 77 top to the **"complex projective line"**. The image point seen by observer Q is now T = (t,0) where $t = (\alpha x+\beta)/(\gamma x + \delta)$ and is also a complex number. Points are in (C,C) = C^2 as just noted.

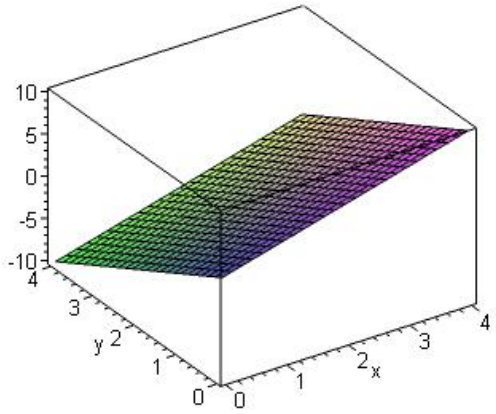
Now of course there are two distinct "complex projective lines" in our picture, one is the objective line y = mx+b (all complex now) and the other is the viewing screen subjective line. As noted below, both these "lines" are more like planes than lines. So the phrase "points on a complex projective line" could refer to either of our two "lines". But when we are talking projective transformations, both X and T = f(X) are points on the subjective complex projective line, so that is probably what one usually means.

We have seen how we can represent these points on the complex projective line either as their "actual" representation X = (x,0) which is a point in CxC, or in their homogeneous coordinate representation which would be $X = (x_1:x_2 ; 0:1)$ say. I am trying to show that the one number x of X = (x,0) is being represented by two coordinates, so $x = (x_1:x_2)$ and $x = x_1/x_2$. So maybe this is the usefulness of the colon notation, so you know that you are talking homogeneous coordinates. It is clear that we can have $x_1 = 0$, or even $x_2 = 0$ so we represent the points x = 0 and ∞. But I think things are undefined if you try to set both to 0, so x = 0/0, This is why you see Ahlfors' p 77 near top comment about the ratios.

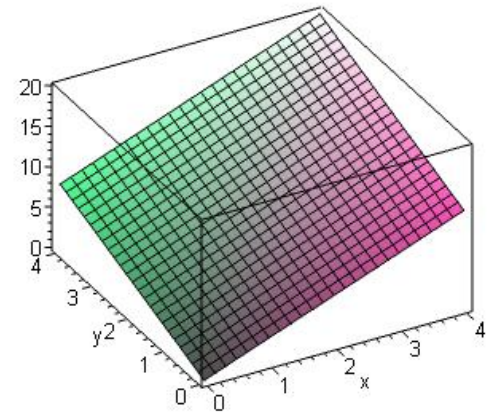
"The ratios $z_1:z_2 \neq 0:0$ are the points of the complex projective line. "

You might be wondering exactly what a "complex line" looks like. We know what a real line y = mx + b looks like when we plot it in the x-y plane. The complex line $w = \alpha z + \beta$ can be plotted in two 3D pictures if we plot the real and imaginary parts of w, and take z = (x,y). Not surprisingly, you find that the "complex line" is a plane in each of these 3D plots, but it is not the same plane. Here is an example where I set $\alpha = 2+3i$ and $\beta = 2$.

```
> plot3d(Re((2+3*I)*(x+I*y)+2), x=0..4,y=0..4);
```




```
> plot3d(Im((2+3*I)*(x+I*y)+2), x=0..4,y=0..4);
```



Question: What does this "complex projective line" have to do with the Riemann Sphere?

The sphere is a mapping from C to a spherical surface in R^3 as best I can tell, and vice versa. All the mapping equations are shown on Ahlfors p 18. If we let z = x+iy, we can think of it as a mapping (x,y,0) $\rightarrow (x_1, x_2, x_3)$

OK, I have a lead now. First of all, we know that if we take the point x,y = **r** in R^2, we can describe the points on a straight line through this point and the origin by $\mathbf{r}_\alpha = \alpha\ \mathbf{r}$ where α is a real number. We can write $\mathbf{r}_\alpha = (\alpha x, \alpha y)$ if we like. Consider the real number t = x/y. For all points on our line, this number is the same (it is the slope of the line really).

Now extend all this to complex. Suppose we take $z_1, z_2 = \mathbf{w}$ in C^2 . We can describe the points on a straight complex line through this point and the C^2 origin by $\mathbf{w}_\alpha = \alpha\ \mathbf{w}$ where α is complex. We can then write $\mathbf{w}_\alpha = (\alpha z_1, \alpha z_2)$ if we like. Consider the complex number $z = z_1/z_2$. For all points on our complex line, that is for all points $(\alpha z_1, \alpha z_2)$, z is the ratio of the coordinates. Notice that if $z_2 = 0$, $z = \infty$.

I am failing to see the connection between a Mobius transformation (linear xform) and the 3D Riemann sphere. Am now scanning the web. // Web not coming through.

I know that Mobius (linear transformation) with complex numbers everywhere:

(1) maps the C plane to the C plane;
(2) Mobius is associated with the projective map of the line, extended to the complex line;
(3) is a group isomorphic to our 2x2 matrix stuff.

Now I also know that we have a map between the sphere and the C plane, so you could think of the Mobius transformation if you like as a mapping from the sphere to the sphere (the sphere surface of course).

[Go back to the complex line idea above where we wrote $\mathbf{w}_\alpha = \alpha\ \mathbf{w}$ where $\mathbf{w} = (z_1, z_2)$. You can I suppose thing of this as a 4-tuple $\mathbf{w} = (z_1, z_2) = (x_1, y_1, x_2, y_2)$ where we show the real and imaginary parts. So now we have a line in some kind of R^4 space. I so see some web reference to Hopf circles that might be related to this.]

(end of huge digression on projective transformations)

****

We return now to Ahlfors page 76 in section 3.1 there.

OK, we have our Mobius thing and its inverse. We know they form a group, and we know this group is isomorphic to SL(2,C) if we normalize the matrices ("special"), and we know that the 2x2 matrices act on the homogeneous coordinates where we write (z_1, z_2) and $z = z_1/z_2$.

He does come up with a name for our group of projective transformations: P(1,C) which I guess is that CP^1 business. The 1 refers to this being projective on the (complex) line, not the (complex) plane which would be P(2,C), etc.

Next, starting page 77 A, he writes three fundamental 2x2 matrices and gives each a name and for each shows the corresponding w = w(z) equation. The **translation** one is very clear. The **rotation** one means we have $w = e^{i\varphi}z$, just a phasor multiplication. If $k \neq 1$, $w = kz$ is called a **homothetic** transformation, a very odd name for multiplication by a constant! The third primitive matrix gives $w = 1/z$, **inversion**. He then shows how the general Mobius can be expressed as a sequence of these primitive transforms, fine.

3.2 The cross ratio (78)

He first writes down a certain transformation

$$w = S(z_2,z_3,z_4)\ z \quad (7)$$

which maps $z_2 \to 1$, $z_3 \to 0$ and $z_4 \to \infty$. This thing IS in fact just a particular Mobius transform (ie, a particular linear transform). So that is an interesting fact: a mapping that takes an arbitrary 3 points in C to these special 3 points. He shows that S is unique.

Now he makes this definition by simply applying the above transform to z_1 :

Cross Ratio = $S(z_2,z_3,z_4)\, z_1 = (z_1 - z_3)(z_2 - z_4) \;/\; (z_1 - z_4)(z_2 - z_3)$

$= [(z_1 - z_3)/(z_1 - z_4)] \,/\, [(z_2 - z_3)/(z_2 - z_4)]$

$= [(z_1 - z_3)/(z_1 - z_4)] * [(z_2 - z_4)/(z_2 - z_3)]$

$=$ " (z_1,z_2,z_3,z_4)" // strange notation

Yes it is "the image of Sz_1" where S is the special linear transform noted.

Notation: Here are various alternate versions of the above:

$(z_1,z_2,z_3,z_4) = S(z_2,z_3,z_4)\, z_1 = (z_1 - z_3)/(z_1 - z_4) * [(z_2 - z_4)/(z_2 - z_3)]$

Do backward cyclic one, then on second line replace z_4 by z :

$(z_4,z_1,z_2,z_3) = S(z_1,z_2,z_3)\, z_4 = (z_4 - z_2)/(z_4 - z_3) * [(z_1 - z_3)/(z_1 - z_2)]$

$(z,z_1,z_2,z_3) = S(z_1,z_2,z_3)\, z = (z - z_2)/(z - z_3) * [(z_1 - z_3)/(z_1 - z_2)]$

$= (z - z_2)/(z - z_3) * (1/Z_{123})$

where $Z_{123} = [(z_1 - z_2)/(z_1 - z_3)]$

He then makes his claim in

Theorem 12: the cross ratio is invariant under linear transformations.

Proof. His proof is very efficient -- I would have written everything out to get a big mess. As usual, A manages to make his proof be hard to follow, so I will expanded his ultra-densepack statements to a human-readable proof.

Warmup Problem: Suppose someone tells you this fact about an LT called R:

$R\,[z_2,z_3,z_4] = [1,0,\infty]$ // ie, it does this mapping on three points

You would conclude that $R = S(z_2,z_3,z_4)$, since $S(z_2,z_3,z_4)$ does this same mapping, and because the LT that does this is unique (could prove that). You would go on to say $(z_1,z_2,z_3,z_4) \equiv S(z_2,z_3,z_4)z_1 = Rz_1$. So your conclusions would be:

(a) $R = S(z_2,z_3,z_4)$

(b) $(w,z_2,z_3,z_4) = Rw$ for any point w

Now onto the real problem:

(1) When S appears with no arguments, it is an abbreviation for $S(z_2,z_3,z_4)$. T = some LT.

(2) Question: what does ST^{-1} do to the points Tz_2,Tz_3,Tz_4 ? Answer:

$ST^{-1}[Tz_2,Tz_3,Tz_4] = S[z_2,z_3,z_4] = [1,0,\infty]$ // last = by definition of S !

(3) Applying the results of our warmup problem, we would conclude that (think $ST^{-1} = R$)

(a) $ST^{-1} = S(Tz_2,Tz_3,Tz_4)$

(b) $(w,Tz_2,Tz_3,Tz_4) = (ST^{-1})\,w$ for any point w

(4) Now select $w = Tz_1$. Then from (b) we can write

$(Tz_1,Tz_2,Tz_3,Tz_4) = (ST^{-1})\,Tz_1 = Sz_1 = S(z_2,z_3,z_4)z_1 = (z_1,z_2,z_3,z_4)$

and therefore the cross ratio is invariant under any LT like T.

Question: How would you find a linear transform that takes z_1, z_2, z_3 to w_1, w_2, w_3 ?

Here is a mapping, with z as variable, which maps z_1, z_2, z_3 to $1,0,\infty$:
$y = Az = S(z_1,z_2,z_3)z = (z,z_1,z_2,z_3)$

Here is a mapping, with w as variable, which maps w_1, w_2, w_3 to $1,0,\infty$:
$y = Bw = S(w_1,w_2,w_3)w = (w,w_1,w_2,w_3)$

The mapping that maps z_1, z_2, z_3 to w_1, w_2, w_3 must be this (with x as variable)
$y = Az$ $w = B^{-1}y \Rightarrow w = (B^{-1}A)z \equiv Cz$

This says $Bw = Az$ which from above says $(w,w_1,w_2,w_3) = (z,z_1,z_2,z_3)$,

$(w,w_1,w_2,w_3) = (z,z_1,z_2,z_3)$ // which **agrees with p 79 A**

Of course we also know that

$(z_1,z_2,z_3,z_4) = [(z_1 - z_3)/(z_1 - z_4)] / [(z_2 - z_3)/(z_2 - z_4)]$

so we substitute painfully. First write

$(z,z_2,z_3,z_4) = [(z- z_3)/ (z- z_4)] / [(z_2- z_3)/ (z_2- z_4)]$

then

$(z,z_1,z_2,z_3) = [(z- z_2)/ (z- z_3)] / [(z_1- z_2)/ (z_1- z_3)] = [(z- z_2)/ (z- z_3)] / k(z_1,z_2, z_3)$

$(w,w_1,w_2,w_3) = [(w- w_2)/ (w- w_3)] / [(w_1- w_2)/ (w_1- w_3)] = [(w- w_2)/ (w- w_3)] / k(w_1,w_2, w_3)$

So set equal to get

$k(w_1,w_2, w_3) [(z- z_2)/ (z- z_3)] = k(z_1,z_2, z_3) [(w- w_2)/ (w- w_3)]$

$W_{123} [(z- z_2)/ (z- z_3)] = Z_{123} [(w- w_2)/ (w- w_3)]$ // new names

$W_{123} (w- w_3) [(z- z_2)/ 1] = Z_{123} (z- z_3) [(w- w_2)/ 1]$

$W_{123} (w- w_3)(z- z_2) = Z_{123} (z- z_3)(w- w_2)$

Let $\alpha = Z_{123}/ W_{123}$ so have

$(w- w_3)(z- z_2) = \alpha(z- z_3)(w- w_2)$

$w (z- z_2) - w_3(z- z_2) = \alpha(z- z_3)w- \alpha w_2(z- z_3)$

$\alpha(z- z_3)w - w (z- z_2) = \alpha w_2(z- z_3) - w_3(z- z_2)$

$w \{ \alpha(z- z_3) - (z- z_2)\} = z \{ \alpha w_2 - w_3 \} + \{w_3z_2- \alpha w_2z_3\}$

$w \{ - z + \alpha(z- z_3) + (z_2)\} = z \{ \alpha w_2 - w_3 \} + \{w_3z_2- \alpha w_2z_3\}$

$w \{ - z + \alpha(z- z_3) + z_2\} = z \{ \alpha w_2 - w_3 \} + \{w_3z_2- \alpha w_2z_3\}$

$w = (Az +B) /(Cz + D)$

$A = \alpha w_2 - w_3$ $\alpha = Z_{123}/ W_{123}$

$B = w_3z_2- \alpha w_2z_3$

$C = \alpha-1$ $Z_{123} = [(z_1- z_2)/ (z_1- z_3)]$

$D = z_2 - \alpha z_3$ $W_{123} = [(w_1- w_2)/ (w_1- w_3)]$

It certainly is not a pleasant result. Probably better to do with the 2x2 matrices. [I repeat this same calculation somewhere, separate document maybe, forgot that I did it here.]

Theorem 13. Claim that (z_1,z_2,z_3,z_4) is real $\Leftrightarrow$ the four points lie on a circle or line.

Two proofs are given and they are both messy as far as I am concerned. First proof is a geometry one, the second one is one of those A proofs where each step is a bigger mystery than the previous step. Let's let these proofs ride and try to see where he is taking us.

Exercise added: Prove theorem 13.

(1) **Lines**. It is pretty clear that (z_1,z_2,z_3,z_4) is real if all four points are on the real axis $y_i=0$ or all on the imaginary axis $x_i=0$, just look at the ratio written out.

Some other line would be written as $y = mx+b$ where m and b are real, so $z_i = (x_i, i\,[\, mx_i+b]\,)$. Now we know:

$$(z_1,z_2,z_3,z_4) = (z_1- z_3)\,(z_2- z_4) \;/\; [\,(z_1- z_4)\,(z_2- z_3)\,]$$

When I install my four collinear points, I get $(z_i-z_j) \equiv \Delta z_{ij} = \alpha\, \Delta x_{ij}$ where $\alpha = 1+im$, so

$$(z_1,z_2,z_3,z_4) = (\alpha\Delta x_{13})\,(\alpha\Delta x_{24}) \,/\, [(\alpha\Delta x_{14})\,(\alpha\Delta x_{23})\,] = (\Delta x_{13})\,(\Delta x_{24}) \,/\, [(\Delta x_{14})\,(\Delta x_{23})\,]$$
$$= (x_1,x_2,x_3,x_4) = \text{real}$$

So this shows that the cross-ratio is real for any four points which lie on *any* line in the C plane.

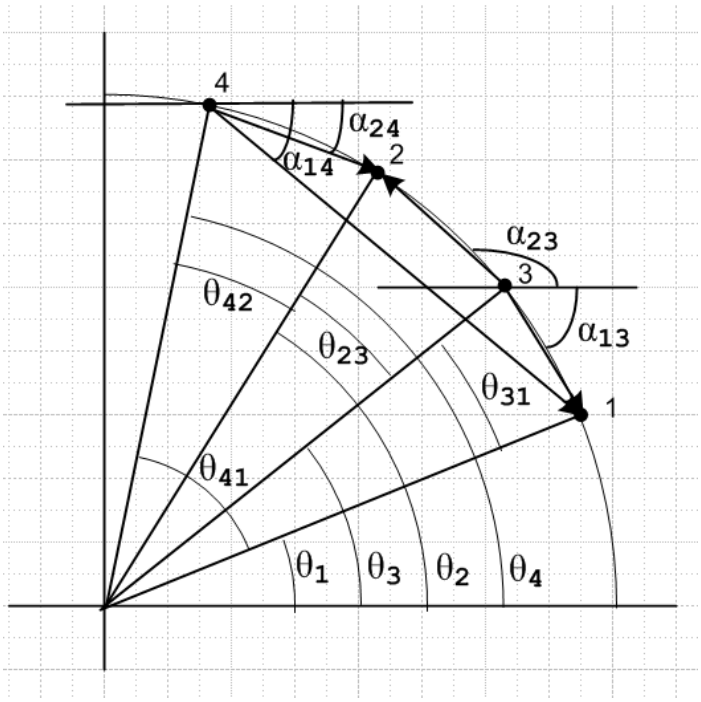
(2) **Circles**. Now let's put all **four points on a circle** of radius R. We might as well take the circle center at 0 since whatever we choose will cancel out in the four differences of z's. It is pretty clear that the R's will factor out and completely cancel, so we might as well just take $z_i = \exp(i\theta_i)$. So we then only have to examine four points on the unit circle centered at the origin.

$$(z_1,z_2,z_3,z_4) = (z_1 - z_3)\,(z_2 - z_4) \;/\; [\,(z_1 - z_4)\,(z_2 - z_3)\,]$$
$$(z_1,z_2,z_3,z_4) = (\exp(i\theta_1)- \exp(i\theta_3))\,(\exp(i\theta_2)- \exp(i\theta_4)) \;/\; [\,(\exp(i\theta_1)- \exp(i\theta_4))\,(\exp(i\theta_2)- \exp(i\theta_3))\,]$$

Now comes a big Visio drawing! We first make some definitions:

$\theta_{31} = \theta_3 - \theta_1$ etc
$\alpha'_{13} = \arg(z_1- z_3)$ etc

I put the four points on a circle in a certain order which makes the picture easier to draw (even then, it is a painful picture) and here it is:



Our goal is to show that the phase of the cross ratio is a multiple of π so that the cross ratio is real when we have all four points on a circle. That means we need to show this:

$\alpha'_{13} + \alpha'_{24} - \alpha'_{14} - \alpha'_{23} = N\pi$ N = some integer

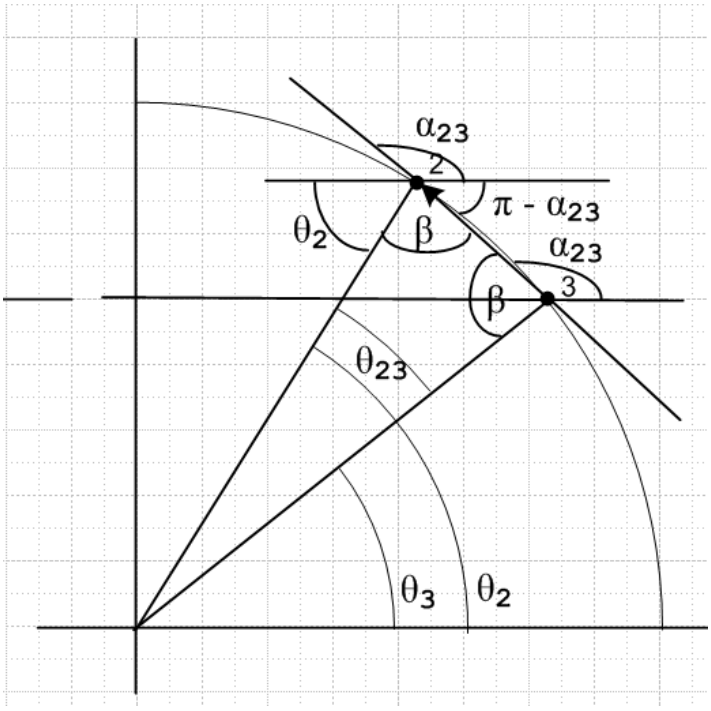
We are about to do some geometry based on the above picture, and it is convenient to have all angles in the picture be positive. So all the unprimed α_{ij} are assumed positive. We then have these relations:

$\alpha'_{13} = -\alpha_{13}$
$\alpha'_{24} = -\alpha_{24}$
$\alpha'_{14} = -\alpha_{14}'$
$\alpha'_{23} = +\alpha_{23}$ // that is to say, three of the α' angles are negative in the picture

Thus, our goal is to show this:

$\alpha_{13} + \alpha_{24} - \alpha_{14} + \alpha_{23} = N\pi$ N = some integer (*)

This is a crucial sign change! This claim (*) is certainly not "obvious" staring at the picture! Let's focus on just one of the triangles in the picture above and add more labels,



The fact that the two points 2 and 3 are on a circle makes this triangle **isosceles** so we can draw in two equal angles β and it is pretty clear that $2\beta + \theta_{23} = \pi$. But the angle we are interested in is α_{23} and we can see at point 2 that $\alpha_{23} = \beta + \theta_2$. These are what we need:

$2\beta + \theta_{23} = \pi$ => $\beta = (\pi - \theta_{23})/2$ so $\alpha_{23} = (\pi - \theta_{23})/2 + \theta_2$

So finally we have one of our angles of interest in terms of angles we actually know. This is the result when the arrow has direction shown, from 3 to 2, since we had $z_2 - z_3$

What would happen if the arrow were reversed? We would then have $\alpha_{32} = \pi - \alpha_{23}$ as the picture shows. Then we would say that

$$\alpha_{32} = \pi - \alpha_{23} = \pi - [(\pi - \theta_{23})/2 + \theta_2]$$

This is in fact the case for the other three triangles in our main figure above. So here we go:

middle thin triangle with arrow reversed *would be* : $\alpha_{32} = \pi - [(\pi - \theta_{23})/2 + \theta_2]$

This is our prototype which we now apply to the other three triangles, making substitutions as needed:

upper thing triangle:	2,3→ 4,2	$\alpha_{24} = \pi - [(\pi - \theta_{42})/2 + \theta_4]$
lower thin triangle:	2,3→3,1	$\alpha_{13} = \pi - [(\pi - \theta_{31})/2 + \theta_3]$
large outer triangle:	2,3 → 4,1	$\alpha_{14} = \pi - [(\pi - \theta_{41})/2 + \theta_4]$

So we can now assemble our accumulated geometric facts:

$\alpha_{23} = (\pi-\theta_{23})/2 + \theta_2$
$\alpha_{24} = \pi - [(\pi-\theta_{42})/2 + \theta_4]$
$\alpha_{13} = \pi - [(\pi-\theta_{31})/2 + \theta_3]$
$\alpha_{14} = \pi - [(\pi-\theta_{41})/2 + \theta_4]$

Now let's compute our "quantity of interest" from equation (*) above:

$\alpha_{13} + \alpha_{24} - \alpha_{14} + \alpha_{23}$

$= \pi - [(\pi-\theta_{31})/2 + \theta_3] + \pi - [(\pi-\theta_{42})/2 + \theta_4] - \pi + [(\pi-\theta_{41})/2 + \theta_4] + (\pi-\theta_{23})/2 - \theta_2$

First, examine the various π factors

$\pi - \pi/2 + \pi - \pi/2 - \pi + \pi/2 + \pi/2 = \pi$

That leaves us with

$\alpha_{13} + \alpha_{24} - \alpha_{14} + \alpha_{23} = \theta_{31}/2 - \theta_3 + \theta_{42}/2 - \theta_4 - \theta_{41}/2 + \theta_4 - \theta_{23}/2 + \theta_2$

$= \theta_{31}/2 - \theta_3 + \theta_{42}/2 - \theta_{41}/2 - \theta_{23}/2 + \theta_2$

$= 1/2 [\theta_{31} - 2\theta_3 + \theta_{42} - \theta_{41} - \theta_{23} + 2\theta_2]$

$= 1/2 [(\theta_{31}) - 2\theta_3 + (\theta_{42}) - (\theta_{41}) - (\theta_{23}) + 2\theta_2]$

$= 1/2 [(\theta_3 - \theta_1) - 2\theta_3 + (\theta_4 - \theta_2) - (\theta_4 - \theta_1) - (\theta_2 - \theta_3) + 2\theta_2]$

$= 1/2 [(- \theta_1) - \theta_3 + (- \theta_2) - (- \theta_1) + (+ \theta_3) + \theta_2]$

$= 1/2 [- \theta_3 + (- \theta_2) + (+ \theta_3) + \theta_2]$

$= 0$

Thus we have shown that

arg (cross) $= \alpha'_{13} + \alpha'_{24} - \alpha'_{14} - \alpha'_{23} = \alpha_{13} + \alpha_{24} - \alpha_{14} + \alpha_{23} = \pi$

which means that the cross ratio is in fact real as claimed.

Now what happens if the points are in a different ordering instead of 1,3,2,4 as we drew it? And what happens if they are not all in the first quadrant? This shows a weakness of the "geometric proof" because it takes some work to convince oneself that these other cases still work right.

First, as we keep our order and let the points start moving around CCW into other quadrants, I would argue that nothing dramatic happens. We do it smoothly and continuously, moving one point at a time. As

a point z_i moves across the top (or bottom) of the circle, its α' angle changes sign, but our equations handle this just fine. If we let all our θ angles go 0 to 2π, those labels stay the same as points swing around. The trapezoid remains uncrossed (convex) all the time. So it seems pretty reasonable (though I really have not proved it) that motion of the points (without reordering) won't affect the conclusion.

Changing point ordering is a little harder. Let's consider a procedure for swapping points 2 and 3 just as an example. In the picture above, let 2 move toward 3 until it is ε away. Then at this point, swing 2 on a little half arc (radius ε) inside around to the other side of 3. In this process, we will have $\alpha_{23} \rightarrow \alpha_{23} + \pi$. In the ε range, none of the other angles like α_{13} will change. So the net result is that one of our four α' angles changes by π, but that then does not affect the realness of the cross ratio. So this procedure handles the swap of any adjacent pair of points. But any reordering can be done by a sequence of pairwise swappings, so QED.

This concludes our proof of Theorem 13. We have shown that if we have 4 points on any line or on any circle, the cross ratio is real.

Theorem 14: Linear transformation maps Circles into Circles.

Here Circle means circle or line = circle of ∞ radius.

Comment: Any Circle in C maps into an actual circle on the Riemann Sphere. Circles in C which are lines of course must be Sphere circles which pass right through the sphere north pole (since it is ∞ and line ends are ∞)

Proof: Start with a Circle as your locus in the domain C plane. We know (z_1,z_2,z_3,z_4) is real if we have any four points on this Circle (Thm 13). But this cross ratio is invariant under any linear transformation (Thm 12). After the transformation we will then have $(z_1',z_2',z_3',z_4')= (z_1,z_2,z_3,z_4)$ where $z_i' = Tz_i$ are the four transformed points. Since (z_1',z_2',z_3',z_4') is then real as well, Thm 13 to the right says that the four points lie on a Circle.

Summary to this point: Any analytic f(z) transformation preserves angles and scale. Linear transformations in addition preserve Circles! The reason is that the cross ratio for a line or circle is real, and the cross ratio is invariant under transformations, hence Circles → Circles.

3.3 Symmetry (80)

I will try this on my own since, as usual, each A sentence is unfriendly to me. Start with points z and $\bar{z}$ being "symmetric" relative to the real line in the usual manner. Now map these points through a linear transform T. The real axis becomes circle C, and the two symmetric points become w = Tz and w* = $T\bar{z}$ where w* is our made-up name for the point symmetric to w. What can we say regarding our little threesome w w* and C? Well, we certainly know that

$z = T^{-1}w$ and $\bar{z} = T^{-1}w^*$ => $T^{-1}w^* = \overline{(T^{-1}w)}$

Aside: We know that , when we generally write some w = Lz and L is a linear transform, we *can find* three points z1,2,3 which L will map into 0,1,∞ . In general z1,2,3 will be THE three points on a domain circle which map into these three particular points on the real line. After all, we know there must be SOME circle that maps into the real line under L, so z1,2,3 will be on THIS circle, whatever it is.

So, now let L = T^{-1}. Then we can write $T^{-1}w^*$ = (w*,z1,z2,z3) using our basic definition p 78 of the cross ratio. Here z1,2,3 are THE three points which T^{-1} maps into 0,1,∞. Obviously it is also true that $T^{-1}w$ = (w,z1,z2,z3) because this equation is true for any "argument" like w* or w. Therefore we have shown that

$$T^{-1}w^* = \overline{(T^{-1}w)} \quad \Rightarrow \quad (w^*,z1,z2,z3) = \overline{(w,z1,z2,z3)}$$

and this is the claim of "Definition" 13. I have defined symmetric points in my own way, and then this fact is something that comes out being true. Now the idea is to imagine that the z1,2,3 are three *arbitrary* points! They define a circle C. We know there exists some transform F that maps these points into 0,1,∞. and of course maps circle C onto the real axis. This transform maps w to some z, and w* to $\overline{z}$. Thus w and w* are a symmetric point pair. (F = L = T^{-1} of the above text). Thus, given any three points, they define a circle C and for any w, the above equation tells us where w* is located, given w.

We say that w and w* are **"reflected in C"**.

On 81 top Ahlfors shows, not to our surprise, that if C is a line, then points w and w* have the expected locations just reflected across this line. They are both on the same "bisecting normal" to the line. Just as they are if the line is the real axis.

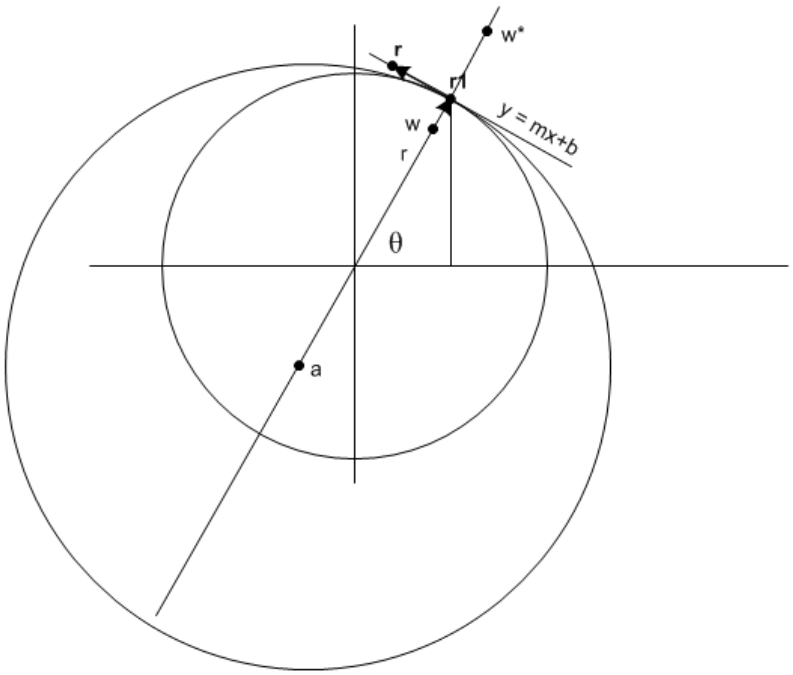
I would prefer to make my own little argument for the line case, but won't bother. Given a line and w, the only point picked out in space that w* could possibly inhabit is the obvious reflection point.

At this point, in mid p 81, I went off and did my own little "exercise" to find how to compute w* if you are given a circle and w. I used the above fact about the cross-ratio in my derivation. Here were my conclusions, which agree with his conclusions on page 81.

(1) the three points a,w and w* are collinear, where a is the center of the circle on which the three points z_1 z_2 z_3 all lie.

(2) the rule I derived |w*- a| |w-a | = R^2 tells you how to find the location of w* if you know where w is, and vice versa. If one distance is > R, the other is < R, which says w and w* are on opposite sides of the circular boundary. If the circle center is a = 0, the result is just |w*| |w| = R^2 which tells you everything you need to know. If w and w* are along the positive real axis, then w*w = R^2 and w* = R^2/w.

(3) Here is how you take the R→ ∞ limit to get the straight line case. Start with a picture like this where the smaller circle has radius r, center at the origin, and is tangent to the line y = mx+b at point r_1. This part of the picture was used by me in a geometry document to compute that the line must be
y = [-cotθ] x + r/sinθ, a fact that is interesting but that won't be needed here. We now make the circle larger by moving the center down along the negative θ ray to location a. As we move the center in this way, the circle gets larger but still passes through point r_1. As a gets very large, the circle approaches the tangent line near point r_1.



For a fixed w inside the circle, we want to see what happens to the symmetry point w* in this limit. So let's parameterize things as follows: (r_1 = a complex number of magnitude r, but α, β and γ are all real and 0)

$w = \alpha r_1 \qquad w^* = \beta r_1 \quad a = -\gamma r_1 \qquad$ where $1>\alpha > 0$ = fixed, $\beta>1$ = unknown, $\gamma > 0 \rightarrow \infty$

We can do this because we know that a and w and w* are collinear. Now consider:

$$|w^* - a|\,|w - a| = R^2 = (r+|a|)^2$$

which is the equation from which we will find w* . Installing things we get

$$|\beta r_1 + \gamma r_1|\,|\alpha r_1 + \gamma r_1| = (r + \gamma |r_1|)^2$$

$$|\beta + \gamma|\,|\alpha + \gamma|\,r^2 = (r + \gamma r)^2$$

$$(\beta + \gamma)(\alpha + \gamma) = (1 + \gamma)^2$$

$$\gamma^2 + \gamma(\alpha+\beta) + \alpha\beta = \gamma^2 + 2\gamma + 1$$

In the limit $\gamma \rightarrow \infty$ we conclude that

$$\alpha + \beta = 2$$

$$\beta = 2 - \alpha \qquad \Rightarrow \qquad (\beta-1) = (1-\alpha)$$

This then is our solution for quantity β. We can now compare these two quantities:

$|r_1 - w|$ and $|w^* - r_1|$

We find that

$|r_1 - w| = |r_1 - \alpha r_1| = (1-\alpha)\, r$
$|w^* - r_1| = |\beta r_1 - r_1| = (\beta-1)r = (1-\alpha)r$

Thus, these two quantities are equal, so we find that w* and w lie along the bisector of y = mx+b which passes through point r_1 and are equal distances from the line. This is what we expect.

(4) I have verified the "**construction**" shown on Ahlfors page 81. I did three pencil Pythagoras theorems for three right triangles, then combined them to get the desired result. Here is how you mechanical build this construction (points w and w* are here called z and z*)

- draw the line that passes through z and a (the circle center)
- draw a perp line to the above line that passes through z
- draw circle tangent lines at the two points where that perp line intersects the circle.
- where these tangent lines intersect is the point z*

Summary: if you take the real axis and z and $z^* = \overline{z}$ as mirror points, and if you run this through a linear transformation, the real axis maps into a Circle, and the points z and $z^* = \overline{z}$ map into a **symmetry point pair** which I call w and w*. The relation between w, w* and R of the circle is as shown above, namely that $|w^* - a|\,|w-a| = R^2$ where a is the circle center. Think of this as D* D = R^2.

We know that a linear transform carries circle C_1 into circle C_2. We could of course write the transform as a product of two transforms one that takes C_1 into the real line, and then one that takes that real line into C_2. If we start with a symmetry point pair w_1, w_1^* relative to C_1 , the first mapping would take this into a symmetry point pair z, $\overline{z}$ relative to the real axis, and then the second mapping would take these symmetry point pair w_2, w_2^* relative to C_2. Theorem 15 then seems pretty obvious.

Question: [p 82 #1] What is the linear transformation that takes the circle z1,2,3 into the circle w1,2,3 ? That is, we need $Tz_i = w_i$ for i = 1,2,3 and generally Tz = w for an arbitrary z. But we know that

$(z,z_1,z_2,z_3) = (Tz,Tz_1,Tz_2,Tz_3)$ // invariance under linear transform
$= (w,w_1,w_2,w_3)$

So you could solve this thing $(z,z_1,z_2,z_3) = (w,w_1,w_2,w_3)$ for $w = f(z, z_i, w_i) = Tz$. [I already did this, see earlier, and summarized in meta notes.]

Question: [p 82 #2] suppose $w_1 = Tz_1$ where z_1 on C (hence w_1 on C') and $w_2 = Tz_2$ where these points are *not* on C and C', We know then that $w_2^* = Tz_2^*$ by definition of the symmetry points. Then

$(z,z_1,z_2,z_2^*) = (Tz,Tz_1,Tz_2,Tz_2^*) = (w,w_1,w_2,w_2^*)$

and we could solve this for w = Tz . This is "another way" to find the T that maps a circle into a circle. All you need is one point on the circles, and one point off!

Definition of "reflection" [p 80 location A] . We know what we mean by reflection in the real axis: take any point z and it reflects into $\bar{z}$. So if you have a little figure above the real axis, that figure gets reflected. You just take each point in the figure and do CC to it.

Reflection in a circle is the same idea. Draw a figure on one side of the circle, and for each z, compute z* on the other side of the circle, so you end up with a reflected figure which of course is distorted relative to the original figure.

You can find a mapping from z to z* which is a linear transformation. For example, consider this mapping:

$z^* = a + R^2/(z-a) \quad \Rightarrow \quad z^*-a = R^2/(z-a) \quad \Rightarrow |z^*-a| = R^2/|z-a|$

The mapping clearing satisfies $|z^*-a||z-a| = R^2$ so that z* is indeed the sym point of z. Now rewrite:

$z^* = [a(z-a) + R^2] /(z-a) = [az + (R^2-a^2)] / (z-a)$ $\qquad M = \begin{pmatrix} a & R^2-a^2 \\ 1 & -a \end{pmatrix}$

which clearly has the form of a linear transformation. Easy to show that $M^2 = R^2 I$, so if you apply this linear transformation twice, you get the identity (ignore constant), just as you would expect.

This is the end of section 3.3 on Symmetry. I will now do a few "extras" just for practice.

Page 82 Exercise 5. What is the most general linear transformation that maps the unit circle into itself?

$w = (az+b)/(cz+d) = (a/c)(z+b/a)/(z+d/c) = (1/\gamma)\,(z-\beta) / (z-\delta)$ $\qquad z = e^{i\varphi}$

Require that

$|w| = 1$, so that $\quad | (z+b/a) |/ | (z+d/c) | = | (c/a) |$

$| (z-\beta) | / | (z-\delta) | = | \gamma |$ $\qquad \beta = -b/a \qquad \delta = -d/c \qquad \gamma = c/a$

But I know this is satisfied if z is on a certain circle. But we need that circle to be the unit circle!

Earlier today I went off and found the characterization of the "certain circle" just mentioned,

$c_x = - (-a_x+\alpha^2 b_x)/ (1-\alpha^2)$ $\qquad | z - a | / |z-b| = \alpha$

$c_y = - (-a_y+\alpha^2 b_y)/ (1-\alpha^2)$

$R^2 = c_x^2 + c_y^2 - (a_x^2 - \alpha^2 b_x^2 + a_y^2 - \alpha^2 b_y^2)/ (1-\alpha^2)$

So I can translate my answer to this problem

$c_x = -(-\beta_x+\alpha^2\delta_x)/(1-\alpha^2)$ $\quad |z-\beta|/|z-\delta| = \alpha$

$c_y = -(-\beta_y+\alpha^2\delta_y)/(1-\alpha^2)$

$R^2 = c_x^2 + c_y^2 - (\beta_x^2 - \alpha^2\delta_x^2 + \beta_y^2 - \alpha^2\delta_y^2)/(1-\alpha^2)$ $\quad \alpha = |\gamma|$

If we want this circle to be the unit circle, we need $c_x = c_y = 0$ for starters, so we get

$\beta_x = \alpha^2\delta_x$

$\beta_y = \alpha^2\delta_y$

or simply

$\beta = \alpha^2\delta$

Then we also want R = 1, so

$-1 = (\beta_x^2 - \alpha^2\delta_x^2 + \beta_y^2 - \alpha^2\delta_y^2)/(1-\alpha^2) = [(\alpha^2\delta_x)^2 - \alpha^2\delta_x^2 + (\alpha^2\delta_y)^2 - \alpha^2\delta_y^2]/(1-\alpha^2)$

$= \alpha^2[\delta_x^2(\alpha^2-1) + \delta_y^2(\alpha^2-1)]/(1-\alpha^2) = \alpha^2[\delta_x^2 + \delta_y^2](-1)$

This seems to say that we need

$\delta_x^2 + \delta_y^2 = 1/\alpha^2$

or more simply

$|\delta| = 1/\alpha$

So I now have these conditions to make the unit circle map into itself:

$\beta = \alpha^2\delta$

$|\delta| = 1/\alpha \qquad \alpha = |\gamma|$

$w = (1/\gamma)(z-\beta)/(z-\delta)$

$= (1/\gamma)(z-\alpha^2\delta)/(z-\delta)$ $\quad$ where $|\delta| = 1/\alpha$

We can write $\delta = (1/\alpha)e^{i\theta}$ or $\alpha\delta = e^{i\theta}$. Then write the above as

$w = (\alpha/\gamma)(z-\alpha^2\delta)/(\alpha z-\alpha\delta)$

$= (\alpha^2/\gamma)(z/\alpha-\alpha\delta)/(\alpha z-\alpha\delta)$

$= (\alpha^2/\gamma)(z/\alpha-e^{i\theta})/(\alpha z-e^{i\theta})$ $\quad \alpha = |\gamma|$

$= (|\gamma|^2/\gamma)\ (z/|\gamma| - e^{i\theta}) / (|\gamma|z - e^{i\theta})$

Now write $\gamma = |\gamma|\ e^{-i\sigma}$ then we have

$= e^{-i\sigma}(|\gamma|^2/|\gamma|)\ (z/|\gamma| - e^{i\theta}) / (|\gamma|z - e^{i\theta})$

$= e^{-i\sigma}\ |\gamma|\ (z/|\gamma| - e^{i\theta}) / (|\gamma|\ z - e^{i\theta})$

$w = e^{-i\sigma}\ \alpha\ (z/\alpha - e^{i\theta}) / (\alpha\ z - e^{i\theta})$

So the $e^{-i\sigma}$ leading factor is pretty obvious, does not affect $|w| = 1$. We then have three free parameters here, α = any positive real number, θ,σ = any angle in $(-\pi,\pi)$. The basic core of the answer is this:

$w = (z - \alpha\ e^{i\theta}) / (\alpha\ z - e^{i\theta})$

This sure is unobvious to me. Of course we know that $z = e^{i\varphi}$ since z must be on the unit circle. Then

$w = (e^{i\varphi} - \alpha\ e^{i\theta}) / (\alpha\ e^{i\varphi} - e^{i\theta}) = (1 - \alpha\ e^{i\kappa}) / (\alpha - e^{i\kappa})$ $\qquad \kappa = \theta - \varphi$

Well, I will now show that in fact $|w| = 1$. We have

$|w|^2 = |1 - \alpha\ e^{i\kappa}|^2 / |\alpha - e^{i\kappa}|^2$

$num = |1 - \alpha\ e^{i\kappa}|^2 = (1 - \alpha\ e^{i\kappa})(1 - \alpha\ e^{-i\kappa}) = 1 + \alpha^2 - \alpha\ (e^{i\kappa} + e^{-i\kappa})$

$den = |\alpha - e^{i\kappa}|^2 = (\alpha - e^{i\kappa})\ (\alpha - e^{-i\kappa}) = \alpha^2 + 1 - \alpha\ (e^{i\kappa} + e^{-i\kappa})$

Thus num = den and $|w| = 1$. So once again, **here is my solution to this problem**:

$w(z) = e^{-i\sigma}\ \alpha\ (z/\alpha - e^{i\theta}) / (\alpha\ z - e^{i\theta})$ where $z = e^{i\varphi}$ on unit circle

σ = arbitrary angle
θ = arbitrary angle
α = arbitrary positive real number

Here is another way to approach this problem. Let $z_1 = e^{iZ1}$ etc. So the z_i are three points on the unit circle. We want to map these into three different points like $w_1 = e^{iW1}$. Then we have

$(z,z_1,z_2,z_3) = (Tz,Tz_1,Tz_2,Tz_3)$ // invariance under linear transform
$= (w,w_1,w_2,w_3)$

which becomes

$(z, e^{iZ1}, e^{iZ2}, e^{iZ3}) = (w, e^{iW1}, e^{iW2}, e^{iW3})$

and we then solve this for w = w(z). So far we have 6 free angle parameters.

$$(z,z_1,z_2,z_3) = (z- z_2)\,(z_1- z_3) \;/\; [\,(z- z_3)\,(z_1- z_2)\,]$$

$$(z, e^{iz1}, e^{iz2}, e^{iz3}) = (z- e^{iz2})\,(e^{iz1}- e^{iz3}) \;/\; [\,(z- e^{iz3})\,(e^{iz1}- e^{iz2})\,]$$

But this is a big mess and I like my previous solution method. The cross-ratio often seems to make something simple much more complicated!

Exercise PL1. Question: do I have a solution to this $(z,z_1,z_2,z_3) = (w,w_1,w_2,w_3)$? Yes! I did this once above (just above Theorem 13) and I did it by the matrix method in the meta notes, so don't do it again here!

Comments: The complex plane is also known as an **Argand Diagram**. You can do an **ellipse** in this way:

$$|z-a| + |z-b| = \text{constant}$$

I think this would be a **hyperbola** half

$$|z-a| - |z-b| = \text{constant}$$

A **circle** of course is

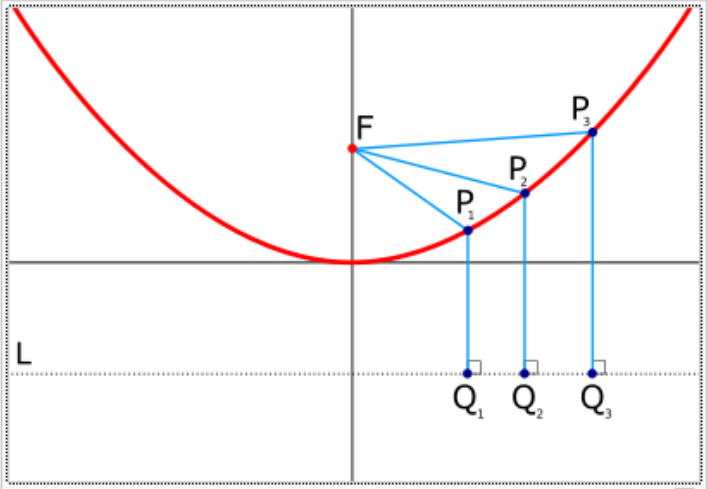
$$|z-a| = \text{constant}$$

I don't know how to do a **parabola** this way. Maybe you can't do it. I know you can say that

$$|z-f| = |z-D(z)| \quad = \text{where D(z) is a point on the directrix.}$$

For a simple vertical parabola through the origin, D(z) = (x, -L) so you get

$$|z-f| = |(x,y)- (x, -L)| = |\,(0,y+L)\,| = |y+L|$$



Parabolic curve showing directrix (L) and focus (F). The distance from a given point P_n to the focus is always the same as the distance from P_n to a point Q_n directly below, on the directrix.

Anyway, today's confusion came from the fact that

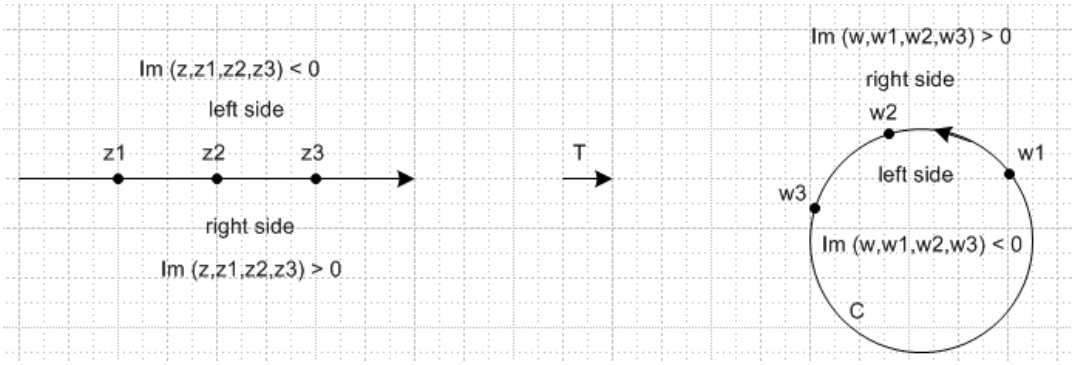
$$|z-a| / |z-b| = \text{constant}$$

ALSO describes a certain circle. The equation $|z-a|\ |z-b| = R^2$ means $|z-a|^2\ |z-b|^2 = R^4$ and this is some kind of quartic equation, not a conic section.

So you can look at the sum, difference, product and quotient of two distances of the form |z-a| being a constant. It turns out that the sum, difference cases give ellipse and hyperbola branch. The quotient gives a certain circle, and the product gives a quartic.

3.4 Oriented Circles (83)

Again Ahlfors has fumbled a chance to be clear and simple. My take would be this. Imagine a mapping that takes the real axis into some circle. Consider three points on the real axis $z1 < z2 < z3$. We can certainly talk about the "increasing direction" of the real axis and draw an arrow on the real axis.



In this picture we show a particular way the mapping T might work. If z lies on the real axis on the left, then certainly (z,z1,z2 z3) = real from Theorem 13. Similarly, the mapped cross ratio (w,w1,w2,w3) is real when the point w = Tz lies on the circle. So maybe we can use Im(cross) to mark the notion of "side". I have shown in the picture how this works out, and now here are the details:

Here is our infamous cross ratio in the format we want here:

$$(z,z_1,z_2,z_3) = [(z- z_2)/ (z- z_3)\ ] / [(z_1- z_2)/ (z_1- z_3)\ ]$$

What happens if z lies off the real axis? Let z = z +ia where z is real (the z_i are all real as just noted). The denominator [ ] is real and is positive, since each factor there is negative, given our chosen orientation of the three points on the real line. So we can ignore this [ ] in our analysis. Then,

$$(z+ia - z_2) / (z+ia - z_3) = [(z+ia - z_2)\ (z-ia - z_3)] / [(z+ia - z_3)\ (z-ia - z_3)]$$

num = $|(z+ia)|^2$ + real stuff + ia{ $(z-z_3) - (z-z_2)$ } = real + ia (z_2-z_3)
den = $|(z+ia - z_3)\ |^2$

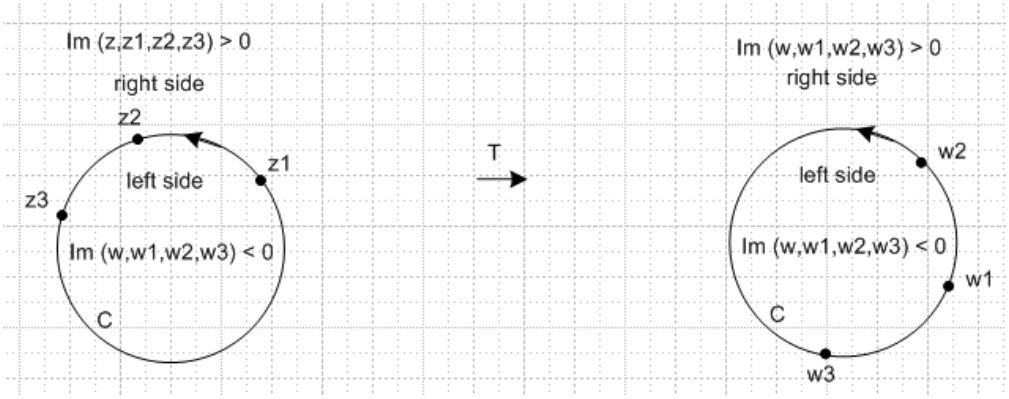
Thus, sign $Im(z,z_1,z_2,z_3)$ = sign[a (z_2-z_3)] . In our picture, we have $(z_2-z_3) < 0$ so if a > 0, our sign is negative. This then determines where I put the labels in the left picture above. The "right side" has positive imaginary part, that is what I have just shown. If we take some point z_4 on the "right side" and map it into some w_4 on the right, it will be outside the circle, which is the "right side" in the range of T.

Ahlfor's comment at point A on page 83 is confusing. He computes a particular value of the cross ratio with z1,2,3 = 1,0,∞ which would I guess be the exact reverse of my ordering above if we put ∞ on the left. With that reverse arrow on the real line, "right" would be above, and he gets "i" which is above, so OK, I guess that is his point.

It seems pretty clear to me that some other T, call it T', might map my real axis above into a circle with the reverse orientation.

At point B he computes the cross in terms of the raw linear transformation parameters a,b,c,d assumed real. This shows that, for a given particular transformation, since ad-bc has one or the other sign, the sign Im(cross) = sign [(ad-bc) Im(z)] Obviously there are only two possibilities.

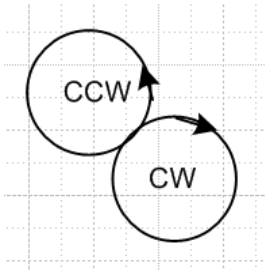
More generally, we can think about circle mapped into circle, and the same discussion will be true. Here is one possibility for circle into circle

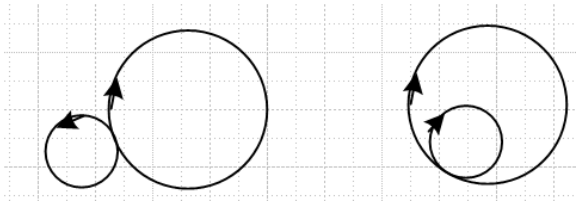


He never uses the terms CW and CCW, but that would seem helpful perhaps.

One comment: we know that in general a T can map z1,2,3 into any w1,2,3 we want. Suppose w1 and w2 were swapped in the above mapping. That swap would just mean the circle on the right is oriented the other way. We get the 1,2,3 order going CW in that case. Same is true for any pair swap, and thus same is true for any permutation of the points! The circle is always the same, but the orientation might get reversed.

At C Ahlfors talks about two circles being tangent. I don't see the point he is making, but I see no problem in how I would analyze this situation. I see no need for mapping into parallel lines. The two tangent domain circles might be "aligned" as shown here,





The resultant two circles in the range would be similarly "aligned" at the touch point, but I don't know what the picture exactly would look like. Here are two possibilities in the range (right above). There are two other cases I could draw with all arrows reversed, and of course either circle or both could be lines.

And again at D on page 83 Ahlfors comes up with yet another vague remark. It adds nothing.

And finally we come to p 84 A. If you look at my previous picture in the range, say, the "point at ∞" is clearly on the "right side" of the circle. So he just wants to say that such a circle has a "positive" orientation, which I would simply call a CCW orientation. He is avoiding the CCW term for some reason. For a line, it is less clear where ∞ lies, but to be consistent, we put it on the right side. So:

positive orientation ∞ on right side ∞ on side where Im(z,z1,z2,z3) > 0

Comment: Once again I have had to rewrite an Ahlfors section to extract the main points.

3.4 Families of Circles (84)

Suppose we write the most general LT as

$$w = (Az+B)/(Cz+D)$$

where we seem to have 8 free real parameters. We can rewrite this with simple algebra as

$$w = (A/C)\,[\, z + (B/A)\,] / [\, z + (D/C)\,] \quad = k\,[\, z - a] / [z - b]$$

Here we see explicitly the 6 true degrees of freedom. So the form shown page 84 B is still a *most general* form of the LT, A does not bother to point that out. Now here are the rules:

(1) In the z plane, if a circle goes through points a and b, then in the w plane it goes through 0 and ∞. But the only Circles hitting 0 and ∞ are straight lines. Therefore:

The above LT maps (circles passing thru both a and b) into (straight lines through the origin) in the range.

These circles passing through a and b in the domain are labeled C_1 in the page 85 picture, "thru". They have no person's name associated with them.

(2) a circle about the origin in the range would be | w | = R . This becomes

$$| z - a| / |z - b| = R/|k|$$

which I now well know is the locus of a circle centered somewhere on the line joining a and b and containing either a or b. [See "a geometry problem.doc"] As you increase R from 0, you transition from circles going around a to circles going around b (I think it is this order). When R/|k| = 1, the circle is the bisector line. These circles in the domain are called "the **circles of Apollonius**".

Apollonius of Perga [**Pergaeus**] (Ancient Greek: Ἀπολλώνιος) (ca. 262 BC–ca. 190 BC) was a Greek geometer and astronomer noted for his writings on conic sections.

The above LT maps (circles of Apo) into (concentric circles around the origin) in the range. The dual set of circles in the domain (those THRU a,b, and the APO ones) form the **circular net**, also known as the **Steiner Circles**. Neither of these bolded terms brings up the right images in the image web search, so maybe not used much now. But some Google books show them, but no picture better than p 85.

Jakob Steiner (1796-1863) search on this name along with Steiner Circles not many hits

You can use the Steiner Circles as an orthogonal coordinate system in the domain if you want. In the range we have an obvious polar coordinate system with circles of radius r, and lines at angle φ. Each point in the domain lies on one of each kind of circle, so could be written (x,y) where somehow x and y label the circles. Right angles are preserved, which is why I call it orthogonal.

What does this look like if a = 0 and b = ∞ ? The C_1 thru circles passing through these two points must be "lines through the origin). The C_2 become concentric circles, so this limit gives you the range system in fact. So he lists properties of the Steiner circles, and then says easy to prove if you go to the limiting case just mentioned, because all these properties are preserved under T.

I see now that this Steiner Circle stuff is nowadays just called **bipolar coordinates**

http://en.wikipedia.org/wiki/Bipolar_coordinates

and the claim is made there that this coordinate system allows a simple separable solution to the problem of the magnetic field of two wires, something maybe I will see soon.

Comment: Look now at the page 85 figure. Consider an arbitrary C_1 type circle, passing through a and b, and thus mapping into 0,∞ hence a line through the origin ("ray"). The inside of this circle back-maps into a half plane attached to the this ray, and the outside back-maps into the other half plane.

Now consider two such C_1 circles. Each one's interior maps into a half plane based on a different ray. The intersection of these two half planes is an angular sector between the rays. This angular sector must be mapping into the intersection of the two circles. So here is a useful map.

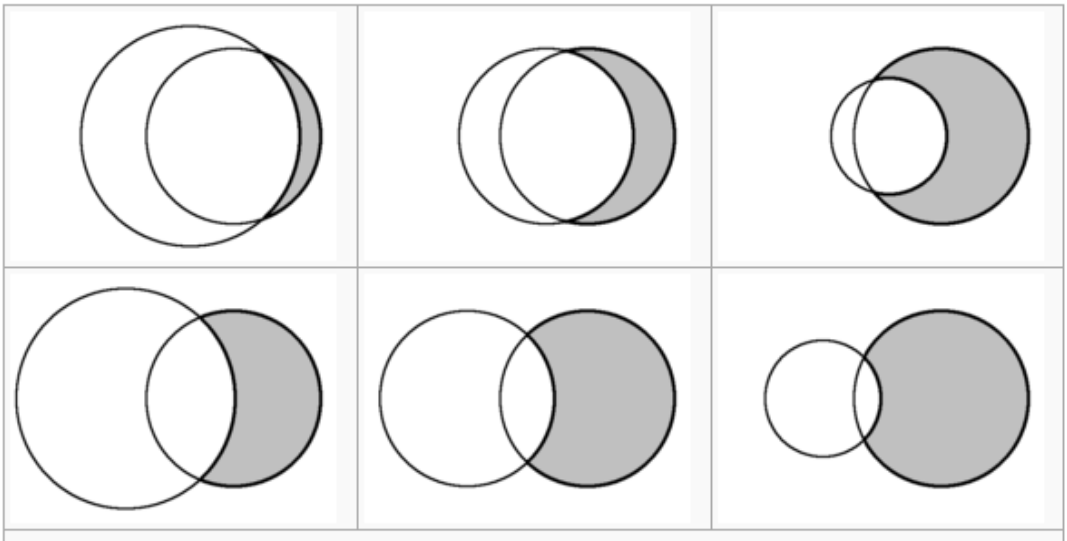
Question/Exercise: what can be said about the "centered" C_1 circle? Which ray does it map into? We really need a "third point" to distinguish rays from each other. For example, $z = \infty$ maps to $w = k$. We can assume k = real since if not, it just rotates the picture. Let b and a also be real with $b > a$. Then the really big "thru" circles with some point near ∞ will pass through k and so are the real axis. So imagine we start with the "really big" upper THRU circle in the page 85 picture. It maps to the real axis in z. As we gradually come down to the center circle, that circle must (by symmetry) map to the imaginary axis. But I don't know which our ray was rotating in the z plane. In any event, as we pass through this centered circle, our ray keeps rotating and we get to the real axis again when we get to the lower huge circles.

Comment resumed: So consider the intersection of two C_1 circles taken from the upper half of the picture. Their intersection maps into an angular sector which is small if the circles are close. As the circles stay on the same half of the Steiner picture but move apart, this sector opens up to a quarter disk (based on exercise above). One intersection of the circles is a "lune" and the other is a "lens" and these intersections map into the complementary angular sectors in the z plane. Here is a nice wiki picture:

Plane geometry [edit]

In plane geometry, a lune is a concave area bounded by two arcs. The corresponding convex shape is a lens.

Formally, a lune is the relative complement of one circle in another (where they intersect but neither is a subset of the other).[1] Alternatively, if A and B are circles, then $L = A - A \cap B$ is a lune.



In plane geometry, the crescent shape formed from two intersecting circles is called a *lune* (in gray).

At page 85 B we start a new topic. A new T form is shown which clearly maps z = a into w = a' and also maps z = b into w = b'. Thus, a circle passing through a,b in the domain would pass through a',b' in the range. These would be the THRU circles or the C_1 circles, so yes, the C_1 circles $\rightarrow C_1'$ circles. It is maybe a little less obvious, but if you take abs value of both sides of p 85 B and set to a constant, then

you get an Apo circle in the domain and an Apo circle in the range. So the claim is $C_2 \rightarrow C_2'$. It certainly seems reasonable that a Steiner circle set in the domain tied to (a,b) becomes a Steiner circle set in the range tied to (a',b'), but the way the various circles is ordered would depend on constant k. If k = 1, the mapping is the identity map so everything lines up as is. You really need some numbers to label the circles in each set, he did not talk much about that. Those could be the r and φ of the concentric thing we had earlier.

Now you could consider the case (a,b) = (a',b') but with $k \neq 1$. Somehow this is a remapping of a Steiner set into itself with some scaling that is not clear to me. In this case a and b would be called **fixed points** of T. He makes these claims:

(1) arg(k) is the angle between a C_1 and its corresponding C_1' (I guess at the points where they intersect, which could be a and b).

(2) And |k| relates to the ratio of course of the factors as 85 B makes obvious and this is how a C_2 is related to a C_2'.

He then makes some more claims:

(3) If k is real and k>0, then $C_1' = C_1$. This agrees with (1) above since then arg(k) = 0. If k is real and k < 0, you just reverse the orientation. This situation k = real is called **hyperbolic**. He talks about the "flow" of points in his usual vague way. If we were to label some points on a C_1 in the domain and also in the range, in the range the points all "flow toward b" as k increases in real value. I have no idea why this fact would ever be useful. We are talking about scaling along the C_1 circles, so to speak. They are reparameterized relative to some scalar parameter, but he does not get into the details.

(4) Of |k| = 1 and you vary k on the unit circle, he claims that you get "flow" along the C_2 circles as you change this phase. This situation |k| = 1 is the **elliptic** case.

You can of course write our general LT as elliptic * hyperbolic.

At page 86 A he observes that a given LT has a *unique* pair of fixed points (up to swapping) and you find them by solving the obvious equation shown. Namely, z = T(z). Notice now the meaning of "fixed points" is more clear. This is quadratic in z which gives the two solutions a and b. If you happen to get a = b, you have the **parabolic** situation I guess regardless of the value of k. He gives the condition for this to happen, obviously $B^2 - 4AC = 0$ in the quadratic.

[87 A] Here we have a horrible typo and the correct equation should be

$$w = z/(z-a) + c$$

It took me a long time to figure this out! You could if you wanted write this as

$$w = (1+c) [z - (ca/(1+c))] / [z-a]$$

which shows that, in terms of our previous discussion, this thing has points (a',b') = (ca/(1+c), a) . The reader is sort of expecting that maybe this new transform will have a convergence of the two points a' and b' and that will be the "degenerate" Steiner, but that is NOT the case! The points stay different.

How do I know he has a typo? If his equation is correct as stated, then

$$w = w/(z-a) + c \qquad \Rightarrow \qquad w = c\,(z-a)/(z - (a+1))$$

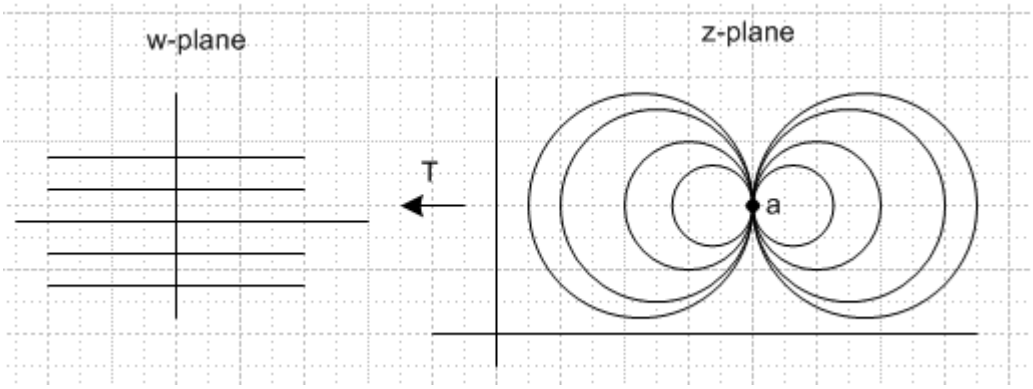
In this case, $w = \infty$ corresponds to $z = a+1$, so any line in the w plane would pass through $z = a+1$, which disagrees with what he claims.

Let's just look at his next few observations:

(1) Straight lines in w-plane: must pass through $w = \infty$ which corresponds to $z = a$. Thus, any straight line in the w plane (not necessarily through the origin) maps back into a circle through a. I agree.

(2) parallel lines in the w-plane: These meet only at $w = \infty$, and thus must map into a pair of circles which are tangent at $z = a$, I agree!

(3) Take as our parallel lines $w = u + iv$ with u varying and v fixed. This is a set of horizontal lines. What do those tangent circles look like? They must ALL be tangent at $z = a$. So here is the picture he should have drawn,



We draw the circles in a horizontal orientation, but the thing could really be at any angle. No doubt the lines above the real axis on the left correspond to one set of circles, and those below to the other.

Then obviously if you were to take all vertical lines in the w-plane, you would get a perpendicular set of z plane circles (since perp in the w plane). So finally we see what he means by his figure page 87, and he just calls this figure "degenerate Steiner circles".

Sure, the p 87 *figure* is what you would get if $b \to a$ in our earlier Steiner picture. BUT, we are no longer talking about any kind of concentric circles and radial arms stuff as we were on page 85 in the w-plane. We are talking a completely different family now and a different transformation! We now care about sets of parallel lines in the w-plane and where they go.

Now consider the family v = constant and u varies, as I have drawn above. I have drawn at point a as if the tangent were ∞ (vertical).

[p 88] Now he wants suddenly to talk about transformations that map the degenerate circles onto themselves. Recall that we did this for the non-degenerates in p 85 B and then required a,b = a',b' so we had a,b as fixed points. So I am at least willing to consider p 88 A which maps a into a'. I agree that p 88 (13) maps a into a. So let's just take (13) as a starting point. But maybe he has his typo here too!

So: I am going to let the tail end of this section ride. I suspect more typos. I see what he wants to do. In this case, he wants to consider mapping from degenerate Steiner → degenerate Steiner, and he wants to again talk about the "flow".

Comment: I can just imagine what happened here. This is a very minor corner of this book, and nobody really wanted to check it very carefully. He did not do it so the bugs did not get fixed. I just looked for web errata and could find nothing. He died in 2007 so not likely to find a university site active where such errata might have been posted. I looked pretty hard, nothing doing.

My second edition is 1966 (317p) , pretty old, and 3rd edition is 1979 for sale now at $151, used at $50/ It is 336 pages, about the same length as second edition. No view inside is provided. Some interesting comments from Amazon however:

This book **has for decades been THE classic graduate level text in Complex Analysis**. It is important to point out that it is not for beginners.

Not only does this book require some previous understanding of Complex Analysis, but it also requires that mysterious ability called "**mathematical maturity**" - the ability to fill in omitted steps and details when following an argument. But, for a person possessing the prerequisites, this is a fine book.

Ahlfors takes very big leaps in his reasoning. He is extremely **hard to follow** at times. You will need a reference and an instructor if this is your first time using Ahlfors (and maybe even if it is your second). If you have already had a complex analysis class along with one or two real analysis classes, you should be fine. Otherwise, you'd be wise to look elsewhere.

I'm not sure why the other reviews are so positive. The book is very thorough and rigorous I'm sure, but the explanations are **terrible**. Everyone I've talked to in my class agrees that it's **extremely difficult to learn from** if you don't already know complex analysis, because the definitions and order of treatment are very unintuitive.

True, Ahlfors was THE book on complex analysis (better known as "theory of functions" then) some thirty to forty years ago. But I can see no point in using it as a textbook now. As there are many good treatments on the subject today, **it is high time we waved a farewell to this book.** I recommend **Cartan's book** published by Dover since it is available for only twelve dollars.

So I am not alone in my observations! Notice the mention of "graduate level" book, which is where I encountered it in 1972 or so.

I finally found this book on McGraw Hill here: 9780070006577

http://www.mcgraw-hill.co.uk/html/0070006571.html

and here is the TOC of the third edition. I have highlighted places where this differs from my second edition, and you can see that the changes are extremely minor! So he did a minor revamp of his book in 1979 to get it back into selling mode, I am sure. No sections were removed a few were added. Certainly there is no need to buy a new edition of this book, the 1966 is just fine.

Chapter 1: Complex Numbers
1 The Algebra of Complex Numbers
1.1 Arithmetic Operations
1.2 Square Roots
1.3 Justification
1.4 Conjugation, Absolute Value
1.5 Inequalities
2 The Geometric Representation of Complex Numbers
2.1 Geometric Addition and Multiplication
2.2 The Binomial Equation
2.3 Analytic Geometry
2.4 The Spherical Representation

Chapter 2: Complex Functions
1 Introduction to the Concept of Analytic Function
1.1 Limits and Continuity
1.2 Analytic Functions
1.3 Polynomials
1.4 Rational Functions
2 Elementary Theory of Power Series
2.1 Sequences
2.2 Series
2.3 Uniform Coverages
2.4 Power Series
2.5 Abel's Limit Theorem
3 The Exponential and Trigonometric Functions
3.1 The Exponential
3.2 The Trigonometric Functions
3.3 The Periodicity
3.4 The Logarithm

Chapter 3: Analytic Functions as Mappings
1 Elementary Point Set Topology
1.1 Sets and Elements
1.2 Metric Spaces
1.3 Connectedness
1.4 Compactness
1.5 Continuous Functions
1.6 Topological Spaces
2 Conformality
2.1 Arcs and Closed Curves
2.2 Analytic Functions in Regions
2.3 Conformal Mapping
2.4 Length and Area
3 Linear Transformations
3.1 The Linear Group
3.2 The Cross Ratio
3.3 Symmetry
3.4 Oriented Circles
3.5 Families of Circles
4 Elementary Conformal Mappings
4.1 The Use of Level Curves
4.2 A Survey of Elementary Mappings
4.3 Elementary Riemann Surfaces

Chapter 4: Complex Integration
1 Fundamental Theorems
1.1 Line Integrals
1.2 Rectifiable Arcs
1.3 Line Integrals as Functions of Arcs
1.4 Cauchy's Theorem for a Rectangle
1.5 Cauchy's Theorem in a Disk
2 Cauchy's Integral Formula
2.1 The Index of a Point with Respect to a Closed Curve

2.2 The Integral Formula
2.3 Higher Derivatives
3 Local Properties of Analytical Functions
3.1 Removable Singularities. Taylor's Theorem
3.2 Zeros and Poles
3.3 The Local Mapping
3.4 The Maximum Principle
4 The General Form of Cauchy's Theorem
4.1 Chains and Cycles
4.2 Simple Connectivity
4.3 Homology
4.4 The General Statement of Cauchy's Theorem
4.5 Proof of Cauchy's Theorem
4.6 Locally Exact Differentials
4.7 Multiply Connected Regions
5 The Calculus of Residues
5.1 The Residue Theorem
5.2 The Argument Principle
5.3 Evaluation of Definite Integrals
6 Harmonic Functions
6.1 Definition and Basic Properties
6.2 The Mean-value Property
6.3 Poisson's Formula
6.4 Schwarz's Theorem
6.5 The Reflection Principle

Chapter 5: Series and Product Developments
1 Power Series Expansions
1.1 Wierstrass's Theorem
1.2 The Taylor Series
1.3 The Laurent Series
2 Partial Fractions and Factorization
2.1 Partial Fractions
2.2 Infinite Products
2.3 Canonical Products
2.4 The Gamma Function
2.5 Stirling's Formula
3 Entire Functions
3.1 Jensen's Formula
3.2 Hadamard's Theorem
4 The Riemann Zeta Function
4.1 The Product Development
4.2 Extension of ?(*s*) to the Whole Plane
4.3 The Functional Equation
4.4 The Zeros of the Zeta Function
5 Normal Families so all these sections are now numbered 6.
5.1 Equicontinuity
5.2 Normality and Compactness
5.3 Arzela's Theorem
5.4 Families of Analytic Functions
5.5 The Classical Definition

Chapter 6: Conformal Mapping, Dirichlet's Problem
1 The Riemann Mapping Theorem
1.1 Statement and Proof
1.2 Boundary Behavior
1.3 Use of the Reflection Principle
1.4 Analytic Arcs
2 Conformal Mapping of Polygons
2.1 The Behavior at an Angle
2.2 The Schwarz-Christoffel Formula
2.3 Mapping on a Rectangle
2.4 The Triangle Functions of Schwarz
3 A Closer Look at Harmonic Functions

3.1 Functions with Mean-value Property
3.2 Harnack's Principle
4 The Dirichlet Problem
4.1 Subharmonic Functions
4.2 Solution of Dirichlet's Problem
5 Canonical Mappings of Multiply Connected Regions
5.1 Harmonic Measures
5.2 Green's Function
5.3 Parallel Slit Regions

Chapter 7: Elliptic Functions
1 Simply Periodic Functions
1.1 Representation by Exponentials
1.2 The Fourier Development
1.3 Functions of Finite Order
2 Doubly Periodic Functions
2.1 The Period Module
2.2 Unimodular Transformations
2.3 The Canonical Basis
2.4 General Properties of Elliptic Functions
3 The Weierstrass Theory
3.1 The Weierstrass p-function
3.2 The Functions ?(*z*) and s(*z*)
3.3 The Differential Equation
3.4 The Modular Function ?(*r*)
3.5 The Conformal Mapping by ?(*r*)

Chapter 8: Global Analytic Functions
1 Analytic Continuation
1.1 The Weierstrass Theory
1.2 Germs and Sheaves
1.3 Sections and Riemann Surfaces
1.4 Analytic Continuations along Arcs
1.5 Homotopic Curves
1.6 The Monodromy Theorem
1.7 Branch Points
2 Algebraic Functions
2.1 The Resultant of Two Polynomials
2.2 Definition and Properties of Algebraic Functions
2.3 Behavior at the Critical Points
3 Picard's Theorem
3.1 Lacunary Values
4 Linear Differential Equations
4.1 Ordinary Points
4.2 Regular Singular Points
4.3 Solutions at Infinity
4.4 The Hypergeometric Differential Equation
4.5 Riemann's Point of View
Index

4. Elementary Conformal Mapping (89)

Ahlfors sounds a bit excited almost that conformal mapping actually finds use in the real world. we are no longer talking about linear transformations, but arbitrary analytic ones.

4.1 The Use of Level Curves (89)

Imagine a mapping w = f(z). There are two obvious things we might want to do:

(1) Assume straight lines $x = x_0$ (vertical lines) or $y = y_0$ (horizontal lines) in the domain, and see what these map into in the range, ie, in terms of $w = u + iv$, ie, in the u-v plane. The curves you get in the u-v plane in this case are called **level curves**. Since angles are preserved, these curves form some kind of orthogonal grid system in the w plane.

(2) Assume straight lines $u = u_0$ (vertical lines) or $v = v_0$ (horizontal lines) in the range, and see what these map into in the domain, ie, in terms of $z = x + iy$, ie, in the x-y plane. I suppose these are also called level curves, etc.

First example is $w = f(z) = z^{\alpha}$. If $\alpha = 1.5$, then unit circle in z-plane maps into 1 1/2 unit circles in the w plane. So a single point in the range might be the image of two points in the domain, so 1:2 mapping, not 1:1 mapping for some unit circle points. If $\alpha > 1$, limit the domain to a wedge is small enough so you stay 1:1 in your mapping. He suggests a domain sector (φ_1, φ_2) as limits.

$w = z^2$

(1) In particular, look at $w = f(z) = z^2$ as a simple case. I already know that $u = x^2 - y^2$ and $v = 2xy$, so what does this all look like? A type (2) level curve above would to take $u = u_0$ = vertical line in range, and back in the domain you get $x^2 - y^2 = u_o$ which is a horizontal hyperbola if $u_0 > 0$ and a vertical one if $u_0 < 0$. So the right-side vertical lines ($u_0 > 0$) map into the horizontal hyperbolas shown page 91 centered at 0,0. The left-side vertical lines are the vertical hyperbolas. So think of both the V and H hyperbolas as one "set" of curves in the mesh. The other set is the horizontal lines $v = v_0$ in the range, so $xy = v_0/2$. Again, for $v_0 > 0$, you get these obvious curves in quadrants I and III, else in II and IV. This is the second set of curves in the page 91 figure. Notice the right angles between the two sets!

(2) Now let's do it the other way. First, take the two equations above and fiddle

$$u = x^2 - y^2 \qquad v = 2xy \qquad \Rightarrow u = -v^2/2x^2 + x^2$$
$$\Rightarrow u = +v^2/2y^2 - y^2$$

The domain level curves $x = x_0$ (vertical lines) map into $u = -v^2/2x_0^2 + x_0^2$ in the range, and these are of course parabolas growing to the left. But $y = y_0$ (horizontal lines) become $u = +v^2/2y_0^2 - y_0^2$ which are parabolas to the right, so we then get the range mesh shown on page 92.

I ask: if you start with $w = z^1$ and gradually raise exponent to $y = z^2$, what happens with all these curves? That would make a great movie. In both cases above, we start with just a plain grid of lines, but they slowly warp into these pictures (I think). The picture on page 91 represents a sort of double bending of things due to the doubling of z^2 in the angle world. That is why z^3 looks like Fig 10.

The parabolas are exactly the "parabolic coordinates" we use doing hydrogen, recall:

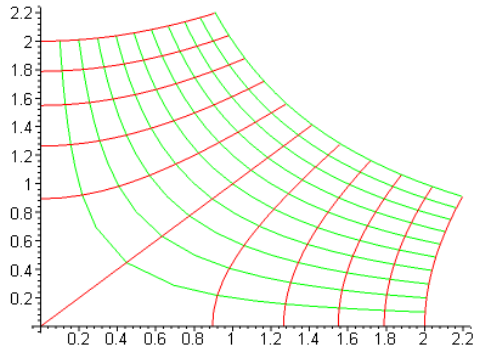
Two-dimensional parabolic coordinates (σ,τ) are defined by the equations

$$x = \sigma\tau$$
$$y = \frac{1}{2}\left(\tau^2 - \sigma^2\right)$$

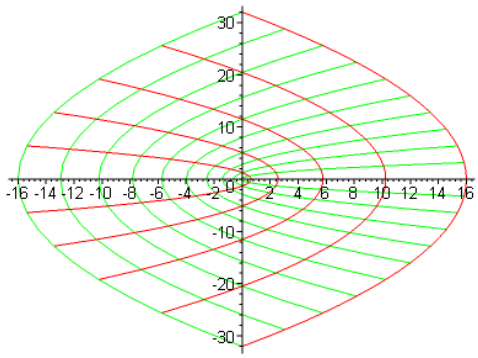
So we now see these coordinates are just the level curves of the function $w = z^2$ (add a constant factor).

Notes added 10.8.09. Maple contains a function "conformal" which takes a rectangular region of the domain plane, treats it as a standard grid, and then plots the resulting "level lines" in the range plane. Here are two plots for the mappings discussed just above, where I have taken a somewhat strange rectangle in the domain plane, one whose corner coordinates are (-4,0) and (4,4),

```
conformal(z^(1/2),z=-4..4+4*I);
```




```
conformal(z^2,z=-4..4+4*I);
```



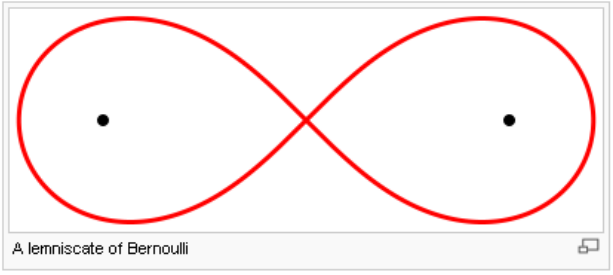
Now staying with $w = z^2$, A suggests we consider a set of circles in the w plane centered at 1. These circles map back into the equations shown in p 91 A. These quartic curves are call **lemniscates** and here they are in wiki,

In mathematics, the **lemniscate of Bernoulli** is an algebraic curve described by a Cartesian equation of the form:

$$(x^2 + y^2)^2 = 2a^2(x^2 - y^2).$$

The curve has a shape similar to the numeral 8 and to the ∞ symbol.

The lemniscate was first described in 1694 by Jakob Bernoulli as a modification of an ellipse, which is the locus of points for which the sum of the distances to each of two fixed *focal points* is a constant. A Cassini oval, by contrast, is the locus of points for which the *product* of these distances is constant. In the case where the curve passes through the point midway between the foci, the oval is a lemniscate of Bernoulli.



A lemniscate of Bernoulli

Bernoulli called it the *lemniscus*, which is Latin for "pendant ribbon".

This lemniscate can be obtained as the inverse transform of a hyperbola, with the inversion circle centered at the center of the hyperbola (bisector of its two foci).

which uses the case k = 1, so this is the image back in the z plane of the unit circle about the point 1 in the w plane! I verified equation p 91 A on scratch paper.

Obviously the orthogonal curve set would come from lines through the point w = 1. These lines are

$u = m(v-1)$ m is the slope

But we know $u = x^2 - y^2$ and $v = 2xy$ so get

$x^2 - y^2 = m(2xy-1) = 2mxy - m$ // compare to p 9 1B

I think he has another typo as marked. In any event, as you vary m continuously, you get some kind of strange rotating set of hyperbolas.

$w = z^3$

Now we look at $w = z^3$. I agree that Figure 10 is reasonable (shows the hyperbola-like forms only, not the orthogonal xy type forms). We now have triple compression of the angle instead of the double in Fig 91. Maybe we do need integral powers in order to be analytic at the origin, only cases he is looking at so far.

Now when we go to z^3, the parabolas of Figure 9 I think get wrapped even more and becomes those **folium of Descartes** things one of which is shown in Fig 11. This is just a certain cubic curve.

$w = e^z = e^x e^{iy}$ so $u = e^x\cos(y)$ and $v = e^x\sin(y)$ w - u_iv

Now we look at $w = e^z$. The usual grid in the z plane goes into circles and rays in the w plane (centered on the origin), and a z plane strip therefore goes into an angular sector. If you choose the strip width to be π, you map into an 180 degree sector which is the right side half plane. So this map can take (strip) → (half plane), which might be a useful conformal map in some problem. Of course we know that LT take the half plane into any circle, so now we have a mapping strip ↔ any circle. I am sure this will come up in electrostatics. ON at the bottom of p 92 he writes the LT that takes right half plane to unit circle, so the combination shown as p 93A provides a map from (strip) → (unit circle at origin).

So this was a very nice tour of some "level curves". [I could not make Maple use anything but Cartesian level curves for the domain plane, but you could first map these to something else, etc etc.]

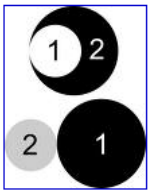
4.2 A survey of elementary mappings (93)

Now we start our study of mapping some region Ω_1 into another region Ω_2. The best approach is to find a mapping which takes each of these to a unit circle (or a half plane), then combine those mappings and you have your answer. So imagine a catalog of region shapes Ω_1 and how to take each shape into a disk or a half plane. The tools of the trade are the simple mappings we studied in the last section: e^z and ln(z) and powers and linear transformations. Since all are analytic and all do *some* Circles → *some* Circles, this method can only work if the boundary of Ω_1 is made up of pieces which are circular or straight.

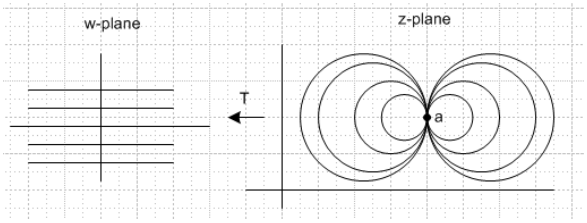
[p 94] Example 1: consider the various lunes and lenses you get by intersecting C_1 type Steiner circles. These are the simplest things having two "pieces" of boundary and each is a circle. We know these map into angular sectors by our linear transformation studied above (see p 84 notes). And we know that we can scale an angular sector into a half-plane angular sector with a power z^α so we then have a prescription

(at least) to take any "lune or lens" into a half plane, and this would include a case where one of the two edges was a straight line, being an extremal Steiner circle.

Example 2: Suppose two circles are exactly tangent. This can happen in 2 ways:

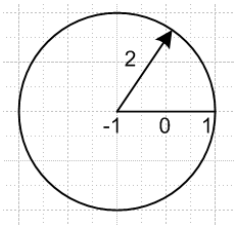


In the first case, we have a black intersection region of interest. But these circles appear in the degenerate Steiner picture page 87 with transform p 87 A (corrected by me). So, we conclude here, arguing as we did earlier about regular Steiner, that the intersection of two tangent circles would map into a strip via this p 97 A special case of a linear transformation (c = 0 does not do much). Recall our picture above:



Example 3: Stare again at Steiner p 85. Consider one of our lune regions formed by the intersection of two C_1 circles. If we were to intersect that with one of the C_2 circles, we would get a "circular triangle with two right angles". I agree. This must map somehow into an angular sector that runs from r = 0 to r = some A, or angular sector from r = A to r=∞. By putting our triangle on the "a side" we can select out the case r = 0 to r=A. We could process that with z^{α} to get then a circle. He says need to go to a half circle and then from there to a half plane.

Example 4: More Ahlfors vagueness and refusal to draw pictures, forcing the reader to try to draw them based on his imprecise words. I think he is starting with this picture on the left,

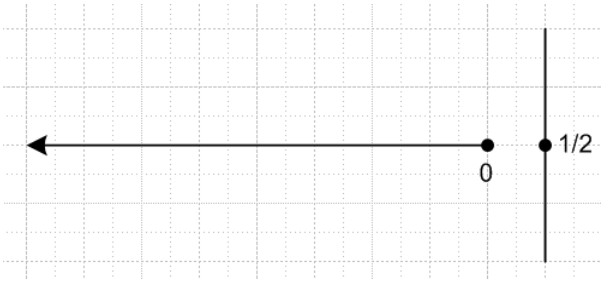


He wants "the complement of a line segment". Within the circular region I show, that complement could be considered a circular "wedge" running from say $0+\varepsilon$ to $2\pi-\varepsilon$ in the limit $\varepsilon \to 0$. The first transform he shows maps the points -1,0,1 to 0,-1,-∞. Thus, the segment has become the negative real axis. What happens to the circle? Let's trace three points on our circle

right point z = 1 → ∞

left point z = -3 → 1/2
top point (-1,2i) → 1/2 - i 1/2

The circle seems to map into a line passing through w = 1/2 and w = 1/2 - i 1/2 which is of course a vertical line, so we now have:



The center of our circle went to 0, so the inside of the circle seems to be the left half plane of our vertical line less the negative real axis. The outside of our circle goes to the other half plane. Thus, one can I think say that "the complement of the segment" is mapped into the entire w plane excluding the negative real axis. We can regard our circle as just a guide.

Now what does $w = z^{1/2}$ do to the above picture if we use it as a second transformation? Well, any point moves along a circle toward the real axis and has its angle cut in half. Let's think $(-\pi,\pi)$. So the entire plane gets morphed into the right half plane. Our vertical line no longer plays any role. (I suppose the vertical line would map into some sort of bent V shape thing with some curve on the edges).

But what happens to the negative real axis in this mapping? It could go to either the positive or negative real axis depending on how we think of its ε. Let's take it to the positive imaginary axis. So now we have I think done the following:

mapped the complement of the segment (-1,1) to the right half plane less the positive imaginary axis.

He then does a third transform which moves the open right half plane to the open circle. I would argue that in fact we have to include part of the circular boundary in our mapped set.

But OK, we get the point. Apart from my technical detail, we have mapped the complement of our finite line segment to the inside of a circle and perhaps some of its boundary. Since the original set is open, you somehow thing the final set should be open, so I am wrong about that boundary piece for some reason. Enough.

If you compose the three transformations just discussed, I agree that you get both results in (15). I did this on scratch, got the lower result first then from it the upper result. Not rocket science.

[p 95] Now he is going to keep going with this mapping to see what some level curves look like. He shows that things look like the figure shown at page bottom. A disk in w-plane maps into *the outside of* an **ellipse** in the z-plane, so we can now map between circles and "outside of ellipses". The "region between hyperbolas" will map into an angular sector in the z-plane, so we have expanded our list of useful region mapping transforms.

Example 5 [p 95 B] Here we assume w = f(z) = a general cubic polynomial. A preliminary transformation changes it to $w = z^3 - 2z$. A second transformation $z = \xi + 1/\xi$ changes it to this form:

$w = \xi^3 + 1/\xi^3$. We know from (15) and the above that a circle in ξ is an ellipse in z. But now he claims that you also get an ellipse in the w plane. We just do the same math we did to get equations on top of page 95 but a different power. So I think we still get ellipses and hyperbolas in this example, but they are different from the ones in the previous example. The rest of the text on page 96 does little for me, just comments on this obscure example, so I skip this text for now.

PL Exercise. Explain why the mapping $w = k(z^n + z^{-n})$ turns circles into ellipses?

Write $z = re^{i\theta}$ then $w = kr^n e^{in\theta} + kr^{-n}e^{-in\theta} = u+iv$. Thus

$u = k(r^n + r^{-n})\cos\theta = a(r)\cos\theta$ $\qquad$ $v = k(r^n - r^{-n})\sin\theta = b(r)\sin\theta$

$u^2/a^2 + v^2/b^2 = 1$ $\qquad$ => $\qquad$ aligned ellipse with $a > b$

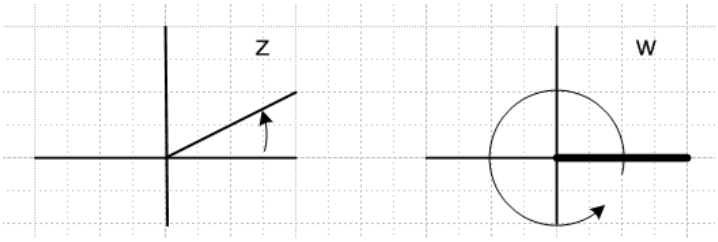
In the case n = 1, we get $a = k(r+1/r)$. If $y = r + 1/r$, then $y' = 1 - r^{-2} = 0$ when r = 1, so $\min(r+1/r) = 2$. If k = 1/2, which is usually the case you see, then min(a) = 1. This that *all* ellipses include the points ±1 regardless of the radius r used in the z plane. When r = 1, the ellipse touches these two points.

Exercises: There are 8 problems here of "find a conformal map that takes some region into some other region". Oddly, I have all these exercises circled from some previous reading of this chapter? As I look back to previous groups of exercises, I see that I have numbered selected ones, as if they were once assigned to me. Pencil underlining is an indication of past reading, by the way, since I am always (with some pencil) using red Razor now. So it does seem that I have "been here before", but I don't remember much! My Math 185 Behrens notes say clearly that we only really covered Chapters 2 and 4 in that course. So I am perplexed as to when I might have read this chapter 3. Maybe when Jackson did conformal mapping? But Jackson was the previous year 1971, Math 185 was 1972. I have no doc files on Alta with Alfors or Ahlfors, so I would presume this reading was done before 1984 when I got a Mac and could make docs.

Idea: probably when I was doing S-matrix theory and we had the stu planes and all those particle creation branch cuts, I went to Ahlfors to learn about such things. That could have been any time in the span 1972-1977 I suppose. In 1977 I came to Utah and stopped worrying about S-matrix stuff.

4.3 Elementary Riemann Surfaces (97)

Example 1: $w = z^n = R^n e^{i\theta n}$. Certainly the angular sector from $\theta = 0$ to $2\pi/n$ in z maps into the entire w plane. But then $2\pi/n$ to $4\pi/n$ does the same thing. He writes this correctly as $\theta = (k-1)2\pi/n$ to $k2\pi/n$ for k = 1,2,3..n. There are then n angular sectors each of which map into the entire w plane, so there are n Riemann sheets in the range. Here might be the case for n = 12:

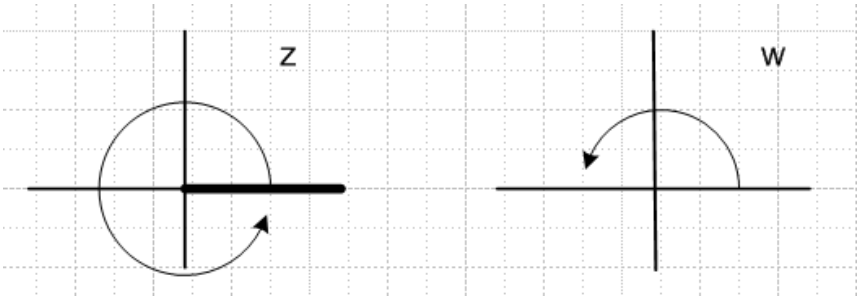


The point w = 0 is a **branch point**. After you go around once as shown, in the w plane you would pass through the **branch cut** onto the second **sheet** as z moved into its second wedge, and so on. Eventually, when you have gone full circle in the z plane, you go from the 12th sheet back to the first sheet in the w plane. The **union of all the Riemann sheets is called the Riemann surface**.

If you were to change the cut to be any curvy line from 0 to infinity, it still works, there are still 12 sheets, and so on. Note that z = 0 touches all 12 sheets. The claim is that this is true for ∞ also.

This example provides an example of what I call a **"range cut"**. There is never going to be a discontinuity across a range cut! The reason is that, just above or just below such a cut, you have the same value of w (ie, that point in the w plane, call it w_1), so the discontinuity is $w_1 - w_1 = 0$.

Example 1A. Consider $w = z^{1/n} = R^{1/n}\, e^{i\theta n/n}$. This is the same as $z = w^n$ so we just reverse the labels on the picture above. Now the branch cut is in the domain plane. Example $w = z^{1/2}$.



Now the domain has 2 sheets and has the branch cut, not the range plane. You have to say which domain plane you are talking about (which branch of the function $w = z^{1/2}$).

This example is an example of what I call a **"domain cut"**. A domain cut always has a non-zero discontinuity. The discontinuity will be $w_1 - w_2 \neq 0$ where w_1 and w_2 are the two different points in w space that correspond to a point in z space just above the cut, and a point in z space just below the cut.

Example 1B: Consider $w = z^{\alpha}$ for some arbitrary positive real number α. In this case, I think you have a branch cut in both planes! If α is rational P/Q, then I suppose P rotations in z brings you Q rotations in w and you then start over.

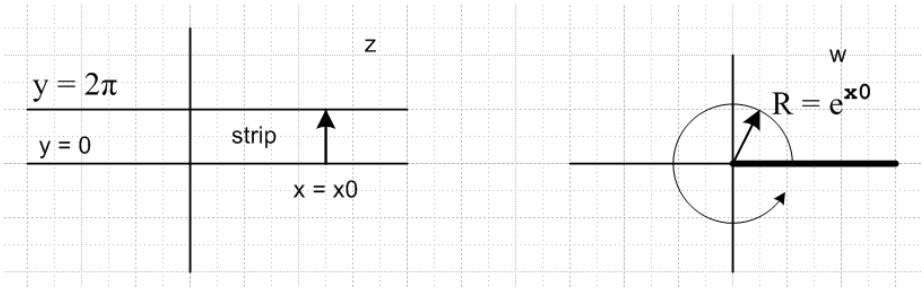
$\theta_w = \theta_z\, \alpha + 2\pi n$ makes $w = z^{\alpha}$

$\theta_w = \theta_z\, P/Q + 2\pi n$

$Q\theta_w = \theta_z\, P + 2\pi m$ $\quad m = nQ$

If α = integer n, the branch cut in the z plane vanishes. And if α = 1/n, the branch cut in the w plane vanishes. If α is not rational, then both phasors go around forever in an infinite screw of sheets and things never return to the starting position. [this is my example, not Ahlfors's example]

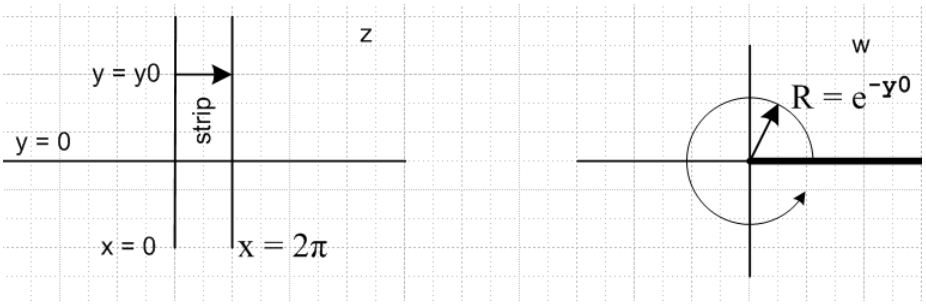
Example 2: $w = e^z = e^x\, e^{iy}$.



Here the one strip shown maps into the full w plane, so you need a branch cut in the range. In this case there are an infinite number of sheets on the Riemann surface, the infinite "screw". Branch point is at w = 0 in the w plane, no branch cuts in the domain plane. [e^{z} is analytic in the entire z plane] [the right side of the strip maps into the outside of the unit circle in w space.]

Example 2A: $w = w = e^{iz} = e^{ix}\, e^{-y}$

Same as above, but slightly different. The strips are now vertical strips.



If you cared about $w = e^{-iz} = e^{y}\, e^{-ix}$, you get a similar mapping but if x increases as the left side arrow shows, then the right circle goes in the reverse direction, and the radius is $R = e^{y0}$. [the upper side of the strip maps into the inside of the unit circle in w space.]

Example 3. $w = (1/2)(z + 1/z)$. First, some math with $z = Re^{i\theta}$:

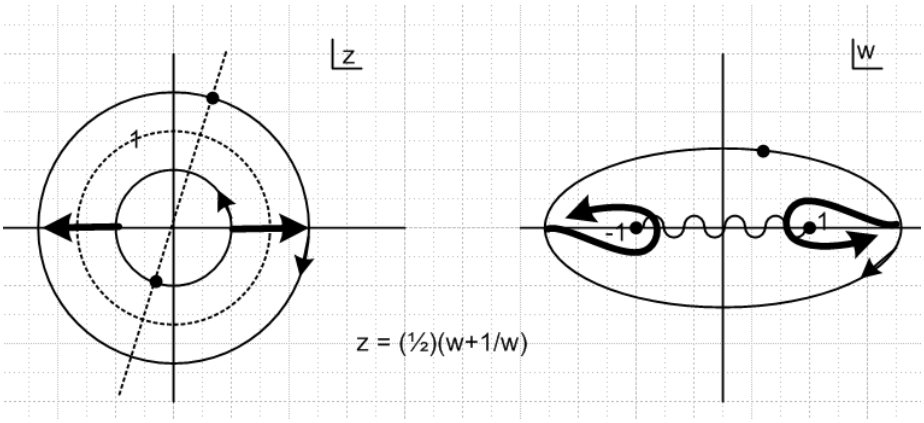
$$w(z) = (z + 1/z)/2 = (Re^{i\theta} + R^{-1}e^{-i\theta})/2 = \{ (R + R^{-1})\cos\theta \}/2 + i \{(R-R^{-1})\sin\theta\}/2 = u+iv$$

Notice that $(u/a)^2 + (v/b)^2 = 1$ where $a = (R + R^{-1})/2$ and $b = (R-R^{-1})/2$ hence a circle maps into an ellipse. If a circle of radius R maps into this ellipse, then obviously so does one of radius 1/R, so two circles map into the same ellipse.

We now consider the question of "which way" we traverse the ellipse. If $R < 1$ for our circle, then we see that $b < 0$ so as θ starts moving in a positive direction around its circle, w(z) gets a neg imag part, so for this circle ellipse traversal is clockwise. But for the corresponding $R' = 1/R > 1$ circle, $b > 0$ and traversal is CCW. We have tried to show this with arrows in the picture below.

So a point on the ellipse maps back to a point on each of two circles. It is easy to show that if one of these points is at angle θ in z space, the other point will be on the opposite point of the other circle.

Here is a picture of the mapping. The outside of the dotted unit circle maps into the entire w plane, and so does the inside of that circle, so in the range we have two full-plane Riemann sheets.



The inverse of the above equation is $z = w \pm \sqrt{w^2-1}$, so that each w point on an ellipse corresponds to two points in z, in agreement with the above. Now, what this means is there is a branch cut as shown in the w plane. If we write $z = w - \sqrt{w^2-1} = w - \sqrt{w-1}\ \sqrt{w+1}$ (this would be the inner circle on the left), we can then start on the right extreme of the ellipse (right extreme of inner circle) and go on a little contour voyage which rotates around the branch point we think is at w = 1. The usual trick is to write $w-1 = Re^{i\theta}$ with small R on the loop around w = 1, and we "come out" of the loop with $z = w + \sqrt{w^2-1}$ which is the extreme right point on the outer circle in the z plane. So the fat loop I show on the right of the w-plane picture corresponds to the fat arrow on the right of the z-plane picture. The same thing happens at w = -1, and I hope I have the sense correct there (I use the same CCW sense of rotation since same meaning of square root).

So, for the function w = (1/2)(z + 1/z), the range has two Riemann sheets, and the same ellipse on the two sheets maps back into two different circles in the z plane. One sheet of the w plane maps to the exterior of the dotted circle in the z plane, while the other sheet maps to the interior of the dotted circle.

(junk material in Appendix A used to be here:)

Example 4. u = (1/2)(w + 1/w) with $w = e^{iz}$, which means u(w(z)) = cos(z). This is obviously a compounding of Example 3 with Example 2A above. First, let's do some math:

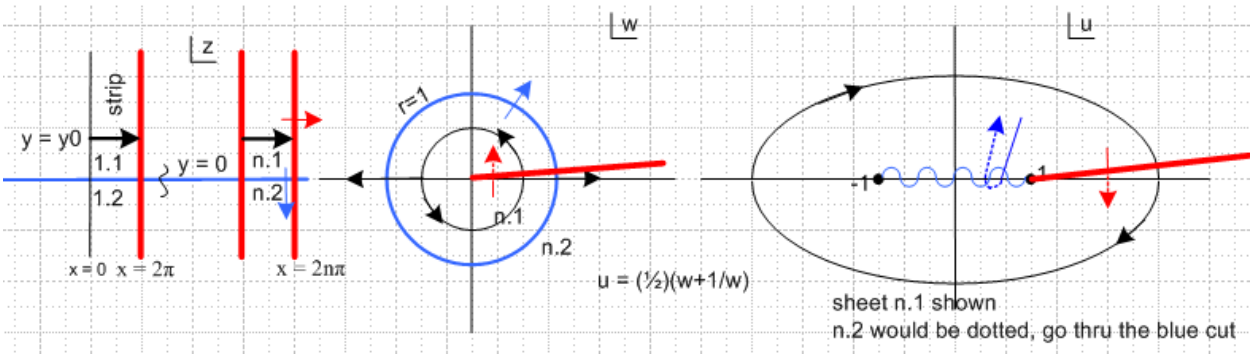
$u(w(z)) = [\ e^{iz} + e^{-iz}]/2 = \cos(z) = [\ e^{i(x+iy)} + e^{-i(x+iy)}]/2$

$= [\ e^{-y}e^{ix} + e^{y}e^{-ix}]/2 = \{(\ e^{-y} + e^{y})\cos x\ \}/2 + i\ \{\ (e^{-y} - e^{y})\sin x\ \}/2$
$= \mathrm{ch}(y)\cos(x) - i\ \mathrm{sh}(y)\sin(x)$

$\Rightarrow\ \mathrm{Im}[\ \cos(z)\ ] = -\ \mathrm{sh}(y)\sin(x)$

Now let's look at all three planes at the same time:

		u(w(z)) = cos(z)
z	$w(z) = e^{iz}$	$u(w) = (w+w^{-1})/2$

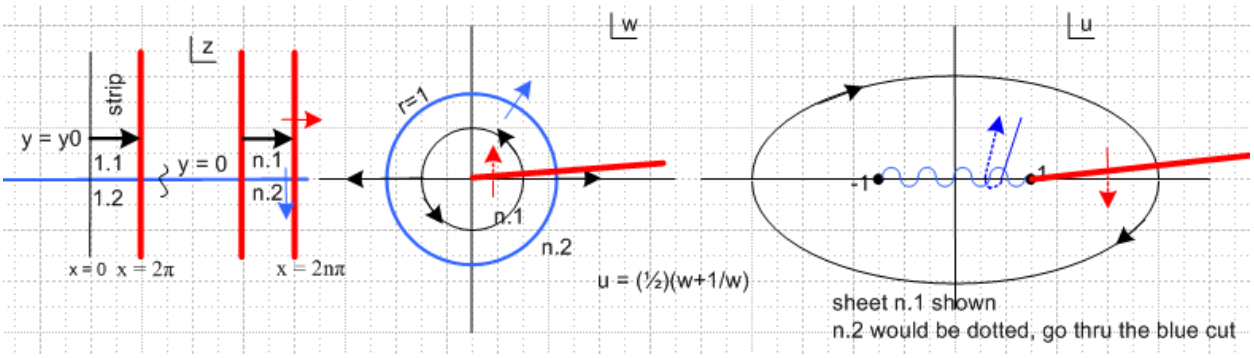


If we do the **black arrow voyage** in z, that becomes the **black circle** in w. Note that the upper half of the strip maps to the inside of the blue circle in w, as we noted above. This circle path then appears as the **black ellipse** path in u. If we were to continue another arrow's worth to the right in the z plane, we do another w circle, but on another sheet of the w plane, and similarly we do an ellipse on another sheet in u. Thus, some cut must exist as shown in red in the u plane. This is the image of the red vertical line(s) that started in z, became the positive real axis cut in w, and then becomes (1,∞) in u.

[**Red cut**: Think about the location of the cut in w space. It is the positive real axis. The w space segment (0→1) maps into (∞→1) in u space. But then the w space segment (1→∞) maps to (1→∞). So the single cut in w maps into two superposed cuts in our u picture, one on each sheet.]

We have concocted a **scheme to label sheets and regions**. We start in z where the strips are numbered 1,2,3....n.... and we show the first and nth strips with black arrows. We then add .1 to indicate the upper half of the strip, and .2 to indicate the lower half strip. So look at the n^{th} strip in z. Its black arrow maps into a circle in the n.1 region of w, and this maps into an ellipse in region (sheet) n.1 in u. Whereas in w this region n.1 consists of the interior of the blue unit circle, in u region n.1 fills an entire sheet of the u plane, so we cannot show the n.2 sheet there. We think of it as a "dotted sheet" that you get to by going through the blue cut, which recall maps back into the blue circle in w and then to the real axis in z.

Suppose we were to **slowly lower the black arrow in z** and let it move across the x axis into the lower half plane there. What happens in the other spaces? The black circles get larger and approach the blue unit circle in w. Meanwhile, the ellipses in u in fact shrink like a rubber band until the limiting one just captures the two points -1,1 and is infinitely thin. Then as our black arrow cross the x axis in z, the circle in w just moves outside the unit circle, and in u the ellipse "falls through the cut" and appears on sheet n.2 which we called the dotted sheet. If we continue, it then expands out on that dotted sheet to a normal size ellipse. [just repeating the above figure for convenience]



Now consider the **three red arrows** shown above (use split screen!) In z we move from region n.1 to region (n+1).1. The corresponding thing happens in w and u, but in both those spaces we move to different sheets so the arrow part of the arrow is drawn dotted (hard to see)

Similarly consider the **three blue arrows**. The interesting thing is what happens in u space As we cross the blue unit circle in w space, we dive into the cut and come back out on the other sheet in u space.

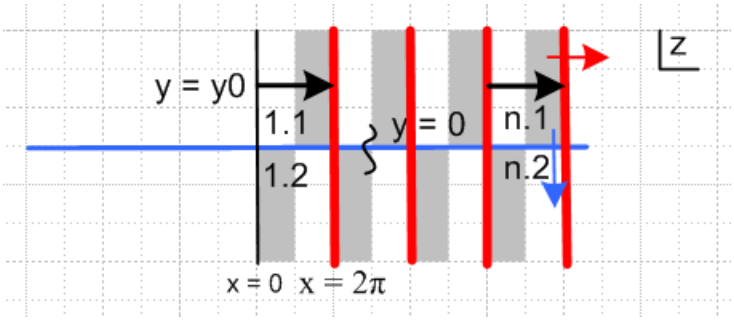
The upshot is that in u space, we label our sheets with notation n.m where n = 1,2,3.... and m = 1,2. (of course we also have n = 0,-1,-2... but I have ignored those for this discussion), so in u space we have a sort of double infinity of sheets. The blue cut is a 2-sheet affair, so if you go through it twice you get back to the sheet you started, but the red cut is an ∞-sheet cut.

So where do the cuts in u space come from? The red cut is the image of the red cut in w space, which arises from the nature of $w = e^{iz}$ each of whose z strips maps into an entire w plane. The blue cut in u space arises from a *region boundary* in w and z space. It is admittedly complicated!

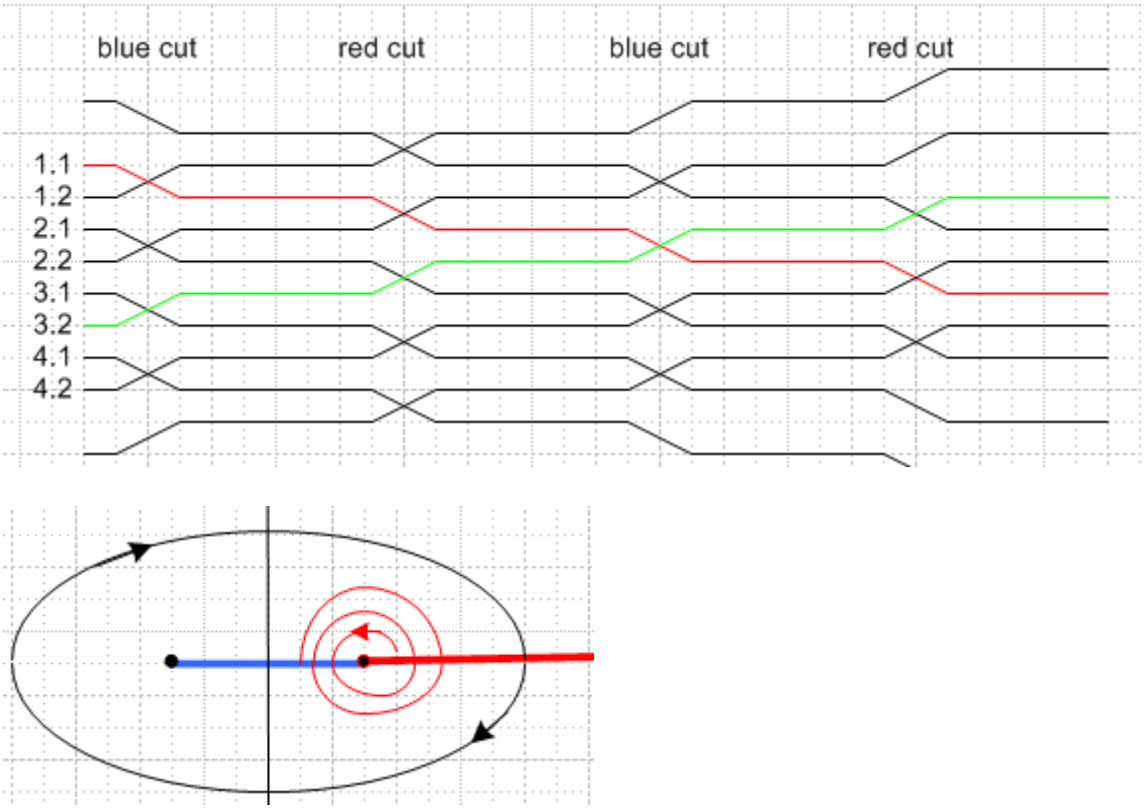
Look again at our first black arrow which maps into the u space ellipse shown. For the first half of this arrow, we are on the lower part of the ellipse, which means Im[cos(z)] < 0. For the second half of this arrow, the opposite is true. This is also obvious from the math above.

$$\cos(x+iy) = \mathrm{ch}(y)\cos(x) - i\,\mathrm{sh}(y)\sin(x)$$

Suppose we mark the regions in gray which have Im[cos(z)] > 0 in this way. In the lower half plane, the gray regions are swapped since y has the opposite sign. We end up with this picture



which (finally) explains what Ahlfors has drawn on **page 99**. Ahlfors has another interesting picture on **page 98** and I draw my own version of it here:



The continuous lines in the upper picture (such as the one I drew in red) show in schematic fashion the path that "an ant" would follow were he to start on some sheet and do a spiral path like that shown in the lower picture. Each time the ant crosses one of the two cuts, he moves to another sheet. So for the red path shown the sheet sequence would be

1.1 pass thru blue cut 1.2 pass thru red cut 2.1 pass thru blue cut 2.2 pass thru red cut, and so on.

The sheet labels are shown on the left edge. The green path shown starts on sheet 3.2 and spirals the other direction. The entire picture is just the union of paths with *selected* starting points and directions.

I used to try to associated the path diagram with "what you would see" if you were an ant outside the multi-sheeted object (say ant is off to the left side and ellipse is tilted back flat). It would be amusing to try to draw some Escher-like 3D "spiral parking garage" picture showing the full sheet structure. Along the blue cut, we have one ramp going up and one intersecting ramp going down, so they magically pass through each other. For the red cut as have an infinite number of ramps going up, and another infinite set going down. Too much for me!

x

Ahlfors wrote a whole book on Riemann Surfaces, so I guess he assumed the reader would know a little more than an actual reader would know, which is why this section was so hard. My pencil word

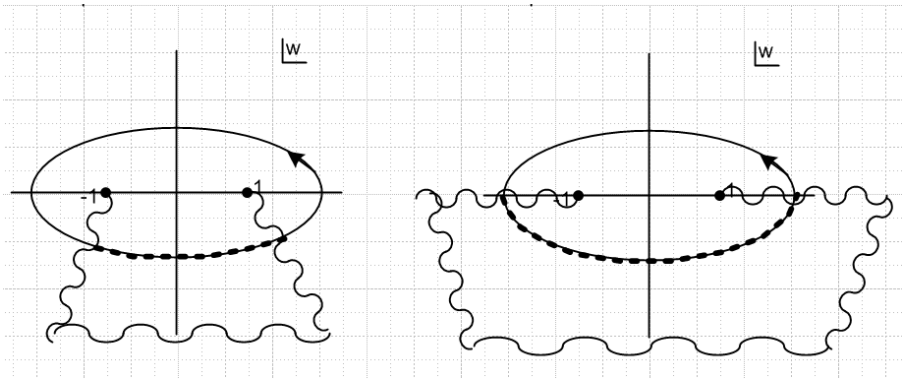
"ugh" is there from my earlier reading in circa 1975. I cannot give him a very good grade for his clarity in this section.

Schedule Comments: I finished my first reading cut of Chapter 3 (p 100) on 8.20.09, but had time to meta-review only the simple Chapter 1 of this book prior to an 11 day trip to Cape Cod from which I returned 8.31.09, having lost all momentum. It then took me from 9.1.09 to 9.12.09 to write my meta-reviews of Chapters 2 and 3, twelve long days. Admittedly I got sidetracked on various important topics, such as " about e and e^x and exp(x) and ln(x)" , how to maintain the math people timeline, and a little on Cantor which I still need to write up, then projective transformations. I added history to the raw notes, and reworked sections that were really bad. I think I had now better return to Stakgold at the point where conformal mapping is brought up, which caused me to digress and read a full 100 pages of Ahlfors. I had no idea Ahlfors's book was going to be so "difficult", as confirmed by web comments.

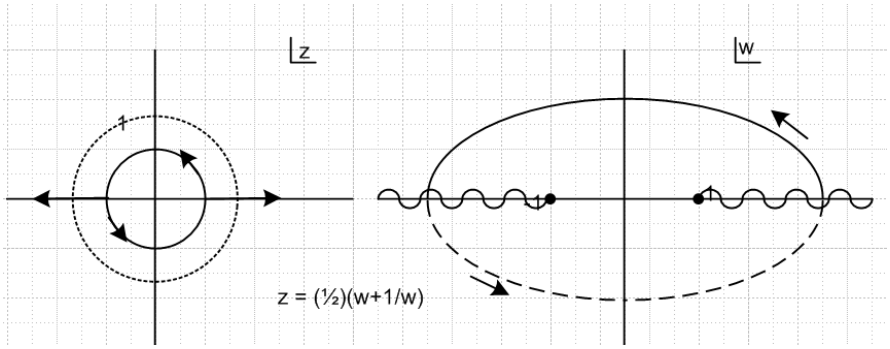
Appendix A: I think this is garbage and will be deleted.

Noted added 1.24.11: I think this whole idea of deforming the cut just creates confusion. Also, I think the label system n.m differs from that used below! I keep it for now, make it blue, and suggest you skip it.

Now let's deform the branch cut in a few steps:



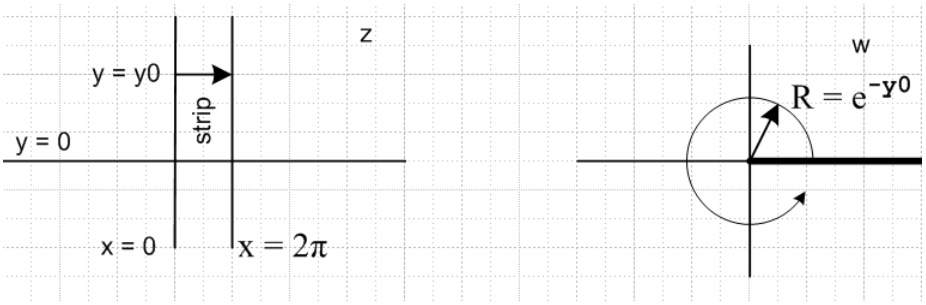
In these pictures, our ellipse is continuous, but we jump onto the second sheet during the dashes, then we come back again to the solid line on the first sheet. So we can then draw this picture:



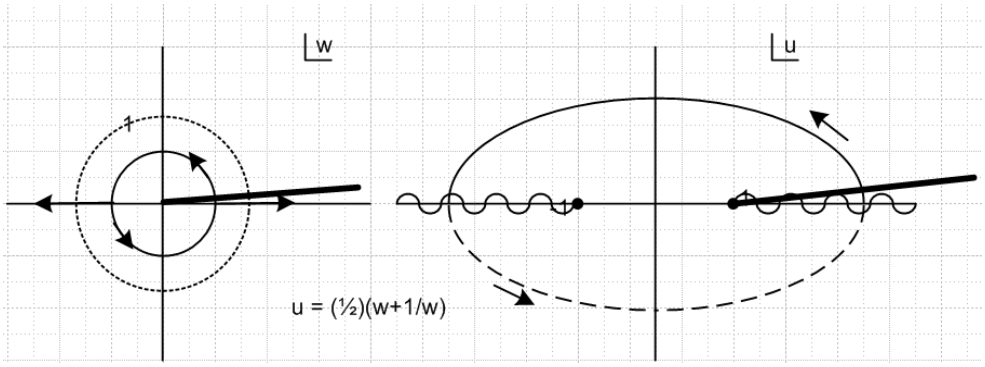
Example 4. $u = \cos(z) = 1/2\ (\ e^{iz} + e^{-iz})$. To understand this example, we need to consider two separate sequential transformations:

$u = (1/2)\ (\ w + 1/w)$ where $w = e^{iz}$

The rightmost transform does this kind of thing'



But now we have to shadow this new solid line branch point into the final plane like so: [think about the location of the cut in w space. It is the positive real axis. The w space segment (0→1) maps into (∞→1) in u space. But then the w space segment (1→∞) maps to (1→∞). So the single cut in w maps into two superposed cuts in our u picture, one on each sheet.]

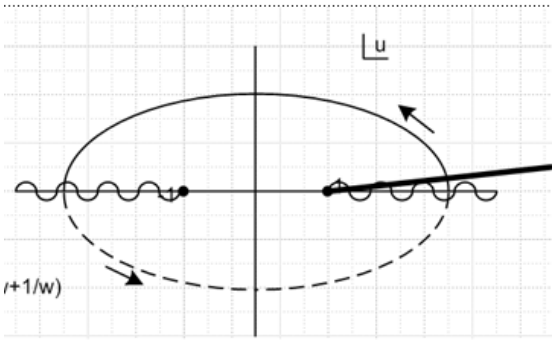


Suppose now we extend our initial strip arrow so it keeps going to the right beyond $x = 2\pi$. Here is the sequence of sheets encountered in the final plane (called w in the picture, but really u).

upper half ellipse on u-sheet 1.1
lower dotted half ellipse on sheet 1.2
upper half ellipse on sheet 2.1
lower dotted half ellipse on sheet 2.2
and so on forever

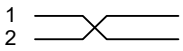
first integer refers to the w plane sheet number
and also refers to the z space strip number

And now for the mysterious page 98 figure. Let's go back to his method for the -1 1 cut



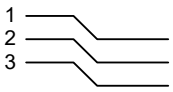
Now imagine tipping this picture back out of the plane of paper and then imagine viewing from the left side, but a little bit skewed so the two branch points -1 and 1 don't lie on top of each other as seen from this direction.

Perhaps we should do this first with a simpler cut situation. Consider $w = z^{1/2}$ which has a branch cut going off to the left. We examine it from our special left vantage point and we see this:



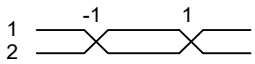
The idea is we start top left on sheet 1, drop down through the cut to sheet 2 on lower right. But that winds around and becomes bottom left which then goes up to sheet 1 again, and there are only 2 sheets.

Now compare this to the log(z) case where the branch cut is an infinite spiral. Our picture is then:



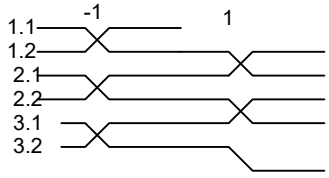
and it goes on forever.

Next, go back to the ellipse without the heavy branch cut so we have only two sheets but we have two branch points. Our picture would then be this:



and again, there are only two sheets. We go 1 to 2 at z = -1, then back 2 to 1 at z = +1 and this goes on forever, so we imagine the right side connects to the left side of this picture.

Now finally we come to the ellipse with the added heavy cut. Our picture is now this:



If we start on 1.1 and move through the left cut, we are on 1.2. But then at the right we pass through both cuts and we end up on 2.1. If we then reappear 2.1 on the left, we step down to 2.2, and so on. We just keep going down down down. On the other hand, if we start on 2.2 at z = -1, we will bump up to 2.1 just

as in our first figure above for $z^{1/2}$. But then when we hit the fat cut on the right, we bump up again to 1.2. So this picture is a combination of our 2nd and 3rd pictures. So this is the "cross section" view of the two branch cuts in the w plane for w = cos(z), and this finally explains the mystery picture on **page 98.**