Math 618

Theory of Functions of a Complex Variable II

Harold P. Boas

notes updated April 30, 2015

There are three general themes for this semester:

- convergence and approximation in the space of holomorphic functions,
- conformal mapping,
- the range of holomorphic functions.

The first item includes infinite products, the Weierstrass factorization theorem, Mittag-Leffler's theorem, normal families, and Runge's approximation theorem. The second item includes the Riemann mapping theorem and the theory of Möbius transformations. The third item includes Picard's theorems. The emphasis this semester is on techniques that are, at least in principle, constructive.

Here is a starting point for the problem of approximation.

Weierstrass approximation theorem, 1885

Theorem. *Every continuous real-valued function on the interval* $[0,1]$ *can be approximated uniformly by polynomials. In other words, the polynomials are dense in the function space* $C[0,1]$*, provided with the uniform norm.*

Question: What about approximation of continuous functions on an arbitrary compact subset of $\mathbb{R}$? The answer is still affirmative, for the Tietze extension theorem produces an extension of the function to a continuous function on $\mathbb{R}$, and in particular on an interval containing the given compact set. Hence the problem reduces to the basic Weierstrass theorem.

Now consider the complex setting. If the approximation is to be by holomorphic polynomials (polynomials in z rather than polynomials in the underlying real coordinates x and y), then there are obstructions.

Example. The function $\overline{z}$ is continuous on $\{ z \in \mathbb{C} : |z| = 1 \}$, the one-dimensional unit circle, and there is no sequence $(p_n(z))$ of polynomials such that $\max_{|z|=1} \left|\overline{z} - p_n(z)\right| \to 0$. Indeed, uniform convergence would imply convergence of the corresponding integrals, but

$$\int_{|z|=1} p_n(z)\, dz = 0, \qquad \text{whereas} \qquad \int_{|z|=1} \overline{z}\, dz = \int_{|z|=1} \frac{1}{z}\, dz = 2\pi i.$$

Example. The function $|z|$ is continuous on the closed unit disk, but there is no sequence $(p_n(z))$ of polynomials such that $\max_{|z|\le 1} \left||z| - p_n(z)\right| \to 0$. Indeed, on the open unit disk, the uniform limit of holomorphic functions must be holomorphic (by Cauchy's integral formula), yet $|z|$ is not holomorphic in the interior of the disk.

The following definition is useful for stating a positive result.

Definition. If S is a subset of $\mathbb{C}$, then a *hole* in S means a bounded component of the complement of S.

The first big theorem on complex approximation is due to the German mathematician Carl Runge (1856–1927), a student of Weierstrass, in a paper in *Acta Mathematica* in 1885 (the same year as the Weierstrass theorem). Runge is known also for the Runge–Kutta method, a numerical method for finding approximate solutions of ordinary differential equations. [The second half of the method is the German mathematician Martin Wilhelm Kutta (1867–1944).]

First version of Runge's theorem

Theorem. *If K is a compact subset of $\mathbb{C}$ (not necessarily connected) having no holes, and if f is a holomorphic function in a neighborhood of K, then f is the uniform limit on K of a sequence of polynomials.*

This theorem has surprising consequences.

Example. There exists a sequence $(p_n(z))$ of polynomials converging *pointwise* everywhere in $\mathbb{C}$, the limit being identically equal to 0 in the open upper half-plane and identically equal to 1 in the closed lower half-plane. The convergence is not uniform on compact sets!

To construct the example, apply Runge's theorem on an increasing sequence (K_n) of compact sets. Let K_n be the union of two closed rectangles, one in the closed lower half-plane with vertices at the points $-n$, n, $n - in$, and $-n - in$, and the other in the open upper half-plane with vertices at $-n + i/n$, $n + i/n$, $n + in$, and $-n + in$. Evidently these rectangles form an increasing sequence whose union is the whole plane.

The piecewise-constant function that equals 0 when $\operatorname{Im} z > 1/(2n)$ and 1 when $\operatorname{Im} z < 1/(2n)$ is holomorphic on an open set containing K_n, a compact set having no holes, so by Runge's theorem there exists a polynomial p_n that approximates this piecewise-constant function uniformly on K_n with error less than $1/n$. In other words, $|p_n(z)| < 1/n$ when z is in the top rectangle, and $|p_n(z) - 1| < 1/n$ when z is in the bottom rectangle.

Consequently, $p_n(z) \to 0$ locally uniformly in the open upper half-plane, and $p_n(z) \to 1$ locally uniformly in the closed lower half-plane. The convergence is not uniform in any neighborhood of a point on the real axis.

The following exercise was assigned in groups.

Exercise. Apply Runge's theorem to show the existence of a so-called *universal* Maclaurin series. Namely, there exists a power series $\sum_{n=0}^{\infty} a_n z^n$ with radius of convergence equal to 1 (that is, $\limsup_{n\to\infty} |a_n|^{1/n} = 1$) with the following property. For every closed disk D disjoint from the closed unit disk, every function $h(z)$ holomorphic in a neighborhood of D, and every positive ε, there exists some N (depending on both h and ε) for which $\sup_{z\in D}\left|h(z)-\sum_{n=0}^{N} a_n z^n\right| < \varepsilon$. In other words, there are partial sums of the series that approximate an arbitrary holomorphic function on a disk outside the disk of convergence. Such a Maclaurin series is "overconvergent" in a strong sense.

A lemma about approximation by special polynomials is useful for solving the exercise. A motivating observation from real analysis is the following question: Can $f : [0,1] \to \mathbb{R}$ be approximated uniformly by polynomials having no x^{17} term?

The answer is yes, for the following reason. Evidently what needs to be shown is that x^{17} itself can be approximated by polynomials having no x^{17} term. Suppose a positive ε has been specified. Take a continuous function that is identically equal to zero in a neighborhood of the origin and then rises linearly to meet the graph of x^{17}. Such a function g can be constructed to approximate x^{17} within $\varepsilon/2$. Next approximate $g(x)/x^{18}$ within $\varepsilon/2$ by a polynomial p. Then $|g(x) - x^{18}p(x)| < x^{18}\varepsilon/2 \le \varepsilon/2$ when $0 \le x \le 1$. Accordingly, $x^{18}p(x)$ approximates x^{17} within ε, and the polynomial $x^{18}p(x)$ manifestly has no x^{17} term.

An alternative argument is to observe that the map sending x to x^3 is a homeomorphism of the unit interval. The function $x^{17/3}$ is continuous, so there is a polynomial $p(x)$ that approximates $x^{17/3}$ within ε. Then $p(x^3)$ approximates x^{17} within ε, and $p(x^3)$ is a polynomial that evidently has no x^{17} term.

A similar argument proves the following generalization of the Weierstrass approximation theorem. The source is Julius Pál, Über eine Anwendung des Weierstrass-schen Satzes von der Annäherung stetiger Funktionen durch Polynome, *Tôhoku Mathematical Journal* **5** (1914) 8–9. Born in Hungary as Gyula Perl, the author later changed his name to sound more Hungarian. Around 1919 (after service in the war), he emigrated to Denmark, after which he dropped the accent mark.

Julius Pál, 1914

Theorem. *In the Weierstrass approximation theorem, the coefficients of any fixed finite number of monomials x, x^2, …, x^N [excluding x^0] can be prescribed arbitrarily.*

There is a constraint on how many monomials can be avoided. Here is a famous theorem proved independently by Herman Müntz [1884–1956] and Otto Szász [1884–1952] (at about the same time as each other).

Müntz–Szász, 1914–1916

Theorem. *A sequence of monomials* $1, x^{n_1}, x^{n_2}, \ldots$ *is dense in the continuous functions on the interval* $[0, 1]$ *if and only if the series* $\sum_j 1/n_j$ *diverges.*

The unit interval is not special: any interval $[0, b]$ will do as well. On an interval $[a, b]$ for which $a > 0$, one can even dispense with the constant function. But the theorem does not extend to an interval like $[-1, 1]$, for the even monomials satisfy the hypothesis but approximate only even functions.

Runge's theorem for general compact sets

Theorem. *If* K *is a compact subset of* $\mathbb{C}$*, and the function* f *is holomorphic on a neighborhood of* K*, then* f *is the uniform limit on* K *of a sequence of rational functions with poles in the holes of* K*. Moreover, within each hole, the position of the pole can be prescribed arbitrarily.*

(If K has no holes, then the approximation is by polynomials, as stated in the first version of the theorem.)

Sketch of the proof of Runge's theorem

There are two main ideas in the proof: (i) approximate Cauchy's integral formula using Riemann sums and (ii) push the poles to new locations. Both ideas go back to Runge.

Proof of step (i). Suppose, then, that f is holomorphic in a neighborhood of a compact set K. The first step is to produce a cycle γ that has winding number 1 around each point of K and that has trace disjoint from K and contained in the open set where f is holomorphic. Intuitively, the cycle γ should be a union of simple closed curves, one for each connected component of K. The precise meaning of the winding number of γ around z is

$$\frac{1}{2\pi i} \int_\gamma \frac{1}{w - z}\, dw.$$

If the compact set K has a simple structure (a closed Jordan region, for example), then the existence of γ is evident: just draw a curve around the boundary of K. But if K is a complicated fractal set, then some work is needed to demonstrate the existence of γ convincingly.

Suppose for the moment that γ has been constructed. By Cauchy's integral formula,

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(w)}{w - z}\, dw \qquad \text{when } z \in K.$$

Since f is uniformly continuous on the trace of γ, and $w - z$ is bounded away from 0 when $z \in K$ and $w \in \gamma$, the integral can be approximated uniformly for such z and w by Riemann sums. These sums are linear combinations of rational functions of z with first-order poles at certain points of γ (the partition points). Thus f is approximated uniformly on K by rational functions with poles off K.

(The error in approximation by a Riemann sum depends on the modulus of continuity of the function. Evidently the dependence on z is uniform when z has distance from γ bounded away from zero.)

The construction of γ can be carried out as follows.

Draw a grid of lines parallel to the coordinate axes with mesh size smaller than $1/\sqrt{2}$ times the distance from K to the boundary of the open set where the function f is holomorphic. Collect all the closed squares of the grid that intersect K, and orient the boundaries counterclockwise. The claim is that γ can be taken to consist of a subset of the oriented edges of these squares: namely, those edges that do not intersect K.

Observe that if z is a point of K not on any of the gridlines, then the sum of the Cauchy integrals of f over the boundaries of all the squares that intersect K equals $f(z)$, for z is inside exactly one of these squares. On the other hand, if an edge of a square intersects K, then there is another square sharing that edge and intersecting K, so the integrals over edges that intersect K cancel out. (There is a special case when the edge intersects K only at an endpoint: then there are four squares touching at the point, and again the integrals cancel out.)

Hence the Cauchy integral over the proposed γ does equal $f(z)$ when z is not on a gridline. By continuity, the integral still equals $f(z)$ even when z is on a gridline. Although γ will be a union of closed curves, that information is not really needed. Viewed merely as a union of edges, the arc γ still gives an integral representing the function, and that integral can be approximated by Riemann sums. □

Proof of pole pushing. The idea behind pole pushing is shown by the following calculation, in which z_0 and z_1 are two arbitrary distinct complex numbers:

$$\frac{1}{z-z_0} = \frac{1}{(z-z_1)-(z_0-z_1)} = \frac{1}{z-z_1}\cdot\frac{1}{1-\dfrac{z_0-z_1}{z-z_1}} = \frac{1}{z-z_1}\sum_{n=0}^{\infty}\left(\frac{z_0-z_1}{z-z_1}\right)^n,$$

with convergence when $|z_0 - z_1| < |z - z_1|$. Thus the rational function $1/(z - z_0)$ can be approximated by rational functions in $1/(z - z_1)$ when z is farther away from z_1 than z_0 is. The convergence is even uniform on sets whose distance from z_1 is strictly greater than the distance from z_1 to z_0.

In particular, if z_0 is in a hole of the compact set K, then a rational function with pole at z_0 can be approximated uniformly on K by rational functions with pole at an arbitrary point of the hole at slightly less than half the distance of z_0 to K. Iterating this observation shows that the pole can be pushed to an arbitrary location inside the hole.

What if z_0 is in the unbounded component of the complement of K? By the preceding reasoning, the pole can be pushed to an arbitrary location in the unbounded component. Suppose the pole has been pushed to a point z_1 outside a disk so large that the disk contains the compact set K. Then compute as follows:

$$\frac{1}{z-z_1} = -\frac{1}{z_1}\cdot\frac{1}{1-\dfrac{z}{z_1}} = -\frac{1}{z_1}\sum_{n=0}^{\infty}\left(\frac{z}{z_1}\right)^n,$$

with convergence when $|z| < |z_1|$, and in particular on K. The partial sums of this series are polynomials. In other words, the pole in the unbounded component of the complement of K can be pushed to infinity. □

Mergelyan's theorem

The hypothesis in Runge's theorem is unnatural: athough the approximation takes place only *on* the compact set, the function being approximated is assumed to be holomorphic in a *neighborhood* of the set. The following improvement of Runge's theorem is due to the Armenian mathematician Sergey Nikitovich Mergelyan (1928–2008). The main idea in the proof (not presented here) is to extend the function in a smooth way to a neighborhood of the compact set and to correct the extended function by solving a $\overline{\partial}$-problem. Then use Runge's theorem to approximate the extended function. The difficulty—which Mergelyan overcame—is to control the size of the correction term.

Mergelyan's theorems, 1951–1952

Theorem. *If K is a compact set with no holes, and f is a continuous function on K that is holomorphic on the interior of K, then there is a sequence (p_n) of polynomials such that* $\max_{z\in K} \left|f(z) - p_n(z)\right| \to 0$.

Theorem. *If K is a compact set with finitely many holes, and f is a continuous function on K that is holomorphic on the interior of K, then f is the uniform limit on K of a sequence of rational functions with poles in the holes.*

More generally, the conclusion holds in the case of infinitely many holes if the diameters of the holes are bounded away from zero.

An example illustrating the second case is the boundary of a rectangle together with a sequence of vertical lines condensing on one side. A Swiss cheese (the subject of a homework assignment) is a counterexample showing that some restriction on the diameters is needed. Vitushkin found a necessary and sufficient condition on the compact set (in terms of capacity) for the conclusion to hold.

Mittag-Leffler's theorem

Magnus Gustaf (Gösta) Mittag-Leffler (1846–1927), founder of *Acta Mathematica* (1882) and namesake of the Mittag-Leffler Institute in the suburbs of Stockholm, was a Swedish mathematician who attended some of Weierstrass's lectures and subsequently generalized the theorem of Weierstrass (to be studied later on) about constructing functions with prescribed zeroes. (His father's name was Leffler, and his mother's name was Mittag; he joined the names himself as an adult, apparently because of his interest in women's rights. His influence made it possible for Sonya Kovalevsky [1850–1891] to be appointed professor of mathematics in Stockholm.)

The theorem of Weierstrass says that there exists an entire function with prescribed zeroes (subject to the zeroes not accumulating). A consequence is the existence of meromorphic functions with prescribed poles (just take the reciprocal of a function with prescribed zeroes). Mittag-Leffler's main contribution was to construct functions not just with prescribed poles but with prescribed singular parts (known as principal parts). Here is one version of the theorem.

Mittag-Leffler's theorem, 1876–1884

Suppose G is an open subset of $\mathbb{C}$ and E is a discrete subset of G. Suppose given, for each point b in E, a holomorphic function p_b on $G \setminus \{b\}$. Then there exists a holomorphic function f on $G \setminus E$ such that for each point b in E, the function $f - p_b$ has a removable singularity at b.

In the statement of the theorem, a discrete set means a set that has no accumulation point in G.

The typical case is that p_b is a finite linear combination of powers of $1/(z-b)$. The theorem guarantees that there exists a function with prescribed isolated singularities and prescribed principal parts at the singularities. A corollary of the theorem is that there exists a meromorphic function with prescribed principal parts.

The theorem is a bit more general, allowing some essential singularities to be prescribed. For instance, the theorem produces a meromorphic function in the plane sharing the whole singular part of $\sin((z-k)^{-1})$ at each integer k. But the theorem does not allow the singular part to be prescribed completely arbitrarily: the Laurent series on a punctured neighborhood of b needs to converge in the whole punctured plane, or at least to admit an analytic continuation to $G \setminus \{b\}$.

Proof. The standard modern proof uses Runge's theorem, although Runge's work actually was motivated by Mittag-Leffler's.

The first step is to exhaust G by a sequence (K_n) of compact sets such that each set is contained in the interior of the next, and no K_n has unnecessary holes. In other words, each hole in the compact set contains a hole in G. This construction will be needed again later in the discussion of normal families and the proof of the Riemann mapping theorem. The construction is considered in the textbook back in Chapter 9 (see Lemma 9.2.0) and then again in Lemma 12.5.

The first try at constructing such sets is $\{\, z \in G \,:\, \operatorname{dist}(z, \partial G) \geq 1/n \,\}$. The difficulty is that this definition could produce an unbounded set if G is unbounded. So intersect the initial set with the closed disk of center 0 and radius n. The resulting set is compact and has no unnecessary holes. (The set might be empty for small values of n.)

The main idea in the proof of the theorem is to build the function as an infinite series. The first try is simply to add together all the functions p_b as b runs over the countable set E. This method certainly works when the set E is finite. In general, however, there will be an infinite series, and the series need not converge. The idea is to add convergence factors that are holomorphic on all of G and hence do not affect the principal parts.

There are only finitely many singular points inside the compact set K_1. Add them together and call that sum f_1. In general, let f_n denote the sum of the functions p_b for b in the set $K_n \setminus K_{n-1}$. Notice that when $n > 1$, the function f_n is holomorphic on K_{n-1}. Use Runge's theorem to find a rational function g_n with poles in $\mathbb{C} \setminus G$ such that $\left| f_n(z) - g_n(z) \right| < 1/2^n$ when $z \in K_{n-1}$.

The required function is $f_1 + \sum_{n=2}^{\infty}(f_n - g_n)$. Indeed, if a compact set K is fixed, then the terms in the tail of the series eventually are holomorphic on K, and the tail of the series converges uniformly on K. Hence the series represents a holomorphic function on $G \setminus E$. Moreover, at a particular point b in E, all the terms of the sum are holomorphic in a neighborhood of b except for one, which has the specified principal part. □

The folllowing theorem is one of the results in Hadamard's 1892 doctoral thesis, "Essai sur l'étude des fonctions, données par leur développement de Taylor."

Hadamard's gap theorem

Suppose (n_k) is a geometrically increasing of exponents. If the gap series $\sum_{k=1}^{\infty} a_k z^{n_k}$ has radius of convergence equal to 1, then the unit circle is a natural boundary.

The hypothesis about the exponents is that there exists a number λ strictly greater than 1 such that $n_{k+1}/n_k \geq \lambda$ for every k. The conclusion means that every boundary point is a singular point: there is no neighborhood of a boundary point to which the function defined by the power series extends holomorphically. Typical examples of gap sequences are (2^k) and $(k!)$.

Elementary considerations suffice to understand the series $\sum_{n=1}^{\infty} z^{n!}$. The radius of convergence is equal to 1, for the series diverges when $z = 1$ and is a subseries of the geometric series (hence converges absolutely when $|z| < 1$). Moreover, this series evidently is unbounded along the radius from 0 to 1, where all the terms are positive. Rotating by a factor $\exp(2\pi i(p/q))$ for natural numbers p and q shows that the series is unbounded along a dense set of radii, so the series cannot continue to a neighborhood of any boundary point. Similar reasoning shows that the series $\sum_{n=1}^{\infty} z^{n!}/2^n$ has a *derivative* that is everywhere noncontinuable, so $\sum_{n=1}^{\infty} z^{n!}/2^n$ itself is everywhere noncontinuable, even though this series converges absolutely and uniformly on the boundary.

A more surprising example is $\sum_{n=1}^{\infty} z^{2^n}/n!$. This series certainly converges when $|z| \leq 1$, since $\sum_{n=1}^{\infty} 1/n!$ converges. On the other hand, the series certainly diverges when $z = 1 + \varepsilon$ and $\varepsilon > 0$, for $n! \leq n^n$, so $(1+\varepsilon)^{2^n}/n! \geq \exp(2^n \log(1+\varepsilon) - n \log n) \to \infty$. Therefore the series has radius of convergence equal to 1. Evidently the series converges absolutely and uniformly on the boundary, where $|z| = 1$. Moreover, the differentiated series converges absolutely and uniformly on the boundary, since $\sum_{n=1}^{\infty} 2^n/n!$ converges (by the ratio test, say). For a similar reason, the kth derivative of the series converges on the boundary for an arbitary value of k. Accordingly, the series represents a class C^∞ function (infinitely differentiable function in the sense of real partial derivatives) on the closed unit disk. By Hadamard's gap theorem, however, the series cannot be extended holomorphically to a neighborhood of any boundary point.

The question of existence of holomorphic functions infinitely differentiable on the closed disk but having the unit circle as natural boundary was in the air at the end of the nineteenth century. A high-level modern way to see that such functions exist is to consider the Riemann mapping function from the unit disk onto a domain bounded by a class C^∞ but nowhere real-analytic curve. The Riemann mapping function is known to extend to be a class C^∞ diffeomorphism on the closure, and the function cannot be holomorphic across a boundary point, else the image

curve would be somewhere real analytic. This technique shows that there is even an *injective* holomorphic function that is smooth on the closure and nowhere continuable.

Mordell's 1927 proof of Hadamard's gap theorem. Louis Mordell (1888–1972) is best known for his work in number theory. The so-called Mordell Conjecture was proved in 1983 by Gerd Faltings, earning Faltings a Fields Medal. The reference for Mordell's proof of Hadamard's theorem is L. J. Mordell, On Power Series with the Circle of Convergence as a Line of Essential Singularities, *Journal of the London Mathematical Society* **2** (1927) 146–148; `doi:10.1112/jlms/s1-2.3.146`.

It suffices to show that the point 1 is a singular point of every gap series satisfying the hypotheses, for if $f(z)$ is a gap series, then so is $f(e^{i\varphi}z)$ for an arbitrary angle φ. Hence if 1 is a singular point of every gap series, then so is the point $e^{i\varphi}$.

Suppose, seeking a contradiction, that 1 is not a singular point of f. Then there is a positive number δ such that f extends to be holomorphic on $D(0,1) \cup D(1,\delta)$.

By hypothesis, the number λ is greater than 1, so there is a natural number p large enough that $\lambda > (p+1)/p$, which implies that $pn_{k+1} > (p+1)n_k$. Let $g(w)$ denote $\frac{1}{2}(1+w)w^p$. Then the smallest power of w in the polynomial expansion of $g(w)^{n_{k+1}}$ exceeds the largest power of w in the polynomial expansion of $g(w)^{n_k}$.

If $\sum_{k=1}^{\infty} a_k z^{n_k}$ is the series expansion of $f(z)$, then the partial sums of the series $\sum_{k=1}^{\infty} a_k (g(w))^{n_k}$ form a subsequence of the partial sums of the Maclaurin series of $(f \circ g)(w)$. Therefore the former series converges whenever the latter series converges. The goal is to show the existence of a point w such that the series for $(f \circ g)(w)$ converges, yet $|g(w)| > 1$. It then follows that the radius of convergence of the original gap series defining f is larger than 1, contrary to the hypothesis.

If $|w| \leq 1$ but $w \neq 1$, then $|g(w)| < 1$. Moreover $g(1) = 1$. Therefore g maps $\overline{D}(0,1)$, the closed unit disk, onto a compact set contained in the open region where f is holomorphic. Consequently, there is a small positive ε such that g maps $\overline{D}(0, 1+\varepsilon)$ into the open region where f is holomorphic. In other words, the composite function $f \circ g$ is holomorphic on $\overline{D}(0, 1+\varepsilon)$, and the radius of convergence of the corresponding Maclaurin series exceeds $1+\varepsilon$. Therefore $\sum_{k=1}^{\infty} a_k (g(1+\varepsilon))^{n_k}$ converges. Since $g(1+\varepsilon)$ is a real number larger than 1, the original gap series has radius of convergence larger than 1, contrary to hypothesis. □

Infinite series and products

Tools for constructing holomorphic functions are series, products, and integrals. Taylor series and Laurent series were covered in Math 617. The next topic is the notion of an infinite product.

Infinite products

The immediate goal is to develop enough theory to make sense of formulas like the following one proved in Chapter 6 of the textbook:

Euler's product formula for the sine function

$$\sin(\pi z) = \pi z \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n^2}\right)$$

What should it mean to say that $\prod_{n=1}^{\infty} b_n$ converges? Apparently, the natural definition would be $\lim_{N\to\infty} \prod_{n=1}^{N} b_n$. That definition will not do, however, because if $b_1 = 0$, then the limit of partial products exists (and equals 0) for completely arbitrary values of the other factors. But the *existence* of a limit ought not to depend on the value of the first term (or on the values of finitely many terms).

One could insist on considering products having no factors equal to 0, but the application to holomorphic functions needs precisely the case in which some factors are equal to 0. On the other hand, if there were infinitely many factors equal to 0, then the limit of partial products could only be 0, and the limit would be independent of the values of the subsequence of nonzero terms.

The standard definition of convergence of infinite products requires that there be only finitely many factors equal to 0 and that the limit of the partial products of the nonzero factors exists and that this limit is not equal to 0. If the limit of nonzero factors exists and equals 0, then the product is said to diverge to 0.

One reason for excluding 0 as a limit is that one would like to pass back and forth between infinite series and infinite products by using the exponential and logarithm functions. Another reason is to preserve the property that a product is equal to zero if and only if some factor is equal to zero.

Example. The infinite product $\prod_{n=1}^{\infty} 1/n$ diverges to 0. The corresponding series $\sum_{n=1}^{\infty} \log(1/n)$ (with the principal branch of the logarithm) diverges to $-\infty$.

Example. The infinite product $\prod_{n=1}^{\infty} \left(1 + \frac{1}{n}\right)$ diverges. Indeed, the partial product $\prod_{n=1}^{k} \left(1 + \frac{1}{n}\right)$ telescopes to the value $k + 1$, which does not approach a finite value.

Example. The product $\prod_{n=1}^{\infty} \left(1 - \frac{1}{n^2}\right)$ converges to 0 for the following reason.

Notice that 0 is an allowed value for the limit, but only if the nonzero factors converge to a nonzero limit. In this example, the first term equals 0. Moreover,

$$\prod_{n=2}^{k} \left(1 - \frac{1}{n^2}\right) = \prod_{n=2}^{k} \frac{(n-1)(n+1)}{n^2}.$$

The product telescopes: each natural number appears twice in the numerator and twice in the denominator, except for special terms at the beginning and the end. The product equals

$$\frac{1}{2} \cdot \frac{k+1}{k},$$

which has limit $1/2$ when $k \to \infty$. Therefore the original infinite product converges to 0.

For a product of nonzero terms b_n to converge to a nonzero limit L, a necessary condition is that $b_n \to 1$. Indeed, if a positive ε less than $|L|$ is specified, then there is a natural number N such that $\left|L - \prod_{n=1}^{k} b_n\right| < \varepsilon$ when $k \geq N$. Write p_k for $\prod_{n=1}^{k} b_n$. Then

$$1 - b_k = 1 - p_k/p_{k-1} = \frac{(p_{k-1} - L) - (p_k - L)}{(p_{k-1} - L) + L},$$

so $|1 - b_k| < 2\varepsilon/(|L| - \varepsilon)$ when $k > N$.

Accordingly, the general term of an infinite product usually is written in the form $1 + a_n$. A necessary condition for convergence of an infinite product is then that $a_n \to 0$. What about sufficient conditions?

Proposition. *If no term* $(1 + a_n)$ *equals* 0*, then the infinite product* $\prod_{n=1}^{\infty}(1 + a_n)$ *and the infinite series* $\sum_{n=1}^{\infty} \log(1 + a_n)$ *(with the principal branch of the logarithm) either both converge or both diverge.*

Example. Consider the product $\prod_{n=1}^{\infty}\left(1 - \frac{z^2}{n^2}\right)$.

If $|z| \leq R$, say, then $|z^2/n^2| \leq R^2/n^2$, and $\sum_{n=1}^{\infty} R^2/n^2$ converges. Therefore the series $\sum_{n=1}^{\infty} |z^2/n^2|$ converges uniformly for z in a compact set by the Weierstrass M-test, so the series $\sum_{n=1}^{\infty} \log\left(1 - \frac{z^2}{n^2}\right)$ converges uniformly on compact sets, and the original infinite product converges uniformly on compact sets. (Exponentiation is a continuous operation, so if the partial sums of the series are close to a limiting value, then the corresponding partial products obtained by exponentiating are closed to the exponential of the limiting value.)

Notice that the discussion so far does not prove Euler's product for the sine function. The infinite product does converge, and multiplying by z gives an entire function having the same zeroes as the sine function. Consequently, the two functions have a ratio that is a zero-free entire function, hence of the form $e^{g(z)}$ for some entire function g. More work is needed to show that g is the 0 function.

Proof of the Proposition. If the partial sums of the infinite series converge, then the exponentials of the partial sums converge; hence the partial products converge (by continuity of the exponential function). The converse argument is more delicate. If the partial products have a limit, then so does the sequence of logarithms of partial products, but the logarithm of a product is not necessarily equal to the sum of the logarithms for a fixed branch of the logarithm.

Suppose that the partial products do converge (to a nonzero limit). Since

$$\log(1 + a_n) = \log|1 + a_n| + i \arg(1 + a_n),$$

what needs to be checked is that both $\sum_{n=1}^{\infty} \log|1 + a_n|$ and $\sum_{n=1}^{\infty} \arg(1 + a_n)$ converge. If the partial products converge, then continuity of the modulus implies that the partial products of the moduli converge, and to a positive real number. Continuity of the real logarithm function implies that $\sum_{n=1}^{\infty} \log|1 + a_n|$ converges.

Now write $1+a_n = |1+a_n|\, e^{i\theta_n}$, where $\theta_n = \arg(1+a_n)$. Since the partial products $\prod_{n=1}^{N}(1+a_n)$ and $\prod_{n=1}^{N}|1+a_n|$ both converge to nonzero limits, it follows that the partial products $\prod_{n=1}^{N} e^{i\theta_n}$ converge, say to some $e^{i\varphi}$. Consequently, there is a sequence of integers m_N such that $\varphi + 2\pi m_N - \sum_{n=1}^{N}\theta_n \to 0$ as $N \to \infty$. But $a_n \to 0$ as $n \to \infty$, so $\theta_n \to 0$, and the integer m_N must eventually stabilize at a constant value (since eventually m_N and m_{N+1} differ by less than 1). Consequently, the series $\sum_{n=1}^{\infty}\theta_n$ converges to some value $\varphi + 2\pi m$. □

A simple sufficient condition for convergence of an infinite product $\prod_{n=1}^{\infty}(1+a_n)$ is that $\sum_{n=1}^{\infty}|a_n|$ converges. The intuitive idea is that $\log(1+a_n) \approx a_n$ when a_n is close to 0, so the hypothesis implies absolute convergence of the series of logarithms, hence convergence of the infinite product. Indeed, since the condition implies that $a_n \to 0$, there is no loss of generality in supposing that $|a_n| < 1/2$, say. Now integrating the geometric series gives a series for the logarithm:

$$\log(1+z) = z - \tfrac{1}{2}z^2 + \tfrac{1}{3}z^3 - \cdots \qquad \text{when } |z| < 1,$$

so $|\log(1+z)| \le |z| + |z|^2 + |z|^3 + \cdots = |z|/(1-|z|)$. Hence $|\log(1+a_n)| \le 2|a_n|$.

This simple sufficient condition for convergence of an infinite product is not necessary.

Example. Consider $\prod_{n=1}^{\infty}\left(1 + \frac{(-1)^n}{n^{2/3}}\right)$.

Now $\log(1+a_n) = a_n - \frac{1}{2}a_n^2 + \cdots$, so

$$\log\left(1 + \frac{(-1)^n}{n^{2/3}}\right) = \frac{(-1)^n}{n^{2/3}} - \frac{1}{2}\cdot\frac{1}{n^{4/3}} + \cdots = \frac{(-1)^n}{n^{2/3}} + O(1/n^{4/3}).$$

The alternating series $\sum_{n=1}^{\infty}(-1)^n/n^{2/3}$ converges, and the remainder series converges absolutely. Hence the infinite product converges (conditionally).

The Weierstrass factorization theorem

Suppose (z_n) is a discrete set of points in $\mathbb{C}$ (no accumulation point), and (m_n) is a sequence of natural numbers. The goal is to construct an entire function having a zero of order m_n at z_n for every n.

The first try is an infinite product of the form $\prod_{n=1}^{\infty}(z-z_n)^{m_n}$, but this attempt fails. Since the factors do not tend to 1, the product diverges.

The second try is an infinite product of the form $\prod_{n=1}^{\infty}\left(1-\frac{z}{z_n}\right)^{m_n}$. This attempt succeeds if z_n tends to infinity fast enough to make the product converge. This method handles, for instance, the construction of a function with simple zeroes at the squares of the natural numbers, since $\sum_{n=1}^{\infty} z/n^2$ converges for every z (and converges uniformly on compact sets, so the limit function is holomorphic). But the method fails to construct a function with a simple zero at each natural number, since $\sum_{n=1}^{\infty}\left(1-\frac{z}{n}\right)$ diverges when $z \neq 0$.

The third try, which succeeds, is to introduce nonvanishing convergence factors. For instance,

$$\prod_{n=1}^{\infty}\left(1-\frac{z}{n}\right)\exp\left(\frac{z}{n}\right)$$

does converge, uniformly on compact sets, since

$$\log\left(1-\frac{z}{n}\right)+\frac{z}{n}=-\frac{1}{2}\cdot\frac{z^2}{n^2}+\cdots=z^2\cdot O(1/n^2).$$

Convergence factors were introduced by Weierstrass and caused a sensation at the time.

Existence of an entire function with prescribed zeroes

Theorem. *Let (z_n) be a sequence of nonzero complex numbers, possibly with repetitions, but with no limit point. There exists an entire function with zeroes precisely at the points of the sequence, the order of each zero being equal to the multiplicity of the point in the sequence.*

The following lemma is useful in establishing the general result.

Lemma 1. *If $|z|\le 1/2$, then the principal branch of the logarithm satisfies the following estimates:*

$$\begin{aligned}|z+\log(1-z)|&\le|z|^2,\\ \left|\left(z+\frac{z^2}{2}\right)+\log(1-z)\right|&\le|z|^3,\\ \left|\left(z+\frac{z^2}{2}+\frac{z^3}{3}\right)+\log(1-z)\right|&\le|z|^4,\end{aligned}$$

and so on.

Proof. By Taylor's theorem,

$$\log(1-z)=-z-\frac{z^2}{2}-\frac{z^3}{3}-\cdots\qquad\text{when } |z|<1.$$

Hence

$$\left|\sum_{n=1}^{k}\frac{z^n}{n}+\log(1-z)\right|=\left|\sum_{n=k+1}^{\infty}\frac{z^n}{n}\right|\le\frac{1}{k+1}\sum_{n=k+1}^{\infty}|z|^n=\frac{1}{k+1}\cdot\frac{|z|^{k+1}}{1-|z|}.$$

Now

$$\frac{1}{k+1}\cdot\frac{1}{1-|z|}\le\frac{2}{k+1}\le 1\qquad\text{when } |z|\le 1/2.$$

Consequently, the required estimate holds. □

A corollary is that

$$\left|1-(1-z)e^{z+\frac{1}{2}z^2+\frac{1}{3}z^3+\cdots+\frac{1}{n}z^n}\right|\le|z|^{n+1}\qquad\text{when } |z|\le 1 \text{ and } n\ge 1,$$

which is Lemma 13.0 in the textbook (page 247). Indeed, the preceding considerations show that the entire function inside the absolute value on the left-hand side is divisible by z^{n+1}. Moreoever,

the derivative of that function is easily seen by explicit computation to have nonnegative coefficients. Hence the modulus of the function divided by z^{n+1} is maximized in the closed unit disk when $z = 1$, where the value is equal to 1.

The expression

$$(1 - z) \exp\left(z + \frac{z^2}{2} + \frac{z^3}{3} + \cdots + \frac{z^n}{n}\right)$$

is known as a Weierstrass elementary factor, denoted $E_n(z)$.

Construction of the convergent Weierstrass product. Behold:

$$\prod_{n=1}^{\infty} \left(1 - \frac{z}{z_n}\right) \exp\left(\frac{z}{z_n} + \frac{1}{2}\left(\frac{z}{z_n}\right)^2 + \cdots + \frac{1}{n}\left(\frac{z}{z_n}\right)^n\right).$$

The claim is that this product converges uniformly on compact sets, in which case the limit is an entire function that evidently has the required zeroes.

Indeed, $|z_n| \to \infty$ by hypothesis, so if z is confined to a compact set, then $|z/z_n|$ is bounded uniformly by $1/2$ when n is sufficiently large. The lemma then implies that the logarithm of the general term of the product has modulus bounded by $1/2^{n+1}$. Since these bounds are the terms of a convergent infinite series, the infinite product converges uniformly on compact sets by the Weierstrass M-test. □

Remark. The proof reveals that the sum in the exponent could be stopped at the term with power $n - 1$ or $n - 17$ instead of the term with power n. This refinement is not especially interesting. The interesting question is whether the sum in the exponent can be stopped at a power that is independent of n. That question is answered by the Hadamard factorization theorem (which we did not have time to cover).

To construct an entire function whose zero set includes the origin, simply multiply the infinite product by a suitable power of z.

Weierstrass factorization theorem for entire functions, 1876

Every entire function $f(z)$ can be expressed in the following form:

$$z^k e^{g(z)} \prod_{n=1}^{\infty} \left(1 - \frac{z}{z_n}\right) \exp\left(\frac{z}{z_n} + \frac{1}{2}\left(\frac{z}{z_n}\right)^2 + \cdots + \frac{1}{m_n}\left(\frac{z}{z_n}\right)^{m_n}\right),$$

where k (possibly 0) is the order of the zero of f at the origin, g is some other entire function, the sequence (z_n) is the list of nonzero zeroes of f (repeated according to multiplicity), and (m_n) is a suitable sequence of natural numbers (it will do to take $m_n = n$).

Corollary. *Every function that is meromorphic in the whole plane can be written as the quotient of two entire functions.*

The corollary appeared as problem 9 on the August 2012 qualifying exam!

Proof. Dividing $f(z)$ by an infinite product with the same zeroes as f (of the same orders) produces a zero-free entire function. Such a function has a holomorphic logarithm, that is, can be written in the form $\exp g$.

Notice that the representation is not unique, for the sequence (m_n) can be replaced by a larger one. Changing this sequence of natural numbers will change the function g in the factorization. □

The extension of Weierstrass's theorem to subsets of the complex plane was done later by other researchers (Picard and Mittag-Leffler). Here is the statement and the proof.

Weierstrass theorem for general regions

Theorem. *Suppose G is an open subset of $\mathbb{C}$, and (z_n) is a sequence of points (possibly with repetitions) in G having no limit point inside G. Then there is a holomorphic function on G having zeroes precisely at the points (z_n) (with order corresponding to the multiplicity of the point in the sequence).*

Example. On an arbitrary open set G, there is a holomorphic function that cannot be analytically continued across any boundary point.

Indeed, take a sequence in G that has every boundary point as an accumulation point, and use the theorem to construct a holomorphic function with zeroes at the points of the sequence. This function cannot extend to a neighborhood of any boundary point, for the zeroes of the function accumulate inside that neighborhood, which would contradict the identity principle.

To build the indicated sequence, take a dense sequence (a_n) in the boundary of G. Create a new sequence (b_n) that contains each a_k infinitely often. For instance, the sequence a_1, a_1, a_2, a_1, a_2, a_3, a_1, a_2, a_3, a_4, … will do. Then take z_n to be a point in G at distance less than $1/n$ from the point b_n.

Proof of the general Weierstrass theorem. The proof in the textbook is the standard modern proof that throws one point to infinity. Here instead is a proof that works directly on the original open set.

The first idea is to split the sequence of points into two parts, depending on whether the points are close to the boundary of G or close to ∞. Namely, view G as the union of the following two disjoint sets:

$$S := \{\, z \in G : |z| \operatorname{dist}(z, \partial G) \geq 1 \,\} \qquad \text{and} \qquad T := \{\, z \in G : |z| \operatorname{dist}(z, \partial G) < 1 \,\}.$$

Observe that those points of the sequence (z_n) that lie in the set S must either be finitely many or tend to ∞. For in the contrary case, there would be infinitely many of these points confined to a bounded set. The definition of S implies that these points would be bounded away from ∂G. The Bolzano–Weierstrass theorem then implies that these points would have a limit point inside G, contrary to hypothesis.

Consequently, the first version of the Weierstrass theorem implies that there is an *entire* function with zeroes precisely at the points of the sequence that lie in S. That entire function of course is holomorphic on G.

All that remains is to construct a holomorphic function on G that has zeroes at the points of the sequence lying in T. The product of this function with the entire function from the preceding paragraph solves the problem.

Accordingly, let (a_n) denote the subsequence of points in the original sequence that happen to lie in T. If there are infinitely many such terms of the sequence, then $\operatorname{dist}(a_n, \partial G)$ must approach 0. If not, there would be a (further) subsequence bounded away from ∂G. The definition of T implies that the subsequence would be bounded, hence would have a limit point inside G, contrary to hypothesis.

Now let b_n be a point of ∂G such that $|a_n - b_n| = \operatorname{dist}(a_n, \partial G)$. Let $E_n(z)$ denote the Weierstrass *elementary factor*:

$$E_n(z) = (1 - z) \exp\left(z + \tfrac{1}{2}z^2 + \cdots + \tfrac{1}{n}z^n\right).$$

The claim is that the following product provides the required holomorphic function:

$$\prod_{n=1}^{\infty} E_n\left(\frac{a_n - b_n}{z - b_n}\right).$$

The argument of E_n takes the value 1 precisely when $z = a_n$, so the nth factor in the infinite product vanishes precisely when $z = a_n$. The singularity of the argument is on the boundary of G, so each factor is holomorphic inside G.

What remains to show is that the infinite product converges uniformly on compact subsets of G. When z is confined to a compact set, then z necessarily is bounded away from ∂G, so the denominator $z - b_n$ is bounded away from 0. On the other hand, $a_n - b_n \to 0$ by construction. Consequently, $|a_n - b_n|/|z - b_n| < 1/2$ for sufficiently large n, so the infinite product converges uniformly on compact sets. □

The metric on $C(K)$

Now I jump back to Chapter 9 in the textbook.

The immediate goal is to define a metric on the space of holomorphic functions on a region G, to understand convergence in this metric, and to characterize compact sets with respect to this metric. The first step is to define a metric on the continuous functions on a compact set.

If K is a compact subset of $\mathbb{C}$, then there is a standard norm on the space $C(K)$ of continuous functions on K: namely, the supremum norm. That is,

$$\|f\|_K = \max\{\, |f(z)| : z \in K \,\}.$$

The norm induces a metric (the distance between f and g is $\|f - g\|_K$), and convergence with respect to this metric is *uniform convergence* (hence the metric is called the *uniform metric*): to say that $f_n \to f$ uniformly on K is precisely the statement that $\|f_n - f\|_K \to 0$.

You know from real analysis that in the metric space $\mathbb{C}$ (more generally in Euclidean space of arbitrary dimension), the compact sets are precisely the sets that are simultaneously closed and bounded (Heine–Borel theorem, named after Eduard Heine [1821–1881] and Émile Borel [1871–1956]).[1]

The corresponding equivalence does not hold in the space $C(K)$. If K is the closed unit disk, then the sequence (z^n) has no subsequence converging to an element of $C(K)$ (hence is not compact), for the sequence converges pointwise to 0 on the open unit disk and is constantly equal to 1 at the point 1; yet the sequence is bounded (being a subset of the closed unit ball of $C(K)$) and closed (since the sequence has no limit point).

You may know that the generalization of the Heine–Borel theorem to general metric spaces says that a subset of a metric space is compact if and only if the set is simultaneously complete (Cauchy sequences converge) and totally bounded (the set can be covered by a finite number of arbitrarily small balls).

The characterization of compact subsets of the metric space $C(K)$ is a famous theorem from the late nineteenth century.

Arzelà–Ascoli theorem

Theorem. *A subset of $C(K)$ is compact if and only if the set is simultaneously closed, pointwise bounded, and equicontinuous.*

More precisely, the hypotheses mean that the set S of functions has the properties that (a) for each point z in K there exists a constant M such that $|f(z)| \leq M$ for every function f in S (the value of M possibly depending on the point z but not depending on the function f), and (b) for every point z in K and for every positive ε there is a positive δ such that $|f(z) - f(w)| < \varepsilon$ whenever $f \in S$ and $|z - w| < \delta$ (the value of δ possibly depending on the point z but not depending on the function f).

Exercise. On a compact set, equicontinuity at every point is equivalent to uniform equicontinuity: the value of δ actually can be taken to be independent both of the point z and of the function f. (The proof is analogous to the proof that a continuous function on a compact set is automatically uniformly continuous.)

Although pointwise boundedness on a compact set is not equivalent to uniform boundedness (think of a sequence of triangle functions with increasingly steep peaks condensing at the origin), the proof of the theorem yields that in the presence of equicontinuity, pointwise boundedness does imply uniform boundedness on compact sets. The theorem can be generalized to continuous functions on a compact Hausdorff space.

The theorem is due to the Italian mathematician Giulio Ascoli (1843–1896) in a paper of 1884 in which he introduced the notion of equicontinuity. It seems that Cesare Arzelà (1847–1912) actually published the idea of equicontinuity a year or so earlier than Ascoli did. Subsequently, in

[1] Apparently, this theorem was in the air in the second half of the nineteenth century, Heine being only one of several mathematicians who used the idea; Borel seems to have made the first explicit statement.

1889 and in 1896, Arzelà (notice the grave accent) clarified, extended, and applied the theorem. So technically it may be Ascoli's theorem, but Arzelà's work popularized the theorem, and Arzelà even had the key concept earlier.

Proof. In a metric space, compactness is the same as sequential compactness. Accordingly, what needs to be shown for the sufficiency of the conditions is that if (f_n) is a pointwise bounded, equicontinuous sequence in $C(K)$, then there is a subsequence that converges uniformly on K (necessarily to an element of $C(K)$, since the uniform limit of continuous functions is continuous). An equivalent statement is that there a subsequence for which Cauchy's criterion for convergence holds uniformly.

Take a dense sequence (z_n) in K. (If K is a finite set, then the proof is easier. To construct the sequence for an infinite set K, cover the set with a mesh of size $1/k$, pick a point of K in each cell that intersects K, increase k, and iterate to produce the dense sequence. For a nice set K, say the closure of an open set, it suffices to take the points of K having both coordinates rational.)

The sequence of complex numbers $(f_n(z_1))$ is bounded (by one of the hypotheses), so the Bolzano–Weierstrass theorem provides an initial increasing sequence $(j_1(n))$ of natural numbers such that the sequence $(f_{j_1(n)}(z_1))$ converges. There is a further subsequence $(j_2(n))$ such that $(f_{j_2(n)}(z_2))$ converges (and $(f_{j_2(n)}(z_1))$ still converges, being a subsequence of $(f_{j_1(n)}(z_1))$). Iterate the procedure. Then the diagonal sequence $(f_{j_n(n)})$ converges at the point z_k for every k. Call this sequence (g_n) for short.

The uniform equicontinuity will force this sequence (g_n) to converge everywhere (and uniformly). Indeed, if a positive ε is specified, then there is a positive δ such that if $|z - w| < \delta$, then $|f(z) - f(w)| < \varepsilon$ for every function f in the original sequence. Now

$$|g_n(z) - g_m(z)| \le |g_n(z) - g_n(z_k)| + |g_n(z_k) - g_m(z_k)| + |g_m(z_k) - g_m(z)|.$$

For each fixed z, there is some point z_k in the specified dense set such that $|z - z_k| < \delta$. Hence the first and third terms in the preceding inequality each can be made less than ε. The middle term will be less than ε when n and m are sufficiently large, in view of the convergence of the sequence of functions at the points of the dense set. Consequently, the diagonal sequence satisfies Cauchy's criterion uniformly on K.

To prove the converse direction of the theorem, first observe that a compact subset of a metric space always is closed. If a compact set of functions fails to be uniformly bounded, then there is a sequence (f_n) of functions in the set and a sequence (z_n) of points in K such that $|f_n(z_n)| > n$. There is an increasing sequence (n_k) such that (z_{n_k}) converges to some point z in K and (f_{n_k}) converges uniformly to some function f in $C(K)$. Then $f_{n_k}(z_{n_k}) \to f(z)$, but also $f_{n_k}(z_{n_k}) \to \infty$. The contradiction shows that a compact set of functions is necessarily uniformly bounded.

If a compact set of functions fails to be uniformly equicontinuous, then there exists some positive ε such that for every natural number n there are points z_n and w_n and a function f_n in the set such that $|z_n - w_n| < 1/n$ but $|f_n(z_n) - f_n(w_n)| \ge \varepsilon$. Compactness implies that there is an increasing sequence (n_k) such that the sequence (z_{n_k}) converges to a point z in K, the sequence (w_{n_k}) converges to a point w in K, and the sequence (f_{n_k}) converges uniformly to a continuous

function f such that $|f(z) - f(w)| \geq \varepsilon$, but $|z - w| \leq 0$. The contradiction shows that a compact set must be uniformly equicontinuous after all. □

The metric on $C(G)$

There is a standard method for bootstrapping the metric on $C(K)$ to a metric on the space of continuous functions on an open set G in $\mathbb{C}$. First notice that $\|f - g\|_K/(1 + \|f - g\|_K)$ defines a bounded metric that determines the same topology (the same convergent sequences) on $C(K)$ as does $\|f - g\|_K$. (To verify the triangle inequality, observe that the real-valued function $x/(1+x)$ on the positive real numbers is both increasing and subadditive.)

A construction from earlier produces an increasing sequence (K_n) of nonempty compact sets that exhaust G. Define $d(f, g)$ as follows:

$$d(f,g) = \sum_{n=1}^{\infty} \frac{\|f - g\|_{K_n}}{1 + \|f - g\|_{K_n}} \cdot \frac{1}{2^n}.$$

Evidently this function d defines a metric on $C(G)$ that is bounded by 1. The metric depends on the choice of the exhaustion, but the topology is independent of the choice. For convenience, fix an exhaustion once and for all.

Convergence with respect to this metric is the same as uniform convergence on every compact subset of G. Indeed, convergence in this metric implies, in particular, convergence on each individual compact set K_n and hence on an arbitrary compact set. Conversely, if convergence happens on each compact set, and a positive ε is fixed, then chopping off the tail at a point where $\sum_{n \geq N} 1/2^n < \varepsilon/2$ and invoking convergence on K_N shows that convergence happens in this metric.

Arzelà–Ascoli revisited

Theorem. *A subset of $C(G)$ is relatively compact (has compact closure) if and only if this set of functions is pointwise bounded and pointwise equicontinuous.*

Proof. An equivalent statement is that every pointwise bounded, equicontinuous sequence (f_n) in $C(G)$ has a subsequence that converges uniformly on every compact subset of G. By the first version of the theorem, there is a subsequence that converges uniformly on K_1, the first set in the exhaustion of G. There is a further subsequence that converges uniformly on K_2, and so on. The diagonal subsequence converges uniformly on every compact subset of G.

For the converse, if the sequence is not bounded at some point, take a subsequence that blows up at the point. No convergent subsequence can exist. And if the sequence fails equicontinuity at a point, then passing to a locally uniformly convergent subsequence gives a contradiction by a 3ε argument. □

Compactness in $H(G)$

The next goal is to characterize compactness in the space $H(G)$ of holomorphic functions on the open set G. The first observation is that $H(G)$ is a *closed* subspace of $C(G)$. In other words, if a sequence of holomorphic functions converges uniformly on compact sets to a (necessarily) continuous limit function, then the limit function is holomorphic. Since holomorphicity can be tested by integration over closed curves, which are compact sets, the desired conclusion follows immediately either from Morera's theorem or from Cauchy's integral formula.

Montel's theorem (one of them)

Theorem. *A set of functions in $H(G)$ is relatively compact if and only if the set of functions is locally bounded.*

The theorem comes from Paul Montel's 1907 thesis, written under the direction of Borel and Lebesgue. A set of functions satisfying the condition of the theorem is called a *normal family* in Montel's terminology. Indeed, Montel published in 1927 a book titled *Leçons sur les familles normales de fonctions analytiques et leurs applications*. Some authors (including Montel himself) allow the term normal family to admit the possibility that the limit of a subsequence is identically equal to ∞.

Proof. For the sufficiency, what needs to be shown is that a locally bounded family of holomorphic functions is equicontinuous at each point. By Cauchy's estimate for the first derivative, the family of derivatives of a locally bounded family is again a locally bounded family (one has to shrink disks, but the property is local, so shrinking is allowable). Hence the functions in the original family are Lipschitz with a Lipschitz constant that is locally bounded independently of the function. Equicontinuity evidently follows.

The converse, that a normal family of holomorphic functions must be locally bounded, follows from a previous observation that pointwise boundedness in the presence of equicontinuity implies local boundedness. □

Example. The set of holomorphic functions mapping G into the unit disk is a normal family. Indeed, the family is not only locally bounded but even bounded. This example will be used in the proof of the Riemann mapping theorem.

Montel's fundamental normality criterion

Theorem. *The family of holomorphic functions mapping G into $\mathbb{C} \setminus \{0,1\}$ (the twice-punctured plane) is a normal family in the extended sense that every sequence of such functions either admits a subsequence that converges uniformly on compact sets to a holomorphic function or admits a subsequence that converges uniformly to ∞.*

Of course the values 0 and 1 could be replaced by two arbitrary distinct complex numbers a and b (the same numbers for all members of the family of functions). Simply make a linear fractional transformation that fixes ∞ and moves the points a and b to 0 and 1.

The name "fundamental criterion" is due to Montel himself (*critère fondamental*). This theorem is quite deep. Indeed, an easy proof of Picard's theorem is a consequence. The proofs of Montel's criterion and Picard's theorem are deferred until later.

Applications of convergence in $H(G)$

Theorem. *If a sequence of holomorphic functions converges normally (uniformly on compact sets), then so does the sequence of derivatives.*

Theorem (Hurwitz)**.** *If G is a connected open set, and (f_n) is a sequence of zero-free holomorphic functions converging uniformly on compact sets to a limit function f, then either f is zero-free or f is identically equal to zero.*

Corollary. *If G is a connected open set, and (f_n) is a sequence of injective holomorphic functions converging uniformly on compact sets to a limit function f, then either f is injective or f is constant.*

Proof that derivatives inherit normality. The derivative of a holomorphic function is represented by an integral, and uniform convergence of the integrands implies convergence of the integrals. □

Proof of Hurwitz's theorem. The second case can occur: consider, for example, the sequence (z^n) on the open unit disk with a puncture at the origin.

If $f(z_0) = 0$, but f is not identically equal to 0, then f has no zeroes in some punctured neighborhood of z_0 (since the zeroes of f are isolated). Therefore if D is a sufficiently small disk centered at 0 whose closure is contained in G, the function f has no zero on the boundary of D. Then

$$\frac{1}{2\pi i}\int_{\partial D}\frac{f_n'(z)}{f_n(z)}\,dz \to \frac{1}{2\pi i}\int_{\partial D}\frac{f'(z)}{f(z)}\,dz.$$

The integral counts the number of zeroes of the function inside D. Since the approximating integrals all are equal to 0, and the limiting integral is equal to 1, a contradiction arises. □

Proof of corollary. Fix a point z_0 in G. The function that sends z to $f_n(z) - f_n(z_0)$ is zero-free on the region $G \setminus \{z_0\}$ by hypothesis. Hurwitz's theorem implies that the limit function $f(z) - f(z_0)$ is either zero-free or constant on $G \setminus \{z_0\}$. Since z_0 is arbitrary, the function f on G takes each value in its range only once, unless the function is constant. □

Vitali's theorem

Theorem. *If (f_n) is a normal family of holomorphic functions on a connected open set G, and if the sequence converges pointwise on a subset of G that has an accumulation point in the interior of G, then the sequence of functions converges normally on all of G.*

The theorem is named for the Italian mathematician Giuseppe Vitali (1875–1932), who is known also for an example of a nonmeasurable set of real numbers. The result is sometimes called the Vitali–Porter theorem, since M. B. Porter discovered the theorem independently at about the same time.

Proof. Every subsequence of (f_n) has a further subsequence that converges normally to a holomorphic limit function. By hypothesis, all of these limit functions agree on a set that has a limit point, so by the identity theorem, all of the limit functions agree identically on G. Call this unique common limit g. If there were a compact set K on which the sequence (f_n) fails to converge uniformly to g, then there would be a positive ε and a subsequence (f_{n_k}) such that $\|f_{n_k} - g\|_K \geq \varepsilon$ for every k. But the subsequence (f_{n_k}) has a further subsequence that does converge uniformly on K to g. This contradiction completes the proof. □

More on Möbius transformations

There is a Möbius transformation taking three arbitrary points to three arbitrary points. But there is not a Möbius transformation taking four general points to four general points. Consequently, there ought to be some invariant of four points that detects which sets of four points are equivalent and which are not. This invariant is the cross ratio, discussed last semester. Here is a reminder.

Cross ratio The cross ratio of four (distinct) complex numbers z_1, z_2, z_3, and z_4 is the quantity

$$\frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_4)(z_3 - z_2)} \qquad \text{or} \qquad \frac{z_1 - z_2}{z_1 - z_4} \Big/ \frac{z_3 - z_2}{z_3 - z_4}.$$

A common shorthand notation is $[z_1, z_2, z_3, z_4]$, not to be confused with homogeneous coordinates on projective space.

The cross ratio is the image of z_1 under the Möbius transformation that takes z_2, z_3, and z_4 to 0, 1, and ∞ respectively, as is evident from inspecting the formula. If one of the points is ∞, then the cross ratio is understood as a limit. For instance, if $z_4 = \infty$, then

$$[z, z_2, z_3, \infty] = \frac{z - z_2}{z_3 - z_2}.$$

Here the point at infinity is fixed by the transformation.

The importance of the cross ratio is that it is invariant under Möbius transformations. In other words, if T is an arbitrary Möbius transformation, then $[z_1, z_2, z_3, z_4] = [Tz_1, Tz_2, Tz_3, Tz_4]$ for all choices of z_1, z_2, z_3, and z_4. The invariance is obvious for translations, since the cross ratio

depends only on differences of points. Similarly, the invariance is evident for dilations and for rotations, since the cross ratio depends only on ratios. The invariance for the inversion that sends z to $1/z$ is an easy computation too, just simplifying compound fractions. Since these basic types of transformations generate all Möbius transformations, the cross ratio is a general invariant.

Symmetry A complex number z and the complex conjugate $\overline{z}$ are symmetric with respect to the real axis. Notice that if z_2, z_3, and z_4 are three distinct points on the real axis, then

$$[\overline{z}, z_2, z_3, z_4] = \overline{[z, z_2, z_3, z_4]},$$

as is evident from the formula for the cross ratio. Now lines and circles are equivalent under Möbius transformations, so a reasonable definition is that points z and z^* are symmetric with respect to a circle if, whenever z_2, z_3, and z_4 are three distinct points on the circle, the following equality holds:

$$[z^*, z_2, z_3, z_4] = \overline{[z, z_2, z_3, z_4]}.$$

The meaning is that if the circle is mapped to the real axis by a Möbius transformation, then the points z and z^* map to points that are complex conjugates.

This notion of symmetry is equivalent to the geometric notion: two points are symmetric with respect to a circle if they lie on the same ray from the center, and the product of their distances from the center equals the square of the radius. Indeed, after a translation and a dilation, the disk can be assumed to be the standard unit disk centered at the origin. If z_2, z_3, and z_4 are three points on the unit circle (hence equal to the reciprocals of their conjugates), and z is any nonzero number, then the invariance of the cross ratio implies that

$$[1/\overline{z}, z_2, z_3, z_4] = [1/\overline{z}, 1/\overline{z_2}, 1/\overline{z_3}, 1/\overline{z_4}] = [\overline{z}, \overline{z_2}, \overline{z_3}, \overline{z_4}] = \overline{[z, z_2, z_3, z_4]}.$$

Thus the symmetric point z^* is equal to $1/\overline{z}$, as claimed.

Automorphisms of the unit disk The knowledge that linear fractional transformations preserve symmetry (in the sense just indicated) gives a way to determine a formula for the automorphism of the unit disk that interchanges 0 and a. Such an automorphism T must also interchange the points ∞ and $1/\bar{a}$ that are symmetric to 0 and a. The invariance of cross ratio implies that

$$[Tz, 0, a, \infty] = [z, a, 0, 1/\bar{a}],$$

or

$$\frac{Tz - 0}{a - 0} = \frac{z - a}{z - 1/\bar{a}} \Big/ \frac{0 - a}{0 - 1/\bar{a}},$$

whence $Tz = \dfrac{a - z}{1 - \bar{a}z}$. A standard name for this transformation T is φ_a.

Proof of the Riemann mapping theorem

Riemann mapping theorem

Theorem. *If G is a simply connected planar region, not the whole plane, then there exists a biholomorphic mapping from G onto the unit disk.*

The theorem stems from Riemann's 1851 thesis. His proof, based on the Dirichlet principle, is justifiable for domains with reasonable boundary, say piecewise smooth. But for nasty boundary, a different method is needed. The modern proof is the work of many hands.

The first complete proof apparently is due to Carathéodory in 1912; he produced the map as the limit of a sequence of maps. The modern proof via an extremal problem is due to Fejér and Riesz, published with permission by Radó in 1922. The square-root trick to cook up an improved mapping apparently is due to Carathéodory and Koebe. Fejér and Riesz make the explicit computation; the method for avoiding the computation seems to be due to Ostrowski and Carathéodory. Carathéodory's proof (given below) avoids taking derivatives.

Remark. The map certainly is not unique, for one can post-compose with an automorphism of the disk. The map can be made unique in various ways. For instance, if a point z_0 in G is chosen that maps to 0, and if the derivative of the mapping is specified to be a positive real number, then the mapping is unique. Indeed, if f and g are two such maps, then $f \circ g^{-1}$ is an automorphism of the unit disk fixing the origin and having positive derivative at the origin. The Schwarz lemma implies that such a map is a rotation, and the positivity of the derivative forces this composite map to be the identity rotation.

Another way to ensure uniqueness is to choose two distinct points z_0 and z_1 in G and demand that z_0 maps to 0 and z_1 maps to a positive real value. Again, if f and g are two such maps, then $f \circ g^{-1}$ is an automorphism of the disk that fixes 0 and maps some positive real number to a positive real number. Hence $f \circ g^{-1}$ is a trivial rotation.

Proof of existence. The outline of the proof is the following. Consider the family of all injective holomorphic functions that map the given simply connected region G into the unit disk, taking a specified point z_0 to 0. The goal is to find a mapping in this normal family that makes the image fill out as much of the disk as possible. Namely, there is an extremal function that maximizes the modulus of the value of the map at a second specified point z_1. This extremal function must be the required holomorphic bijection, else a new function could be constructed that increases the value at z_1.

There are numerous steps to fill in.

First of all, are there any injective holomorphic functions mapping G into the unit disk? If G were the whole plane, then there would be no nonconstant maps (by Liouville's theorem), hence the exclusion of the plane is necessary in the statement of the theorem.

If G is a *bounded* region, then there are lots of injective maps into the unit disk: translate G to move z_0 to the origin, then shrink by a suitable dilation with a factor less than 1.

What if G is unbounded? Since G is not the whole plane, there is at least one point b in the complement of G. If b is an interior point of $\mathbb{C} \setminus G$, then an inversion with respect to b maps G into

a bounded region, and the previous case can be invoked. So the hard case is that the complement of G has empty interior: the region G could be the plane with a slit, for instance.

If b is a point in the complement of G, then $z - b$ is a zero-free holomorphic function on G (a simply connected region), so there is a holomorphic branch of $\sqrt{z-b}$ on G. Evidently this square root is injective (if not, the square would not be), so what needs to be shown is that the image of G under $\sqrt{z-b}$ omits some disk, thus reducing to the previous case.

Now if c is a point in the image of $\sqrt{z-b}$ (and c is necessarily different from zero, since $z-b$ is zero-free on G), then the point $-c$ is not in the image: for if both $\sqrt{z_2-b} = c$ and $\sqrt{z_3-b} = -c$, then squaring shows that $z_2 - b = z_3 - b$, a contradiction. Since $\sqrt{z-b}$ is an open map, a whole neighborhood of c is in the image, so a neighborhood of $-c$ is not in the image. Thus the previous case produces an injective map of G into the unit disk. Composing with a suitable automorphism of the disk will send the specified point z_0 to the origin.

Fix a point z_1 in G different from z_0. Take a sequence in the family for which the modulus of the value of the function at z_1 approaches the least upper bound of all such values. There is a subsequence converging normally to a holomorphic limit function f. Evidently $f(z_0) = 0$, and $|f(z_1)|$ achieves the extreme value in the family. In particular, $f(z_1)$ is different from $f(z_0)$. Therefore the limit function is not constant, so f is injective, being the limit of injective holomorphic functions. The function f a priori maps into the closed unit disk, but by the maximum principle the image lies in the open unit disk. Thus f is indeed an extremal function in the family.

What remains is to show that the extremal function is surjective. The argument is by contradiction. Suppose a nonzero point c in the disk is not in the image of f. The goal is to produce a contradiction by finding a new function in the family whose value at z_1 has modulus larger than $|f(z_1)|$.

Notice that $c \neq 0$, since $f(z_0) = 0$. Under the hypothesis that c is not in the image of f, the function $\varphi_c \circ f$ is zero-free in G, hence has a holomorphic square root, call it g. Here φ_c is the standard self-inverse disk automorphism that swaps c and 0: namely, $\varphi_c(z) = (c - z)/(1 - \overline{c}z)$. This function g is injective, for otherwise the square would not be injective. Now g maps the region G into the unit disk, but g does not belong to the specified family of functions, for g is not normalized at z_0. Indeed, $\varphi_c \circ f(z_0) = c$, so $g(z_0) = \sqrt{c}$ for one of the two possible values of the square root.

Set h equal to $\varphi_{\sqrt{c}} \circ g$. Then h again maps G into the unit disk, and now $h(z_0) = 0$. What remains to show (to reach the desired contradiction) is that $|f(z_1)| < |h(z_1)|$. The plan now is to unwind the definitions to relate f to h.

On the one hand, $g^2 = \varphi_c \circ f$, so $f = \varphi_c \circ g^2$. On the other hand, $g = \varphi_{\sqrt{c}} \circ h$, so $g^2 = (\varphi_{\sqrt{c}} \circ h)^2 = \varphi^2_{\sqrt{c}} \circ h$. Therefore $f = (\varphi_c \circ \varphi^2_{\sqrt{c}}) \circ h$. Now the function $\varphi_c \circ \varphi^2_{\sqrt{c}}$ maps the unit disk to itself, fixing the origin. By the Schwarz lemma, $|\varphi_c \circ \varphi^2_{\sqrt{c}}(z)| \leq |z|$ for every point z in the unit disk. Moreover, if equality holds in the Schwarz lemma for even one nonzero point, then the function has to be a rotation. But the map $\varphi_c \circ \varphi^2_{\sqrt{c}}$ evidently is not a rotation, since this map is two-to-one (because of the square). Therefore $|\varphi_c \circ \varphi^2_{\sqrt{c}}(z)| < |z|$ for every nonzero point z in

the disk, with strict inequality.

In particular,

$$|f(z_1)| = |\varphi_c \circ \varphi^2_{\sqrt{c}}(h(z_1))| < |h(z_1)|,$$

so the function h violates the extremality of f. The contradiction shows that the map f must be surjective after all. □

Harmonic functions

The Cauchy–Riemann equations imply that the real part u of a holomorphic function f is harmonic (satisfies Laplace's equation). Namely, if $f = u + iv$, and if $u_x = v_y$ and $u_y = -v_x$, then $u_{xx} + u_{yy} = v_{yx} - v_{xy} = 0$. (Since the holomorphic function f has continuous derivatives of all orders, so do the functions u and v, whence the mixed second-order partial derivatives of v match.)

Is every real-valued harmonic function u the real part of some holomorphic function f? The general answer is negative, for there is a topological obstruction.

Exercise. The function $\log |z|^2$ is well defined and harmonic on $\mathbb{C} \setminus \{0\}$, the punctured plane, but there is no holomorphic function f on the punctured plane such that $\operatorname{Re} f(z)$ equals $\log |z|^2$.

The answer is affirmative in a simply connected domain. Fix a base point (x_0, y_0) in the domain, and define a harmonic conjugate function v as follows:

$$v(x, y) = \int_{(x_0,y_0)}^{(x,y)} u_1(s, t)\, dt - u_2(s, t)\, ds.$$

The integral is well defined—independent of the path—because the harmonicity of u implies that the integrand is a closed differential form. (The integral over a closed loop is zero by Green's theorem.) The fundamental theorem of calculus implies that if v is defined by this formula, then $v_1 = -u_2$ and $v_2 = u_1$.

This formula really is the fundamental theorem of calculus. Namely, the Cauchy–Riemann equations imply that

$$\begin{aligned} dv &= \frac{\partial v}{\partial x}\, dx + \frac{\partial v}{\partial y}\, dy \\ &= -\frac{\partial u}{\partial y}\, dx + \frac{\partial u}{\partial x}\, dy, \end{aligned}$$

so the indicated line integral simply expresses v as the integral of the derivative.

Experience with the Cauchy integral suggests that a more powerful way to recover f from u would be an integral over the boundary of a region. The first step is to see how to recover u itself from a boundary integral. In the case of the unit disk, the formula is named for Siméon Denis Poisson (1781–1840), a contemporary of Cauchy.

Here is a magic trick for deriving the Poisson integral, an integral representation for harmonic functions in a disk. The value of a holomorphic function at the center of a disk is the average of the boundary values, and taking the real part shows that harmonic functions too satisfy the mean-value property. Accordingly, if u is continuous on the closed unit disk and harmonic on the open disk, then

$$u(0) = \frac{1}{2\pi}\int_0^{2\pi} u(e^{i\theta})\,d\theta = \frac{1}{2\pi i}\int_{|w|=1} u(w)\,\frac{dw}{w}.$$

Compose with the disk automorphism φ_a to say that

$$u(a) = u\circ\varphi_a(0) = \frac{1}{2\pi i}\int_{|w|=1} u(\varphi_a(w))\,\frac{dw}{w}.$$

Make a change of variable, replacing w by $\varphi_a(w)$ and remembering that φ_a is self-inverse. Locally,

$$\frac{dw}{w} = d(\log w),$$

so using that w and $\overline{w}$ are reciprocals on the boundary shows that

$$\frac{d\varphi(w)}{\varphi(w)} = \left(\frac{1}{w-a} + \frac{\overline{a}}{1-\overline{a}w}\right) w\,\frac{dw}{w} = \left(\frac{w}{w-a} + \frac{\overline{a}}{\overline{w}-\overline{a}}\right)\frac{dw}{w}$$

when $|w| = 1$, and the expression in parentheses simplifies to $\dfrac{1-|a|^2}{|w-a|^2}$. Thus

$$u(a) = \frac{1}{2\pi}\int_0^{2\pi} u(e^{i\theta})\frac{1-|a|^2}{|e^{i\theta}-a|^2}\,d\theta, \tag{1}$$

which is the Poisson integral representation for harmonic functions. Moreover,

$$\frac{1-|a|^2}{|e^{i\theta}-a|^2} = \operatorname{Re}\frac{w+a}{w-a} \qquad \text{when } w = e^{i\theta},$$

so the real-valued harmonic function $u(a)$ is the real part of the holomorphic function

$$\frac{1}{2\pi i}\int_{|w|=1} u(w)\frac{w+a}{w-a}\cdot\frac{dw}{w} \qquad \text{when } |a| < 1.$$

This formula, named for Hermann Amandus Schwarz (1843–1921), explicitly determines a holomorphic function (up to an additive purely imaginary constant) from the real part.

Notice that the Poisson kernel

$$\frac{1}{2\pi}\cdot\frac{1-|a|^2}{|e^{i\theta}-a|^2}$$

is a positive function whose integral from 0 to 2π is equal to 1 (as follows from (1) when u is identically equal to 1). Accordingly, the Poisson integral (1) exhibits the value $u(a)$ as a weighted average of the boundary values of u.

This interpretation of the Poisson integral immediately yields a local maximum principle for real-valued harmonic functions: if a harmonic function on a closed disk attains a maximum at an interior point a, then the function reduces to a constant. Indeed, if the weighted average $u(a)$ is a maximum, then the values of u on the boundary must all be equal to $u(a)$. Invoking (1) again at a different interior point shows that u is constantly equal to $u(a)$ everywhere in the disk. Considering the negative of u shows that harmonic functions satisfy a minimum principle too.

The Dirichlet problem for the disk

The preceding discussion shows that the Poisson integral reproduces harmonic functions on the unit disk. A little more work shows that the Poisson integral solves the problem of finding a harmonic function on the disk with prescribed boundary values.

Suppose that a continuous, real-valued function u is given on the boundary circle. Define the Poisson integral $P[u]$ at a point a in the disk to be

$$\frac{1}{2\pi}\int_0^{2\pi} u(e^{i\theta})\frac{1-|a|^2}{|e^{i\theta}-a|^2}\,d\theta.$$

This integral defines a function of a inside the disk that is harmonic because the kernel is the real part of a holomorphic function. Question: is the limit of this function when a tends to a boundary point equal to the original function u at that boundary point?

The answer is affirmative when u is continuous on the boundary. Indeed, let $e^{i\psi}$ be a specified boundary point. The Poisson integral reproduces constant functions, so the difference between the Poisson integral of u at a and the constant $u(e^{i\psi})$ is

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{1-|a|^2}{|e^{i\theta}-a|^2}\left(u(e^{i\theta})-u(e^{i\psi})\right)\,d\theta.$$

Fix a positive ε, and invoke the continuity of u to choose a positive δ such that $|u(e^{i\theta})-u(e^{i\psi})|<\varepsilon$ when $|\theta-\psi|<\delta$. Split the integral into the part for which $|\theta-\psi|<\delta$ and the part for which $|\theta-\psi|\geq\delta$. The integral over the first part is at most

$$\frac{\varepsilon}{2\pi}\int_{|\theta-\psi|<\delta} \frac{1-|a|^2}{|e^{i\theta}-a|^2}\,d\theta,$$

which by the positivity of the Poisson kernel does not exceed

$$\frac{\varepsilon}{2\pi}\int_0^{2\pi} \frac{1-|a|^2}{|e^{i\theta}-a|^2}\,d\theta, \qquad \text{or} \qquad \varepsilon,$$

the inequality being independent of the value of a inside the disk. The integral over the second part tends to 0 when a tends to $e^{i\psi}$ since the Poisson kernel converges to 0 uniformly on that piece. Accordingly, the limit of the value of the Poisson integral of u at a tends to $u(e^{i\psi})$ when a tends to $e^{i\psi}$.

The same argument yields a more general local result. If u is merely (Lebesgue) integrable on the boundary, then the Poisson integral of u approaches the value of u on the boundary at every point where the boundary function is continuous.

Uniqueness of the solution of the Dirichlet problem in the disk is easy. The difference of two solutions is a harmonic function with boundary value identically equal to zero; by the maximum and minimum principles, such a function is identically equal to zero inside the disk.

More on the mean-value property

Morera's theorem gives a way to characterize holomorphic functions by integration instead of by differentiation. There is an analogous way to characterize harmonic functions via integration.

The claim is that if u is a continuous real-valued function with the property that for every point z in a domain there is a positive radius r (depending on z) such that the average of u on every circle centered at z of radius less than r equals $u(z)$, then u is necessarily harmonic.

Harmonicity is a local property, so there is no loss of generality in supposing that the domain of u is a disk and that u is continuous on the closure of the disk. Scaling and translation do not affect the problem, so there is no loss of generality in taking the disk to be the unit disk centered at 0.

A key observation is that the mean-value property implies a maximum principle: the function u must attain its maximum on the boundary of the disk. Since u is continuous on a compact set, a maximum is attained somewhere. If there is an interior maximum, then the mean value on every small circle centered at that point equals the maximum, so u must be constantly equal to the central value on small circles. Hence u is locally equal to the maximal value. A connectedness argument now shows that u is constantly equal to the maximal value. So the maximum is taken on the boundary in any case.

Let v denote the Poisson integral of the boundary value of u. (In this discussion, the symbols u and v do not denote harmonic conjugates!) The solution of the Dirichlet problem shows that v matches u on the boundary.

The function v, being harmonic inside the disk, satisfies both the mean-value property and the maximum principle. The difference $u - v$ satisfies the mean-value property since both u and v do. Hence $u - v$ attains its maximum on the boundary. This boundary value equals 0, so $u - v \leq 0$ inside the disk. But the same argument applies to the difference $v - u$, so $v - u \leq 0$. The two inequalities combine to show that $u - v$ is identically equal to 0.

Accordingly, the function u is harmonic because u matches a known harmonic function.

Variations on the maximum principle

The local maximum principle says that a nonconstant harmonic function on a connected open set cannot have a local maximum. A corresponding minimum principle holds too (simply consider the negative of the function). This principle follows from the mean-value property of harmonic functions, for an average cannot equal the maximum unless the quantity being averaged is constant.

An apparent corollary would be a global maximum principle saying that the maximum of a harmonic function must occur on the boundary, but this statement is false unless an additional hypothesis is added. Indeed, the harmonic function $\mathrm{Re}(z)$ restricted to the right-hand half-plane is equal to 0 on the boundary but is unbounded.

A correct global statement is that if G is a *bounded* open set, and u is a continuous function on the closure of G that is harmonic on the interior of G, then u attains a maximum value on the boundary of G. Indeed, the closure of G is compact, so the continuous function u attains a maximum somewhere on the closure of G. By the local maximum principle, this maximum must be taken on the boundary of G if u is not a constant function; and if u is constant, then the maximum is taken on the boundary (as well as everywhere else).

An apparently more general statement is that if the domain G is bounded, the function u is harmonic, the limit $\lim_{z\to b} u(z)$ exists for every point b of the boundary of G, and this limit is no larger than some number M (independent of b), then $u(z) \le M$ for every point z in G. This statement actually is no more general than the previous one, for the hypotheses imply that the function u extends to be continuous on the closure of G. Indeed, define u on the boundary to be equal to the limit that exists by hypothesis. If b is a specified boundary point of G, and a positive ε is prescribed, then there is a neighborhood V of b such that the values of u in $V \cap G$ differ from $u(b)$ by at most ε. The existence of the limit of u at nearby boundary points then implies that $u(b')$ differs from $u(b)$ by at most ε when $b' \in V \cap \overline{G}$. Consequently, the extended u is continuous at an arbitary boundary point b. The previous version of the maximum principle implies that the extended u attains a maximum on the boundary, and this maximum is at most M.

The next refinement is to relax the requirement that the limit of u exist at the boundary. Suppose the domain G is bounded, the function u is harmonic in G, and $\limsup_{z\to b} u(z) \le M$ for every point b in the boundary of G. Then $u(z) \le M$ for every point z in G. Indeed, the compactness of the boundary of G implies that if a positive ε is prescribed, then there exists a neighborhood V of the boundary of G such that $u(z) < M + \varepsilon$ when $z \in G \cap \overline{V}$. Apply the previous version of the maximum principle to the open set $G \setminus \overline{V}$ to deduce that $u(z) \le M + \varepsilon$ when $z \in G$. Now let ε go to 0.

The following example shows that the maximum principle can break down if there is an exceptional boundary point at which the lim sup is not under control. Suppose G is the unit disk, and $u(z) = \mathrm{Re}\,\frac{1+z}{1-z} = \frac{1-|z|^2}{|1-z|^2}$. The linear fractional transformation sending z to $\frac{1+z}{1-z}$ takes the unit disk to the right-hand half-plane and the boundary circle to the imaginary axis. Accordingly, the limit of $u(z)$ exists and equals 0 at every boundary point of the disk except the point where $z = 1$ (the limit does not exist at this point). The function u is unbounded, so the boundary value 0 is not a maximum!

Nonetheless, there is a maximum principle in the presence of an exceptional point if an additional hypothesis is added. Suppose that the domain G is bounded, the function u is harmonic in G and *bounded*, and $\limsup_{z\to b} u(z) \le M$ for every point b in the boundary of G with one possible exception. Then $u(z) \le M$ for every point z in G. (After the fact, one can deduce that there is no exceptional boundary point after all.) For the proof, let b_0 denote the exceptional boundary

point, let ε be an arbitrary positive number, and consider the function

$$u(z) + \varepsilon \log \left| \frac{z - b_0}{\operatorname{diam} G} \right|,$$

the value of which does not exceed $u(z)$. Since this modified harmonic function has limit $-\infty$ at b_0, the version of the maximum principle without exceptional point applies. Consequently,

$$u(z) + \varepsilon \log \left| \frac{z - b_0}{\operatorname{diam} G} \right| \le M \qquad \text{when } z \in G.$$

Now let ε go to 0. (The same argument handles a finite number of exceptional points.)

The preceding discussion considers two-dimensional limits at boundary points. In the special case of the unit disk, one-dimensional limits along radii are natural to consider. Suppose u is harmonic and bounded on the open unit disk, and $\lim_{r \to 1-} u(re^{i\theta}) \le 0$ for every angle θ. The following argument shows that $u(a) \le 0$ for every point a in the open unit disk.

When a is fixed, and r is a radius strictly between $|a|$ and 1, represent $u(a)$ by the Poisson integral on a disk of radius r: namely,

$$u(a) = \frac{1}{2\pi} \int_0^{2\pi} u(re^{i\theta}) \frac{r^2 - |a|^2}{|re^{i\theta} - a|^2} \, d\theta.$$

The hypothesis that u is bounded means that the bounded convergence theorem (or the dominated convergence theorem) applies to justify taking the limit as $r \to 1$ inside the integral. Consequently, $u(a)$ equals a weighted average over the unit circle of a nonpositive function, so $u(a) \le 0$.

When u is bounded, the preceding argument works just as well in the presence of a finite number of exceptional values of θ for which the radial limit fails to exist. If you are willing to admit the Lebesgue integral, then you can allow even a set of measure zero of exceptional values of θ.

On the other hand, the hypothesis of boundedness of u cannot be relaxed. Observe that

$$\operatorname{Im} \left(\frac{1+z}{1-z} \right)^2 = \operatorname{Im} \frac{(1 + z - \overline{z} - |z|^2)^2}{|1 - z|^4} = \frac{4(\operatorname{Im} z)(1 - |z|^2)}{|1 - z|^4}.$$

This harmonic function on the unit disk is identically equal to 0 on the real axis, so the radial limit exists and equals 0 at every boundary point (including the point where $z = 1$). Even though the radial limit is everywhere equal to 0, the function is not bounded. Indeed, along the line where $\operatorname{Re} z + \operatorname{Im} z = 1$, the function blows up as z approaches 1.

Another special geometry often encountered is a strip, say the vertical strip where $-\pi/2 < \operatorname{Re}(z) < \pi/2$. The function that takes z to $\sin(z)$ maps this strip bijectively to the plane with slits along the real axis from 1 to $+\infty$ and from $-\infty$ to -1. Consequently, $e^{-(\sin z)^2}$ is a holomorphic function of z that is bounded on the two sides of the strip but unbounded in the strip. This example shows again that the maximum principle breaks down on unbounded regions if there is no control over the function at infinity. A perhaps even more dramatic example is $\operatorname{Re} \cos(z)$, a harmonic function that is identically equal to zero on the sides of the strip, yet is strictly positive (and unbounded) inside the strip. The three-lines theorem, to be considered later, is a version of the maximum principle for a strip.

Normal families of harmonic functions

One of Montel's theorems says that a family of holomorphic functions on a domain is normal (relatively compact in the topology on continuous functions) if and only if the family is locally bounded. Does the parallel statement hold for harmonic functions? That the answer is affirmative can be seen in several ways.

A useful first observation is that normality is a local property: if a sequence of functions has the property that every point admits a neighborhood for which there is a subsequence that converges uniformly on compact subsets of that neighborhood, then there is a subsequence that converges uniformly on every compact subset of the domain. Indeed, the domain can be covered by countably many of the indicated neighborhoods (which can be assumed to be disks without loss of generality). There is a subsequence converging on the first neighborhood, a subsequence of the first subsequence converging on the second neighborhood, and so on. The diagonal subsequence then converges uniformly on every compact set.

Since normality is a local condition, answering the original question on disks suffices. On a disk, every harmonic function is the real part of a holomorphic function, so a natural idea is to try to deduce the statement for harmonic functions as a corollary of the statement for holomorphic functions. Normality implies local boundedness, so the question is whether local boundedness implies normality.

If (u_j) is a sequence of harmonic functions on a disk, then there is a sequence (f_j) of holomorphic functions such that $\operatorname{Re} f_j = u_j$ for each j. If the sequence of harmonic functions is locally bounded, then so is the sequence $(\exp(f_j))$, since $|\exp(f_j)| = \exp(u_j)$. A subsequence $(\exp(f_{j_k}))$ converges normally to a holomorphic function g. (Notice that there is no claim here about convergence of (f_{j_k}), for the imaginary part of f_j is not under control.) The local boundedness of the sequence of harmonic functions implies that the limiting function g is nowhere equal to zero. Continuity of the absolute-value function implies that the sequence $(\exp(u_{j_k}))$ converges normally to $|g|$. Continuity of the real logarithm function implies that the sequence (u_{j_k}) converges normally to $\log|g|$. Accordingly, local boundedness of a family of harmonic functions implies normality.

An alternative method is to look back at the discussion of the Poisson integral to see that there is an explicit integral representation to produce a holomorphic function on a disk with specified real part. The integral representation reveals that the modulus of the holomorphic function is bounded on a slightly smaller disk by the maximum of the absolute value of the harmonic function on the original disk. Consequently, local boundedness of a family of harmonic functions on a disk implies local boundedness of the corresponding family of holomorphic functions. The family of holomorphic functions is normal by Montel's theorem, so the family of harmonic functions is normal.

A third argument is to go back to the Arzelà–Ascoli theorem. What needs to be shown is that a locally bounded family of harmonic functions is equicontinuous at every point. The proof in the case of holomorphic functions uses the Cauchy integral. The same argument applies to harmonic functions if the Cauchy integral is replaced by the Poisson integral. Namely, if u is harmonic in a disk, and z_1 and z_2 are two points in the disk, then the difference $u(z_1) - u(z_2)$ equals the integral

over the boundary circle of $u(w)(P(w, z_1) - P(w, z_2))$, where P denotes the Poisson kernel. This integral is bounded by the maximum of $|u|$ times the maximum of $|P(w, z_1) - P(w, z_2)|$ for w on the boundary circle. The explicit formula for the Poisson kernel reveals that the latter expression is bounded by a constant (independent of the integration variable w) times $|z_1 - z_2|$. Accordingly, a locally bounded family of harmonic functions is not only equicontinuous but even equi-Lipschitz.

Remark. Problem 9 on the August 2010 qualifying examination is an analogue for harmonic functions of Vitali's theorem.

Harnack's principle

Related to the preceding discussion of normal families of harmonic functions is a convergence principle for monotonic sequences of (real-valued) harmonic functions.

Proposition. *An increasing sequence of harmonic functions on a connected open set converges uniformly on compact subsets either to $+\infty$ or to a harmonic function.*

The lemma is named for Axel Harnack (1851–1888), a Baltic German mathematician. A corresponding statement holds for a decreasing sequence of harmonic functions, since the negative of a harmonic function is again a harmonic function.

The proof depends on Harnack's inequality for positive harmonic functions. Namely, if u is harmonic and nonnegative in the unit disk, and $0 < r < 1$, then

$$u(0)\frac{1-r}{1+r} \le u(re^{i\theta}) \le u(0)\frac{1+r}{1-r}.$$

Indeed, since u is nonnegative and has the mean-value property, the Poisson integral representation shows that

$$u(re^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi} u(e^{i\varphi})\frac{1-r^2}{|re^{i\theta} - e^{i\varphi}|^2}\,d\varphi \le \frac{1}{2\pi}\int_0^{2\pi} u(e^{i\varphi})\frac{1-r^2}{(1-r)^2}\,d\varphi = u(0)\frac{1+r}{1-r}.$$

The other inequality follows in the same way, using that $|re^{i\theta} - e^{i\varphi}|^2 \le (1+r)^2$. (Strictly speaking, one should integrate over a slightly smaller circle and take the limit.)

Proof of Harnack's principle. Replacing the increasing sequence (u_n) by $(u_n - u_1)$ reduces to the case of nonnegative functions, so Harnack's inequality is in force. Suppose the domain contains the unit disk. The increasing sequence $(u_n(0))$ of real numbers either tends to $+\infty$ or is a Cauchy sequence. In the former case, Harnack's inequality implies that the sequence (u_n) converges uniformly on compact sets to $+\infty$. In the latter case, the same reason implies that the sequence is uniformly Cauchy on compact subsets of the disk. The continuous limit function is represented by the Poisson integral and so is harmonic. The generalization from convergence on disks to convergence on general connected open sets is a routine compactness argument. □

Dirichlet problem on general domains

Dirichlet problem

The Poisson integral solves the Dirichlet problem on a disk. The corresponding problem in a general region is not always solvable.

Example. In the punctured disk $\{ z \in \mathbb{C} : 0 < |z| < 1 \}$, there is no harmonic function u such that u has boundary value 0 on the outer boundary and boundary value 1 on the inner boundary.

Indeed, the maximum principle implies that the harmonic function u is bounded between 0 and 1, and the version of the maximum principle with an exceptional boundary point implies that u is bounded above by 0, hence is constantly equal to 0. Therefore the function u does not have the required limit at the origin.

(If you forgot about the maximum principle with an exceptional point, apply the usual maximum principle to $u(z) + \varepsilon \log |z|$, where ε is an arbitrary positive number. Then let ε go to 0.)

This example reveals the basic obstruction to solvability of the Dirichlet problem: thinness of the boundary. An upcoming theorem shows that the Dirichlet problem is solvable when the boundary has no isolated points.

The method to be considered is due[2] to the German mathematician Oskar Perron (1880–1975), who is noted for beautiful expository books, especially one on continued fractions. He is remembered too for the Perron integral, for a formula in analytic number theory, and for the Perron–Frobenius theorem in linear algebra about eigenvalues of matrices with positive entries (a result that has applications to internet search engines).

Subharmonic functions

A key tool in Perron's method for solving the Dirichlet problem is a class of functions known as *subharmonic functions*. The philosophy is that holomorphic functions and harmonic functions are inconveniently rigid: the values of the function on an open set determine the values of the function everywhere. Subharmonic functions are more flexible, enabling cut-and-paste operations. Yet there is a way to get from subharmonic functions to harmonic functions through taking envelopes.

Roughly speaking, subharmonic functions sit underneath harmonic functions in the same way that convex functions sit underneath affine linear functions. Like convex functions, subharmonic functions need not be everywhere differentiable. In fact, subharmonic functions need not be continuous (although continuous ones will do for a basic solution to the Dirichlet problem).

The natural context for subharmonic functions is the class of real-valued upper semicontinuous functions. A function u (with arbitrary domain in a topological space) taking values in $[-\infty, \infty)$ is called upper semicontinuous if any of the following equivalent conditions holds:

- $\limsup_{z \to z_0} u(z) \leq u(z_0)$ for every point z_0 in the domain of u.

[2] Oskar Perron, Eine neue Behandlung der ersten Randwertaufgabe für $\Delta u = 0$, *Math. Z.* **18** (1923), no. 1, 42–54.

- Reinterpretation of the preceding statement: For every number M larger than $u(z_0)$, there is a neighborhood of z_0 such that $u(z) < M$ when z is in that neighborhood. (When $u(z_0) \neq -\infty$, the number M can be written conveniently in the form $u(z_0) + \varepsilon$.)

- The set $\{ z : u(z) < c \}$, the inverse image of $[-\infty, c)$ under u, is open for every real number c.

The word "upper" in the definition corresponds to the upper half of the inequality that characterizes continuity. What the condition says about the graph of the function is that the dot at a discontinuity fills in at (or above) the high point.

A reason for allowing the value $-\infty$ but excluding the value $+\infty$ is that upper semicontinuous functions arise naturally as limits of decreasing sequences of continuous finite-valued functions. Such limits can attain the value $-\infty$ but not the value $+\infty$.

Proposition. *An upper semicontinuous function is bounded above on every compact set and attains the least upper bound.*

Proof. The hypothesis implies that every point z in the compact set K has a neighborhood on which the function u is bounded above by $u(z) + 1$. Finitely many such neighborhoods cover K. Hence u is bounded above on K.

If the least upper bound M is not attained, then the compact set K is covered by the sequence of open sets of the form $\{ z : u(z) < M - \frac{1}{n} \}$ (where n runs through the natural numbers), but there is no finite subcover. The contradiction shows that the bound M must be attained after all. □

If G is an open set in $\mathbb{C}$, then an upper semicontinuous function u is called subharmonic if for every disk in G and for every harmonic function v on the disk, the difference $u - v$ satisfies the (local) maximum principle: namely, the function $u - v$ cannot have a strict local maximum and can attain a weak local maximum at a point only if $u - v$ is constant in a neighborhood of the point. Thus if $u \leq v$ on the boundary of the disk, then $u \leq v$ in the interior of the disk.

This property evidently is local. The property needs to hold merely on all sufficiently small disks. In other words, for every point z_0 there should be a radius r_0 such that the property holds on each disk $D(z_0, r)$ when $0 < r < r_0$.

Example. If f is holomorphic, then $\log |f|$ is subharmonic. (The function is defined to be equal to $-\infty$ at zeroes of f.)

Indeed, if v is harmonic, then $\log |f| - v$ evidently cannot attain a local maximum at a zero of f (except in the trivial case that f is identically equal to 0). Away from the zeroes of f, there is a locally defined branch of $\log f$, so $\log |f|$ is harmonic, and so is the difference $\log |f| - v$. Hence there cannot be a local maximum unless the function is constant.

Example. If $u(x, y) = \min(0, x^2 - y^2)$ in $\mathbb{C}$, then u is *not* subharmonic.

Indeed, if $v(x, y)$ is the harmonic function $x^2 - y^2$, then $u(x, y) - v(x, y)$ is equal to 0 when $x^2 - y^2 \leq 0$ and is equal to the negative quantity $-(x^2 - y^2)$ when $x^2 - y^2 > 0$. Hence $u - v$ attains a maximal value of 0 but is not constant in a neighborhood of any point at which $x = y$, violating the maximum principle.

The initial definition of subharmonicity appears hard to verify. For functions having some regularity, there are equivalent properties that are more easily checked.

If u is continuous, then an equivalent property is the local sub-mean-value property. In other words, for each point z_0 there is a radius r_0 such that

$$u(z_0) \le \frac{1}{2\pi} \int_0^{2\pi} u(z_0 + re^{i\theta})\, d\theta \qquad \text{when } 0 < r < r_0.$$

(This property can be used when u is merely upper semicontinuous, not necessarily continuous, if you are willing to accept the Lebesgue integral. You need to go back to the theory of the Poisson integral and check that the Poisson integral of a merely upper semicontinuous function produces a harmonic function whose lim sup at the boundary sits below the boundary value.)

If u satisfies the sub-mean-value property, then so does $u - v$ when v is harmonic. Hence $u - v$ satisfies the local maximum principle (if the average value at the center of some disk is maximal, then the integrand must be constant on the disk). Conversely, if $u - v$ satisfies the local maximum principle for every harmonic v, then in a small disk let v be the Poisson integral of u. The maximum principle implies that the value of u at the center is at most the value of the Poisson integral of u at the center, which equals the average of the values of u around the boundary circle. Hence u has the sub-mean-value property. The same argument shows that if u has the local sub-mean-value property, then u has the global sub-mean-value property on every disk whose closure lies inside the domain of u.

If u is twice continuously differentiable, then an equivalent condition to subharmonicity is that $\Delta u \ge 0$, where Δ is the Laplace operator. For the proof, suppose first that $\Delta u > 0$ with strict inequality. If v is harmonic, then $\Delta(u - v) = \Delta u > 0$. Hence $u - v$ cannot have a local maximum, because at a local maximum, the second derivatives $\partial^2/\partial x^2$ and $\partial^2/\partial y^2$ of a function must be negative or zero. So $u - v$ does indeed satisfy the local maximum principle.

Next suppose only that $\Delta u \ge 0$. The goal is to show that if v is a harmonic function in a small disk, say in $D(0, r)$, and if $u \le v$ on the boundary of the disk, then $u \le v$ inside the disk. If ε is an arbitrary positive number, then $u(z) + \varepsilon|z|^2$ has strictly positive Laplacian, and $u(z) + \varepsilon|z|^2 \le v(z) + \varepsilon r^2$ on the boundary of the disk, so the previous case implies that $u(z) + \varepsilon|z|^2 \le v(z) + \varepsilon r^2$ inside the disk. Now let ε go to zero.

Conversely, suppose that a twice continuously differentiable function u is subharmonic. Why is $\Delta u \ge 0$? In the contrary case, Δu would be negative on some open set. By what was just proved, the function $-u$ would be subharmonic on that set. Then both $u - v$ and $-u - (-v)$ would satisfy the maximum principle for every harmonic function v. Setting v equal to the Poisson integral of u on a small disk implies that u is equal to its local Poisson integral, that is, u is harmonic. Hence Δu cannot be negative after all.

Example. Here are some standard ways to produce subharmonic functions.

- $|f|$ when f is holomorphic. (The subharmonicity is easy to check from the mean-value property.)

- $|f|^p$ when p is a positive number and f is holomorphic. (At zeroes of f, the sub-mean-value property is immediate. Away from zeroes of f, there is a local holomorphic branch of f^p, so the subharmonicity follows from the preceding example.)

- $u \circ f$, where u is subharmonic and f is holomorphic. (When u is twice continuously differentiable, compute that $\Delta(u \circ f) = |f'|^2 (\Delta u) \circ f$. In general, approximate u by smooth subharmonic functions, which can be done by convolving with a mollifier.)

- $\alpha u_1 + \beta u_2$, where u_1 and u_2 are subharmonic, and α and β are *nonnegative* real numbers. (This case is clear from the sub-mean-value property.)

- $\max(u_1, u_2)$ (pointwise maximum), where u_1 and u_2 are subharmonic. (This case is clear from the sub-mean-value property.)

- More generally, suppose $\{u_t\}_t$ is a family of subharmonic functions, and consider the pointwise supremum $\sup_t u_t(z)$. In general, this envelope need not be upper semicontinuous, but if the envelope is upper semicontinuous, then the envelope is subharmonic.

  [Aside: Here is an example of failure of upper semicontinuity of the envelope. The function $(1/n) \log |z|$ is subharmonic and negative in the unit disk for each natural number n. The pointwise supremum of this sequence of functions equals 0 on the punctured disk but $-\infty$ at the center, hence is not upper semicontinuous. On the other hand, for a family that is locally bounded above, the upper semicontinuous regularization of the envelope is subharmonic.]

  To see why the envelope is subharmonic, apply the sub-mean-value property. If a positive ε is specified, and a point z_0 is specified, then there is some parameter value t_0 such that

$$\begin{aligned} \sup_t u_t(z_0) &\leq u_{t_0}(z_0) + \varepsilon \leq \frac{1}{2\pi} \int_0^{2\pi} u_{t_0}(z_0 + re^{i\theta})\, d\theta + \varepsilon \\ &\leq \frac{1}{2\pi} \int_0^{2\pi} \sup_t u_t(z_0 + re^{i\theta})\, d\theta + \varepsilon. \end{aligned}$$

  Letting ε go to 0 shows that the upper envelope has the sub-mean-value property.

- $\log(1 + |z|)$ is subharmonic. In principle, the subharmonicity can be verified by computing second derivatives, but the calculation is nasty. Here is a trick. Observe that

$$\log(1 + |z|) = \sup_\theta \log |1 + e^{i\theta} z|,$$

  by the triangle inequality. For each fixed θ, the function $\log |1 + e^{i\theta} z|$ is subharmonic, being the logarithm of the modulus of a holomorphic function. The envelope is not only upper semicontinuous but even continuous. Hence the preceding example shows that $\log(1 + |z|)$ is subharmonic.

- If u is subharmonic in a region, and D is a closed disk in the region, build a new function by replacing u inside D by the Poisson integral of the value of u on ∂D. Then u satisfies the mean-value property at points inside D and the sub-mean-value property at points outside D. What about points on ∂D? The original function satisfies the sub-mean-value property at these points, and the Poisson integral is at least as large as u inside D, so the average value of the new function increases. Hence the sub-mean-value property can only improve. Thus local "Poissonization" of a subharmonic function produces a new subharmonic function.

Three-lines theorems

Since the modulus of a holomorphic function is a subharmonic function, many versions of the maximum principle are most naturally stated in the context of subharmonic functions. Here is one example that appears in applications.

Theorem. *Suppose u is subharmonic in a strip $\{\,(x,y) \in \mathbb{R}^2 \,:\, a < x < b\,\}$, and u is bounded above. Let $M(x)$ denote $\sup\{\,u(x,y) \,:\, y \in \mathbb{R}\,\}$. Then $M(x)$ is a convex function of x on the interval (a,b).*

The word "convex" is understood in the usual sense of real analysis: namely, if x_1 and x_2 are two arbitrary points in the interval (a,b), and t is a real number between 0 and 1, then

$$M(tx_1 + (1-t)x_2) \le tM(x_1) + (1-t)M(x_2).$$

The geometric content of the inequality is that the graph of M lies below each chord: convex functions are "sublinear."

The reason for the name "three lines" is that bounds on the function on two lines control the size of the function on any third line in between.

Proof. Since subharmonicity is a property that is preserved by translations and by dilations, there is no loss of generality in supposing that $a = -\pi/2$ and $b = \pi/2$. Suppose x_1 and x_2 are two numbers such that $-\pi/2 < x_1 < x_2 < \pi/2$. What needs to be shown is that if p is a first-degree polynomial such that $M(x_1) \le p(x_1)$ and $M(x_2) \le p(x_2)$, then $M(x) \le p(x)$ whenever $x_1 < x < x_2$.

View $p(x)$ as a harmonic function that is independent of y. For an arbitrary positive ε, consider the function

$$u(x,y) - p(x) - \varepsilon \operatorname{Re}\cos(x+iy). \tag{2}$$

Since the real part of the cosine is strictly positive in the strip where $|x| < \pi/2$, the indicated function (2) is negative on the vertical lines where $x = x_1$ and $x = x_2$. Moreover, the real part of $\cos(x+iy)$ equals $\cos(x)\cosh(y)$, which tends to $+\infty$ uniformly with respect to x between x_1 and x_2 when $|y| \to \infty$. Accordingly, for sufficiently large R, the function (2) is negative on the horizontal line segments where $y = \pm R$ and $x_1 \le x \le x_2$.

The function (2) is the difference between a subharmonic function and a harmonic function, so the maximum principle for bounded regions implies that for every sufficiently large R, the expression (2) is negative on the rectangular region where $x_1 \le x \le x_2$ and $|y| \le R$. Letting R tend to infinity shows that the expression (2) is negative on the whole strip where $x_1 \le x \le x_2$. Letting ε tend to zero shows that $u(x, y) \le p(x)$ when $x_1 \le x \le x_2$. Taking the supremum over y shows that $M(x) \le p(x)$ when $x_1 \le x \le x_2$, as claimed. □

Remark. The proof reveals that the hypothesis of boundedness of u can be relaxed to a hypothesis that u does not grow too fast at infinity. For instance, if there are positive constants A and B, with B strictly less than 1, such that $u(x, y) \le Ae^{B|y|}$ when $-\pi/2 < x < \pi/2$, then boundedness of u on two lines implies boundedness on the region between the two lines. For an interval (a, b), the requirement is that $B < \pi/(b - a)$. Generalizations along these lines are part of so-called Phragmén–Lindelöf theory.

Corollary. *Suppose f is holomorphic, not identically zero, and bounded in a vertical strip. Let $M(x)$ denote* $\sup\{\, |f(x + iy)| \,:\, y \in \mathbb{R} \,\}$. *Then* $\log M(x)$ *is a convex function; equivalently, if x_1 and x_2 are real numbers in the strip, and* $0 < t < 1$, *then*

$$M(tx_1 + (1 - t)x_2) \le M(x_1)^t M(x_2)^{1-t}.$$

Proof. Apply the preceding theorem to the subharmonic function $\log |f|$ and exponentiate the convexity inequality. □

Perron's method

Suppose φ is a given function on the boundary of a bounded region. Consider the class of all subharmonic functions in the region whose boundary values do not exceed those of φ. Take the pointwise supremum of all such subharmonic functions. If there is a solution of the Dirichlet problem, then this construction must yield the solution.

Indeed, the putative solution is in the class. Moreover, the putative solution is an upper bound for all subharmonic functions with the given boundary values.

The question, then, is whether the envelope actually does solve the Dirichlet problem. The counterexample mentioned earlier (the punctured disk) shows that some information about the boundary has to come into play. The essential element turns out to be the existence or non-existence of subharmonic peak functions. A *peak function* at a boundary point z_0 of a region G is a negative function u on G such that $\lim_{z \to z_0} u(z) = 0$ and $\limsup_{z \to w} u(z) < 0$ when $w \in \partial G \setminus \{z_0\}$.

Theorem (Solvability of the Dirichlet problem)**.** *If G is a bounded region in $\mathbb{C}$ such that G admits a subharmonic peak function at each boundary point, and if φ is a continuous function on the boundary of G, then there exists a harmonic function u on G such that* $\lim_{z \to w} u(z) = \varphi(w)$ *for every point w in the boundary of G.*

Moreover, if $\mathcal{F}$ is the Perron family consisting of all subharmonic functions v on G such that $\limsup_{z \to w} v(z) \le \varphi(w)$ *for every w in the boundary of G, then* $u(z) = \sup_{v \in \mathcal{F}} v(z)$ *for every z in G.*

The proof has two parts. The first part is to show that the envelope of the Perron family is harmonic. That conclusion holds even without the hypothesis of the existence of peak functions. The second part is to show that peak functions force the envelope of the Perron family to have the right boundary values.

In Perron's method, a needed fact is that if φ is the boundary value of a function u that is subharmonic in a neighborhood of the closed disk, then the Poisson integral of φ is at least as large as u inside the disk. Since the previous discussion about the Poisson integral used continuity of the boundary values, some further argument is needed to handle subharmonic boundary values.

The necessary proposition is that every upper semicontinuous function on a compact set (or on any set where the function is bounded above) is the limit of a decreasing sequence of continuous functions. Namely, let $\varphi_n(w)$ be $\sup_t\{\varphi(t)-n|t-w|\}$. (To see the point of this definition, consider the case of a function that is constant except for a jump at one point.) When $t = w$, the expression in brackets equals $\varphi(w)$, so $\varphi_n(w) \geq \varphi(w)$. Moreover, for each fixed t the expression in brackets decreases as n increases, so the sequence $\{\varphi_n\}$ is decreasing. If M is an arbitrary number larger than $\varphi(w)$, then by upper semicontinuity there is a neighborhood of w such that $\varphi(t) < M$ for t in the neighborhood. On the other hand, the quantity $|t - w|$ is bounded away from 0 outside the neighborhood, and φ is bounded above, so $\varphi(t) - n|t - w| \to -\infty$ uniformly outside the neighborhood when $n \to \infty$. It follows that $\varphi_n(w) < M$ for large n. Since M is arbitrary, the limit to which the decreasing sequence $\{\varphi_n(w)\}$ converges is $\varphi(w)$. What remains to see is that φ_n is continuous. For arbitrary points w_1 and w_2, the triangle inequality implies that

$$\varphi(t) - n|t - w_1| \geq \varphi(t) - n|t - w_2| - n|w_1 - w_2| \qquad \text{for each } t,$$

so $\varphi_n(w_1) \geq \varphi_n(w_2) - n|w_1 - w_2|$. Interchanging w_1 and w_2 then shows that φ_n is a Lipschitz function with Lipschitz constant equal to n. In particular, φ_n is continuous.

Returning to the Poisson integral, suppose that v is the Poisson integral of the boundary value of a subharmonic function u. Approximate the boundary value by a decreasing sequence $\{u_n\}$ of continuous functions. Let v_n be the Poisson integral of u_n. Then v_n has the boundary values of u_n, so v_n is a harmonic function than exceeds u on the boundary, whence v_n exceeds u inside the disk. The functions v_n decrease inside the disk by the maximum principle. By the monotone convergence theorem for integrals, the functions v_n converge to v, which therefore dominates u inside the disk.

This argument has a further implication. By Harnack's principle, the limiting function v is harmonic and not identically $-\infty$ (unless u is identically $-\infty$). Consequently, a subharmonic function (not identically $-\infty$) is integrable on each circle (that is, the integral is not $-\infty$). For similar reasons, subharmonic functions are area-integrable.

Return to the solution of the Dirichlet problem

Proof of the harmonicity of the Perron envelope. Suppose that G is a bounded domain, and φ is a bounded function on the boundary. (For this part of the proof, the continuity of φ is not needed.) The goal is to show that the envelope u of the Perron family is harmonic.

Recall that a function v belongs to the Perron family if and only if v is subharmonic, and $\limsup_{z\to w} v(z) \le \varphi(w)$ for every point w in the boundary of G. If M is a constant larger than the upper bound on φ, then every function v in the Perron family has the property that $v - M$ is negative near the boundary of G and hence is negative everywhere inside G (by the maximum principle; the boundedness of the domain G is used here). Therefore every function in the Perron family is bounded above by M. Hence u, the envelope, is bounded above by M.

It suffices to verify harmonicity—a local property—on an arbitrary disk $D(z_0, r)$ whose closure is contained in G. Let $\{v_n\}$ be a sequence of subharmonic functions in the Perron family such that the sequence $\{v_n(z_0)\}$ increases up to $u(z_0)$. Replacing each v_k by $\max\{v_1, \ldots, v_k\}$ ensures that the sequence $\{v_n\}$ is increasing at each point of G.

Next replace each v_k with its "Poissonization" inside $D(z_0, r)$ to ensure that v_k is harmonic inside the disk. The modified sequence $\{v_n\}$ now is an increasing sequence in the Perron family, and inside $D(z_0, r)$ this sequence is an increasing sequence of harmonic functions that converges at z_0 to $u(z_0)$. By Harnack's principle, the limit of the sequence $\{v_n\}$ is a harmonic function v^* inside $D(z_0, r)$.

The proof is not finished, for what is known so far is that u, the Perron envelope, matches v^*, a harmonic function, at one point. Does u match v^* at other points of $D(z_0, r)$ besides z_0?

Suppose z_1 is an arbitrary point of $D(z_0, r)$. Repeat the preceding construction to obtain an increasing sequence $\{u_n\}$ in the Perron family such that the sequence $\{u_n(z_1)\}$ converges to $u(z_1)$. Replacing each u_k by $\max(u_k, v_k)$ gives a new increasing sequence of subharmonic functions in the Perron family that converges to u at both points z_0 and z_1. Poissonizing as before produces a harmonic limit function u^* in $D(z_0, r)$ that matches u at both z_0 and z_1.

By construction, $v^* - u^* \le 0$ in $D(z_0, r)$, and $v^*(z_0) = u(z_0) = u^*(z_0)$. By the maximum principle, the harmonic function $v^* - u^*$ is identically equal to 0 in $D(z_0, r)$. Consequently, $v^*(z_1) = u^*(z_1) = u(z_1)$. Since z_1 is arbitrary, the function v^* is a harmonic function in $D(z_0, r)$ that equals the envelope u in all of $D(z_0, r)$. Thus the envelope is harmonic (in all of G, since z_0 is arbitrary). □

Proof that peak functions imply the right boundary values. Suppose now that the boundary function φ is continuous at z_0 and that there is a subharmonic peak function at z_0. The claim is that the Perron envelope function u has limit $\varphi(z_0)$ at z_0. There is no loss of generality in supposing that $\varphi(z_0) = 0$. (Simply subtract $\varphi(z_0)$ from all functions.)

Fix a positive ε. The goal is to find a neighborhood of z_0 such that $-\varepsilon < u(z) < \varepsilon$ when z is a point of G lying in the neighborhood. Since φ is continuous at z_0, there is a radius r such that $-\varepsilon/2 < \varphi(z) < \varepsilon/2$ when z is a point of ∂G for which $|z - z_0| < r$.

Let ψ be a subharmonic function peaking at z_0. The intersection of the boundary of G with the set $\{ z \in \mathbb{C} : |z - z_0| \ge r \}$ is compact, and each point z of this compact set has a neighborhood N_z such that the upper semicontinuous function ψ is negative on $N_z \cap G$. Taking a finite subcover shows that there is an open neighborhood U of the compact set $\{ z \in \mathbb{C} : |z - z_0| \ge r \} \cap \partial G$ such that the function ψ has a negative upper bound on $U \cap G$, say $-\delta$.

Let M be a large positive constant such that $M\delta$ exceeds the supremum of $|\varphi|$ on ∂G. If w is a point of ∂G at distance at least r from z_0, then $\limsup_{z\to w} M\psi(z) \le -M\delta < \varphi(w)$. On

the other hand, if w is a point of ∂G within distance r from z_0, then $\limsup_{z\to w} M\psi(z) \le 0 < \varphi(w) + \varepsilon/2$. Therefore the function $M\psi - \varepsilon/2$ belongs to the Perron family associated to the boundary function φ. Accordingly, $M\psi(z) - \varepsilon/2 \le u(z)$ for every point z in G. By the definition of peak function, $\lim_{z\to z_0} M\psi(z) = 0$, so there is a neighborhood of z_0 in which $-\varepsilon/2 < M\psi(z)$. In this neighborhood, $-\varepsilon < u(z)$.

Similarly, $\limsup_{z\to w} M\psi(z) < -\varphi(w)$ when w is a point of ∂G at distance at least r from z_0, and $\limsup_{z\to w} M\psi(z) \le 0 < -\varphi(w) + \varepsilon/2$ when w is a point of ∂G within distance r from z_0. Consequently, if v is an arbitrary member of the Perron family, then $\limsup_{z\to w}(v + M\psi - \varepsilon/2) < 0$ for every point w in ∂G. Since $v + M\psi - \varepsilon/2$ is subharmonic, the maximum principle implies that $v + M\psi - \varepsilon/2$ is negative everywhere inside G. Thus $v < -M\psi + \varepsilon/2$ inside G. Taking the pointwise supremum over functions v in the Perron family shows that $u \le -M\psi + \varepsilon/2$. Since $\lim_{z\to z_0} -M\psi(z) = 0$, there is a neighborhood of z_0 in which $-M\psi(z) < \varepsilon/2$. In this neighborhood, $u(z) < \varepsilon$.

In conclusion, there is a neighborhood of z_0 such that $-\varepsilon < u(z) < \varepsilon$ when z is a point of G in the neighborhood. Since ε is arbitrary, $\lim_{z\to z_0} u(z) = 0$, as claimed. □

Remark on barriers

The term "barrier" was introduced by Henri Lebesgue in his note "Sur le problème de Dirichlet" in *Comptes rendus hebdomadaires des séances de l'Académie des sciences* **154** (1912) 335–337. For Lebesgue, a barrier was a harmonic peak function (more precisely, a family of functions obtained from a harmonic peak function). He obtained a peak function at a boundary point z_0, under the hypothesis of solvability of the Dirichlet problem, by finding a harmonic function with boundary values matching the distance to z_0; evidently such a harmonic function is positive except at z_0, where the function takes the extreme value 0.

Lebesgue's main point in the note was to provide an algorithm for solving the Dirichlet problem under the assumption that there is a solution. Suppose given a continuous function on the boundary of a bounded open set in the plane. Extend the function arbitrarily to a continuous function on the closed region, say by the Tietze extension theorem. (For Lebesgue, the continuous function was given initially on the closed region.) Execute the following algorithm.

Replace the value of the function at each point by the average value over the largest disk centered at the point and contained in the region (two-dimensional average over the disk, not one-dimensional average over the boundary circle). Repeat the averaging process for the new function that arises, and iterate.

The sequence of averages converges uniformly on the closed region to the solution of the Dirichlet problem, assuming the existence of a harmonic barrier at each boundary point. (The same argument works assuming the existence of a subharmonic peak function at each boundary point.)

Construction of peak functions

When do subharmonic peak functions exist? Examples in the homework assignment reveal that there cannot be a subharmonic peak function at the center of a punctured disk.

But subharmonic peak functions do exist at reasonable boundary points. The construction is easy at points where there is a supporting line, straightforward at points that are accessible from the exterior by a line segment, and difficult for boundary points about which all that is known is that the point is not a singleton boundary component.

Example. If G is a convex domain, in the sense that at each boundary point there is a supporting line that intersects the (open) domain at no other point, then there is a harmonic peak function. Indeed, a translation puts the boundary point at the origin, and a rotation makes the imaginary axis the supporting line, with the domain lying in the right-hand half-plane. If the domain is strongly convex (no boundary point besides the origin lies on the imaginary axis), then $-\operatorname{Re} z$ is a peak function. If the domain is only weakly convex, then $-\operatorname{Re}\sqrt{z}$ is a peak function.

Example. Suppose z_0 is a boundary point of a domain with the property that there is a line segment lying in the complement of the domain with one endpoint at z_0. Then there is a peak function at z_0.

In particular, a domain bounded by a finite number of smooth curves admits peak functions at all boundary points. The boundary curves can even have cusps. Moreover, the region can have some straight slits.

To construct the peak function, let z_1 be a second point on the indicated line segment. Use the linear fractional transformation $(z - z_0)/(z - z_1)$ to send z_0 to 0 and z_1 to ∞, and make a rotation to ensure that the line segment maps to the negative part of the real axis. If z_0 is the only point of the original line segment that lies on the boundary of the region, then use $\sqrt{z}$ to map into the right-hand half-plane, and take the negative of the real part of the image as the peak function. If the original line segment touches the boundary of the region at more than one point, then use a fourth root instead of a square root.

The goal now is to prove the much more general statement that if z_0 is a boundary point of G, and the connected component K of the complement of G containing z_0 contains at least one other point, then there is a subharmonic peak function at z_0. A linear fractional transformation makes it possible to put the point z_0 at 0 and a second point of K at ∞.

The complement of K is then a simply connected region containing G, so it is possible to define a holomorphic branch of $\log(z)$ on G. Notice that if G has a spiral structure, then the imaginary part of $\log(z)$ could take values in an unbounded set. This phenomenon causes a technical complication in the proof.

Fix a radius r. The immediate goal is to construct a subharmonic function u_r that is bounded between -1 and 0, takes the value -1 on the part of G outside $D(0, r)$, and has limit 0 at 0. This function is not yet the required peak function, for u_r might approach 0 at some points inside $D(0, r)$ other than the origin. But the convergent series $\sum_{n=1}^{\infty} 2^{-n} u_{1/n}(z)$ will serve as the peak function.

Having fixed r, observe that a branch of $\log(z/r)$ on G has negative real part on $G \cap D(0, r)$ that

tends to $-\infty$ when z approaches the origin. Moreover, $\log(z/r)$ maps $G \cap \partial D(0,r)$ bijectively to an open subset of the imaginary axis, that is, to a union of disjoint open intervals in the imaginary axis of total length at most 2π. The construction depends on those intervals.

There are (at most) countably many intervals, say I_k with center ic_k and length δ_k. The easy case occurs when all the intervals lie in a bounded subset of the imaginary axis, say between $-iM$ and $+iM$, where $M > 0$. Then

$$(2 + M^2)\,\mathrm{Re}\,\frac{1}{z-1}, \qquad \text{or} \qquad (2 + M^2)\frac{(x-1)}{(x-1)^2 + y^2},$$

is a negative harmonic function in the left-hand half-plane that tends to 0 when $x \to -\infty$. At points of the intervals I_k on the imaginary axis, this function tends to a negative limit that is smaller than -1 (by the choice of M). Composing this function with $\log(z/r)$ gives a negative harmonic function v in $G \cap D(0,r)$. The function that equals -1 in $G \setminus D(0,r)$ and $\max(-1, v)$ in $G \cap D(0,r)$ is the required function u_r.

Now consider the general case that the intervals I_k are not contained in a bounded region. For each positive number c, the series

$$\sum_{k=1}^{\infty} 2\delta_k\,\mathrm{Re}\,\frac{1}{z - \delta_k - ic_k}, \qquad \text{or} \qquad \sum_{k=1}^{\infty} 2\delta_k \frac{x - \delta_k}{(x-\delta_k)^2 + (y-c_k)^2}, \tag{3}$$

converges absolutely and uniformly on the half-plane where $x \le -c$. Indeed,

$$\left| 2\delta_k \frac{x - \delta_k}{(x-\delta_k)^2 + (y-c_k)^2} \right| \le \frac{2\delta_k |x - \delta_k|}{(x-\delta_k)^2} = \frac{2\delta_k}{|x| + \delta_k} < \frac{2}{|x|}\delta_k,$$

and the series $\sum_k \delta_k$ converges (to a value no larger than 2π). Thus the series (3) represents a negative harmonic function in the left-hand half-plane that tends to 0 when $x \to -\infty$. Since each term of the series is negative, the series is smaller than any one particular term. When $x + iy$ approaches a point in I_k from within the left-hand half-plane, the series approaches a value (possibly $-\infty$) no larger than

$$\frac{-2\delta_k^2}{\delta_k^2 + (y - c_k)^2}.$$

Since $|y - c_k| \le \delta_k/2$ when $iy \in I_k$, the preceding fraction is at most $-8/5$, hence less than -1. As in the preceding case, composing (3) with $\log(z/r)$ gives a harmonic function v in $G \cap D(z,r)$ that is less than -1 near $G \cap \partial D(0,r)$, so taking the maximum with the constant -1 and extending to $G \setminus D(0,r)$ with the constant -1 gives the required subharmonic function u_r.

Analytic continuation, part I

Another application of the Poisson integral is the following useful tool for extending holomorphic functions from a domain to a larger domain.

Theorem (Schwarz reflection principle)**.** *Suppose G is a connected open set that is symmetric with respect to the real axis, and f is a holomorphic function defined in the open upper half of G, say G^+. If* $\operatorname{Im} f(z)$ *approaches* 0 *whenever z approaches a point of the intersection of G with the real axis, then there exists a holomorphic function defined in all of G that agrees with f in G^+. This "analytic continuation" of f is unique: when z lies in the lower half of G, the value of the extended function at z equals $\overline{f(\overline{z})}$.*

The intuitive formulation of the hypothesis is that f maps a segment of the real axis into the real axis. The theorem can be stated that way if f is known ahead of time to be continuous onto the real axis. The more general statement above is needed in some applications. The conclusion of the theorem shows that, somewhat surprisingly, the hypothesis about $\operatorname{Im} f$ forces $\operatorname{Re} f$ to extend continuously to the real axis.

Notice that the function $1/z$ appears to map the real axis to the real axis but fails to extend from the upper half-plane to the whole plane. The problem is that $\operatorname{Im}(1/z)$ does not have a finite limit when $z \to 0$. The theorem is still applicable but shows only that $1/z$ extends holomorphically from the upper half-plane to the punctured plane.

Proof of the Schwarz reflection principle. The first observation is that when z lies in the lower half of G, the function that sends z to $\overline{f(\overline{z})}$ is holomorphic. The Cauchy–Riemann equations provide one way to verify the holomorphicity. Alternatively, use that holomorphicity is a local property, and argue as follows. If z_0 is a point in G^+, then $f(z)$ admits a local power series expansion when z is near z_0 of the form $\sum_{n=0}^{\infty} a_n(z - z_0)^n$. Now if z is close to $\overline{z_0}$ in the lower half of G, then the value $\overline{f(\overline{z})}$ is represented by the convergent series $\sum_{n=0}^{\infty} \overline{a_n}(z - \overline{z_0})^n$, hence corresponds to a holomorphic function.

The proof is easy to complete under the extra hypothesis that f is continuous onto the real axis. In this situation, $f(z) = \overline{f(\overline{z})}$ when z lies on the real axis, so the extended function is at least continuous in G, holomorphic in the open upper half of G, and holomorphic in the open lower half of G. Holomorphicity in a neighborhood of the real axis follows from Morera's theorem. Indeed, the integral of f over a simple closed curve in G can be rewritten by adding and subtracting a piece of contour over the real axis. This trick produces two closed contours, one in the closed upper half-plane and one in the closed lower half-plane. Bump each contour into the appropriate open half-plane to see that the integral equals zero.

More work is needed to finish the proof under the weaker hypothesis that $\operatorname{Im} f$ approaches zero on the real axis, for this hypothesis implies continuity of $\operatorname{Im} f$ but does not immediately imply continuity of $\operatorname{Re} f$ onto the real axis. At any rate, $\operatorname{Im} f$ is a continuous function in G satisfying the mean-value property on sufficiently small disks centered at points off the real axis, and the mean-value property holds in small disks centered at points on the real axis for the following reason. The value of $\operatorname{Im} f$ equals 0 at the center of such a disk, and the values of $\operatorname{Im} f$ on the top half of the disk are the negatives of the values on the bottom half, so the integral around the boundary of the disk equals 0 by symmetric cancellation. Accordingly, the function $\operatorname{Im} f$ is harmonic throughout the domain G.

Continuing to work in a disk centered at a point on the real axis, observe that the harmonic function $\operatorname{Im} f$ is the imaginary part of some holomorphic function g in the disk. In the open upper half of the disk, the holomorphic function $f - g$ has vanishing imaginary part and so is equal to some real constant c. The function $g + c$ provides a holomorphic extension of f from the open upper half disk to the whole disk. Moreover, the holomorphic function $g(z) - \overline{g(\overline{z})}$ is zero on the real axis, hence identically zero by the identity principle. Thus the extension to the lower half disk is uniquely determined by the symmetry property that $f(z) = \overline{f(\overline{z})}$. □

Mapping by linear fractional transformations leads to generalizations of the reflection principle. For example, if f is holomorphic in the unit disk, and $|f(z)| \to 1$ when $|z| \to 1$, then f extends holomorphically to a circularly symmetric region and has the property that $f(1/\overline{z}) = 1/\overline{f(z)}$.

The modular group

A famous group that is important in both number theory and geometry is the so-called *modular group*, which is the group of linear fractional transformations that can be represented using integer coefficients corresponding to a matrix with determinant equal to 1. More explicitly, every such transformation sends z to $\dfrac{az+b}{cz+d}$, where $ad - bc = 1$. A concrete example is $\dfrac{3z+2}{4z+3}$. This group is sometimes designed $PSL(2, \mathbb{Z})$, where $\mathbb{Z}$ indicates the integers, the letter L stands for linear, the letter S stands for special (determinant equal to 1), and the letter P stands for projective (since the coefficients are unique only up to change of sign).

Since the coefficients are, in particular, real numbers, every such transformation maps the extended real line to the extended real line. Therefore the upper half-plane maps either to the upper half-plane or to the lower half-plane. When the determinant is positive, the upper half-plane actually maps to the upper half-plane. Indeed, notice that

$$\operatorname{Im} \frac{az+b}{cz+d} = \operatorname{Im} \frac{ad\,z + bc\overline{z} + ac|z|^2 + bd}{|cz+d|^2} = \frac{(ad-bc)\operatorname{Im} z}{|cz+d|^2}.$$

Thus the modular group is a subgroup of the group of biholomorphisms of the upper half-plane.

The modular group has a *fundamental domain*, a region that contains exactly one point from each orbit. (An orbit is the set to which the group elements move a certain point.) The fundamental domain is not unique. A standard choice is the set of points in the upper half-plane for which $|z| > 1$ and $|z + \overline{z}| < 1$ (the second condition says that $-1/2 < \operatorname{Re} z < 1/2$). A part of the boundary needs to be included in the fundamental domain (which means that the set is actually not a "domain" in the usual sense of that word in complex analysis). A standard choice is to include the right-hand ray where $\operatorname{Re} z = 1/2$ and $\operatorname{Im} z \geq \sqrt{3}/2$, as well as the closed arc of the unit circle from $e^{\pi i/3}$ to i.

A convenient way to verify the above claims about the fundamental domain is to show that the modular group is generated by two elements, the translation sending z to $z + 1$ and the inversion sending z to $-1/z$. (Notice that the simple inversion $1/z$ maps the upper half-plane to the lower

half-plane, so the extra reflection by -1 is needed.) The corresponding matrices are

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

The second transformation has square equal to the identity (since the square of the matrix is the negative of the identity matrix, and that matrix represents the same linear fractional transformation as does the identity matrix). The first transformation followed by the second transformation is a group element whose cube equals the identity. These two relations turn out to be the only independent ones, so the group has the presentation $\{ (s,t) : s^2 = 1, (st)^3 = 1 \}$.

An alternative set of generators is the pair of transformations defined by the matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, perhaps a more natural choice from the point of view of symmetry. Since

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

the two pairs of alleged generators do generate the same group.

Here is a strategy for proving that the fundamental domain is as claimed.

- Show that distinct group elements move the fundamental domain to sets whose interiors do not intersect. Equivalently, show that a nontrivial group element moves the interior of the fundamental domain to a disjoint set: no interior point of the fundamental domain can move to another interior point of the fundamental domain. (On the boundary, the point i is fixed by $z \mapsto -1/z$, and $e^{\pi i/3}$ is fixed by $z \mapsto 1 - (1/z)$, and $e^{2\pi i/3}$ is fixed by $z \mapsto -1 - (1/z)$.)
- Show that an arbitrary point in the upper half-plane can be moved into the fundamental domain by some composition of generators.
- Show that the alleged generators really do generate the whole group.

In class, you provided details to fill out the preceding plan, approximately as follows.

Suppose that a nontrivial group element maps some interior point z_0 of the fundamental domain to an interior point (possibly the same one). There is no loss of generality in supposing that the second point has imaginary part at least as large as the imaginary part of the first point. (Interchange the roles of the two points, if necessary.) The formula for the imaginary part of a linear fractional transformation reveals that $|cz_0 + d|^2 \leq 1$.

Expanding, remembering that c and d are real, shows that

$$c^2|z_0|^2 + d^2 + 2|cd|\,\mathrm{Re}(z_0) \leq 1.$$

But $|z_0|^2 > 1$, and $-1/2 < \mathrm{Re}(z_0) < 1/2$, so as long as $c \neq 0$, the following inequality is strict:

$$c^2 + d^2 - |cd| < 1, \qquad \text{or} \qquad (|c| - |d|)^2 + |cd| < 1.$$

Consequently, $|cd| < 1$; but c and d are integers, so either $c = 0$ or $d = 0$.

If $c = 0$, then the determinant condition implies that $ad = 1$, so either $a = d = 1$ or $a = d = -1$; in either case, the linear fractional transformation is the identity, contrary to assumption. If $d = 0$, then the determinant condition implies that $|c| = 1$, so $1 \geq |cz_0+d|^2 = |z_0|^2$, contradicting that z_0 lies in the interior of the fundamental domain (which requires that $|z_0|^2 > 1$). In summary, the supposition that an interior point of the fundamental domain moves to another interior point via some nonidentity group element leads to a contradiction.

Next consider how to move an arbitrary point z into the fundamental domain by some composition of alleged generators. The first observation is that under all possible motions, there is one that achieves the maximum value of the imaginary part. To see why, observe again that the imaginary part of $\dfrac{az+b}{cz+d}$ equals $(\operatorname{Im} z)/|cz+d|^2$. Now $|cz+d| \geq |\operatorname{Im}(cz+d)| = |c| \operatorname{Im}(z)$. Accordingly, if $|c| \operatorname{Im}(z) > 1$, then the imaginary part of the image of z is less than $\operatorname{Im}(z)$. Therefore only the finitely many values of the integer c for which $|c| \leq 1/\operatorname{Im}(z)$ are candidates when seeking to maximize the imaginary part of the image of z. For any particular value of c, the quantity $(\operatorname{Im} z)/|cz+d|^2$ tends to zero when $|d| \to \infty$, so only finitely many values of d are candidates when seeking to maximize the imaginary part of the image of z. From the finite number of choices for c and d, evidently some choice achieves the maximum value for the imaginary part of the image. Composing with a translation puts the image point within the strip where the real part has absolute value not exceeding $1/2$.

Now the only issue is whether this image point could be inside the unit circle. But $\operatorname{Im}(-1/z) = (\operatorname{Im} z)/|z|^2$, so if $|z| < 1$, then the imaginary part can be increased by applying another transformation, contradicting the maximality. Therefore the image point is either in the interior of the fundamental domain or on the boundary. The two vertical boundaries are equivalent under translation, and the two circular arcs are equivalent under $z \mapsto -1/z$. Therefore an arbitrary point of the upper half-plane can be moved into the fundamental domain by a composition of alleged generators.

Finally, why do the alleged generators actually generate the modular group? Pick an arbitrary group element and a point z_0 in the interior of the fundamental domain. The specified group element moves that point somewhere. Compose with a suitable product of generators to move the image back into the fundamental domain. The composition is a group element that moves a point of the fundamental domain to another point of the fundamental domain. If the new image point is in the interior of the fundamental domain, then the first part of the argument shows that the composite function is the identity; that is, the specified group element is a product of generators. If the new image point is on the boundary of the fundamental domain, then the open mapping principle implies that an interior point close to z_0 maps to an interior point of the fundamental domain, reducing to the preceding case. Thus every group element is a composition of the alleged generators.

The third step can alternatively be handled directly, without using the other two steps. Namely, if translation and inversion generate a proper subgroup, then consider among the group elements outside the subgroup one that minimizes the value of $|c|$. That minimal value cannot be 0, for the determinant condition implies that if $c = 0$, then $a = d = 1$ or $a = d = -1$, so the transformation

reduces to the form $z \pm b$, which is a translation (thus in the subgroup). When $c \neq 0$, the Euclidean algorithm provides integers k and r, where $|r| < |c|$, such that $a = kc + r$. Then

$$\begin{pmatrix} 1 & -k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} r & b - kd \\ c & d \end{pmatrix}, \quad \text{so} \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ -r & -b + kd \end{pmatrix},$$

contradicting the miminality of c.

The congruence subgroup

Consider the subgroup of the modular group consisting of elements whose matrices are congruent modulo 2 to the identity matrix. Examples are

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}.$$

In general, the matrices in question have the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with a and d odd but b and c even. The claim is that a scheme similar to the preceding one shows that these two matrices generate the congruence subgroup, and a fundamental domain is the set of points in the strip in the upper half-plane where $\mathrm{Re}\,(z)$ has absolute value less than 1 and z lies outside the circles of radius $1/2$ with centers at $\pm 1/2$. The fundamental domain for the whole modular group has three boundary curves, but the fundamental domain for the congruence subgroup has four boundary curves (two of which should be included as part of the fundamental domain). Details are available in the textbook.

The modular function

By the Riemann mapping theorem, the right-hand half of the fundamental domain of the congruence subgroup can be mapped to the upper half-plane, with the boundary mapping to the real axis, and the points 0 and 1 fixed. Schwarz reflection across the imaginary axis extends the function to be holomorphic on the whole fundamental domain. Now the function extends to be holomorphic on the whole upper half-plane either by iterated Schwarz reflection or by defining the function to be invariant under the action of the congruence subgroup. In other words, the modular function λ has the property that $\lambda(g(z)) = \lambda(z)$ for every element g of the congruence subgroup. The modular function is locally injective on the open upper half-plane and defines an infinite-sheeted covering of $\mathbb{C} \setminus \{0, 1\}$, the twice-punctured plane.

Analytic continuation along curves

A standard, general method for extending the domain of a holomorphic function (a process traditionally called "analytic" continuation rather than "holomorphic" continuation) is continuation along a curve. Suppose $\gamma : [0, 1] \to \mathbb{C}$ is a continuous function (an arc, possibly self-intersecting),

and f_0 is a convergent power series centered at $\gamma(0)$. An alternative terminology is that f_0 is a *germ* of a holomorphic function at $\gamma(0)$. A germ at a point is an equivalence class of holomorphic functions, two functions being equivalent if they agree in some neighborhood of the point.

An *analytic continuation along* γ is a family (f_t), each f_t being a germ of a holomorphic function at $\gamma(t)$, or equivalently a convergent power series in powers of $z - \gamma(t)$, satisfying the following local compatibility condition. For every t there exists a positive δ such that if $|s-t| < \delta$, then

1. $|\gamma(s) - \gamma(t)|$ is less than the radius of convergence of f_t, and

2. the power series defined by f_t, which according to the preceding condition converges in a neighborhood of $\gamma(s)$, determines a germ at $\gamma(s)$ that equals f_s.

The intuition is that there is a chain of overlapping disks covering the curve such that on each disk there is an analytic continuation of the function from the preceding disk.

Remark. 1. The radius of convergence of f_t is a continuous function of t. Indeed, if s is close to t, then $\gamma(s)$ is close to $\gamma(t)$, say $|\gamma(t) - \gamma(s)| < \varepsilon$. Accordingly, the radius of convergence of f_s is at least the radius of convergence of f_t minus ε. By symmetry, the radius of convergence of f_t is at least the radius of convergence of f_s minus ε. Put the two inequalities together.

2. Consequently, there is a positive lower bound for the radius of convergence of f_t as t runs over the interval $[0, 1]$.

3. Analytic continuation of f_0 along a specified curve γ is unique. For suppose (f_t) and (g_t) are two continuations. Then $f_t = g_t$ for t close to 0 (within the radius of convergence of f_0). Suppose T is the supremum of values of t for which $f_t = g_t$. If s is so close to T (from below) that $\gamma(s)$ and $\gamma(T)$ differ by less than the minimal radius of convergence, then the equality of f_s and g_s implies the equality of f_t and g_t for some values of t larger than T. Hence equality holds for all t.

Example. If γ_1 and γ_2 are two curves with the same endpoints, that is, $\gamma_1(0) = \gamma_2(0)$ and $\gamma_1(1) = \gamma_2(1)$, then the analytic continuation of f_0 along γ_1 is *not* necessarily equal to the analytic continuation of f_0 along γ_2.

Suppose, for instance, that $\gamma_1(t) = \exp(\pi i t)$ and $\gamma_2(t) = \exp(-\pi i t)$. Let f_0 be $\sqrt{z}$, defined in polar coordinates near 1 as $r^{1/2}e^{i\theta/2}$, with θ close to 0. Analytic continuation of f_0 along γ_1 has to have the angle changing continuously through positive values, which means that the continued value of $\sqrt{z}$ near -1 is $\exp(i\pi/2)$, or i. On the other hand, the analytic continuation of f_0 along γ_2 has the angle changing continuously through negative values, giving a value of $\sqrt{z}$ near -1 equal to $\exp(-i\pi/2)$, or $-i$.

The question arises of when analytic continuation along two different curves does lead to a unique value. The following theorem gives an answer.

Theorem (Monodromy theorem)**.** *Suppose γ_1 and γ_2 are two curves with the same endpoints in a region G. Suppose f_0 is a convergent power series at $\gamma_1(0)$ (which equals $\gamma_2(0)$). Suppose f_0 admits analytic continuation along* every *curve in G. If γ_1 and γ_2 are homotopic in G, then analytic continuation of f_0 along γ_1 matches the analytic continuation of f_0 along γ_2.*

What "homotopic" (strictly speaking, "fixed endpoint homotopic") means is that γ_1 can be continuously deformed into γ_2 within G. More formally, there exists a continuous function $F : [0,1] \times [0,1] \to G$ such that $F(0,t) = \gamma_1(t)$, $F(1,t) = \gamma_2(t)$, $F(s,0) = \gamma_1(0)$, and $F(s,1) = \gamma_1(1)$.

Notice that in the preceding example with the square-root function, the two curves are not homotopic in $\mathbb{C} \setminus \{0\}$, which is the region in which f_0 admits unrestricted analytic continuation.

Example. You know from Math 617 that a zero-free holomorphic function in a simply connected domain admits a holomorphic logarithm. A new way to deduce this property is to apply the monodromy theorem. Locally (in a small disk), a nonzero quantity has a logarithm, since the only ambiguity in determining a logarithm is the choice of the argument (angle). Analytic continuation to an overlapping disk is possible by matching the choice of angle at one point in the intersection of the two disks. Accordingly, unrestricted analytic continuation is possible. If the region is simply connected, then continuation along two different paths joining the same points leads to the same value, so a globally defined function appears.

Example. A problem on the January 2011 qualifying examination asks for a proof that the power series $\sum_{n=1}^{\infty} z^n/n^2$ can be analytically continued to $\mathbb{C} \setminus [1, \infty)$.

The idea is that $f'(z) = \sum_{n=1}^{\infty} z^{n-1}/n$, so $zf'(z) = \sum_{n=1}^{\infty} z^n/n = \log \frac{1}{1-z}$. Thus $f'(z)$ extends to be holomorphic on $\mathbb{C} \setminus [1, \infty)$ (notice that there is a removable singularity at 0).

Consequently, f admits unrestricted analytic continuation along curves in $\mathbb{C} \setminus [1, \infty)$: simply integrate the globally defined derivative to continue f along a curve. Since the region is simply connected, all curves are homotopic to each other. Hence f can be analytically continued to the whole region.

Proof of the monodromy theorem. If two parametrized curves are close to each other (so close that the separation $|\gamma_1(t) - \gamma_2(t)|$ is always less than the minimum of the radius of convergence of the function along γ_1), then the analytic continuation is the same along both curves. (The argument is the same as before: look at the supremum of values of t for which the analytic continuation at t matches on both curves.) Now look at the supremum of values of s for which analytic continuations along $F(0,t)$ and $F(s,t)$ match at the endpoint where $t = 1$. By the argument just indicated, the value of s can be bumped slightly. Hence analytic continuations along $F(0,t)$ and $F(1,t)$ match at the terminal point. □

Remark. The hypothesis that the two curves are homotopic is automatic if the region is simply connected. The hypothesis that the germ admits analytic continuation along every curve in the region is a hypothesis that is not obvious how to check in general.

Proof of Montel's fundamental normality criterion using the modular function

Since normality is a local property, there is no loss of generality in supposing that the common domain of the functions in the family is the unit disk. Each function f in the family has range contained in $\mathbb{C} \setminus \{0, 1\}$, the twice-punctured plane. The modular function λ is locally injective, so there is a local inverse of λ defined in a neighborhood of $f(0)$. The local inverse is highly nonunique, but there is a canonical way to single out a choice of local inverse: the one that maps $f(0)$ into the standard fundamental domain. The composition of f with this local inverse admits unrestricted analytic continuation within the unit disk: if the function has been continued to a neighborhood of a point z_0, and D is a disk overlapping the neighborhood, define $\lambda^{-1} \circ f$ in D by choosing a branch of λ^{-1} compatible with the existing choice in a neighborhood of $f(z_0)$. By the monodromy theorem, a global function $\widetilde{f}$ appears that maps the unit disk into the upper half-plane and has the property that $\lambda \circ \widetilde{f} = f$.

This construction with the modular function can be carried out for each function in the given family, producing a new family that of functions mapping the disk into the upper half-plane. A further composition with the Cayley transform (the function sending z to $(z - i)/(z + i)$, which maps the upper half-plane bijectively to the unit disk) produces a family of functions mapping the disk into the disk. This new family is bounded, hence normal by a simpler theorem of Montel (the one about local boundedness characterizing normality). Consequently, corresponding to a sequence (f_n) in the original family is a subsequence (n_k) such that the composition of the Cayley transform with $\widetilde{f}_{n_k}$ is a sequence converging normally on the unit disk to a holomorphic function (possibly a finite constant).

If the limit function is a nonconstant holomorphic function, then the range is contained in the open unit disk (by the maximum principle). Composing with the inverse Cayley transform and then with the modular function λ shows that the sequence (f_{n_k}) converges normally to a holomorphic function. The same argument applies when the initial limiting function is a constant that lies in the interior of the unit disk.

The difficult case occurs when the first limiting function is a constant of modulus 1. Undoing the Cayley transform shows that the sequence $(\widetilde{f}_{n_k})$ converges normally either to a boundary point of the upper half-plane or to ∞. In particular, the sequence $(\widetilde{f}_{n_k}(0))$ of complex numbers converges to a point of the extended boundary of the upper half-plane. Therefore the sequence $(f_{n_k}(0))$ of complex numbers cannot remain in a compact subset of $\mathbb{C} \setminus \{0, 1\}$.

Passing to a further subsequence shows that there is no loss of generality in supposing that the sequence $(f_{n_k}(0))$ converges either to 0 or to 1 or to ∞. The three cases are essentially equivalent, for the three points can be permuted by linear fractional transformations that map the twice-punctured plane to itself.

Suppose first that 1 is the accumulation point. Since the unit disk is simply connected, and the functions in the family omit the value 0, there is a holomorphic function g_{n_k} such that $g_{n_k}^2 = f_{n_k}$ and $g_{n_k}(0) \to -1$. Since g_{n_k} evidently omits the values 0 and 1, the previous analysis applies to the sequence (g_{n_k}) and shows that this sequence admits a subsequence converging normally

to a holomorphic function. Accordingly, the sequence (f_{n_k}) admits a subsequence converging normally to a holomorphic function.

If instead the point 0 is the accumulation point, apply the same argument to $1 - f_n$. And if ∞ is the accumulation point, apply the same argument to $1/f_n$ (which moves the accumulation point to 0). This final case is the one in which the extended sense of normality arises (allowing the point at ∞ as a limit).

Proof of Picard's great theorem

The theorem says that if f is holomorphic in a punctured disk, and there is an essential singularity at the puncture (that is, the Laurent series has infinitely many terms with negative exponents), then every complex number—with one possible exception—is in the range of the function. The same conclusion holds when the disk is shrunk, so an immediate consequence is that every value—with one possible exception—is taken infinitely often.

An exceptional value can occur: the function $e^{1/z}$ has an essential singularity at the origin and takes every nonzero value infinitely often in every punctured neighborhood of the origin. On the other hand, the function $\sin(1/z)$ has an essential singularity at the origin and takes every complex value infinitely often.

Picard's "little theorem" says that a transcendental (nonpolynomial) entire function takes every complex value—with one possible exception—infinitely often. Since an entire function that is not a polynomial can be viewed as having an essential singularity at infinity, the little theorem is a corollary of the great theorem.

The function e^z has 0 as an exceptional value. The function ze^z also has 0 as an exceptional value, since the value 0 is taken once but not infinitely often.

To prove the great theorem, suppose without loss of generality that the essential singularity is at 0. Seeking a contradiction, suppose there are two distinct complex numbers a and b that f takes only finitely many times. Shrinking the neighborhood reduces to the case that these two values are not taken at all. And considering the function $(f - a)/(b - a)$ reduces to the case that the omitted values are 0 and 1. Dilating the independent variable shows that the punctured neighborhood can be taken to be the punctured unit disk.

Define f_n via $f_n(z) = f(z/n)$, and consider the family (f_n) in the punctured disk. By Montel's fundamental normality criterion, this family is normal in the extended sense. There are two cases.

First suppose there is a subsequence (f_{n_k}) converging normally to a holomorphic function. The circle of radius $1/2$ is a compact set on which the subsequence is bounded, say by M. Accordingly, there is a sequence of annuli with outer radius $1/2$ and inner radius approaching 0 such that f is bounded by M on the boundary, hence on the whole annulus (by the maximum principle). Therefore f is bounded by M on the union of the annuli, which is the whole punctured disk of radius $1/2$. Then the singularity is removable, contrary to the hypothesis.

Second, suppose there is a subsequence converging normally to ∞. Considering the sequence of reciprocals gives a sequence converging normally to 0. By the preceding argument, the reciprocal of the original function has a removable singularity, and the singularity is removed by

setting the value at the origin to be 0. Hence the original function has a pole, again contrary to the hypothesis.

Thus the assumption that f omits two values contradicts the hypothesis that the singularity is essential. The proof of Picard's theorem is complete.