Lectures on the Theory of Functions of a Complex Variable

Vol. 2
Geometric Theory

Giovanni Sansone, Johan Gerretsen

P. Noordhoff

GIOVANNI SANSONE - JOHAN GERRETSEN

LECTURES ON THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE

II. GEOMETRIC THEORY

WOLTERS-NOORDHOFF PUBLISHING GRONINGEN 1969
THE NETHERLANDS

PREFACE

One of the most important events in the history of complex analysis was B. Riemann's discovery in the 1850s, viz., that each simply connected region which is not the whole complex plane can be mapped one-to-one and conformally onto the interior of a circle and that consequently two of such regions are conformally equivalent. First of all it became apparent that holomorphic functions are intimately related to certain geometrical structures, and at the same time their discovery had a stimulating effect when it became clear that Riemann's proof showed a serious logical gap, the removal of which appeared to be far from obvious.

Later developments revealed that many striking properties of complex functions and the mappings defined by them can be seen in their true perspectives, if in geometrical formulation non-euclidean metrics are used. G. Julia's investigations with respect to the generalisation of Schwarz's lemma serve as a classic example. In chapter nine the reader is introduced to the theory of these metrics and their geometrical background.

In the 1920 s A. Bloch discovered that each function holomorphic in the interior of the unit circle, the derivative of which has the value 1 at the origin, represents a one-to-one mapping of a subregion onto another region which covers an open circular disc, the radius of which is not smaller than a universal constant. This theorem opened up the possibility for finding an "elementary" proof of the famous theorem of E. Picard, which is a refinement of the Casorati-Weierstrass theorem. The problems connected with Bloch's theorem have been treated in chapter nine as well, because the most elegant proof known for this theorem is based on a generalization of Schwarz's lemma due to L. Ahlfors, which has a very general metric as its underlying principle.

In chapter ten the theory of conformal mapping comes under review in some detail, while special attention is being paid to the treatment for finding the mapping functions in actual situations. The end of this chapter consists of a proof of Riemann's mapping theorem and of some elementary considerations on the difficult problem of the correspondence between the boundaries of regions mapped onto each other. Only that which is needed further on has been mentioned here.

Chapter eleven has been devoted to the theory of univalent functions, one of the most fascinating parts of the theory of functions of a complex variable. Proceeding from simple geometrical considerations a large number of remarkable properties of univalent functions can be found, i.e. that for each power series $z+a_2z^2+a_3z^3+\ldots$, convergent for $|z|<1$, the estimate $|a_2| \leqq 2$ holds, while an example shows that this estimate is the best possible. This example lead L. Biederbach to the conjecture that for each $n \geqq 2$ the estimate $|a_n| \leqq n$ should hold, which conjecture has only been proved for $n = 3$ and $n = 4$ so far, beyond the elementary case $n = 2$. The classic methods, however, are completely inadequate to obtain this result. An important step forward became possible through the theory of K. Löwner. Much attention is paid to this beautiful theory, not only because with the help of it Bieberbach's conjecture can be proved for $n = 3$, but especially since the coefficient problem for the inverse functions turns out to be completely solvable by means of this theory. Besides the formulation of a strong form of the rotation theorem, of which Bieberbach could only give a preliminary result, has become feasible as well. The original paper of Löwner is highly schematic; he relies on investigations by C. Carathéodory, which bear on sequences of regions and a convergence theorem related to those. In chapter eleven due attention has been paid to it. A more analytic formulation of Löwner's theory is to be found in the well-known monograph on multivalent functions by W. K. Hayman, in which, however, the theory of Lebesgue integration plays an essential part.

One of the most important means in the theory of complex functions is the analytic continuation, and that is dealt with in chapter twelve. Closely connected with it is the beautiful concept of a Riemann surface.

In applications of the analytic continuation along curves the deformation of those curves is a technique often used. Leaving out a discussion of its modern definition did not seem acceptable to us, also because that technique will be used extensively again later on; thus it has been indicated how simple topological concepts are applied in the theory of functions.

The theory of Riemann surfaces has only been touched upon, since using these surfaces as a mode of proof has restricted meaning for our purposes. Any detailed treatment of the theory of the Riemann surface requires an extensive knowledge of topology, real analysis and Hilbert-space theory and would take up a disproportionate part of the present text. Good books exist on the subject and its present interest is apparent from the fact that a contempory abstract method, viz., the theory of sheaves, offers a unique approach to the classic problems as well.

It is self-evident that an introductory theory of algebraic functions could not be dispensed with, the more so, since the Riemann surfaces related to them can easily be classified. Moreover this served to illustrate the famous uniformization problem, the treatment of which can be placed within the framework of a textbook without difficulties, due to an ingenious idea of B. L. van der Waerden.

The problem of the uniformization is closely related to the theory of automorphic functions. A discussion of this theory is to be found in chapter thirteen. Extensive use has been made of the concept of the isometric circles, introduced by L. R. Ford, through which the theory became accessible for the application of elementary means.

A beautiful illustration of the theory of automorphic functions is provided by the inverses of the so-called triangle functions of H. A. Schwarz, which are brought up in chapter fourteen. The relations between the theory of functions and certain geometrical figures is beautifully exemplified especially by the polyhedral functions of F. Klein.

E. Picard's classic proof of his famous theorem is based on the theory of modular functions. Owing to the work of C. Carathéodory the theory of these functions and the theorems connected with them have been essentially improved and space has been accorded to the subject in this chapter.

Schwarz's functions are obtained in relation to the problem of mapping a triangle bounded by circular arcs or straight line segments onto the interior of a circle (or onto a half plane). These mapping functions satisfy the wellknown hypergeometric differential equation. However, before we treat the theory connected, in chapter fifteen, detailed attention is paid to the theory of homogeneous linear differential equations in as far as it is relevant. A dominating part is played here by the theory of L. Fuchs. We also availed ourselves of the opportunity to discuss some important examples, in particular the Bessel functions and the Legendre functions. Obviously, the stress has been laid on the most general equation of the Fuchsian type, which does not have accessory parameters yet, viz., Riemann's differential equation.

By means of a suitable linear fractional transformation Riemann's equation can be put in the form of the hypergeometric differential equation and the last chapter has been devoted to the many aspects arising with the study of this equation. We also thought fit to pay attention to the hypergeometric polynomials and those polynomials attainable from them through confluence. Their great importance in practical applications is well-known. Moreover, the fact that the theory is clearly selfcontained and dominated by an equation of the Rodrigues type, encountered in the theory of the Legendre polynomials, renders study of these objects attractive.

No extensive treatment of the general problem of mapping defined by hypergeometric functions has been given, for it can hardly be summarized yet and requires the discussion of many details. However, some special cases are considered, partly with the intention of describing the elegant method of L. Ahlfors and H. Grunski, through which an upper estimate of Bloch's constant can be obtained which is close to the lower bound.

Some mathematicians believe that this upper estimate is the exact value of Bloch's constant, but this conjecture can up to now neither be proved nor disproved.

In spite of the fact that in volume 2 a great variety of subjects have been considered, the book is not of an encyclopaedic nature. The choice of the subjects has been made rather subjectively, but not without due observance of some coherence. At the end of the preface of volume 1 we indicated which considerations directed us when we were writing the book. This observation applies equally well to volume 2, and we hope that the students will show the same favourable appreciation of this volume as has been shown for the first volume.

Florence (Italy) G. SANSONE

Groningen (The Netherlands) J. GERRETSEN

CONTENTS

CHAPTER 9. APPLICATIONS OF GENERAL METRICS TO THE THEORY OF FUNCTIONS.

CHAPTER 10. CONFORMAL MAPPING OF SIMPLY CONNECTED REGIONS

CHAPTER 11. UNIVALENT FUNCTIONS

CHAPTER 12 ANALYTIC FUNCTIONS. RIEMANN SURFACES .

CHAPTER 13. AUTOMORPHIC FUNCTIONS

CHAPTER 14. THE SCHWARZIAN TRIANGLE FUNCTIONS AND THEIR INVERSES

CHAPTER 15. LINEAR HOMOGENEOUS DIFFERENTIAL EQUATIONS

CHAPTER 16. THE HYPERGEOMETRIC DIFFERENTIAL EQUATION

CHAPTER 9

APPLICATIONS OF GENERAL METRICS TO THE THEORY OF FUNCTIONS

9.1. – Topological considerations

9.1.1. – EXTENSION OF THE NOTION OF REGION

In the beginning of section 3.12.1 we remarked that a function $f(z)$ can be interpreted geometrically as a mapping. By stressing the mapping properties of a function we consider it from a geometric point of view, which has many advantages. It is often possible to obtain a lot of information about a function by studying the geometric correspondence given by it.

In this connection it will appear convenient to extend the notion of region to sets of points including the point at infinity. This point is considered as an interior point of a set if a neighbourhood of this point (being the set of points outside a sufficiently large circumference around the origin together with the point at infinity itself) is included in the set. The difference between ordinary points and the point at infinity may often be eliminated by introducing the notion of chordal distance (section 1.1.5).

First we wish to state a theorem of a general topological character. Let $f(z)$ denote a single-valued function defined throughout the extended plane (or, which amounts to the same, throughout the complex sphere, see section 1.1.3). If $\mathfrak{S}$ is any set of points then $f^{-1}(\mathfrak{S})$ denotes the set of all points which have images under the mapping as given by f belonging to $\mathfrak{S}$. We shall say that $f^{-1}(\mathfrak{S})$ is the *original* of $\mathfrak{S}$ with respect to the mapping f. After these preliminaries we assert

The function f, defined throughout the extended plane, is chordally continuous at every point if and only if the original with respect to f of every open set is open.

The term "chordally continuous" has been explained in section 1.2.1.

The condition is necessary. Let f be chordally continuous throughout the extended plane. Let, further, $f^{-1}(\mathfrak{A})$ denote the original of an open set $\mathfrak{A}$ and a a point of this set. Then $b = f(a)$ is a point of $\mathfrak{A}$ and there is a neighbourhood $\mathfrak{B}$ of b included in $\mathfrak{A}$. Since f is chordally continuous there is a neighbourhood $\mathfrak{U}$ of a such that $f(\mathfrak{U})$ is included in $\mathfrak{B}$. Hence all points of $\mathfrak{U}$ are mapped onto points of $\mathfrak{A}$ and, consequently, $\mathfrak{U}$ is a subset of $f^{-1}(\mathfrak{A})$. Thus we see that this set is open.

Assume now that f is such that the original of every open set is open. Let a be an arbitrary point and $b = f(a)$. If $\mathfrak{V}$ is an ε-neighbourhood of b, then $f^{-1}(\mathfrak{V})$ is open, by hypothesis. Hence there is a neighbourhood $\mathfrak{U}$ of a included in $f^{-1}(\mathfrak{V})$ and, as a consequence, $f(\mathfrak{U})$ is a subset of $\mathfrak{V}$. This proves the continuity of f at $z = a$.

It should be noticed that *the same theorem holds if we replace everywhere open sets by closed sets.* This is an immediate consequence of the fact that in the extended plane the complement of an open set is closed and conversely.

When proceeding we need the following notion: We shall say that a non-empty open set is *connected* if it is not the union of two disjoint open non-empty sets. Similarly a non-empty closed set will be called connected if it is not the union of two disjoint closed non-empty sets.

This notion of connectedness is somewhat more general than the one introduced at the end of section 1.2.4. We shall refer to it as *arcwise connectedness.* In the following case the notions have the same content:

A non-empty open set in the open plane is connected if and only if the set is a region, i.e., if it is open and arcwise connected. By the *open plane* we understand the whole complex plane without the point at infinity.

Let $\mathfrak{R}$ denote a connected open set not including the point at infinity. Choose a point c in $\mathfrak{R}$. Let $\mathfrak{A}$ denote the subset of $\mathfrak{R}$ whose points can be joined to c by a polygon in $\mathfrak{R}$ and let $\mathfrak{B}$ denote the remaining points of $\mathfrak{R}$. If a is a point of $\mathfrak{A}$ then there is a neighbourhood of a included in $\mathfrak{R}$. All points of this neighbourhood can be joined to a by a straight line segment and from there to c by a polygon. Hence the whole neighbourhood is already contained in $\mathfrak{A}$, i.e., $\mathfrak{A}$ is open. If b is a point of $\mathfrak{B}$ then it has a neighbourhood contained in $\mathfrak{R}$. If a point of this neighbourhood could be joined to c by a polygon then also b could be joined to c, contrary to the definition of $\mathfrak{B}$. Hence $\mathfrak{B}$ is also open and $\mathfrak{R}$ is the union of the disjoint open sets $\mathfrak{A}$ and $\mathfrak{B}$. Since c belongs to $\mathfrak{A}$ this set is not empty. Hence $\mathfrak{B}$ is empty, for $\mathfrak{R}$ is supposed to be connected and it follows that $\mathfrak{A}$ and $\mathfrak{R}$ coincide. It is now clear that each pair of points of $\mathfrak{R}$ can be joined by a polygon in $\mathfrak{R}$ by way of c, i.e., $\mathfrak{R}$ is a region.

Conversely, let $\mathfrak{R}$ be a region and the union of two open disjoint sets $\mathfrak{A}$ and $\mathfrak{B}$. Take a point a in $\mathfrak{A}$ and a point b in $\mathfrak{B}$. Since a and b can be joined by a polygon there must be a straight line segment connecting a point a_1 of $\mathfrak{A}$ and a point b_1 of $\mathfrak{B}$. We bisect this segment and note that one of the halves has one of its end points – say a_2 – in $\mathfrak{A}$ and the other – say b_2 – in $\mathfrak{B}$. Continuing this process we obtain a sequence $a_1, a_2, \ldots$ of points of $\mathfrak{A}$ and a sequence $b_1, b_2, \ldots$ of points of $\mathfrak{B}$ converging to the same limit c of the segment. If c belongs to $\mathfrak{A}$ then a neighbourhood of

c also belongs to $\mathfrak{A}$, for $\mathfrak{A}$ is open. From a certain index upwards all points b_n belong to this neighbourhood and hence to $\mathfrak{A}$. This is impossible. By the same argument we see that c is not a point of $\mathfrak{B}$ either. But c belongs to $\mathfrak{R}$, the union of $\mathfrak{A}$ and $\mathfrak{B}$, and thus we arrive at a contradiction. This leads to the conclusion that $\mathfrak{R}$ is connected.

This theorem enables us to generalize the notion of region to sets in the extended plane:

A *region* is an open connected set. It is clear that the extended plane is a region. In fact, it is an open set and a decomposition into two open disjoint sets is not possible, for the complement of an open set is closed and the only open and closed sets are the empty set and the extended plane itself. Finally we wish to prove the following theorem

The union of two connected closed sets having at least one point in common is connected.

Let $\mathfrak{C}$ denote a closed set, being the union of two connected closed sets $\mathfrak{C}_1$ and $\mathfrak{C}_2$ whose intersection is not empty. Assume, moreover, that $\mathfrak{C}$ is the union of two closed disjoint sets $\mathfrak{B}_1$ and $\mathfrak{B}_2$ which are not empty. Every point of $\mathfrak{C}_1$ belongs either to $\mathfrak{B}_1$ or to $\mathfrak{B}_2$. Hence $\mathfrak{C}_1$ is the union of the intersections of $\mathfrak{C}_1$ with $\mathfrak{B}_1$ and $\mathfrak{B}_2$. Since the intersection of two closed sets is closed it follows that one of the intersections is empty, i.e., $\mathfrak{C}_1$ must coincide with $\mathfrak{B}_1$ or with $\mathfrak{B}_2$. If $\mathfrak{C}_1$ coincides with $\mathfrak{B}_1$ then $\mathfrak{C}_2$ must coincide with $\mathfrak{B}_2$, for, by hypothesis, $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are not empty. But this means that $\mathfrak{C}_1$ and $\mathfrak{C}_2$ are disjoint, contrary to assumption.

9.1.2 – Invariance of the region

By a *topological mapping* or *homeomorphism* of the extended plane onto itself we understand a mapping which is one-to-one and in both directions chordally continuous throughout the extended plane.

A region is invariant under a topological mapping of the extended plane onto itself.

Since the inverse of a topological mapping f is everywhere chordally continuous we deduce from the first theorem of the previous section that the image of a region $\mathfrak{R}$ is an open set $\mathfrak{S} = f(\mathfrak{R})$. Assume that $\mathfrak{S}$ is not connected, i.e., the union of two open sets $\mathfrak{A}$ and $\mathfrak{B}$. Then $f^{-1}(\mathfrak{A})$ and $f^{-1}(\mathfrak{B})$ are open and disjoint. Intersecting these sets by $\mathfrak{R}$ we find a decomposition of $\mathfrak{R}$, unless one of the sets $\mathfrak{A}$ or $\mathfrak{B}$ is empty. Thus we see that $\mathfrak{S}$ is also a region.

The theorem of the invariance of a region stated in section 3.12.2 may be generalized in the following way:

If f is meromorphic in a region $\mathfrak{R}$ and does not reduce to a constant, the image of $\mathfrak{R}$ as given by the function is also a region.

In order to prove this theorem we make a preliminary remark. Assume that $\mathfrak{R}$ contains the point at infinity, but does not coincide with the extended plane. Then we can find a meromorphic (and at the same time topological) mapping which transforms the region $\mathfrak{R}$ into a region $\mathfrak{R}^*$, which leaves the point at infinity outside. In fact, let $z = a$ denote a point not belonging to $\mathfrak{R}$. Then the function

$$\frac{1}{z-a} \tag{9.1–1}$$

gives rise to a topological mapping of the extended plane onto itself which interchanges the points $z = a$ and $z = \infty$. This function is meromorphic.

Assume first that $\mathfrak{R}$ is the extended plane and f meromorphic throughout $\mathfrak{R}$. If $\mathfrak{S} = f(\mathfrak{R})$ does not coincide with the extended plane then we may replace $\mathfrak{S}$ by a set $\mathfrak{S}^*$ which does not contain the point at infinity on applying the above mentioned mapping. The correspondence between $\mathfrak{R}$ and $\mathfrak{S}^*$ is described by a function which is holomorphic throughout $\mathfrak{R}$. Hence this function is a constant (section 3.2.2). But this means that $\mathfrak{S}^*$, and with it $\mathfrak{S}$, is a single point, i.e., f is constant, contrary to hypothesis. As a consequence $\mathfrak{S}$ is also the extended plane and, therefore, a region.

Next we consider a region $\mathfrak{R}$ not coinciding with the entire extended plane. If necessary we replace $\mathfrak{R}$ by a region $\mathfrak{R}^*$ on applying the above transformation. If $\mathfrak{S} = f(\mathfrak{R})$ is the extended plane we are through. Assume, therefore, that $\mathfrak{S}$ does not coincide with the extended plane. Again we may suppose that $\mathfrak{S}$ is a set which leaves the point at infinity outside. Now the correspondence between $\mathfrak{R}^*$ and $\mathfrak{S}$ is described by a function which is holomorphic throughout $\mathfrak{R}^*$. From the theorem of section 3.12.2 we conclude that $\mathfrak{S}$ is a region again and this concludes the proof of the theorem.

9.1.3 – Simply connected regions

Our next task will be the extension of the notions of simple connectivity. To this end we shall prove

A region $\mathfrak{R}$ not containing the point at infinity is simply connected if and only if its complement in the extended plane is connected.

Assume first that the complement $\mathfrak{R}'$ of the region $\mathfrak{R}$ is connected. Let C denote a cycle in $\mathfrak{R}$. As we pointed out in section 2.5.2 we may replace C by a polygon L, taking into account formula (2.1–4). Proceeding as in that section we see that we may decompose the plane into a finite number of triangles and of convex sets tending to infinity. The winding number $\Omega_L(z)$ is constant in the interior of each of these parts of the plane. For a point z not on L at the boundary of such a part the winding number is the same as that of the interior. Hence the function $\Omega_L(z)$ has only a finite number of integral values. Let $\mathfrak{R}'_n$ denote the set of points of $\mathfrak{R}'$ for which $\Omega_L(z) = n$. We assert that $\mathfrak{R}'_n$ is closed. For if a point a of $\mathfrak{R}'$ is an accumulation point of $\mathfrak{R}'_n$, then in every neighbourhood of a there are points of $\mathfrak{R}'_n$. The value $\Omega_L(a)$ exists, for a is in $\mathfrak{R}'$. Hence $\Omega_L(a) = n$, because of the continuity of $\Omega_L(z)$. Thus we see that we can decompose $\mathfrak{R}'$ into a finite number of closed disjoint sets and this number must be unity, for $\mathfrak{R}'$ is connected. Hence $\Omega_L(z)$ is constant throughout $\mathfrak{R}'$. On the other hand $\Omega_L(z)$ is zero for points which are sufficiently near $z = \infty$. Hence $\Omega_L(z) = 0$ throughout $\mathfrak{R}'$ and $C \sim 0$ in $\mathfrak{R}$.

For the proof of the necessity of the condition mentioned in the theorem we assume that the complement $\mathfrak{R}'$ of $\mathfrak{R}$ is the union of two disjoint closed sets $\mathfrak{A}$ and $\mathfrak{B}$. One of these sets contains the point at infinity and the other is consequently bounded. Let $\mathfrak{A}$ be the bounded component. By ρ we denote the distance between $\mathfrak{A}$ and $\mathfrak{B}$. Now we cover the whole plane with a net of squares $\mathfrak{Q}$ of side $<\rho/\sqrt{2}$. It is possible to choose the net such that a preassigned point a of $\mathfrak{A}$ lies in the interior of a square. The boundary curve of a square $\mathfrak{Q}$ is denoted by Q which may be oriented in the counter clockwise sense. Consider now the cycle

$$C = \sum_{\nu} Q_{\nu}, \tag{9.1–2}$$

where the sum ranges over all squares of the net which have a point in common with $\mathfrak{A}$, (fig. 9.1–1). Because a is included in one and only one of these squares it follows that $\Omega_C(a) = 1$. If the cancellations are

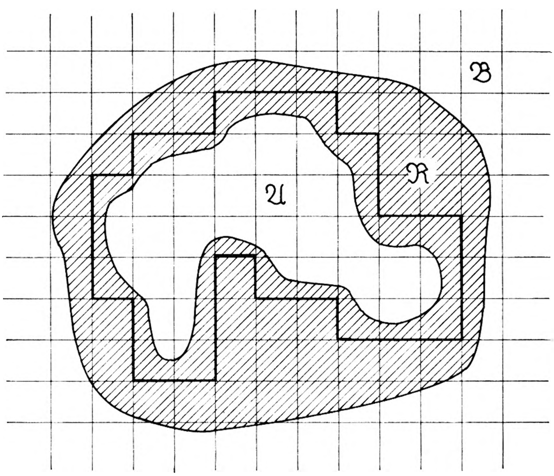


Fig. 9.1–1. The connectivity of a region.

carried out it is clear that C does not meet $\mathfrak{A}$, for any side which meets $\mathfrak{A}$ is a common side of two squares contributing to the sum (9.1–2). Since the two squares induce opposite orientations for the common side it does not appear in the reduced expression for C. Hence C is a cycle of $\mathfrak{R}$ which is not homologous to zero. This concludes the proof of the theorem.

Accordingly we may define: A region in the extended plane is *simply connected* if its complement is connected. Proceeding as in the proof of the first theorem of section 9.1.2. we easily can prove

A simply connected region is an invariant under topological mapping of the extended plane onto itself.

Indeed, connectedness of a closed set is a topological invariant.

It should be noticed that in a simply connected set as generalized above not every cycle is necessarily homologous to zero. Thus, for instance, the exterior of a circumference is simply connected, for the closed disc bounded by the circumference is connected. But the winding number of any larger concentric circumference with respect to the centre is not zero.

9.1.4 – A MONODROMY THEOREM

We wish to prove a useful theorem which is a particular case of a general statement to be considered in section 12.2.3.

In section 1.11.2 we pointed out that all solutions of the equation

$$\exp w = z \tag{9.1–3}$$

do not constitute the values of a single-valued function. By restricting, however, the values of z to a principal region (section 1.2.2.) it is possible to define a function satisfying (9.1–3) and which is holomorphic in this region.

We proceed to investigate a more general problem, viz. the solution of the equation

$$\exp w = f(z), \tag{9.1–4}$$

where $f(z)$ is holomorphic throughout a certain region $\mathfrak{R}$. We recall that $f(z)$ is regular at $z = \infty$ if $f(1/z)$ is regular at $z = 0$. As we shall see the restriction that $\mathfrak{R}$ is simply connected will enable us to obtain a single valued function satisfying (9.1–4), provided $f(z)$ has no zeros within the region. Thus

If $f(z)$ is regular at every point of a simply connected region $\mathfrak{R}$ and vanishes nowhere at any point of $\mathfrak{R}$, then we can find a function $w(z)$ also regular at every point of $\mathfrak{R}$ and which satisfies the equation (9.1–4).

The theorem is trivial in the case that $\mathfrak{R}$ coincides with the extended plane, for a function regular at every point, the point at infinity included, is necessarily a constant c (section 3.2.2.). Then we may take $w(z) = \log c$, where the logarithm is not necessarily a principal value.

Henceforth we assume that $\mathfrak{R}$ is not the extended plane. The theorem is rather obvious if the region $f(\mathfrak{R})$ is included in a principal region, for then we can take the composite function $w(z) = \log f(z)$. In all other cases we may proceed as follows. By a preliminary topological transformation of the type (9.1–1) we may assume that $\mathfrak{R}$ does not contain the point at infinity. For if we have proved the theorem for this case then the general statement follows easily. Since $f(z)$ is holomorphic in $\mathfrak{R}$ and vanishes nowhere in $\mathfrak{R}$ the function $f'(z)/f(z)$ is also holomorphic throughout $\mathfrak{R}$. In view of the theorem of section 2.11.4 (which is valid for every region leaving the point at infinity outside, as follows by analyzing the proof of the theorem of section 2.10.1) this function is the derivative of another function $h(z)$. That is to say, there is a function $h(z)$ such that

$$h'(z) = \frac{f'(z)}{f(z)}. \tag{9.1–5}$$

Next we consider the function

$$g(z) = \frac{f(z)}{\exp h(z)}.$$

Its derivative is

$$g'(z) = \frac{f'(z)-f(z)h'(z)}{\exp h(z)} = 0$$

throughout $\mathfrak{R}$. Hence it is equal to a constant c (section 2.11.3) and $c \neq 0$, for $f(z)$ vanishes nowhere in $\mathfrak{R}$.

Accordingly we may introduce the function

$$w(z) = h(z)+\log c,$$

where $\log c$ is not necessarily a principal value of the logarithm. It follows that

$$\exp w(z) = c \exp h(z) = f(z).$$

The function $w(z)$ is not uniquely determined, for besides $w(z)$ also $w(z)+2n\pi i$, where n is a fixed integer, satisfies the equation. There are, however, no other solutions. In fact, if

$$\exp w_1(z) = \exp w_2(z),$$

then

$$\exp\left(w_1(z)-w_2(z)\right) = 1.$$

By differentiation we find

$$w_1'(z)-w_2'(z) = 0.$$

Hence $w_1(z)-w_2(z)$ is a constant and this must be an integral multiple of $2\pi i$. Thus

The solutions of (9.1–4) *are determined up to an integral multiple of* $2\pi i$.

All solutions of (9.1–4) constitute the *general logarithm* $\log f(z)$ of the function $f(z)$. Any single valued function satisfying this equation will be called a *branch* of the logarithm of $f(z)$ and denoted by

$$\log f(z). \tag{9.1–6}$$

This symbol has a definite meaning if z ranges through a simply connected region, $f(z)$ has no zeros in the region and one of the possible values of (9.1–5) is assigned.

As a particular result we mention

In any simply connected region which does not contain the origin and the point at infinity a branch of $\log z$ *can be defined.*

It is clear that a branch of $\log f(z)$ may be written as

$$\log f(z) = \log f(z_0) + \int_{z_0}^{z} \frac{f'(\zeta)}{f(\zeta)}\, d\zeta, \tag{9.1–7}$$

where z_0 is a point of the region and the integration is performed along any path in the region connecting z_0 and z.

It is easy to state similar theorems for other functions which are intimately related to the logarithm.

A branch

$$f^{\lambda}(z) \tag{9.1–8}$$

of the *general power* of $f(z)$ is defined as

$$\exp\left(\lambda \log f(z)\right). \tag{9.1–9}$$

Here λ may denote any complex number.

By a branch

$$\arg f(z) \tag{9.1–10}$$

of the *general argument* $\arg f(z)$ is understood the function

$$\operatorname{Im} \log f(z). \tag{9.1–11}$$

The various branches of the argument differ by integral multiples of 2π.

9.2 – Conformal mapping

9.2.1 – The symbolic partial differentiation

A function $f(z)$ of a complex variable is a function of two independent variables x and y, if $z = x+iy$:

$$f(z) = u(x, y)+iv(x, y). \tag{9.2–1}$$

Let us assume that $f(z)$ possesses continuous partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$, where $\partial f/\partial x$ stands for $\partial u/\partial x + i\partial v/\partial x$ and $\partial f/\partial y$ for $\partial u/\partial y + i\partial v/\partial y$. Then there is a total differential

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy. \tag{9.2–2}$$

Observing that

$$dz = dx + idy, \qquad d\bar{z} = dx - idy,$$

we have

$$dx = \tfrac{1}{2}(dz + d\bar{z}), \qquad idy = \tfrac{1}{2}(dz - d\bar{z}). \tag{9.2–3}$$

Inserting this into (9.2–2) we readily find

$$df = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\,\frac{\partial f}{\partial y}\right) dz + \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\,\frac{\partial f}{\partial y}\right) d\bar{z}. \tag{9.2–4}$$

It is natural to write this as

$$\boxed{df = \frac{\partial f}{\partial z}\,dz + \frac{\partial f}{\partial \bar{z}}\,d\bar{z},} \tag{9.2–5}$$

with

$$\begin{aligned} \frac{\partial f}{\partial z} &= \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\,\frac{\partial f}{\partial y}\right), \\ \frac{\partial f}{\partial \bar{z}} &= \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\,\frac{\partial f}{\partial y}\right). \end{aligned} \tag{9.2–6}$$

The conditions of monogenity (1.3–7) can now be condensed into the simple equation

$$\boxed{\frac{\partial f}{\partial \bar{z}} = 0.} \tag{9.2–7}$$

Then the derivative is

$$f'(z) = \frac{\partial f}{\partial z}. \tag{9.2–8}$$

For the sake of illustration we calculate $\partial \arg z/\partial \bar{z}$, where z ranges throughout the principal region. Observing that

$$\log z = \log |z| + i \arg z = \tfrac{1}{2} \log z\bar{z} + i \arg z$$

and taking into account the fact that $\log z$ is holomorphic we readily find

$$0 = \frac{1}{2}\,\frac{z}{z\bar{z}} + i\,\frac{\partial \arg z}{\partial \bar{z}} = \frac{1}{2\bar{z}} + i\,\frac{\partial \arg z}{\partial \bar{z}},$$

whence

$$\frac{\partial \arg z}{\partial \bar{z}} = \frac{-1}{2i\bar{z}}.$$

Thus we see that $\arg z$ is not holomorphic.

In many cases the symbolic partial derivative may be used to simplify calculations as we shall presently.

Since

$$\overline{\frac{\partial f}{\partial x}} = \frac{\partial \bar{f}}{\partial x}, \qquad \overline{\frac{\partial f}{\partial y}} = \frac{\partial \bar{f}}{\partial y},$$

it follows easily from (9.2–6)

$$\frac{\partial \bar{f}}{\partial z} = \overline{\frac{\partial f}{\partial \bar{z}}}, \qquad \frac{\partial \bar{f}}{\partial \bar{z}} = \overline{\frac{\partial f}{\partial z}}. \tag{9.2–9}$$

We wish to apply this result to the following problem. Let $f(z)$ be holomorphic within a certain open set $\mathfrak{A}_z$ and let $\mathfrak{A}_{\bar{z}}$ denote the symmetric set with respect to the real axis. Now we define a function

$$g(z) = \overline{f(\bar{z})}, \tag{9.2–10}$$

where $\bar{z}$ runs through $\mathfrak{A}_z$. We assert that $g(z)$ is holomorphic in $\mathfrak{A}_{\bar{z}}$.

In fact, putting $w = \bar{z}$, we have

$$\frac{\partial g(z)}{\partial \bar{z}} = \frac{\partial \overline{f(\bar{z})}}{\partial \bar{z}} = \frac{\partial \overline{f(w)}}{\partial w} = \overline{\frac{\partial f(w)}{\partial \bar{w}}} = 0.$$

Finally we observe that if z is a differentiable function of a real variable t and we put $g(t) = f(z(t))$, then

$$\frac{dg}{dt} = \frac{\partial f}{\partial z}\frac{dz}{dt} + \frac{\partial f}{\partial \bar{z}}\frac{d\bar{z}}{dt}. \tag{9.2–11}$$

For many applications expressions for the Laplace operator and the Cauchy-Riemann equations in polar form turn out to be very useful. It is not difficult to obtain them in a straight forward manner by elementary computations, but we prefer to derive them by making use of the symbolic method as discussed before.

From (9.2–6) we obtain at once

$$\frac{\partial^2 f}{\partial z \partial \bar{z}} = \frac{1}{4}\left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right)$$

and thus the Laplace operator occurring in (1.4–3) appears as

$$\boxed{\triangle f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 4\,\frac{\partial^2 f}{\partial z \partial \bar{z}}.} \tag{9.2–12}$$

It is now easy to write down a transformation formula, if we introduce a new variable $w = w(z)$ such that $w(z)$ is regular at a given point. For, taking into account (9.2–7) and (9.2–9) we find, if we write $f(z)$ instead of $g(w(z))$

$$\frac{\partial f}{\partial z} = \frac{\partial g}{\partial w}\,\frac{\partial w}{\partial z} + \frac{\partial g}{\partial \bar{w}}\,\frac{\partial \bar{w}}{\partial z} = \frac{\partial g}{\partial w}\,\frac{\partial w}{\partial z}.$$

Similarly

$$\frac{\partial f}{\partial \bar{z}} = \frac{\partial g}{\partial w}\,\frac{\partial w}{\partial \bar{z}} + \frac{\partial g}{\partial \bar{w}}\,\frac{\partial \bar{w}}{\partial \bar{z}} = \frac{\partial g}{\partial \bar{w}}\,\frac{\partial \bar{w}}{\partial \bar{z}}.$$

Hence

$$\frac{\partial^2 f}{\partial z \partial \bar{z}} = \frac{\partial^2 g}{\partial w \partial \bar{w}}\,\frac{\partial w}{\partial z}\,\frac{\partial \bar{w}}{\partial \bar{z}} + \frac{\partial g}{\partial w}\,\frac{\partial^2 w}{\partial z \partial \bar{z}}.$$

Since

$$\frac{\partial^2 w}{\partial z \partial \bar{z}} = \frac{\partial}{\partial z}\,\frac{\partial w}{\partial \bar{z}} = 0,$$

we have in an obvious notation

$$\boxed{\triangle_z f = |w'(z)|^2 \triangle_w g.} \tag{9.2–13}$$

Next we introduce polar coordinates by $z = re^{i\theta}$.

Then, evidently, if we set $g(r, \theta) = f(z) = f(re^{i\theta})$,

$$r\,\frac{\partial g}{\partial r} = re^{i\theta}\,\frac{\partial f}{\partial z} + re^{-i\theta}\,\frac{\partial f}{\partial \bar{z}} = z\,\frac{\partial f}{\partial z} + \bar{z}\,\frac{\partial f}{\partial \bar{z}},$$

$$\frac{\partial g}{\partial \theta} = ire^{i\theta}\,\frac{\partial f}{\partial z} - ire^{-i\theta}\,\frac{\partial f}{\partial \bar{z}} = i\left(z\,\frac{\partial f}{\partial z} - \bar{z}\,\frac{\partial f}{\partial \bar{z}}\right).$$

Solving for $\partial f/\partial z$ and $\partial f/\partial \bar{z}$

$$\begin{aligned} z\,\frac{\partial f}{\partial z} &= \frac{1}{2}\left(r\,\frac{\partial g}{\partial r} - i\,\frac{\partial g}{\partial \theta}\right), \\ \bar{z}\,\frac{\partial f}{\partial \bar{z}} &= \frac{1}{2}\left(r\,\frac{\partial g}{\partial r} + i\,\frac{\partial g}{\partial \theta}\right), \end{aligned} \tag{9.2–14}$$

whence

$$z\bar{z}\,\frac{\partial^2 f}{\partial z \partial \bar{z}} = \frac{1}{4}\left(r\,\frac{\partial}{\partial r}\left(r\,\frac{\partial g}{\partial r}\right) + \frac{\partial^2 g}{\partial \theta^2}\right)$$

and so

$$\boxed{\triangle f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial g}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 g}{\partial \theta^2}\,.} \tag{9.2–15}$$

Let finally $f(z)$ be regular at a given point. Then $\partial f/\partial \bar{z} = 0$. If we put

$$f(re^{i\theta}) = Re^{i\Phi}$$

we have in view of (9.2–14)

$$0 = r\frac{\partial R}{\partial r} + irR\frac{\partial \Phi}{\partial r} + i\frac{\partial R}{\partial \theta} - R\frac{\partial \Phi}{\partial \theta},$$

whence

$$\boxed{\frac{r}{R}\frac{\partial R}{\partial r} = \frac{\partial \Phi}{\partial \theta}, \quad \frac{1}{R}\frac{\partial R}{\partial \theta} = -r\frac{\partial \Phi}{\partial r}\,.} \tag{9.2–16}$$

These are the Cauchy-Riemann equations in polar form.

9.2.2 – Conformal mapping

Consider a correspondence as given by

$$w = f(z) \tag{9.2–17}$$

in a neighbourhood of $z = z_0$, where $f(z)$ possesses continuous partial derivatives. It is not yet assumed that $f(z)$ is holomorphic near z_0. To $z = z_0$ corresponds $w = w_0 = f(z_0)$ and we suppose that z_0 and w_0 are finite. Let $z(t)$, where t is a real variable, denote a curve passing through z_0 such that $z_0 = z(t_0)$. We make the assumption that $z(t)$ is differentiable at $t = t_0$ and that the derivative at this point is different from zero.

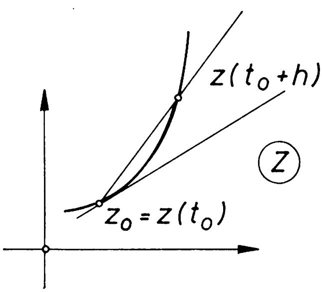


Fig. 9.2–1. The tangent of a curve at a given point.

The set of points

$$z = z(t_0)+t\,\frac{z(t_0+h)-z(t_0)}{h},$$

where h is a number $\neq 0$ and t a real variable is a straight line passing through z_0, (fig. 9.2–1). Making $h \to 0$ we obtain the line

$$z = z_0+tz'(t_0) = z_0+at, \tag{9.2–18}$$

where $a \neq 0$. It is called the *tangent* of the curve at $z = z_0$. The pair of numbers (z_0, a) is called a *direction* at $z = z_0$. Its geometric equivalent is a half ray issuing from z_0 and being represented by (9.2–18) with $t > 0$. Hence every curve through z_0, being differentiable there, determines a direction, provided the derivative is not zero. We shall not distinguish between the directions (z_0, a) and $(z_0, \lambda a)$, $\lambda > 0$. This is geometrically clear, for they determine the same half rays.

Transforming the curve $z(t)$ by means of (9.2–17) we obtain the curse

$$w = f(z(t)), \tag{9.2–19}$$

passing through w_0. It possesses also a direction number b given by (according to 9.2–11).

$$b = \frac{\partial f}{\partial z}\,a+\frac{\partial f}{\partial \bar{z}}\,\bar{a}, \tag{9.2–20}$$

provided that $\partial f/\partial z$ and $\partial f/\partial \bar{z}$ do not vanish simultaneously at $z = z_0$.

If two curves $z_1(t)$ and $z_2(t)$ are given through $z_0 = z_1(t_0) = z_2(t_0)$, they give rise to two directions (z_0, a_1) and (z_0, a_2). By the *angle* at z_0 between these curves we understand

$$\arg\frac{a_2}{a_1}. \tag{9.2–21}$$

As a consequence the angle between their images is

$$\arg\frac{b_2}{b_1} = \arg\frac{\dfrac{\partial f}{\partial z}\,a_2+\dfrac{\partial f}{\partial \bar{z}}\,\bar{a}_2}{\dfrac{\partial f}{\partial z}\,a_1+\dfrac{\partial f}{\partial \bar{z}}\,\bar{a}_1}. \tag{9.2–22}$$

This result simplifies considerably if $f(z)$ is regular at $z = z_0$. For then $\partial f/\partial \bar{z} = 0$, hence $\partial f/\partial z \neq 0$. In this case we have

$$\frac{b_2}{b_1} = \frac{a_2}{a_1}. \tag{9.2–23}$$

This means that the angle between the image curves is equal to that between the original curves at the corresponding points. We express this by saying that the mapping is *isogonal* at $z = z_0$.

It may also occur that $\partial f/\partial z = 0$. In view of (9.2–9) this means that $\bar{f}(z)$ is regular at $z = z_0$. Then we find

$$\frac{b_2}{b_1} = \frac{\bar{a}_2}{\bar{a}_1} \tag{9.2–24}$$

and we shall say that the mapping is *anti-isogonal* at $z = z_0$, (fig. 9.2–2). Isogonality is restored if we first reflect the angle at z_0 with respect to a horizontal line passing through z_0.

Next we assume, conversely, that the mapping as given by (9.2–17), where $f(z)$ has continuous partial derivatives, is isogonal at $z = z_0$. By equating the right hand sides of (9.2–22) and (9.2–23) we readily find

$$(\bar{a}_1 a_2 - a_1 \bar{a}_2)\frac{\partial f}{\partial \bar{z}} = 0$$

and since a_1 and a_2 can be chosen arbitrarily, it follows that $\partial f/\partial \bar{z} = 0$, i.e. that $f(z)$ is regular at $z = z_0$. In the same way we conclude that $\partial f/\partial z = 0$ if the mapping is anti-isogonal.

A related property of the mapping is derived by considering the derivative of the arc length of a curve through z_0 and comparing it with its image. The derivative of the arc length of $z(t)$ at $t = t_0$ is $|a| = |z'(t_0)|$. The same quantity for the image curve is $|b|$ and we have the relation

$$|b| = \left|\frac{df}{dt}\right| = \left|\frac{\partial f}{\partial z}a + \frac{\partial f}{\partial \bar{z}}\bar{a}\right| = |a|\left|\frac{\partial f}{\partial z} + \frac{\partial f}{\partial \bar{z}}\frac{\bar{a}}{a}\right|. \tag{9.2–25}$$

The expression

$$\left|\frac{\partial f}{\partial z} + \frac{\partial f}{\partial \bar{z}}\frac{\bar{a}}{a}\right| \tag{9.2–26}$$

is called the *distortion* at the given point in the direction (z_0, a). It measures the infinitesimal change of scale at this point provoked by the mapping. *A necessary and sufficient condition that this mapping be independent of the direction number a is expressed by the fact that* $\partial f/\partial \bar{z} = 0$ *or* $\partial f/\partial z = 0$. In fact, the point represented by the number (9.2–26) between the bars moves along a circle with centre $\partial f/\partial z$ and radius $|\partial f/\partial \bar{z}|$. In order that its modulus is independent of a, either the radius must vanish, or the centre must be the origin. This proves the assertion.

A mapping having the same non-zero distortion in all directions issuing from this point is usually called *conformal*. It behaves like a homothetic transformation in an infinitesimal neighbourhood of the centre z_0.

It is either isogonal or anti-isogonal. Summing up we may say

A function, regular at a point z_0, such that its derivative does not vanish, is conformal and preserves the sense of the angles. The distortion is $|f'(z_0)|$.

We conclude this section by making some remarks about the point at infinity. A half ray issuing from $z = 0$ may also be considered as a half ray issuing from $z = \infty$. If it represents a direction $(0, a)$ at the origin then we shall say that it represents a direction $(\infty, 1/a)$ at $z = \infty$. The angle at $z = \infty$ defined by two half rays representing at the origin the directions $(0, a_1)$, $(0, a_2)$ is by definition

$$\psi = \arg \frac{1/a_2}{1/a_1} = \arg \frac{a_1}{a_2} = -\varphi,$$

if $\varphi = \arg(a_2/a_1)$. Hence it is the negative of the angle at $z = 0$. This is in accordance with the fact that the mapping carries the two half rays into

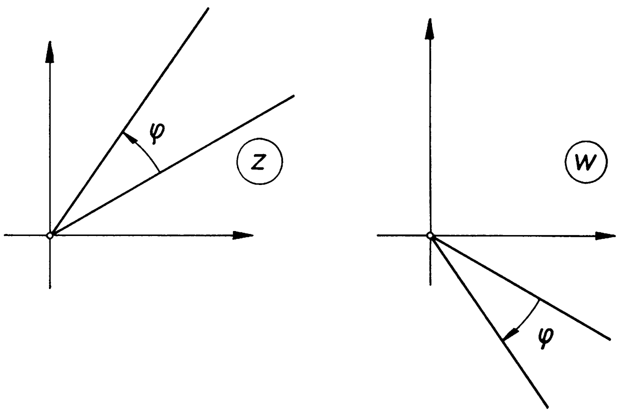


Fig. 9.2–2. Isogonality of the mapping $w = 1/z$ at $z = \infty$.

two others making an angle $-\varphi$ at the origin (fig. 9.2–2). Hence this mapping is also isogonal at $z = \infty$.

Accordingly we shall say that a function $f(z)$ defines a mapping conformal at $z = \infty$ if the mapping as given by $f(1/z)$ is conformal at $z = 0$. In addition a function $f(z)$ having a singularity at $z = z_0$ defines a conformal mapping at this point if this is the case with the function $1/f(z)$.

Some counter-examples may illustrate the previous considerations. The function $w = z^2$ defines a mapping which transforms the region $\operatorname{Re} z > 0$, $\operatorname{Im} z > 0$, i.e., the first quadrant, into the half plane $\operatorname{Im} w > 0$. The function is holomorphic throughout the z-plane. It is not conformal at $z = 0$, for $f'(z) = 2z = 0$ at the origin. Conversely the function $w = \sqrt{z}$, transforms the principal region into the right half plane. It is not conformal at $z = 0$ for there the function is not regular.

9.3 – Automorphisms of the extended plane

9.3.1 – A lemma

A one-to-one bicontinuous mapping of a region $\mathfrak{R}$ (in the extended plane) onto itself is called a *homeomorphism* of $\mathfrak{R}$. If $f_1(z)$ and $f_2(z)$ define homeomorphisms then evidently $f_2(f_1(z))$ also does. This is called the *product* of the homeomorphisms f_1 and f_2 and denoted by $f_2 f_1$. The inverse of a homeomorphism f is again a homeomorphism and denoted by f^{-1}. The homeomorphism $f^{-1}f = ff^{-1}$ leaves each point of $\mathfrak{R}$ at rest. It is the *identity*.

The product of two homeomorphisms is associative:

$$f_3(f_2 f_1) = (f_3 f_2) f_1,$$

for each member stands for $f_3(f_2(f_1(z)))$. The following statement is now clear:

A family of homeomorphisms of a region $\mathfrak{R}$ such that with each element also the inverse belongs to the family and with each two elements also their product is a group. It is clear that the identity belongs to the group.

It is understood that the algebraic structure of the family is given by the product rule as defined above.

A group of homeomorphisms of $\mathfrak{R}$ is called *transitive* if there is always an element in the group which transforms an arbitrarily given point z_1 of $\mathfrak{R}$ into any other given point z_2 of $\mathfrak{R}$.

More generally we may consider a group G of one-to-one transformations of a set $\mathfrak{S}$ of arbitrary things onto itself. The subgroup of all transformations of G leaving a given element a invariant is called the *subgroup of isotropy* associated with a. Next we wish to establish the following useful lemma:

Let G denote a group of one-to-one transformations of a set $\mathfrak{S}$ onto itself and H a transitive subgroup. Assume, moreover, that the subgroup of isotropy associated with a certain element a is contained in H. Then H coincides with the whole group G.

Since H is transitive we can find a transformation h of H which transforms a in $g(a)$, where g is an arbitrarily given element of G, i.e., $h(a) = g(a)$. Hence a is invariant under the transformation $h^{-1}g$ and this transformation, being an element of the subgroup of isotopy associated with a belongs to H, by hypothesis. As a consequence $g = (hh^{-1})g = h(h^{-1}g)$ also belongs to H and this proves the assertion. We shall have the opportunity to apply this lemma many times.

9.3.2 – Univalent functions

A function providing a one-to-one mapping of a certain open set $\mathfrak{A}$

is called *univalent* or *simple*. Some authors also use the German word "schlicht", which has no adequate translation in English.

At each point of an open set where the univalent function $f(z)$ is regular the derivative is different from zero.

Suppose that $f'(z) = 0$ and let $w_0 = f(z_0)$. Then the function $f(z)-w_0$ has a zero of order $k > 1$ at $z = z_0$. From the theorem of section 3.12.5 we deduce that around $z = z_0$ and around $w = w_0$ we can describe two circles such that to any point $w \neq w_0$ inside the second circle correspond precisely k different points inside the first circle for which $f(z)$ takes the value w. This is in contradiction with the assumption of univalence.

The theorem remains true if $z_0 = \infty$ in the following sense: The derivative of $f(1/z)$ tends to a finite limit different from zero as $z \to 0$. In fact, $f(z)$ has the Laurent expansion

$$f(z) = a_0 + \frac{a_1}{z} + \ldots.$$

Hence $f(1/z)$ is regular at $z = 0$ and univalent. This implies $a_1 \neq 0$.

The converse of the above theorem need not be true. Thus, for instance, the derivative of z^2 is different from zero if $|z| > 0$, but the function is not univalent in this region. *If, however, $f(z)$ is regular at $z = z_0$ and $f'(z_0) \neq 0$, then it is univalent in a sufficiently small neighbourhood of z_0.* This is again a direct consequence of the theorem of section 3.12.5. If $z_0 = \infty$ we must assume that $f(1/z)$ has a derivative which tends to a finite number different from zero as $z \to 0$.

A univalent function which is meromorphic throughout an open set $\mathfrak{A}$ cannot have other singularities than simple poles in $\mathfrak{A}$.

Indeed, if $z = z_0$ is a pole of order n then

$$f(z) = (z-z_0)^{-n}h(z),$$

where $h(z)$ is regular at $z = z_0$ and different from zero. Hence $1/f(z) = (z-z_0)^n/h(z) = (z-z_0)^n g(z)$ is regular at $z = z_0$ and has a zero of order n there. Since $1/f(z)$ is again univalent we conclude that $n = 1$. The theorem remains true in the case that $z_0 = \infty$.

In view of the preceding considerations we may state

A simple map as given by a meromorphic function is conformal and preserves the sense of the angles.

9.3.3 – The group of automorphisms of the extended plane

A particular example of a homeomorphism is an *automorphism* of an open set $\mathfrak{A}$ which is a mapping of $\mathfrak{A}$ onto itself provided by a meromorphic univalent function. It is clear, in the light of the previous considerations, that the product of two automorphisms is again an automorphism, as well as the inverse of an automorphism. Hence the automorphisms of $\mathfrak{A}$ constitute a group.

It is our aim to investigate the group of all automorphisms of the extended plane. The problem of determining all these automorphisms has a very simple solution.

First we observe that a *linear fractional transformation*, i.e. a transformation of the type

$$w = \frac{az+b}{cz+d}, \quad \begin{vmatrix} a & b \\ c & d \end{vmatrix} \neq 0, \tag{9.3–1}$$

is an automorphism of the extended plane. It is clear that the determinant (9.3–1) must be different from zero, for in the contrary case the denominator and the numerator would be proportional and hence w constant for general values of z. The transformation is uniquely invertible, the inverse is

$$z = \frac{dw-b}{-cw+a}. \tag{9.3–2}$$

To $z = -d/c$ corresponds $w = \infty$ and to $z = \infty$ the point a/c.

The product of two transformations as given by

$$w = \frac{a_1 z+b_1}{c_1 z+d_1}, \qquad w = \frac{a_2 z+b_2}{c_2 z+d_2} \tag{9.3–3}$$

is the transformation

$$w = \frac{a_2 \dfrac{a_1 z+b_1}{c_1 z+d_1} + b_2}{c_2 \dfrac{a_1 z+b_1}{c_1 z+d_1} + d_2} = \frac{a_3 z+b_3}{c_3 z+d_3}, \tag{9.3–4}$$

with

$$\begin{aligned} a_3 &= a_2 a_1 + b_2 c_1, & b_3 &= a_2 b_1 + b_2 d_1, \\ c_3 &= c_2 a_1 + d_2 c_1, & d_3 &= c_2 b_1 + d_2 d_1. \end{aligned} \tag{9.3–5}$$

Hence

$$\begin{vmatrix} a_3 & b_3 \\ c_3 & d_3 \end{vmatrix} = \begin{vmatrix} a_2 & b_2 \\ c_2 & d_2 \end{vmatrix} \begin{vmatrix} a_1 & b_1 \\ c_1 & d_1 \end{vmatrix} \neq 0. \tag{9.3–6}$$

This rule of composition can be formulated in a more concise way. We agree to write (9.3–1) symbolically as

$$w = \mathsf{A}z. \tag{9.3–7}$$

Here A symbolizes an *operator* which assigns to every number z a uniquely determined number w. This operator is characterized by the matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{9.3–8}$$

which shall also be denoted by the same symbol A. Now the effect of the transformation is not influenced by multiplying all coefficients occurring in (9.3–1) by the same number different from zero. We agree that A may represent any of this class of matrices. In many cases it is convenient to assume that the determinant of (9.3–1) is unity. Then the linear fractional transformation is called *unimodular* and from (9.3–6) follows that the product of two unimodular transformations is unimodular again. The same is true for the inverse transformation.

It follows from (9.3–5) that the matrix representing the transformation (9.3–4) is the product of the matrices representing the tranformation (9.3–3), respectively, i.e.,

$$\begin{bmatrix} a_3 & b_3 \\ c_3 & d_3 \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}. \tag{9.3–9}$$

Hence the law of composition is essentially the multiplication law for matrices: the rows of the first matrix are multiplied by the columns of the second matrix according to (9.3–5).

The inverse of the transformations (9.3–7) is written as

$$z = \mathsf{A}^{-1} w. \tag{9.3–10}$$

with

$$\mathsf{A}^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{9.3–11}$$

In the unimodular case we always have

$$\mathsf{A}\mathsf{A}^{-1} = \mathsf{A}^{-1}\mathsf{A} = \mathsf{E}, \tag{9.3–12}$$

where E denotes the identity operator represented by

$$\mathsf{E} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{9.3–13}$$

In view of the results obtained we may state

The family of all automorphisms represented by linear fractional transformations is a group.

Next we wish to establish the fact that this group is already the whole group of automorphisms of the extended plane. To this end we investigate the subgroup of isotropy of the group of all automorphisms with fixed point $z = \infty$ first.

Let $f(z)$ provide an automorphism of the open z-plane. It is an integral function, for it is holomorphic throughout the z-plane. Since it is univalent it maps the region $|z| < 1$ onto a region which has no point in common with the image of the region $|z| > 1$. The image of $|z| < 1$ is again a region, hence there is a point having a neighbourhood which has no points in common with the image of $|z| > 1$. It follows that $z = \infty$ cannot be an essential singular point of $f(z)$, for if it were an essential singular point, then by the Casorati–Weierstrass theorem (section 3.2.1) $f(z)$ would take values inside a given neighbourhood of b corresponding to points in the region $|z| > 1$. As a consequence $f(z)$ is a polynomial and since it is univalent it must be of degree unity. Hence

The group of all automorphisms of the open z-plane consists of the linear transformations

$$w = az+b. \tag{9.3–14}$$

Let now G denote the group of all automorphisms of the extended plane and H the subgroup of all linear fractional transformations. The subgroup of isotropy associated with $z = \infty$ is evidently contained in H. Since in (9.3–1) we can prescribe the values of z and w and then determine values of a, b, c and d suitably, the group H is transitive. In view of the lemma of section 9.3.1 we may infer that

The group of all automorphisms of the extended plane consists of the linear fractional transformations (9.3–1).

9.3.4 – The fixed points of an automorphism of the extended plane

A point z of the extended plane is called a *fixed point* of an automorphism (9.3–1) if it satisfies the equation

$$z = \frac{az+b}{cz+d},$$

or

$$cz^2-(a-d)z-b = 0. \tag{9.3–15}$$

This is a quadratic equation and it possesses no more than two roots, provided its coefficients do not all vanish simultaneously. In this latter case $b = c = 0$, $a = d$ and (9.3–4) represents the identity. If $c = 0$ and $a \neq d$, then (9.3–15) is linear and we agree that it has a root $z = \infty$. If also $a = d$ then we agree that $z = \infty$ is a root counted doubly. As a consequence

An automorphism with more than two different fixed points is the identity.

This result enables us to write down directly an expression for an automorphism which transforms three different points z_1, z_2, z_3 into three other different points w_1, w_2, w_3, viz.,

$$\frac{w-w_2}{w-w_1} : \frac{w_3-w_2}{w_3-w_1} = \frac{z-z_2}{z-z_1} : \frac{z_3-z_2}{z_3-z_1}. \tag{9.3–16}$$

In fact, if W denotes either member of this equation then W is a linear fractional function of z as well as of w. To $z = z_1, z_2, z_3$ correspond the values $W = \infty$, 0, 1 and, consequently, to $w = w_1, w_2, w_3$.

By the *cross-ratio* of four points we understand

$$(z_1, z_2, z_3, z_4) = \frac{z_4-z_2}{z_4-z_1} : \frac{z_3-z_2}{z_3-z_1} = \frac{z_3-z_1}{z_4-z_1} : \frac{z_3-z_2}{z_4-z_2}. \tag{9.3–17}$$

It is at once clear that

$$(z_1, z_2, z_3, z_4) = (w_1, w_2, w_3, w_4) \tag{9.3–18}$$

if w_4 corresponds to z_4 because of (9.3–16). Hence

The cross ratio is an invariant for linear fractional transformations and hence for automorphisms of the extended plane.

9.3.5 – Conjugate automorphisms

If we perform the automorphism

$$z^* = \mathsf{P}z \tag{9.3–19}$$

then the relation

$$w = \mathsf{A}z \tag{9.3–20}$$

becomes

$$w^* = \mathsf{P}w = \mathsf{PA}z = \mathsf{PAP}^{-1}z^*. \tag{9.3–21}$$

Two automorphisms A and PAP^{-1} are called *conjugate*.

The relation of conjugacy is *reflexive*, i.e., every automorphism is conjugate to itself. This follows from

$$\mathsf{A} = \mathsf{EAE}^{-1}.$$

The relation is *symmetric*, for

$$\mathsf{B} = \mathsf{PAP}^{-1}$$

implies

$$\mathsf{A} = \mathsf{P}^{-1}\mathsf{BP} = \mathsf{P}^{-1}\mathsf{B}(\mathsf{P}^{-1})^{-1}.$$

The relation is *transitive*, for if

$$\mathsf{B} = \mathsf{PAP}^{-1}, \qquad \mathsf{C} = \mathsf{QBQ}^{-1},$$

then

$$\mathsf{C} = \mathsf{QPAP}^{-1}\mathsf{Q}^{-1} = (\mathsf{QP})\mathsf{A}(\mathsf{QP})^{-1}.$$

Hence the relation of conjugacy is an *equivalence relation* and we can divide all automorphisms into disjoint classes of mutually conjugate automorphism.

Next we wish to prove that two conjugate automorphisms have a characteristic in common. By the *trace* of the matrix (9.3–8) we understand the number

$$\mathrm{tr}\,\mathsf{A} = a+d. \tag{9.3–22}$$

If A is unimodular it is determined within sign.

Let P be represented by the unimodular matrix

$$\mathsf{P} = \begin{bmatrix} p & q \\ r & s \end{bmatrix}.$$

If

$$\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} = \begin{bmatrix} p & q \\ r & s \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} s & -q \\ -r & p \end{bmatrix},$$

we find by straight forward computation

$$a'+d' = a+d, \tag{9.3–23}$$

i.e.,

$$\mathrm{tr}\,\mathsf{PAP}^{-1} = \mathrm{tr}\,\mathsf{A}.$$

By the *fundamental number* of an automorphism represented by the unimodular matrix (9.3–8) we shall understand the modulus of the trace, i.e. $|a+d|$. Hence

Two conjugate automorphisms have the same fundamental number.

As we shall see in the next section it plays an important part in the problem of classifying automorphisms.

9.3.6 – Classification of the automorphisms

It is always possible to perform a transformation which carries two given different points into 0 and ∞ respectively. An automorphism with these latter points as fixed points is given by

$$w = \kappa z \tag{9.3–24}$$

where κ is a constant different from unity (for we wish to exclude the identity). We can normalize it by writing

$$w = \frac{\pm z\sqrt{\kappa}}{\pm 1/\sqrt{\kappa}} \tag{9.3–25}$$

and the trace appears to be equal to $\pm(\sqrt{\kappa}+1/\sqrt{\kappa})$. Hence

$$\boxed{(a+d)^2-2 = \kappa+\frac{1}{\kappa}.} \tag{9.3–26}$$

An automorphism with a fixed point of multiplicity two is conjugate to an automorphism with a double fixed point at infinity. It can be represented by

$$w = z+p = \frac{z+p}{1} = \frac{-z-p}{-1}, \qquad p \neq 0. \tag{9.3–27}$$

Hence

$$a+d = \pm 2$$

and (9.3–26) is also valid for this case if we agree to take $\kappa = 1$ in this formula.

We are now sufficiently prepared to list the various types of automorphisms:

i) First we assume that κ occurring in (9.3–26) is real and positive. The expression on the right of (9.3–26) takes its minimum value equal to 2 for $\kappa = 1$. Hence for $\kappa \neq 1$ we have $(a+d)^2 > 4$ and $|a+d|$ is real and > 2. An automorphism of this type is called *hyperbolic*. In this case (9.3–24) represents a *stretching* issuing from the origin. The number κ is a measure of the change of scale connected with it.

ii) Secondly we assume that κ is the number $e^{i\varphi}$, $\varphi \neq 2n\pi$, where n is an integer. Then

$$\kappa + \frac{1}{\kappa} = e^{i\varphi} + e^{-i\varphi} = 2\cos\varphi = 4\cos^2 \tfrac{1}{2}\varphi - 2$$

and

$$(a+d)^2 = 4\cos^2 \tfrac{1}{2}\varphi < 4.$$

Hence again $a+d$ is real, but now $|a+d| < 2$. An automorphism of this type is called *elliptic*. In this case the transformation (9.3–24) represents a *rotation* about the origin through an angle φ.

iii) The case $\kappa = 1$ corresponds to an automorphism having a doubly counted fixed point. An automorphism of this type is called *parabolic*. The particular transformation (9.3–27) represents a *translation*.

iv) Finally we assume that κ is a number $re^{i\varphi}$, $r > 0$, $r \neq 1$, $\varphi \neq 2n\pi$, where n is an integer. Then

$$\kappa + \frac{1}{\kappa} = re^{i\varphi} + \frac{1}{r}e^{-i\varphi} = \left(r + \frac{1}{r}\right)\cos\varphi + i\left(r - \frac{1}{r}\right)\sin\varphi.$$

This denotes either a complex or a negative real number. In both cases $a+d$ is not real. An automorphism of this kind is called *loxodromic*. The transformation (9.3–24) is evidently the product of a stretching and a rotation. This operation is often called an *elation*.

Summing up we have:

The automorphism as given by the unimodular linear fractional transformation

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

is

1) *hyperbolic, if* $a+d$ *is real and* $|a+d| > 2$;
2) *elliptic, if* $a+d$ *is real and* $|a+d| < 2$;
3) *parabolic, if* $a+d$ *is real and* $|a+d| = 2$;
4) *loxodromic, if* $a+d$ *is complex.*

Since each of these cases excludes the remaining ones there are no other types of automorphisms of the extended plane.

9.3.7 – Rotations of the complex sphere

By means of stereographic projection a rotation of the complex sphere about its centre induces a homeomorphism of the extended plane. It is characterized by the property that the chordal distance of two points remains invariant. In particular two diametral points on the sphere are carried into two other diametral points. The chordal distance between the image z_1 and z_2 of two diametral points is 2. Hence (cf. 1.1–15)

$$|z_1 - z_2| = \sqrt{1+z_1\bar{z}_1}\sqrt{1+z_2\bar{z}_2},$$

or

$$(z_1 - z_2)(\bar{z}_1 - \bar{z}_2) = (1+z_1\bar{z}_1)(1+z_2\bar{z}_2).$$

This is equivalent to

$$0 = (1+z_1\bar{z}_2)(1+\bar{z}_1 z_2) = |1+z_1\bar{z}_2|^2.$$

Hence

$$\boxed{z_1\bar{z}_2 = -1.} \qquad (9.3\text{–}28)$$

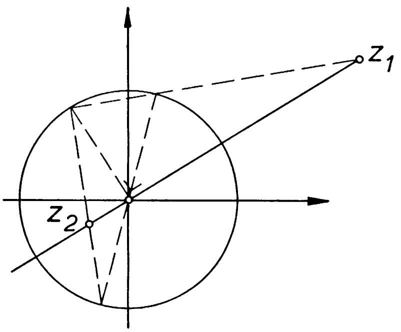


Fig. 9.3–1. Construction of the images of diametral points in the plane.

For the sake of brevity we shall call two points in the extended plane having a chordal distance equal to 2 also diametral points. A construction of diametral points is pictured in fig. 9.3–1.

Now we wish to determine those automorphisms of the extended plane which carry diametral points into diametral points. If

$$w_1 = \frac{az_1+b}{cz_1+d}, \tag{9.3–29}$$

we must have

$$\frac{-1}{\bar{w}_2} = \frac{-a/\bar{z}_2+b}{-c/\bar{z}_2+d},$$

or

$$w_2 = \frac{\bar{d}z_2-\bar{c}}{-\bar{b}z_2+\bar{a}}. \tag{9.3–30}$$

Since (9.3–29) and (9.3–30) represent the same automorphism there must be a number ρ such that

$$a = \rho\bar{d}, \quad b = -\rho\bar{c}, \quad c = -\rho\bar{b}, \quad d = \rho\bar{a}.$$

It follows that

$$ad-bc = \rho^2(\bar{a}\bar{d}-\bar{b}\bar{c}).$$

Since the transformation is assumed to be unimodular we have $\rho^2 = 1$ and we can take $\rho = 1$. Thus we find the so-called *unitary* transformations

$$\boxed{w = \frac{az+b}{-\bar{b}z+\bar{a}}, \qquad a\bar{a}+b\bar{b} = 1.} \tag{9.3–31}$$

We assert that *these transformations leave the chordal distance between two arbitrary points invariant*, i.e., they are *isometric.*

It is clear that the automorphisms represented by (9.3–31) constitute a transitive group. In the usual way with the help of the lemma of section 9.3.1 we may show that this group coincides with the group of all isometric automorphisms. The subgroup of isotropy of this latter group with the invariant point $z = 0$ is given by

$$z^* = \kappa z$$

with $|\kappa| = 1$. Hence $\kappa = e^{i\varphi}$ and all these transformations are characterized by the matrix

$$\begin{bmatrix} e^{\frac{1}{2}i\varphi} & 0 \\ 0 & e^{-\frac{1}{2}i\varphi} \end{bmatrix}.$$

This matrix is unitary.

Finally we wish to show that *the group of isometric automorphisms is isomorphic to the group of rotations of the complex sphere.*

The isometric automorphisms induce rotations of the sphere, for the chordal distance of two points in the extended plane is the euclidean distance of the corresponding points on the sphere. Consider now the rotations of the sphere around the vertical axis. They are given by

$$\begin{aligned} \xi^* &= \xi \cos \varphi - \eta \sin \varphi \ , \\ \eta^* &= \xi \sin \varphi + \eta \cos \varphi \ , \\ \zeta^* &= \zeta . \end{aligned} \tag{9.3–32}$$

It follows that

$$\frac{\xi^* + i\eta^*}{1+\zeta^*} = \frac{\xi + i\eta}{1+\zeta} (\cos \varphi + i \sin \varphi),$$

or, in view of (1.1–12),

$$z^* = ze^{i\varphi}.$$

Hence these rotations are induced by a subgroup of those induced by the isometric automorphisms and they constitute the subgroup of isotropy of the group of all rotations which have the north pole as a fixed point. This proves the assertion.

It should be noticed that isometric automorphisms do not induce a reflection of the sphere with respect to a plane through the centre, for they tend continuously to the identity if $a \to 1$, $b \to 0$.

9.3.8 – The invariant axis and the eulerian angles

It is well-known that a rotation of a sphere around its centre is a rotation through a certain angle ω around a line passing through the centre, the *invariant axis* of the rotation. This may be verified by analyzing more closely the isometric automorphisms.

First we observe that they are elliptic, for $a+d = a+\bar{a}$ is real and

$$(a+\bar{a})^2 - 4 = (a+\bar{a})^2 - 4(a\bar{a}+b\bar{b}) = (a-\bar{a})^2 - 4b\bar{b} < 0.$$

The fixed points are obtained from equation (9.3–15) which now takes the form

$$\bar{b}z^2 + (a-\bar{a})z + b = 0. \tag{9.3–33}$$

Since $|a+\bar{a}| < 2$, we may put

$$a+\bar{a} = 2 \cos \tfrac{1}{2}\omega, \qquad 0 < \omega < 2\pi. \tag{9.3–34}$$

Hence ω is uniquely determined. The cases $\omega = 0$ and $\omega = 2\pi$ are of no interest, for they correspond to the identity (for parabolic transformations do not occur). Hence we can write

$$a = \cos \tfrac{1}{2}\omega + i\gamma \sin \tfrac{1}{2}\omega, \tag{9.3–35}$$

where γ is a uniquely determined real number. In addition we put

$$ib = (\alpha + i\beta) \sin \tfrac{1}{2}\omega \tag{9.3–36}$$

and, evidently, α and β are also uniquely determined. Inserting the expressions (9.3–35) and (9.3–36) into (9.3–33) we get, since $\sin \frac{1}{2}\omega \neq 0$,

$$(\alpha - i\beta)z^2 + 2\gamma z - (\alpha + i\beta) = 0. \tag{9.3–37}$$

In addition we deduce from (9.3–35) and (9.3–36)

$$\begin{aligned} 1 = a\bar{a} + b\bar{b} &= \cos^2 \tfrac{1}{2}\omega + \gamma^2 \sin^2 \tfrac{1}{2}\omega + (\alpha^2 + \beta^2) \sin^2 \tfrac{1}{2}\omega \\ &= 1 + (\alpha^2 + \beta^2 + \gamma^2 - 1) \sin^2 \tfrac{1}{2}\omega. \end{aligned}$$

Since $\sin^2 \frac{1}{2}\omega \neq 0$, we must have

$$\alpha^2 + \beta^2 + \gamma^2 = 1. \tag{9.3–38}$$

Solving (9.3–37) we find the roots

$$z_{1,2} = \frac{-\gamma \pm \sqrt{\gamma^2 + \alpha^2 + \beta^2}}{\alpha - i\beta} = \frac{\pm 1 - \gamma}{\alpha - i\beta},$$

provided the denominator is not zero. This denominator vanishes, however, if $\alpha = \beta = 0$ and this implies $b = 0$. But this case is directly accessible, for then we have a rotation about the vertical axis.

Let us take

$$z_1 = \frac{1 - \gamma}{\alpha - i\beta} = \frac{\alpha + i\beta}{1 + \gamma},$$

and

$$z_2 = \frac{-1 - \gamma}{\alpha - i\beta} = \frac{-\alpha - i\beta}{1 - \gamma}.$$

It appears that the fixed points z_1 and z_2 correspond to the diametral points $\{\alpha, \beta, \gamma\}$ and $\{-\alpha, -\beta, -\gamma\}$ on the sphere.

Since $a + \bar{a} = 2 \cos \frac{1}{2}\omega$ is the trace of the transformation we find that

$$\kappa + \frac{1}{\kappa} = 2 \cos \omega. \tag{9.3–39}$$

Hence we may take

$$\kappa = e^{i\omega} \tag{9.3–40}$$

and since the stereographic projection is isogonal we may interpret ω as the angle of rotation as seen at the image of z_1 on the sphere.

By geometric arguments it can be shown that any rotation of the sphere can be considered as the product of three rotations. A first rotation around the ζ-axis through an angle ψ, a second rotation around the ξ-axis through an angle θ and a third rotation around the ζ-axis again through

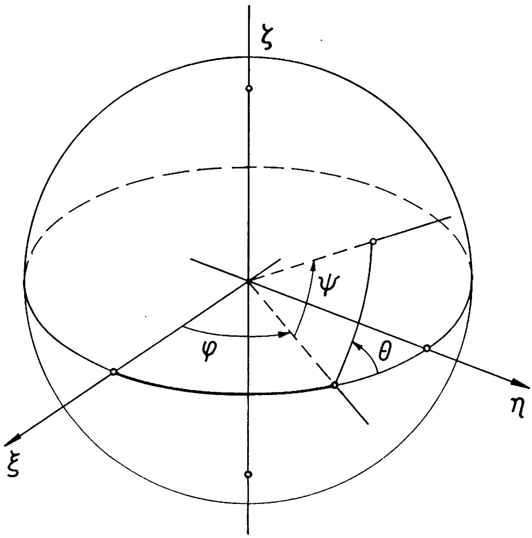


Fig. 9.3–2. The Eulerian angles.

an angle φ, (fig. 9.3–2). The angles ψ, θ and φ are called the *Eulerian angles* of the resulting rotation. We wish to investigate how they appear in the image under stereographic projection onto the extended plane.

Rotations around the ζ-axis through angles φ and ψ respectively are represented by automorphisms with matrices

$$\mathsf{D}_\varphi = \begin{bmatrix} e^{\frac{1}{2}i\varphi} & 0 \\ 0 & e^{-\frac{1}{2}i\varphi} \end{bmatrix}, \quad \mathsf{D}_\psi = \begin{bmatrix} e^{\frac{1}{2}i\psi} & 0 \\ 0 & e^{-\frac{1}{2}i\psi} \end{bmatrix} \qquad (9.3\text{–}41)$$

respectively (section 9.3–7). A rotation around the ξ-axis is induced by an automorphism whose fixed points are $+1$ and -1 and, therefore, is given by the transformation

$$\frac{w-1}{w+1} = e^{i\theta}\,\frac{z-1}{z+1}$$

whence

$$w = \frac{z(1+e^{i\theta})+(1-e^{i\theta})}{z(1-e^{i\theta})+(1+e^{i\theta})} = \frac{z\cos\frac{1}{2}\theta - i\sin\frac{1}{2}\theta}{-iz\sin\frac{1}{2}\theta+\cos\frac{1}{2}\theta}\,.$$

This automorphism is represented by the matrix

$$\mathsf{D}_\theta = \begin{bmatrix} \cos\frac{1}{2}\theta & -i\sin\frac{1}{2}\theta \\ -i\sin\frac{1}{2}\theta & \cos\frac{1}{2}\theta \end{bmatrix} \tag{9.3–42}$$

and the resulting rotation by

$$\mathsf{D}_\varphi \mathsf{D}_\theta \mathsf{D}_\psi = \begin{bmatrix} e^{\frac{1}{2}i(\varphi+\psi)}\cos\frac{1}{2}\theta & -ie^{\frac{1}{2}(\varphi-\psi)}\sin\frac{1}{2}\theta \\ -ie^{-\frac{1}{2}i(\varphi-\psi)}\sin\frac{1}{2}\theta & e^{-\frac{1}{2}i(\varphi+\psi)}\cos\frac{1}{2}\theta \end{bmatrix}. \tag{9.3–43}$$

It is an automorphism of the type (9.3–31) with

$$a = e^{\frac{1}{2}i(\varphi+\psi)}\cos\tfrac{1}{2}\theta, \qquad ib = e^{\frac{1}{2}i(\varphi-\psi)}\sin\tfrac{1}{2}\theta. \tag{9.3–44}$$

These expressions are called the *Cayley-Klein parameters* of the rotation.

Conversely it is possible to find Eulerian angles such that the Cayley-Klein parameters have prescribed values, provided that $a\bar{a}+b\bar{b} = 1$. This follows from the fact that

$$a\bar{a} = \cos^2\tfrac{1}{2}\theta, \qquad b\bar{b} = \sin^2\tfrac{1}{2}\theta.$$

Since $0 < \frac{1}{2}\theta < \pi$ the angle θ is uniquely determined. From (9.3–44) follow $\varphi+\psi$ and $\varphi-\psi$ and we are ready.

From (9.3–35) and (9.3–36) we may derive expressions for the parameters α, β, γ, ω of the rotation in terms of the Cayley-Klein parameters, viz.:

$$\begin{aligned} \cos\tfrac{1}{2}\omega &= \tfrac{1}{2}(a+\bar{a}), \\ \alpha\sin\tfrac{1}{2}\omega &= \tfrac{1}{2}i(b-\bar{b}), \\ \beta\sin\tfrac{1}{2}\omega &= \tfrac{1}{2}(b+\bar{b}), \\ \gamma\sin\tfrac{1}{2}\omega &= -\tfrac{1}{2}i(a-\bar{a}). \end{aligned} \tag{9.3–45}$$

Taking into account (9.3–44) we deduce from these equations the following

$$\begin{aligned} \cos\tfrac{1}{2}\omega &= \cos\tfrac{1}{2}(\varphi+\psi)\cos\tfrac{1}{2}\theta, \\ \alpha\sin\tfrac{1}{2}\omega &= \cos\tfrac{1}{2}(\varphi-\psi)\sin\tfrac{1}{2}\theta, \\ \beta\sin\tfrac{1}{2}\omega &= \sin\tfrac{1}{2}(\varphi-\psi)\sin\tfrac{1}{2}\theta, \\ \gamma\sin\tfrac{1}{2}\omega &= \sin\tfrac{1}{2}(\varphi+\psi)\cos\tfrac{1}{2}\theta, \end{aligned}$$

expressing the parameters of the rotation in terms of the Eulerian angles.

9.4 – Möbius geometry

9.4.1 – The Möbius plane

The group of all automorphisms of the extended plane impose on it a certain geometric structure which is called the *Möbius geometry*. The extended plane considered as a support of the Möbius geometry is called the *Möbius plane*. Two configurations of points in the Möbius plane are said to be *equivalent* if there is an automorphism which carries one of the configurations into the other. If we wish to emphasize the geometric side of the automorphisms we also refer to them as *Möbius transformations.* Only those geometric assertions are relevant which remain equally valid for all entities which are equivalent under Möbius transformations.

In the Möbius plane we cannot distinguish between circles and straight lines. This is a consequence of the following theorem.

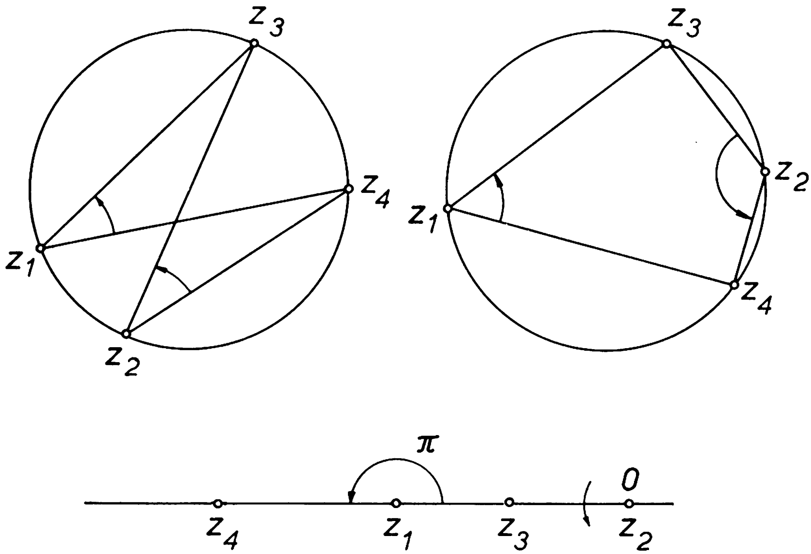


Fig. 9.4–1. Invariance of the circle with respect to a Möbius transformation.

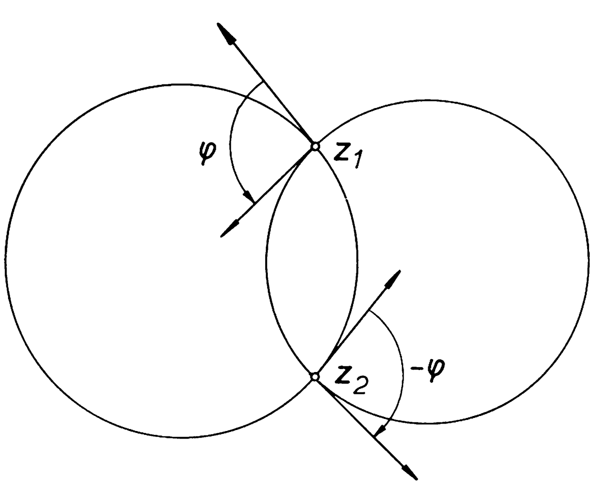


Fig. 9.4–2. The angles at the intersection points of two oriented circles.

A Möbius transformation transforms a circle or a straight line into a circle or a straight line.

A geometric proof runs as follows. If z_1, z_2, z_3 and z_4 are four points on a circle or a straight line then

$$\arg\frac{z_3-z_1}{z_4-z_1} \quad \text{and} \quad \arg\frac{z_3-z_2}{z_4-z_2}$$

are either equal or their sum is π, (fig. 9.4–1). In the case of a straight line each of these angles is 0 or π. In all cases the cross ratio (z_1, z_2, z_3, z_4) is a real number. Since the converse is also true the theorem follows from the invariance of the cross-ratio (section 9.3.3). A straight line shall be considered as a circle through the point at infinity.

Since the Möbius transformations are conformal, preserving the sense of angles, the notion of angle belongs to the Möbius geometry. In this respect the following remark deserves mention. Consider two circles passing through the finite points z_1 and z_2, (fig. 9.4–2). Imposing an orientation on two arcs connecting these points, the tangent half-rays are uniquely determined. It is clear that the angle between the half rays issuing from z_1 is the negative of the angle formed by the half rays at z_2. This is in accordance with the fact that a Möbius transformation having z_1 and z_2 as fixed points rotates the half rays at z_1 in a sense which is opposite to the sense of rotation at the other point. It is, therefore, natural to define the angle between two half-rays at $z = \infty$ as the negative of that at the finite vertex (see also section 9.2.2).

9.4.2 – Inversion

We encountered the mapping $w = 1/z$ several times. For many purposes the mapping

$$w = 1/\bar{z} \tag{9.4–1}$$

deserves mention. Combined with a reflection with respect to the real axis it yields the former mapping. Let $z = re^{i\theta}$. Then $w = e^{i\theta}/r$. Hence corresponding points are on the same ray issuing from the origin in such a way that the product of their distances from the origin is equal to unity (fig. 9.4–3). The transformation (9.4–1) is called an *inversion* with respect to the unit circle. It is an anti-isogonal mapping.

We may enlarge the circle and translate its centre to a point $z = a$, i.e., we may transform (9.4–1) by the transformation

$$w = Rz + a. \tag{9.4–2}$$

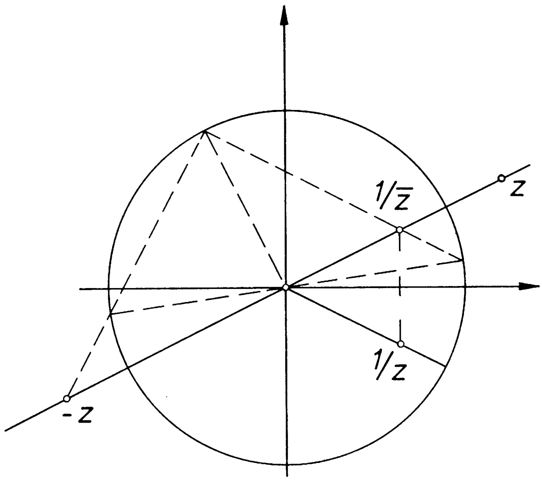


Fig. 9.4–3. Inversion of a point with respect to a unit circle.

Then we get a transformation conjugate to (9.4–1):

$$w = a + \frac{R^2}{\bar{z} - \bar{a}}. \tag{9.4–3}$$

This is an inversion with respect to a circle around $z = a$ having a radius R. If $z = a + re^{i\theta}$ then $w = a + R^2 e^{-i\theta}/r$ and we see that w and z are two points on a ray issuing from the centre of the circle in such a way that the product of their distances from the centre is equal to the square of the radius.

Let us now perform the transformation

$$w = i\,\frac{1+z}{1-z}. \tag{9.4–4}$$

It is easily seen that the unit circle is transformed into the real axis, for the points $z = -1, -i, 1$ correspond to $w = 0, 1, \infty$ respectively. Transforming the transformation (9.4–1) according to (9.4–4) we get

$$w = \bar{z}. \tag{9.4–5}$$

This is a reflection with respect to the real axis. For this reason the transformation (9.4–1) is called a *reflection with respect to the unit circle* and, more generally, any inversion is called a reflection. Two corresponding points are said to be *symmetric* with respect to the circumference which defines the inversion.

A characteristic property of two points symmetric with respect to a straight line is the following. All circles which pass through these points are orthogonal to the line and circles passing through one point and being orthogonal to the line pass through the other point. Since orthogonality is preserved under Möbius transformation we have

All circles passing through two points which are symmetric with respect to a circle are orthogonal to the circle. All circles which pass through a point and are orthogonal to a given circle also pass through the point which is symmetric with respect to this circle.

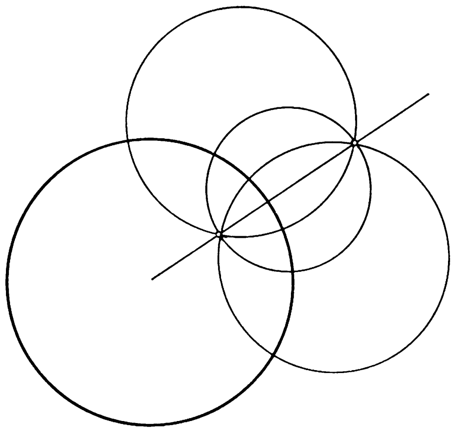


Fig. 9.4–4. Invariant characterization of the symmetry of two points with respect to a circle.

This is an invariant characterization of symmetric points, (fig. 9.4–4). Hence if a Möbius transformation carries a circle C_1 into a circle C_2, then points which are symmetric with respect to C_1 are carried into points which are symmetric with respect to C_2.

9.4.3 – Pencils of circles

We wish to prove the following theorem

There are infinitely many circles which are orthogonal to two given circles C_1 and C_2.

Let us first assume that C_1 and C_2 intersect in two different points. By a suitable Möbius transformation we can carry one of them to infinity and we thus obtain two straight lines through a finite point z_0. All circles with z_0 as centre are orthogonal to these lines and they correspond to circles orthogonal to C_1 and C_2. Secondly we suppose that C_1 and

C_2 have one point in common, i.e., that they are tangent at this point. Bringing this point to infinity we obtain two parallel lines and all lines orthogonal to them correspond to circles orthogonal to the given circles.

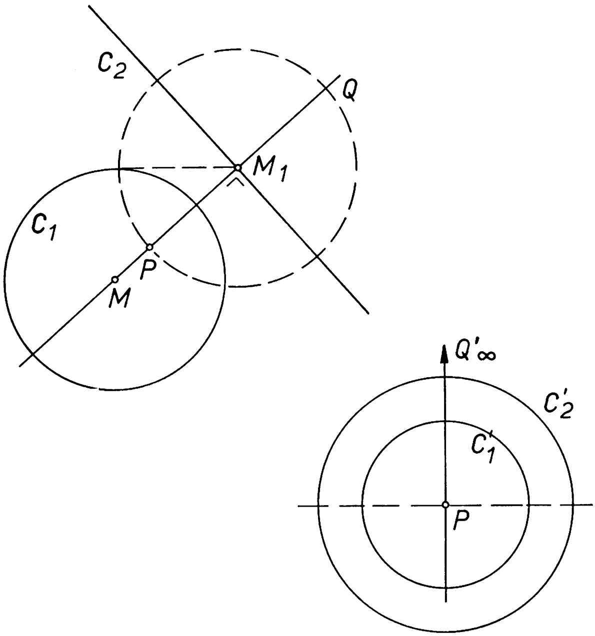


Fig. 9.4–5. Proof of the existence of circles orthogonal to two non intersecting circles.

The case that the circles do not intersect remains to be investigated. Then we may bring a point of one of these circles to infinity and we obtain a line and a circle which have no points in common, (fig. 9.4–5). Let M denote the centre of the circle and M_1 the foot of the perpendicular through M on the line. The circle with centre M_1 and of radius equal to the segment on a tangent through M_1 at the circle determined by M_1 and the point of contact is orthogonal to the circle and the line. Bringing now one of the points where this latter circle cuts the line MM_1 to infinity we obtain two orthogonal lines and the line and the circle transform into circles which meet these lines under right angles. Thus we see that the given circles can be transformed into two concentric circles. The lines through the common centre correspond to circles orthogonal to the given circles.

The set of all circles orthogonal to two given circles is called a *pencil*.

From the above considerations it is clear that there are three types. First a pencil of circles which can be transformed into a system of concentric circles. A pencil of this type is called *elliptic*. No two circles of the pencil have a point in common and there are two points symmetric with respect to any circle of the pencil. They are the *limiting points* of the pencil.

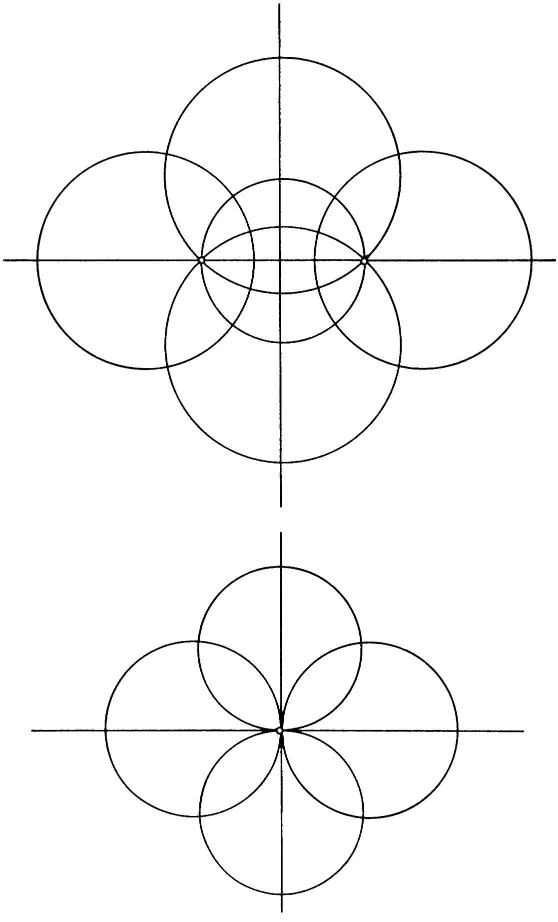


Fig. 9.4–6. Pencils of circles.

Secondly a pencil of circles which can be transformed into a system of parallel lines. The circles touch each other at the same point. A pencil of this type is called *parabolic*.

Thirdly a pencil of circles which can be transformed into a system of lines through a finite point. A pencil of this type is called *hyperbolic* All circles have two points in common, the *base points* of the pencil.

From the standard types (i.e. pencils of straight lines or of concentric circles) it is also clear that

A given pencil is always connected with a second pencil such that each circle of one pencil is orthogonal to all circles of the other. If one pencil is elliptic then the other is hyperbolic. The limiting points of the first pencil are the base points of the second pencil. If one pencil is parabolic, then the other is also parabolic, (fig. 9.4–6).

In addition we find

Through every point of the plane which is neither a limiting point nor a point common to all circles of a pencil there passes exactly one circle of the pencil.

From the standard forms (9.3–24) and (9.3–27) we deduce:

A hyperbolic transformation leaves all circles of a hyperbolic pencil invariant, the base points being the fixed points of the transformation.

An elliptic transformation leaves all circles of an elliptic pencil invariant, the limiting points being the fixed points of the transformations.

A parabolic transformation leaves all circles of a parabolic pencil invariant. The fixed point of the transformation is the point of common contact of the circles.

In all these cases the circles of the orthogonal pencil are interchanged, i.e., the pencil as a whole remains unaltered.

Finally we shall establish

If two circles do not intersect, but intersect a third circle, then there is exactly one circle orthogonal to all three.

Let C_1, C_2 and C_3 denote these circles and assume that C_2 and C_3 intersect as do C_1 and C_3. Hence C_1 and C_2 do not intersect. By a suitable

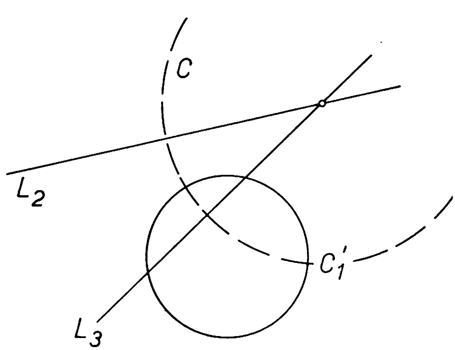


Fig. 9.4–7. Existence of a circle orthogonal to three circles.

transformation (if necessary) we may bring one of the intersections of C_2 and C_3 to infinity. They become straight lines L_2 and L_3 and C_1 another proper circle C_1', (fig. 9.4–7). Since L_2 has no point in common with C_1' the intersecting point of L_2 and L_3 is outside C_1'. This point is centre of a circle orthogonal to C_1'. Transforming back we obtain the desired circle.

9.5 – Hyperbolic geometry

9.5.1 – The automorphisms of the unit circle

It is our aim to construct a plane geometry which can be described by the same postulates as the ordinary Euclidean geometry, except the postulate of parallels.

This geometry has played an important part in the development of function theory. Our main task will consist of defining a group of transformations which will enable us to introduce the notion of congruence with the same properties as the corresponding notions in ordinary geometry. As we shall see this group will appear as the group of automorphisms of the unit circle.

On the transformation

$$w = \frac{az+b}{cz+d}, \qquad ad-bc = 1, \tag{9.5–1}$$

we impose the condition that $z\bar{z} = 1$ implies $w\bar{w} = 1$. Hence the circumference $|z| = 1$ remains invariant as a whole. In addition we wish that $z\bar{z} < 1$ implies $w\bar{w} < 1$, that is to say, the points in the interior are carried into points also in the interior.

The first condition yields

$$(az+b)(\bar{a}\bar{z}+\bar{b}) = (cz+d)(\bar{c}\bar{z}+\bar{d}),$$

or, taking into account $z\bar{z} = 1$,

$$a\bar{b} = c\bar{d} \tag{9.5–2}$$

and

$$a\bar{a}-c\bar{c} = d\bar{d}-b\bar{b}. \tag{9.5–3}$$

Now we have

$$w\bar{w}-1 = \frac{(a\bar{a}-c\bar{c})(z\bar{z}-1)}{|cz+d|^2}$$

and the second condition gives rise to

$$a\bar{a}-c\bar{c} > 0. \tag{9.5–4}$$

It follows that $a \neq 0$ and from (9.5–2) also $d \neq 0$. For from (9.5–2) we deduce $\bar{a}b = \bar{c}d$, and $d = 0$ implies $b = 0$, in contradiction to (9.5–3) and (9.5–4). From (9.5–2) follows

$$\frac{c}{a} = \frac{\bar{b}}{\bar{d}} = k, \tag{9.5–5}$$

say, and from (9.5–4) follows $|k| < 1$. Hence (9.5–3) may be written as

$$a\bar{a}(1-k\bar{k}) = d\bar{d}(1-k\bar{k}),$$

whence

$$|a| = |d|, \qquad |b| = |c|. \tag{9.5–6}$$

Now we write

$$w = \frac{az+b}{cz+d} = \frac{a}{d}\,\frac{z+b/a}{1+zc/d}.$$

Then, according to (9.5–6), $a/d = e^{i\theta}$. Put $z_0 = -b/a$. In view of (9.5–5) and (9.5–6): $z_0 = -b\bar{a}/a\bar{a} = -\bar{c}d/d\bar{d} = -\bar{c}/\bar{d}$, $z_0 = -k(b/c)$, whence $|z_0| = k < 1$.

Hence

The transformations

$$\boxed{w = e^{i\theta}\,\frac{z-z_0}{1-z\bar{z}_0},} \qquad |z_0| < 1, \tag{9.5–7}$$

represent automorphisms of the interior of the unit circle. They transform the circumference into itself.

Introducing the numbers

$$\alpha = e^{\frac{1}{2}i\theta}/\sqrt{1-z_0\bar{z}_0}, \qquad \beta = -\alpha z_0,$$

the transformations appear in the form

$$\boxed{w = \frac{\alpha z+\beta}{\bar{\beta}z+\bar{\alpha}},} \qquad \alpha\bar{\alpha}-\beta\bar{\beta} = 1, \tag{9.5–8}$$

which are quite similar to the isometric automorphisms (9.3–31) of the extended plane.

It is clear that the automorphisms (9.5–7) or (9.5–8) constitute a transitive group. Next we investigate the subgroup of isotropy of all automorphisms of the interior of the unit circle with fixed point at the point $z = 0$. If $z \to f(z)$ is such an automorphism, then $|z| < 1$ entails $|f(z)| < 1$. But Schwarz's lemma (section 2.21.1) implies $|f(z)| \leqq |z|$. Since the same argument holds for the inverse automorphism we also have $|z| \leqq |f(z)|$. Hence $|z| = |f(z)|$, and then Schwarz's lemma states that $f(z) = e^{i\theta}z$. Thus we see that the subgroup of isotropy is included in the group of transformations (9.5–7) and according to the lemma of section 9.3.1 this group is the group of all automorphisms of the interior of the unit circle.

Finally we remark that there are no loxodromic automorphisms of the interior of the unit circle. This is a direct consequence of the fact that the trace of the matrix of the transformations (9.5–8) is $\alpha+\bar{\alpha}$, i.e., real.

9.5.2 – The hyperbolic plane

A model of the *hyperbolic geometry* is obtained if we consider the interior points of the unit circle around the origin as "points". This interior is called the *hyperbolic plane*. The circumference of the unit circle shall be denoted by Ω. The part of a circle orthogonal to Ω (a diameter not being excluded) which is in the interior of Ω shall be considered as a hyperbolic "straight line", (fig. 9.5–1). This can be motivated as follows. A point within Ω is paired with its symmetric point outside. All circles through these points constitute a pencil of circles orthogonal to Ω. Through a point different from the given point there passes just one member of the pencil, (section 9.4.3).

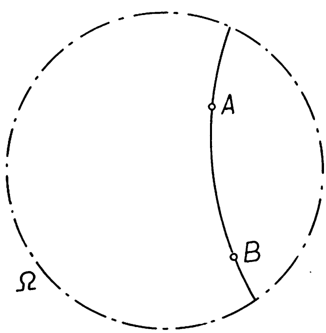


Fig. 9.5–1. The hyperbolic plane is the interior of the circumference Ω.

We have a group of automorphisms of the hyperbolic plane at our disposal. This group can be extended with the so-called reflections with respect to a hyperbolic line. If C_1 is a circle cutting Ω orthogonally then an inversion with respect to C_1 leaves Ω invariant but interchanges the regions into which the interior of Ω is divided by C_1. Such an inversion will be called a reflection. The group of automorphisms and reflections is taken as the group of *congruent transformations*. Two configurations in the hyperbolic plane are said to be *congruent* if one can be derived from the other by a congruent transformation. Since the congruent transformations constitute a group the relation "congruent" is reflexive, symmetric and transitive.

A hyperbolic line divides the hyperbolic plane into two half planes. Hence a point on a hyperbolic line divides this line into two half lines. Each half line determines a point on Ω, *the ideal point* of the half line.

Let A and B denote two points on a hyperbolic line. The first point A is the endpoint of a half line passing through B while B is the endpoint of a half line passing through A. The intersection of these half lines is called the *segment* with *end points* A and B.

With this new concepts it is not difficult to build up that part of the Euclidean geometry which is independent of the postulate of parallels. As regards the angles we notice that angles equal in Euclidean sense are also hyperbolically equal. This is a consequence of the fact that hyperbolic congruent transformations preserve Euclidean angles.

For the sake of illustration we prove

Through a point there passes just one perpendicular on a given line.

If C_1 is the circle on which the given hyperbolic line lies then there is a pencil of circles orthogonal to C_1 and Ω. This pencil is elliptic, its limiting points are the ideal points of the given line. Hence through an arbitrary point inside Ω there passes exactly one member of the pencil.

9.5.3 – Parallels and hyperparallels

The Euclidean axioms of parallels state that through a point not on a given line there is exactly one line which has no point in common with the given line. The situation in hyperbolic geometry is quite different. In order to clear this up we assume that the point is the centre O of Ω. This can always be achieved by a suitable congruent transformation. Hence the line does not pass through O and O is exterior to the circle on which the hyperbolic line lies. The tangents of this circle at the points where it intersects Ω pass through O. Now it is easy to see that these tangents separate the hyperbolic lines through O and intersect the given line from those through O which do not meet this line. The tangents through O have the ideal points with the line in common, (fig. 9.5–2).

Lines of the first category through O are lines which meet the given line at a hyperbolic point. The two tangents through O are called *parallels*. They have an ideal point in common with the given line. The lines of the third category are called *hyperparallels* to the given line. Otherwise stated: Two lines which do not intersect and are not parallel are hyperparallel. They have an interesting property which has no counterpartin Euclidean geometry.

Two hyperparallels have exactly one common perpendicular.

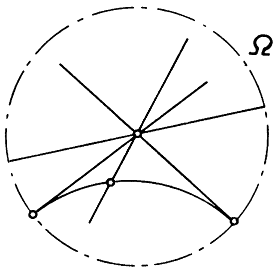


Fig. 9.5–2. Intersecting line, parallels and a hyperparallel through a point with respect to a given line.

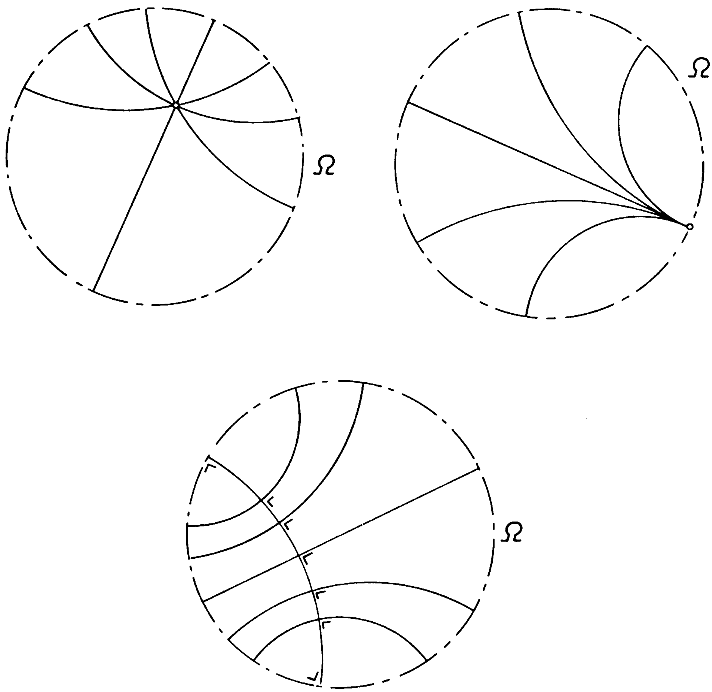


Fig. 9.5–3. The three types of pencils of lines.

This is a direct consequence of the last theorem of section 9.4.3.

In hyperbolic geometry we may distinguish three types of pencils of lines, (fig. 9.5–3):

1) All lines through a hyperbolic point.

2) All lines parallel to a given line (tending to the same ideal point).

3) All lines orthogonal to a given line.

They are defined by means of the three types of pencils of circles.

There are also three types of direct congruent transformations, i.e., automorphisms of the hyperbolic plane.

First the elliptic automorphisms. Each has a fixed point in the interior of Ω. They are called the *rotations* around their fixed point.

Secondly the parabolic automorphisms. The fixed point of each is necessarily on Ω and they are called *parallel displacements.*

Thirdly the hyperbolic automorphisms. The fixed points of each are on Ω and there is just one hyperbolic straight line connecting these fixed points. These transformations are called *translations* and the line mentioned above associated with a translation is called the *axis* of the translation.

In the Euclidean geometry parallel displacements and translations coincide and they have infinitely many axes.

9.5.4 – Cycles

By a cycle we understand an orthogonal trajectory of a pencil of hyperbolic lines. These lines are called the *diameters* of the cycle, (fig. 9.5–4).

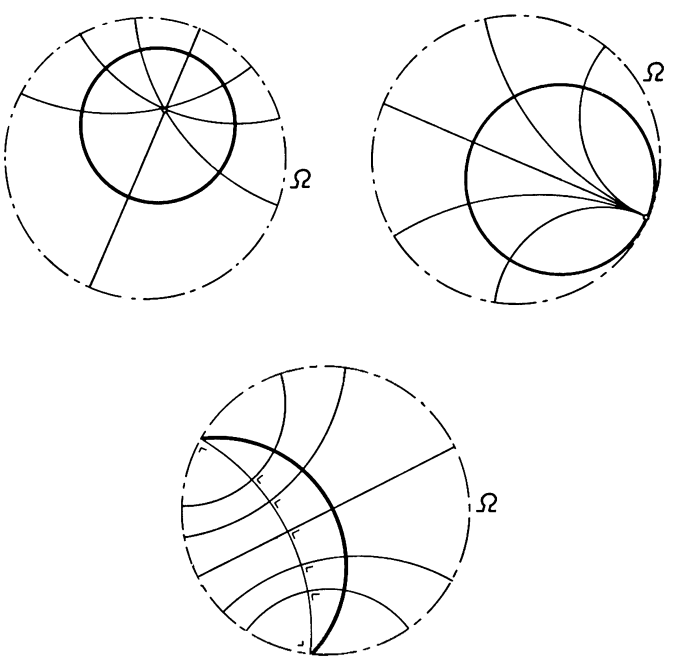


Fig. 9.5–4. Circles in hyperbolic geometry.

If the pencil consists of lines passing through a hyperbolic point then a cycle associated with it is a hyperbolic circle with this point as its centre. A circle in our model is also a circle in the Euclidean sense, but in general it has not the same centre. This is only the case for circles around the origin.

If the pencil consists of lines through the same ideal point an associated cycle is called a *horicycle.* It is a Euclidean circle which touches Ω.

All horicycles are congruent.

First we observe that all horicycles through O are congruent, for they are interchanged by a rotation about O. Secondly it is clear that by a transformation (9.5–7) any horicycle can be carried into a horicycle through O, for we may take z_0 as a point of the horicycle.

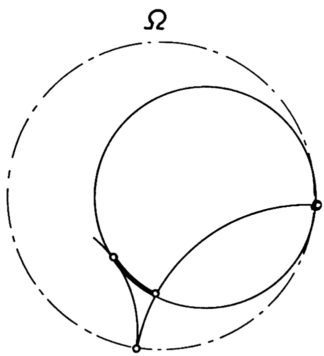


Fig. 9.5–5. A fundamental arc.

Finally we have a pencil of straight lines orthogonal to a given line. Their orthogonal trajectories are invariant for the translations with this line as their axis. They are called *hypercycles* with axis the given line. The distances of the points of a hypercycle are evidently equal; a hypercycle is a locus of points equidistant to a given line.

The cycles have many properties in common. Thus, for instance,

At a point of a given cycle there is always a tangent perpendicular to the diameter through this point.

If z_0 is a point of the cycle we can always perform a transformation which brings this point to O. Then the desired tangent corresponds to the Euclidean straight line through O which touches the transformed cycle.

A remarkable figure is a *fundamental arc,* an arc of a horicycle such that the tangent in one end point is parallel to the diameter through the other end point, (fig. 9.5–5). It is not difficult to prove that all fundamental arcs are congruent (see e.g. section 9.5.8).

9.5.5 – Hyperbolic metric

Now we turn to the problem of introducing a measure for angles and line segments.

As regards the measure of an angle there is no difficulty, for we can define it as its Euclidean measure. As we know this is an invariant for hyperbolic congruent transformations.

In order to introduce the notion of length we shall need an invariant intimately connected to the line on which the segment lies. We observe that a segment gives rise to two ideal points and the two end points of the segment are associated with a quadruple. In order to avoid ambiguities we shall make the following agreement. Let z_1, z_2 denote the end points of a given segment. By $z_{1\Omega}$ we denote the ideal point of the half ray issuing

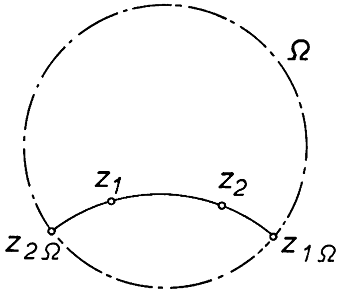


Fig. 9.5–6. The hyperbolic distance of two points.

from z_1 and passing through z_2, by $z_{2\Omega}$ the ideal point of the half ray issuing from z_2 and passing through z_1, (fig. 9.5–6). Then, evidently, the cross ratio

$$(z_1, z_2, z_{1\Omega}, z_{2\Omega}) = \frac{z_{1\Omega}-z_1}{z_{1\Omega}-z_2} : \frac{z_{2\Omega}-z_1}{z_{2\Omega}-z_2} \qquad (9.5\text{–}9)$$

is an invariant, (section 9.3.4). We shall prove that it is always greater than one. In fact, it is possible to transform the segment in such a way that $z_1 = r > 0$, $z_2 = 0$, $z_{1\Omega} = -1$, $z_{2\Omega} = 1$. Then

$$(z_1, z_2, z_{1\Omega}, z_{2\Omega}) = (r, 0, -1, 1) = \frac{1+r}{1-r} > 1. \qquad (9.5\text{–}10)$$

This proves the assertion.

Next we consider a point z_3 on the same line such that z_2 is between z_1 and z_3. This means that z_2 is on the half line issuing from z_1 and pointing to $z_{1\Omega}$ and also on the half line issuing from z_3 and pointing to $z_{2\Omega}$. An easy calculation shows that

$$(z_1, z_2, z_{1\Omega}, z_{2\Omega})(z_2, z_3, z_{1\Omega}, z_{2\Omega}) = (z_1, z_3, z_{1\Omega}, z_{2\Omega}). \qquad (9.5\text{–}11)$$

Since the notion of arc length is to be additive it is natural to define the *distance* between the points z_1 and z_2 (or, which amounts to the same, the length of the segment between these points) by the formula

$$\boxed{\operatorname{dist}(z_1, z_2) = \log(z_1, z_2, z_{1\Omega}, z_{2\Omega}).} \tag{9.5–12}$$

It follows from (9.5–10) that dist $(z_1, z_2) > 0$. Hence, if z_2 is between z_1 and z_3

$$\operatorname{dist}(z_1, z_2) < \operatorname{dist}(z_1, z_3). \tag{9.5–13}$$

In particular, if in (9.5–9) we let tend $z_2 \to z_{1\Omega}$ the distance increases beyond any bound. Thus we may say that the ideal points have an infinite distance to any other point.

The triangle inequality

$$\operatorname{dist}(z_1, z_3) \leqq \operatorname{dist}(z_1, z_2) + \operatorname{dist}(z_2, z_3), \tag{9.5–14}$$

holds for the measure defined above, equality occurring only if z_2 is between z_1 and z_3. A direct verification of this inequality is not easy. But we may recall that the analogous result in Euclidean geometry can be obtained without reference to the axiom of parallels. Hence it must also be valid in hyperbolic geometry. (See, however, also section 9.5.8).

It should be noticed that the expression on the right of (9.5–12) may be multiplied by a positive constant without the fundamental properties of distance being affected. This is, however, only a matter of scaling and not very important.

It is not very satisfactory that in (9.5–12) the ideal points $z_{1\Omega}$ and $z_{2\Omega}$ occur, for it is not always a simple matter to find them when z_1 and z_2 are given. There is, however, a possibility to express dist (z_1, z_2) in terms of z_1, z_2 only.

Let us consider an automorphism which carries z_1, z_2 to z_1^*, z_2^* respectively. According to (9.5–7) we have

$$\frac{z_2^* - z_1^*}{1 - z_2^* \bar{z}_1^*} = e^{i\theta} \frac{z_2 - z_1}{1 - z_2 \bar{z}_1}, \tag{9.5–15}$$

whence

$$\frac{|z_2^* - z_1^*|}{|1 - z_1^* \bar{z}_2^*|} = \frac{|z_2 - z_1|}{|1 - z_1 \bar{z}_2|}. \tag{9.5–16}$$

Thus we have another invariant. In the particular case considered above, where $z_1 = r$, $z_2 = 0$ this expression turns out to be equal to r and from

(9.5–10) we deduce

$$\operatorname{dist}(z_1, z_2) = \log \frac{1 + \dfrac{|z_1 - z_2|}{|1 - z_1 \bar{z}_2|}}{1 - \dfrac{|z_1 - z_2|}{|1 - z_1 \bar{z}_2|}} . \tag{9.5–17}$$

This is the desired formula.

9.5.6 – The formula of Lobatčevskij

In the hyperbolic geometry a point is the endpoint of two half rays pointing to the ideal points of an assigned line not through this point. The half of the angle between these half rays is called the *parallel angle*

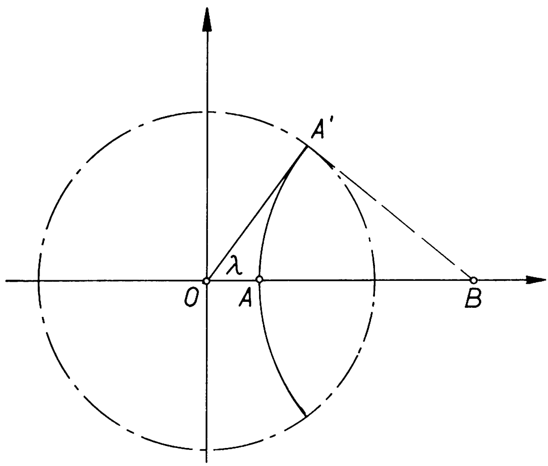


Fig. 9.5–7. The parallel angle.

associated with this point and by elementary geometry it can be shown that that it is uniquely determined by the distance of the point from the line. Lobatčevskij, the founder of hyperbolic geometry, has discovered a simple relation between this angle and the distance of the point to the line.

Without loss of generality we may assume that the point coincides with O and that the parallels issuing from O are symmetric with respect to the positive real axis. Denoting by λ the parallel angle it is easily seen that the Euclidean distance between O and the given line is, (fig. 9.5–7),

$$r = \frac{1}{\cos\lambda} - \tan\lambda = \frac{1 - \sin\lambda}{\cos\lambda} .$$

Hence

$$\frac{1+r}{1-r} = \frac{\cos\lambda - \sin\lambda + 1}{\cos\lambda + \sin\lambda - 1} = \operatorname{ctn} \tfrac{1}{2}\lambda.$$

If p denotes the hyperbolic distance from O to the line we have in view of (9.5–10),

$$e^p = \frac{1+r}{1-r} = \operatorname{ctn} \tfrac{1}{2}\lambda,$$

or, writing $\pi(p)$ for λ,

$$\boxed{\tan \tfrac{1}{2}\pi(p) = e^{-p}.} \tag{9.5–18}$$

This is the famous formula of Lobatčevskij. Alternative expressions are

$$\tan \pi(p) = \frac{2e^{-p}}{1-e^{-2p}} = \frac{1}{\sinh p},$$

$$\sin \pi(p) = \frac{2e^{-p}}{1+e^{-2p}} = \frac{1}{\cosh p},$$

$$\cos \pi(p) = \frac{1-e^{-2p}}{1+e^{-2p}} = \tanh p.$$

Thus we see that in hyperbolic geometry there is a coupling between the measure of angles and the measure of line segments.

9.5.7 – The linear element and the element of area

Dividing both members of (9.5–17) by $|z_1 - z_2|$ and making z_2 tend to z_1, we get

$$\frac{\operatorname{dist}(z_1, z_2)}{|z_1 - z_2|} \to \frac{2}{1-|z_1|^2}, \tag{9.5–19}$$

the denominator on the right being positive. It is, therefore, convenient to consider

$$\boxed{ds_h = \frac{2|dz|}{1-|z|^2}} \tag{9.5–20}$$

as the linear element at the point z, in the hyperbolic plane. It is easy to verify directly that the expression on the right is a differential invariant for hyperbolic motions. It is sufficient to check this for automorphisms. Let z_0^*, z_0^* correspond to z_0, z, then the transformation is

$$\frac{z^* - z_0^*}{1 - z^* \bar{z}_0^*} = e^{i\theta} \frac{z - z_0}{1 - z\bar{z}_0}. \tag{9.5–21}$$

If z moves along a curve $z(t)$ which passes through $z_0 = z(t_0)$ then by dividing by $t-t_0$ and if $t \to t_0$ we get

$$\frac{dz^*/dt}{1-|z_0^*|^2} = e^{i\theta} \frac{dz/dt}{1-|z_0|^2},$$

the derivatives being evaluated at $t = t_0$. This proves the assertion.

According to (9.5–10) and (9.5–12) the hyperbolic distance of a point to the origin is

$$r_h = \log \frac{1+r}{1-r}, \tag{9.5–22}$$

where r denotes its Euclidean distance. This is equivalent to

$$r = \tanh \tfrac{1}{2} r_h. \tag{9.5–23}$$

If we represent z by its polar form $z = re^{i\theta}$ an easy calculation shows that in Euclidean polar coordinates the formula for the line element appears as

$$ds_h^2 = 4 \frac{dr^2 + r^2 d\theta^2}{(1-r^2)^2}. \tag{9.5–24}$$

Making use of (9.5–23) an easy calculation shows that the formula for the line element in hyperbolic polar coordinates takes the form

$$ds_h^2 = dr_h^2 + \sinh^2 r_h d\theta^2. \tag{9.5–25}$$

We shall define the element of hyperbolic area as the product of line elements in two orthogonal directions. Taking these as the radial and lateral directions we find from (9.5–24)

$$dA_h = \frac{4r\,dr\,d\theta}{(1-r^2)^2}, \tag{9.5–26}$$

which is the area element in Euclidean polar coordinates. From (9.5–25) we have

$$dA_h = \sinh r_h dr_h d\theta, \tag{9.5–27}$$

the element of area in hyperbolic polar coordinates.

9.5.8 – Poincaré's model of the hyperbolic plane

For many purposes the simple model of the hyperbolic plane discussed above is not always convenient. Poincaré has introduced another model which can be derived from the model described by the interior of the unit circle in a simple way.

Poincaré's model, (fig. 9.5–8), is obtained when we apply a transformation which changes the interior of the unit circle in a w-plane into the upper half of the z-plane by performing the transformation

$$w = \frac{z-i}{z+i}. \tag{9.5–28}$$

Let $z = x+iy$. Then

$$|w|^2 = \frac{x^2+(y-1)^2}{x^2+(y+1)^2}.$$

Hence $|w| = 1$ for $y = 0$, i.e., if z is real, then w is on the circumference. If, however $y > 0$ then $|w| < 1$.

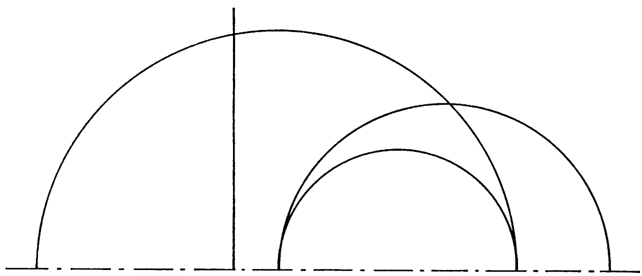


Fig. 9.5–8. Poincaré's model of the hyperbolic plane with straight lines.

The automorphisms of the upper half plane can be found on applying (9.5–28). But it is easier to derive these automorphisms directly. Since they leave the real axis invariant they must be

$$z^* = \frac{az+b}{cz+d}, \qquad ad-bc = 1, \tag{9.5–29}$$

where a, b, c and d are real. In order to show that they are all automorphisms we need only to determine the subgroup of isotropy with a suitably chosen point as a fixed point of the group of all automorphisms. In the w-plane we consider the automorphisms leaving the origin invariant. They are represented by

$$w^* = e^{i\theta}w.$$

The origin corresponds to $z = i$ in the z-plane. The inverse of (9.5–28) is

$$z = i\,\frac{1+w}{1-w}. \tag{9.5–30}$$

Hence

$$z^* = i\,\frac{1+e^{i\theta}w}{1-e^{i\theta}w} = i\,\frac{(z+i)+(z-i)e^{i\theta}}{(z+i)-(z-i)e^{i\theta}}$$
$$= \frac{z(e^{i\theta}+1)-i(e^{i\theta}-1)}{iz(e^{i\theta}-1)+e^{i\theta}+1} = \frac{z\cos\frac{1}{2}\theta+\sin\frac{1}{2}\theta}{-z\sin\frac{1}{2}\theta+\cos\frac{1}{2}\theta},$$

and these transformations are of the type (9.5–29). The group of automorphisms can be extended to a group of congruent transformations just as in section 9.5.2.

On applying the transformation (9.5–28) it is not difficult to obtain the line element in Poincaré's model. But we can also proceed in the following way. We seek a differential invariant for the transformations (9.5–29). An easy computation yields

$$dz^* = \frac{dz}{(cz+d)^2}$$

and

$$z^* - \bar{z}^* = \frac{z-\bar{z}}{(cz+d)(c\bar{z}+d)}$$

Hence

$$\frac{dz\,d\bar{z}}{|z-\bar{z}|^2}$$

is an invariant of the desired kind. Thus we may arrive at the following result. Putting $z = x+iy$, we have

The line element in Poincaré's model is given by

$$ds_h^2 = \frac{dx^2+dy^2}{y^2}\,. \tag{9.5–31}$$

In order to prove this conjecture we verify formula (9.5–12). For the line we take a vertical through the origin, i.e., the upper half of the imaginary axis and on it two points $(0, y_1)$, $(0, y_2)$.

The length of the segment with these points as end points is

$$\int_{y_1}^{y_2} \frac{dy}{y} = \log \frac{y_2}{y_1}\,.$$

But

$$(y_1, y_2, y_{1\Omega}, y_{2\Omega}) = \frac{\infty - y_1}{\infty - y_2} : \frac{0-y_1}{0-y_2} = \frac{y_2}{y_1}\,.$$

Thus, we see that (9.5–31) is the right formula. We may connect the points y_1 and y_2 by a curve $x = x(t)$, $y = y(t)$ with $x(t_1) = x(t_2) = 0$, $y(t_1) = y_1$, $y(t_2) = y_2$. If this curve does not coincide with the vertical axis then its arc length is

$$\int_{t_1}^{t_2} \frac{\sqrt{x'^2+y'^2}}{y^2}\,dt > \int_{t_1}^{t_2} \frac{y'\,dt}{y} = \int_{y_1}^{y_2} \frac{dy}{y}$$

This is essentially the triangle inequality.

The element of area is

$$dA_h = \frac{dx\,dy}{y^2}. \tag{9.5–32}$$

For the sake of illustration we wish to make the following application. We consider a horicycle. A standard form for it is a line parallel to the

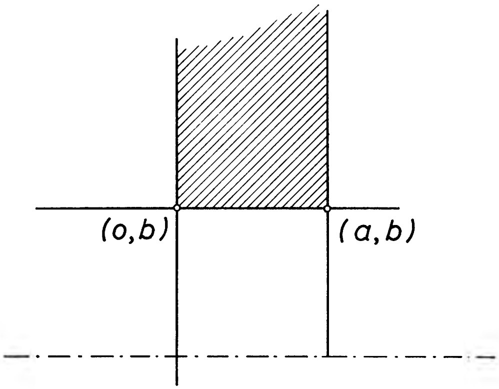


Fig. 9.5–9. Area of a sector of a horicycle.

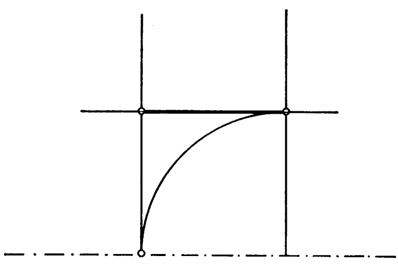


Fig. 9.5–10. The fundamental arc in Poincaré's model.

real axis, (fig. 9.5–9), for this is orthogonal to all vertical lines. Assume that an arc between the points $(0, b)$ and (a, b) is taken. Its length is

$$\int_0^a \frac{dx}{b} = \frac{a}{b}.$$

The area of the sector of this horicycle is the hyperbolic area of the half strip above the segment and is equal to

$$\int_0^a dx \int_b^\infty \frac{dy}{y^2} = \frac{a}{b}.$$

Hence

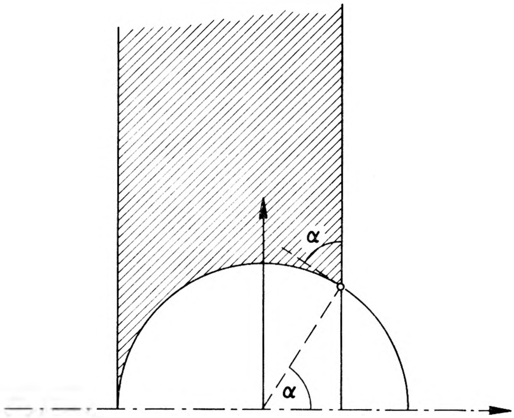
The area of a sector of a horicycle is equal to the length of its bounding arc.

A particular case arises when $b = a$. Then the tangent at one of the end points is parallel to the diameter through the other end point, i.e., the arc is a fundamental arc (fig. 9.5–10). Thus (section 9.5.4)

The length of a fundamental arc is unity.

9.5.9 – The area of a triangle

The problem of finding the area of an arbitrary triangle can be solved easily if we know the solution for a particular case. We consider a



Fig, 9.5–11. Area of a doubly asymptotic triangle.

so-called *doubly asymptotic triangle*, a triangle having two ideal vertices. Notwithstanding the fact that two vertices are on infinite distance of the third, the triangle still has a finite area. This may be seen in the following way. We make use of Poincaré's model and take the triangle in the standard form as shown in fig. 9.5–11. Let α denote the interior angle at the finite vertex. By a suitable transformation we may suppose that one side is on the unit circle around the origin and that the point $z = -1$ is one of the ideal vertices. The area is

$$\int_{-1}^{\cos\alpha} dx \int_{\sqrt{1-x^2}}^{\infty} \frac{dy}{y^2} = \int_{-1}^{\cos\alpha} \frac{dx}{\sqrt{1-x^2}} \qquad (9.5\text{–}33)$$

$$= (\tfrac{1}{2}\pi - \arccos x)\Big|_{-1}^{\cos\alpha} = \pi - \alpha$$

In particular we have a *trebly asymptotic triangle* if $\alpha = 0$. Thus

The area of a trebly asymptotic triangle is finite and equal to π.

It appears that all trebly asymptotic triangles have the same area. This is in accordance with the fact that all triangles of this kind are congruent.

Consider next a triangle ABC with interior angles α, β, γ respectively. Let C_Ω denote the ideal point of the half ray AB, A_Ω that of BC and B_Ω that of CA, (fig. 9.5–12). Then

$$\Delta ABC + \Delta AB_\Omega C_\Omega + \Delta BC_\Omega A_\Omega + \Delta CA_\Omega B_\Omega = \Delta A_\Omega B_\Omega C_\Omega.$$

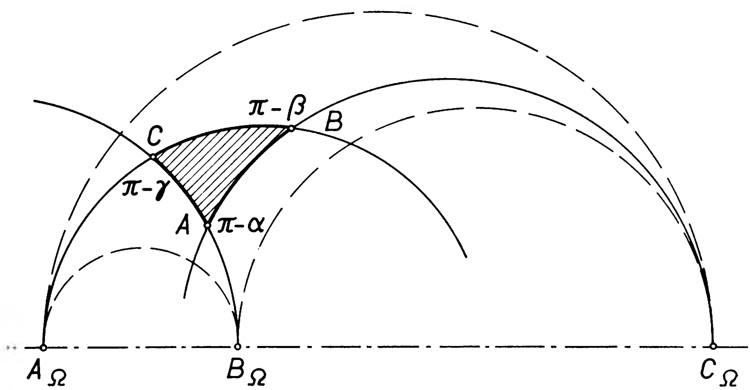


Fig. 9.5–12. Area of a triangle in the hyperbolic plane.

Since the interior angles of the doubly asymptotic triangles are $\pi-\alpha$, $\pi-\beta$ and $\pi-\gamma$ respectively, we find, denoting the area of ΔABC by Δ

$$\Delta = \pi-(\alpha+\beta+\gamma). \tag{9.5–33}$$

The number on the right is called the *defect* of the triangle. Hence

The area of a triangle is equal to its defect.

In addition we have found that in hyperbolic geometry the sum of the angles of a triangle is less than π.

9.6 – Elliptic and absolute geometry

9.6.1 – Elliptic geometry

There is another approach to non-Euclidean geometry which consists in considering the extended plane endowed with the metric of the complex sphere. In this case it is natural to take the stereographic projections of the large circles on the sphere as "straight" lines. They become circles (or straight lines) which cut the unit circle diametrally. If we retain the whole extended plane there would be the difficulty that two lines always have two points in common, viz. the images of two diametral points on the sphere. We can eliminate this difficulty by identifying the points which are images of diametral points on the sphere. Thus we obtain a model of the so-called *elliptic plane.* In space it is the surface of the unit sphere with identified diametral points. In the extended plane we may take for it the points of the interior of the unit circle completed with the circumference on which diametral points are identified, (fig. 9.6–1).

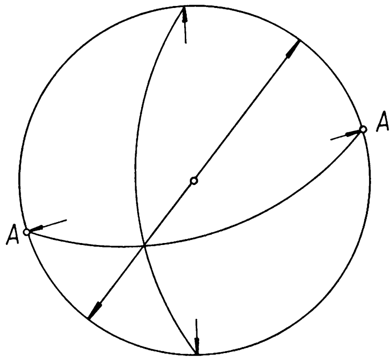


Fig. 9.6–1. Model of the elliptic plane with straight lines.

It is clear that in elliptic geometry there are no parallels. The group of congruent transformations is provided by the isometric homeomorphisms of the extended plane completed with reflections with respect to the elliptic lines.

Let us represent the points of the complex sphere by

$$\begin{aligned} \xi &= \cos\varphi \sin\vartheta \\ \eta &= \sin\varphi \sin\vartheta, \\ \zeta &= \cos\vartheta. \end{aligned} \tag{9.6–1}$$

Then the corresponding point in the z-plane is (see 1.1–10)

$$z = \frac{\xi + i\eta}{1+\zeta} = \frac{\sin\vartheta}{1+\cos\vartheta}\, e^{i\varphi} = e^{i\varphi} \tan \tfrac{1}{2}\vartheta. \tag{9.6–2}$$

The line element on the sphere is

$$ds^2 = \sin^2\vartheta\, d\varphi^2 + d\vartheta^2. \tag{9.6–3}$$

An easy calculation shows

$$dz\,d\bar{z} = \tan^2 \tfrac{1}{2}\vartheta\, d\varphi^2 + \tfrac{1}{4} \sec^4 \tfrac{1}{2}\vartheta\, d\vartheta^2$$

and

$$1 + z\bar{z} = 1 + \tan^2 \tfrac{1}{2}\vartheta = \sec^2 \tfrac{1}{2}\vartheta.$$

Hence

The line element in the elliptic plane may be taken as

$$ds_e^2 = \frac{4|dz|^2}{(1+|z|^2)^2} = 4\,\frac{dr^2+r^2\,d\theta^2}{(1+r^2)^2}, \qquad z = re^{i\theta}. \tag{9.6–4}$$

Next we evaluate the elliptic distance of a point z to the origin. From (9.6–4) we deduce

$$r_e = 2\int_0^r \frac{d\rho}{1+\rho^2} = 2\arctan r,$$

hence

$$r = \tan \tfrac{1}{2} r_e. \tag{9.6–5}$$

Comparing this with (9.6–2) we see that $r_e = \vartheta$.

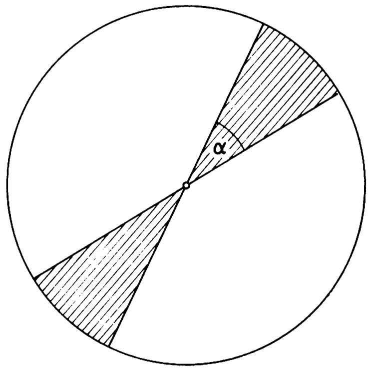


Fig. 9.6–2. Area of a 1-gon in the elliptic plane.

We can also relate the elliptic distance with the chordal distance, for it is clear from the result just obtained that

$$\chi(z_1, z_2) = 2\sin\tfrac{1}{2} r_e, \tag{9.6–6}$$

where r_e is the elliptic distance between z_1 and z_2. In particular, when $\chi(z_1, z_2) = 2$, i.e., when z_1 and z_2 are diametral, then $r_e = \pi$. Thus

All elliptic straight lines have the finite length π.

Inserting (9.6–5) into (9.6–4) we obtain

$$ds_e^2 = dr_e^2 + \sin^2 r_e\, d\theta^2. \tag{9.6–6}$$

The surface element is

$$dA_e = \sin r_e\, dr_e\, d\theta. \tag{9.6–7}$$

It is now possible to develop elliptic geometry along the same lines as hyperbolic geometry. The similarity between the pertinent formulas is striking.

The problem of finding the area of a triangle in elliptic geometry is solved very easily. First we consider a part of the elliptic plane inclosed by two straight lines. By a suitable transformation we may assume that the common point coincides with the origin, (fig. 9.6–2). Let α denote the interior angle. The area is

$$\int_0^\alpha d\theta \int_0^\pi \sin r_e dr_e = 2\alpha.$$

If $\alpha = \pi$ the figure covers the whole elliptic plane. Hence

The area of the elliptic plane is the finite number 2π.

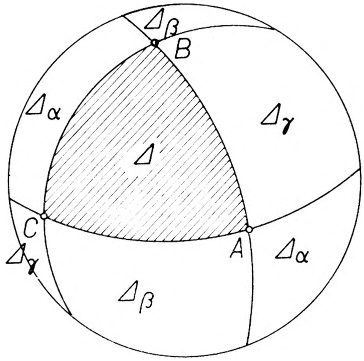


Fig. 9.6–3. The area of a triangle in the elliptic plane.

Given an arbitrary triangle we can complete it to a figure considered above in three ways, (fig. 9.6–3). The triangle together with the added regions fill the whole plane completely. Denoting the complements Δ_α, Δ_β, Δ_γ respectively, we have

$$\begin{aligned} 2\pi &= \Delta+\Delta_\alpha+\Delta_\beta+\Delta_\gamma \\ &= \Delta+\Delta_\alpha+\Delta_\beta+\Delta+\Delta_\gamma+\Delta-2\Delta \\ &= 2\alpha+2\beta+2\gamma-2\Delta, \end{aligned}$$

whence

$$\Delta = \alpha+\beta+\gamma-\pi. \qquad (9.6\text{–}8)$$

The number on the right is called the *excess* of the triangle. Hence

The area of a triangle is equal to its excess.

In addition we have found that in elliptic geometry the sum of the angles of a triangle exceeds π.

9.6.2 – Absolute geometry

The striking similarity between the hyperbolic and elliptic metric suggests the idea of investigating a more general geometry which includes the above mentioned systems as special cases. It is natural to consider a geometric model which arises from the Möbius plane by identifying two points which satisfy a relation

$$Kz_1\bar{z}_2+1 = 0. \tag{9.6–9}$$

Here K is an arbitrary, but fixed, real number. By "straight" lines we understand circles passing through points related by (9.6–9). The transformations

$$w = \frac{az+b}{-K\bar{b}z+\bar{a}}, \qquad a\bar{a}+Kb\bar{b} = 1, \tag{9.6–10}$$

completed with the reflections with respect to straight lines constitute the group of congruent transformations. A geometric system defined in this way will be called *absolute geometry.*

We introduce a metric by means of the line element

$$ds_a^2 = \frac{4dz\,d\bar{z}}{(1+K|z|^2)^2} = 4\frac{dr^2+r^2\,d\theta^2}{(1+Kr^2)^2}, \qquad z = re^{i\theta}. \tag{9.6–11}$$

The distance r_a of a point z to the origin is then

$$r_a = 2\int_0^r \frac{dr}{1+Kr^2} = \frac{2}{\sqrt{K}}\arctan r\sqrt{K}, \tag{9.6–12}$$

and an alternative form of the line element is

$$ds_a^2 = dr_a^2+\frac{1}{K}\sin^2(r_a\sqrt{K})d\theta^2. \tag{9.6–13}$$

The element of area takes the form

$$dA_a = \frac{1}{\sqrt{K}}\sin(r_a\sqrt{K})dr_a\,d\theta. \tag{9.6–14}$$

The formulas listed in this section change into those of elliptic geometry if $K = 1$ and of hyperbolic geometry if $K = -1$. In general we must distinguish between the cases $K > 0$ and $K < 0$. All geometries with $K > 0$ do not differ essentially of elliptic geometry. It is only a matter of scaling. The same can be said about the case $K < 0$. In order to interpret the parameter K we recall that on a sphere of radius R the circumference of a circle with spherical radius ϑ is

$$2\pi R\sin\vartheta = 2\pi R\sin\frac{r}{R} = \frac{2\pi}{\sqrt{K}}\sin(r\sqrt{K}), \tag{9.6–15}$$

where r is the length of the spherical radius and $K = 1/R^2$ the Gaussian curvature of the sphere. From (9.6–13) we obtain for the length of the circumference of a circle in absolute geometry with (absolute) radius r_a

$$\frac{2\pi}{\sqrt{K}} \sin (r_a\sqrt{K}). \tag{9.6–16}$$

By this reason the number K is called the *curvature* of the absolute plane.

9.6.3 – The area of a triangle

As regards the area of a triangle we may expect the general formula

$$K\Delta = \alpha+\beta+\gamma-\pi. \tag{9.6–17}$$

It is not satisfactory that for $K = 1$ and $K = -1$ this formula is obtained by essentially different methods. In the hyperbolic case we made use of the theory of parallels and in the elliptic case of the fact that the area of the whole elliptic plane is finite.

It is possible, however, to give a proof without an appeal to the theory of parallels, that is, by an absolute method. To this end we make use of the principles of the calculus of variations.

In the absolute plane we consider a curve given by

$$r_a = r_a(t), \qquad \theta = \theta(t). \tag{9.6–18}$$

The arc length between the points t_1 and t_2 is, according to (9.6–13),

$$s_a = \int_{t_1}^{t_2} \sqrt{\dot{r}_a^2 + \frac{1}{K} \sin^2 (r_a\sqrt{K})\dot{\theta}^2}\, dt = \int_{t_1}^{t_2} \dot{s}_a dt, \tag{9.6–19}$$

where the dots represent differentiation with respect to t. We consider the variational problem

$$\delta s_a = 0, \tag{9.6–20}$$

end points being fixed. The Euler equations are

$$\frac{d}{dt}\frac{\partial \dot{s}_a}{\partial \dot{r}_a} - \frac{\partial \dot{s}_a}{\partial r_a} = 0, \qquad \frac{d}{dt}\frac{\partial \dot{s}_a}{\partial \dot{\theta}} - \frac{\partial \dot{s}_a}{\partial \theta} = 0, \tag{9.6–21}$$

where $\dot{s}_a$ is considered as a function of the independent variables r_a, θ, $\dot{r}_a$, $\dot{\theta}$. Performing the differentiation we get

$$\frac{d}{dt}\frac{\dot{r}_a}{\dot{s}_a} - \frac{1}{\sqrt{K}} \frac{\sin (r_a\sqrt{K}) \cos (r_a\sqrt{K})}{\dot{s}_a} \dot{\theta}^2 = 0, \quad \frac{d}{dt}\left(\frac{\sin^2 (r_a\sqrt{K})}{\dot{s}_a}\dot{\theta}\right) = 0. \tag{9.6–22}$$

The extremals of this variational problem are called *geodesics*. They satisfy a system of ordinary differential equations of the second order and, consequently, through a point there passes exactly one geodesic which has a prescribed direction there. If we have verified that the straight lines are solutions of (9.6–22), then we may infer that the straight lines are the geodesics. We may restrict ourselves to lines through the origin. Then, however, $\dot{\theta} = 0$ and $r_a = s_a$, i.e., $\dot{r}_a = \dot{s}_a$. It is clear that the Eulerian equations are trivially satisfied.

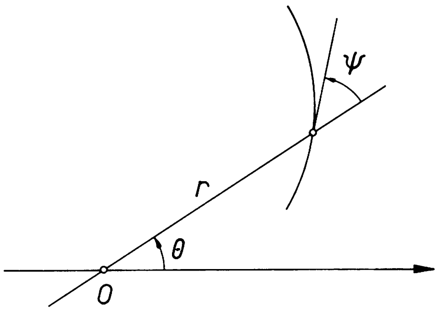


Fig. 9.6–4. The angle ψ.

It is our aim to characterize the geodesics in a more manageable form. To this end we introduce the angle ψ between the radius vector and the tangent at a given point, (fig. 9.6–4). If we exclude the case of a geodesic through the origin we may take the angle θ as parameter instead of t. It is well-known that

$$\tan \psi = \frac{r}{\dot{r}}, \tag{9.6–23}$$

where r is the Euclidean distance of the point under consideration from the origin. It follows that

$$\cos \psi = \frac{\dot{r}}{\dot{s}} = \frac{\dot{r}_a}{\dot{s}_a} \tag{9.6–24}$$

$$\sin \psi = \sqrt{\frac{\dot{s}_a^2 - \dot{r}_a^2}{\dot{s}_a^2}} = \frac{1}{\sqrt{K}} \frac{\sin (r_a\sqrt{K})}{\dot{s}_a} \dot{\theta}, \tag{9.6–25}$$

if we agree to take $\dot{s}_a > 0$ and $0 \leqq \psi \leqq \pi$. In view of (9.6–24) and the first equation (9.6–22) we have

$$-\sin \psi \frac{d\psi}{d\theta} = \frac{1}{\sqrt{K}} \frac{\sin (r_a\sqrt{K}) \cos (r_a\sqrt{K})}{\dot{s}_a} \dot{\theta}$$

and comparing this with (9.6–25) we get

$$\frac{d\psi}{d\theta} = -\cos (r_a\sqrt{K}). \tag{9.6–26}$$

Now it is an easy matter to derive the formula for the area of a triangle in absolute geometry. Assume that the vertex C is at the origin. Along the side AB the radius vector is a function of θ. At A the value of ψ is $\pi-\alpha$ and at B its value is β, (fig. 9.6–5). Hence, according to (9.6–14)

$$\Delta = \frac{1}{\sqrt{K}}\int_0^\gamma d\theta \int_0^{r_a} \sin(\rho\sqrt{K})d\rho = \frac{1}{K}\int_0^\gamma (1-\cos(r_a\sqrt{K}))d\theta$$

$$= \frac{1}{K}\left(\gamma + \int_{\pi-\alpha}^{\beta} d\psi\right) = \frac{1}{K}(\alpha+\beta+\gamma-\pi),$$

which is the desired result.

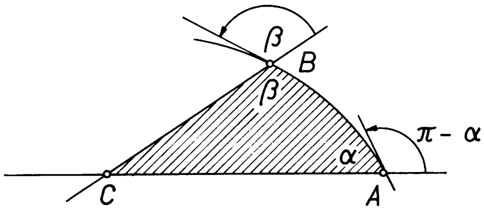


Fig. 9.6–5. The area of a triangle in absolute geometry.

In the foregoing considerations we tacitly assumed $K \neq 0$. But a geometry with $K = 0$ also makes sense. Now (9.6–9) expresses that every finite point of the plane must be paired with the point at infinity and then the straight lines are the ordinary straight lines of Euclidean geometry. In fact, this *parabolic geometry* is the same as Euclidean geometry and by passage to the limit $K \to 0$ we obtain the usual formulas. The parabolic metric differs from the original Euclidean metric only by a scaling factor. Thus, for instance the parabolic distance

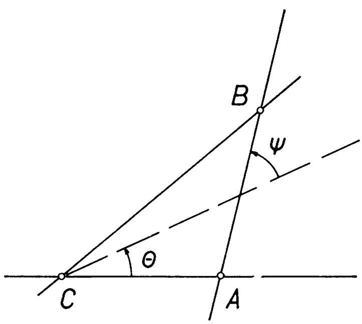


Fig. 9.6–6. Proof of (9.6–17) for the case $K = 0$.

to the origin is related to the original Euclidean distance by the equation

$$r = \lim_{K\to 0} \frac{\tan \frac{1}{2}(r_a\sqrt{K})}{\sqrt{K}} = \tfrac{1}{2}r_p,$$

where we have written r_p instead of r_a.

The proof of (9.6–17) for the case $K = 0$ is very easy. The formula (9.6–26) takes the form

$$\frac{d\psi}{d\theta} = -1,$$

hence $\psi+\theta$ is constant, (fig. 9.6–6). At the point A we have $\psi = \pi-\alpha$, $\theta = 0$, hence $\psi+\theta = \pi-\alpha$. At B we have $\psi = \beta$, $\theta = \gamma$; it follows that $\beta+\gamma = \pi-\alpha$ or $\alpha+\beta+\gamma = \pi$. This concludes the proof.

9.7 – Blaschke's theorems

9.7.1 – Extension of the identity principle

The identity principle stated in section 2.11.2 may be formulated in the following way. Let

$$a_1, a_2, \ldots, \tag{9.7–1}$$

denote a sequence of points within a certain region $\mathfrak{R}$ converging to a point a_0 of the region. If these points are the zeros of a holomorphic function, then this function is identically zero throughout $\mathfrak{R}$.

This assertion is not necessarily true if a_0 is a point of the boundary of the region. Consider e.g. the function

$$\sin \frac{1}{1-z}. \tag{9.7–2}$$

This function is holomorphic in the interior of the unit circle and it has infinitely many zeros, e.g.,

$$1-\frac{1}{n\pi}, \qquad n = 1, 2, \ldots \tag{9.7–3}$$

These tend to the point $z = 1$ on the boundary.

It is our aim to investigate whether there are restrictive conditions for (9.7–1) in order to make the identity principle valid in the case that the sequence tends to a point of the boundary. We confine ourselves to the case that the region is the interior of the unit circle. To this end we introduce the mapping

$$w = \frac{1-z/a}{1-z\bar{a}}\,|a|, \; |a| < 1, \tag{9.7–4}$$

which, in accordance with (9.5–7), represents an automorphism of the interior of the unit circle and as an automorphism of the extended plane leaves the circumference $|z| = 1$ invariant. This fact may be verified directly, for if $z\bar{z} = 1$ then

$$\frac{|a-z|}{|1-z\bar{a}|} = \frac{|z-a|}{|z\bar{z}-z\bar{a}|} = \frac{1}{|z|}\frac{|z-a|}{|\bar{z}-\bar{a}|} = 1.$$

The expression on the right of (9.7–4) is called a *Blaschke factor*.

Let now $f(z)$ denote a function holomorphic within the unit circle and having infinitely many zeros. If this function does not vanish identically its zeros have no accumulation point in the unit circle and, therefore, constitute an enumerable set. We may arrange them in order of increasing moduli, multiple zeros being repeated as many times as their multiplicity indicates. In addition we assume that the function is bounded. Without loss of generality we may suppose that $|f(z)| < 1$, for if $|f(z)| < M$ we consider the function $f(z)/M$.

Let (9.7–1) represent the sequence of zeros of $f(z)$. If $z = 0$ is a zero of multiplicity k, we consider the function $z^{-k}f(z)$, that is to say, the numbers (9.7–1) are all different from zero.

Continuing we consider the functions

$$g_n(z) = \prod_{\nu=1}^{n} \frac{1-z/a_\nu}{1-z\bar{a}_\nu}|a_\nu|, \qquad n = 1, 2, \ldots \tag{9.7–5}$$

The expressions on the right are Blaschke factors. Hence every $g_n(z)$ is holomorphic throughout the interior of the unit circle and $|g_n(z)| = 1$, if $|z| = 1$.

Let ε denote an arbitrary positive number and keep n fixed. Since $g_n(z)$ is uniformly continuous on the disc $|z| \leqq 1$, we can find a circle C_ε about the origin inside $|z| = 1$ such that $|g_n(z)| > 1-\varepsilon$ for z on the circumference of C_ε, provided the radius of C_ε is sufficiently close to 1. On the other hand we have $|f(z)| < 1$. Hence, for the same values of z

$$\left|\frac{f}{g_n}\right| < \frac{1}{1-\varepsilon} \tag{9.7–6}$$

and, as a consequence of the maximum principle of section 2.13.3, this equation holds also inside C_ε. Since ε is arbitrary, we even have

$$\left|\frac{f}{g_n}\right| \leqq 1 \tag{9.7–7}$$

for all z within the unit circle. Equality is only possible if f/g_n is a constant, (section 2.13.3). Taking $z = 0$, we get

$$|g_n(0)| \geqq |f(0)|,$$

whence

$$0 < |f(0)| \leqq |g_n(0)| = \prod_{\nu=1}^{n} |a_\nu| < 1. \qquad (9.7\text{–}8)$$

Since this result is valid for every integer n, we may infer that the infinite product of the absolute values of the zeros of $f(z)$ is convergent.

In view of the theorem of section 4.1.2 (applied to the case that all factors are positive) we have the equivalent statement:

The series

$$\sum_{\nu=1}^{\infty} (1-|a_\nu|) \qquad (9.7\text{–}9)$$

is convergent.

Now we may state an extension of the identity principle in the following form:

A function $f(z)$, *holomorphic and bounded in the region* $|z| < 1$ *which vanishes at the points of a sequence* (9.7–1) *such that the series* (9.7–9) *is divergent, vanishes identically throughout the region.*

The example (9.7–2) considered above does not contradict this assertion, though the series

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu\pi}$$

is divergent. The function, however, is not bounded, for $z = 1$ is an essential singularity.

The theorem proved in this section is due to W. Blaschke.

9.7.2 – Canonical representation

The infinite product

$$\lim_{n\to\infty} g_n(z) = \prod_{\nu=1}^{\infty} \frac{1-z/a_\nu}{1-z\bar{a}_\nu} |a_\nu| \qquad (9.7\text{–}10)$$

deserves a closer consideration. Let us write

$$1+f_n(z) = \frac{1-z/a_n}{1-z\bar{a}_n} |a_n|, \qquad n = 1, 2, \ldots,$$

where $a_1, a_2, \ldots$, are defined as in the previous section. If $a_n = |a_n| e^{i\theta_n}$, then

$$f_n(z) = \frac{(|a_n|-1)(1+ze^{-i\theta_n})}{1-\bar{a}_n z}.$$

Assume now that $|z| \leqq r < 1$. Taking into account (1.1–9) we readily find

$$|f_n(z)| \leqq (1-|a_n|) \frac{1+r}{1-r},$$

i.e., the series $\sum_{\nu=1}^{\infty}|f_\nu(z)|$ is dominated by the series (9.7–9). If this series is convergent then the series $\sum_{\nu=1}^{\infty} f_\nu(z)$ is uniformly convergent on every disc $|z| \leqq r < 1$ and hence (9.7–10) represents a function $g(z)$, holomorphic throughout the region $|r| < 1$. From (9.7–7) we deduce that

$$\left|\frac{f}{g}\right| = \lim_{n\to\infty}\left|\frac{f}{g_n}\right| \leqq 1,$$

provided that $f(z)$ does not vanish at the origin. If, however, f has a zero of multiplicity k there, then

$$\left|\frac{z^{-k}f(z)}{g(z)}\right| \leqq 1.$$

The function $f/z^k g$ has no zeros within the unit circle. It may, therefore, be represented as exp $G(z)$, with Re $G(z) \leqq 0$, (section 9.1.4). Thus we find the following *canonical representation* for $f(z)$

$$\boxed{f(z) = z^k e^{G(z)} \prod_{\nu=1}^{\infty} \frac{1-z/a_\nu}{1-\bar{a}_\nu z}|a_\nu|,} \qquad \operatorname{Re} G(z) \leqq 0. \quad (9.7\text{–}11)$$

which exhibits all zeros of $f(z)$.

9.7.3 – An extension of Vitali's theorem

The identity principle in the version of section 9.7.1 gives rise to an extension of Vitali's theorem also due to Blaschke.

Let the functions of the sequence

$$F_0(z), \qquad F_1(z), \ldots \qquad (9.7\text{–}12)$$

be holomorphic throughout the region $|z| < 1$ *and uniformly bounded. The sequence has a finite limit at the points of a subset* (9.7–1) *which is such that the series* (9.7–9) *is divergent. Under these assumptions the sequence* (9.7–12) *is convergent throughout the interior of the unit circle and uniformly convergent on every closed subset of this interior.*

In view of Vitali's theorem (section 2.22.1) it is sufficient to show that the sequence (9.7–12) is convergent at an arbitrary point within the unit circle. Suppose that the series is not convergent at a certain point z_0. Then the sequence $F_0(z_0)$, $F_1(z_0)$, . . ., has at least two different accumulation points b_1 and b_2. Then there is a subsequence $F_{10}(z_0)$, $F_{11}(z_0)$, $F_{12}(z_0)$, . . . tending to b_1 and a subsequence $F_{20}(Z_0)$, $F_{21}(z_0)$, $F_{22}(z_0)$, . ., tending to b_2. In view of the theorem of section 2.22.2 we can select from the sequence $F_{10}(z)$, $F_{11}(z)$, $F_{12}(z)$, . . . a subsequence $f_{10}(z)$, $f_{11}(z)$, . . ., converging to a holomorphic function $f_1(z)$, and from the sequence

$F_{20}(z)$, $F_{21}(z)$, $F_{22}(z)$, . . ., a subsequence $f_{20}(z)$, $f_{21}(z)$, . . ., converging to a holomorphic function $f_2(z)$. In particular $f_1(z_0) = b_1$, $f_2(z_0) = b_2$. Since the given sequence (9.7–12) is convergent at each of the points $a_1, a_2, \ldots$, the same is true for the subsequences and hence $f_1(z) - f_2(z)$ vanishes at all these points. From the identity principle of section 9.7.1 follows that $f_1(z) = f_2(z)$ identically, i.e., $b_1 = b_2$. Thus we arrived at a contradiction and we may conclude that (9.7–12) is convergent throughout the interior of the unit circle.

9.8 – Schwarz's lemma

9.8.1 – Jensen's lemma

Using simple geometry we may give various extensions of Schwarz's lemma. First we shall make an application of formula (9.7–11). We retain the assumptions of sections 9.7.1.

Let $H(r)$ denote the maximum of $\exp G(z)$ on a circle of radius $r < 1$. Then for any z on this circle

$$|f(z)| \leqq H(r)r^k \prod_{\nu=1}^{\infty} \frac{|z-a_\nu|}{|1-z\bar{a}_\nu|}. \tag{9.8–1}$$

The ratio

$$\frac{|z-a|}{|z-1/\bar{a}|}, \qquad |a| < 1, \tag{9.8–2}$$

represents the ratio of the distances of the point z to the points a and $1/\bar{a}$ respectively. If z describes the circle of radius r this ratio attains its maximum at z_0 on the half ray opposite the one passing through a, (fig. 9.8–1).

In order to prove this we denote the points a and $1/\bar{a}$ by A and B respectively. We assume that the point z, denoted by P, inside the unit circle is not on the line AB. Through P passes an Apollonian circle with respect to A and B. Its centre O' is between O and A. If P_0 is the point z_0, then $P_0O' = P_0O + OO' = PO + OO' > PO'$ and it follows that the Apollonian circle through P cuts the line AB in a point Q between P_0 and O. The assertion follows from $P_0A/P_0B = (P_0Q+QA)/(P_0Q+QB) > QA/QB = PA/PB$.

The value of the maximum of (9.8–2) is

$$\frac{r+|a|}{r+1/|a|},$$

evidently < 1. Hence

$$|f(z)| \leqq H(r)r^k \prod_{\nu=1}^{\infty} \frac{r+|a_\nu|}{1+r|a_\nu|}, \tag{9.8–3}$$

and, on account of the maximum principle, this is true for all z within the circle $|z| = r < 1$. Retaining only the first h zeros and omitting $H(r)$ (which is $\leqq 1$), we obtain the less strong, however more manageable form

$$\boxed{|f(z)| \leqq |z|^k \prod_{\nu=1}^{n} \frac{r+|a_\nu|}{1+r|a_\nu|},} \qquad |z| \leqq r < 1. \tag{9.8–4}$$

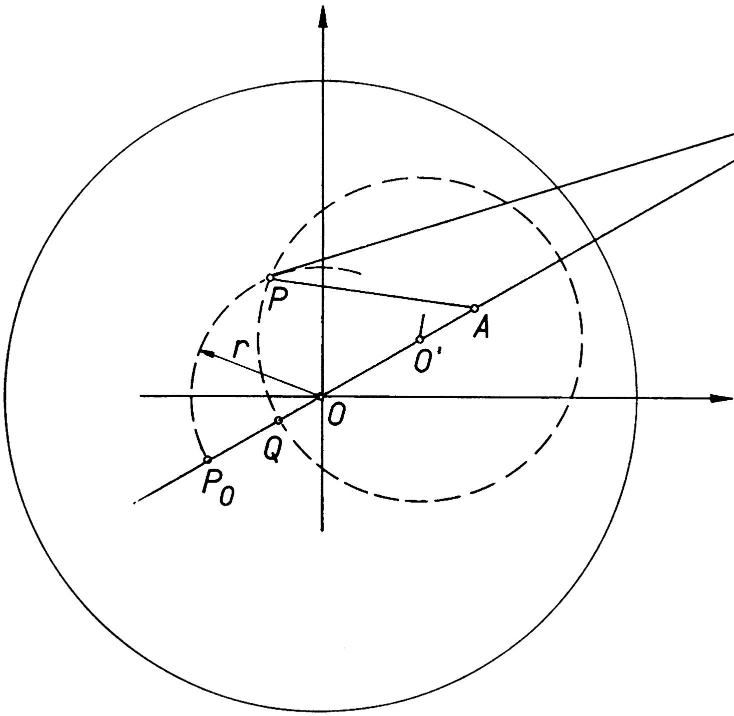


Fig. 9.8–1. Determination of the maximum of (9.8–1) for $|z|$ = constant.

This is Jensen's lemma. In the case that $k = 1$ we obtain Schwarz's lemma by omitting the product on the right.

9.8.2 – Invariant statement of Schwarz's lemma

In the original statement of Schwarz's lemma we assumed that $f(0) = 0$. It follows that $|f(z)| \leqq |z|$ and, moreover, from (2.21–5) $|f'(0)| \leqq 1$. Equality can occur only if $|f(z)| = |z|$.

Now we omit the condition that $f(0) = 0$. Denoting by z_0 any point within the unit circle we may apply the transformations

$$z^* = \frac{z-z_0}{1-z\bar{z}_0}, \qquad w = \frac{f(z)-f(z_0)}{1-f(z)\overline{f(z_0)}}. \tag{9.8–5}$$

The function $w(z^*)$ obtained by eliminating z satisfies the conditions of Schwarz's lemma, hence

$$|w| \leqq |z^*|,$$

or

$$\left|\frac{f(z)-f(z_0)}{1-f(z)\overline{f(z_0)}}\right| \leqq \frac{|z-z_0|}{|1-z\bar{z}_0|}. \tag{9.8–6}$$

Differentiating $w(z^*)$ with respect to z, we get

$$\frac{dw}{dz^*}\,\frac{1-z_0\bar{z}_0}{(1-z\bar{z}_0)^2} = \frac{1-f(z_0)\overline{f(z_0)}}{(1-f(z)\overline{f(z_0)})^2}f'(z).$$

If $z = z_0$ then $z^* = 0$. Hence $|dw/dz^*| \leqq 1$ at $z = z_0$, i.e.,

$$|f'(z_0)| \leqq \frac{1-|f(z_0)|^2}{1-|z_0|^2}, \qquad |z_0| < 1. \tag{9.8–7}$$

In these more general cases equality can only occur if $f(z)$ is an automorphism of the unit circle.

A very interesting formulation of Schwarz's lemma in its more general form is based on the notion of hyperbolic distance. Observing that

$$\log\frac{1+r}{1-r}$$

is an increasing function of r if $0 \leqq r < 1$, the inequality (9.8–6) expresses that (taking into account (9.5–17))

$$\operatorname{dist}\big(f(z), f(z_0)\big) \leqq \operatorname{dist}(z_0, z).$$

Now we may enunciate Schwarz's lemma in the following geometric form due to G. Pick.

If $f(z)$ is holomorphic within the circle $|z| < 1$ and such that $|f(z)| < 1$ and if z_1, z_2 denote any two points inside the unit circle then, measuring the distances in the hyperbolic metric,

$$\operatorname{dist}\big(f(z_1), f(z_2)\big) \leqq \operatorname{dist}(z_1, z_2). \tag{9.8–8}$$

Equality can occur only if $f(z)$ is an automorphism of the interior of the unit circle.

An alternative statement is

If z_0 and $f(z_0)$ are the hyperbolic centres of two circles having the same hyperbolic radius, then to a point z inside the first circle corresponds a point $f(z)$ inside the second circle. If z is on the circumference of the first circle then $f(z)$ is in the interior or on the circumference of the second circle. If it is also on the circumference of the second circle for a certain value of z then z is on the circumference of the first circle and for every value of z

on this circumference the point $f(z)$ is on the second circumference. In this case $f(z)$ is an automorphism of the interior of the unit circle.

This result enables us to obtain a remarkable inequality. Let z denote an arbitrary point within the unit circle and C a hyperbolic circle around $f(0)$ with radius dist $(0, z)$. If d denotes the smallest Euclidean distance of C to the circumference $|z| = 1$, we have evidently

$$(0, |z|, 1, -1) = (|f(0)|, 1-d, 1, -1)$$

whence

$$\frac{1-|f(0)|}{1+|f(0)|} \cdot \frac{2-d}{d} = \frac{1+|z|}{1-|z|}. \tag{9.8–9}$$

Since $f(z)$ is within or on the boundary of C we have

$$1-|f(z)| \geqq d.$$

Solving for d from (9.8–9) we obtain

$$1-|f(z)| \geqq (1-|z|)\frac{1-|f(0)|}{1+|z||f(0)|},$$

or, since $|z| < 1$, $|f(0)| \geqq 0$,

$$\frac{1-|f(z)|}{1-|z|} \geqq \frac{1-|f(0)|}{1+|f(0)|}. \tag{9.8–10}$$

In the case that $f(0) = 0$ this reduces to Schwarz's lemma.

9.8.3 – Julia's theorem

Consider two hyperbolic circles C and C', with hyperbolic centres $1-u$ and $1-u'$ on the real axis ($u < 1$, $u' < 1$) and equal hyperbolic radii. Let the first circle cut the real axis in a point $x < 1-u$ and the second in a point $x' < 1-u'$. The equality of radii is expressed by

$$(x, 1-u, 1, -1) = (x', 1-u', 1, -1) \tag{9.8–11}$$

which is equivalent to

$$\frac{1-x'}{1+x'} = \frac{(2-u)u'}{(2-u')u} \cdot \frac{1-x}{1+x}. \tag{9.8–12}$$

Consider next two sequences of such circles $C_1, C_2, \ldots$ and C_1', C_2', ..., centred at $1-u_1$, $1-u_2, \ldots$ and $1-u_1'$, $1-u_2', \ldots$ respectively, such that for each n the circles C_n and C_n' have equal hyperbolic radii. In addition we make the assumption

$$\lim_{n\to\infty} u_n = \lim_{n\to\infty} u_n' = 0, \quad \lim_{n\to\infty} \frac{u_n'}{u_n} = \alpha, \tag{9.8–13}$$

where α is finite, and all circles C_n, $n = 1, 2, \ldots$ pass through the same point x on the real axis. Then the circles C'_n pass through the points x'_n determined by

$$\frac{1-x'_n}{1+x'_n} = \frac{(2-u_n)u'_n}{(2-u'_n)u_n} \cdot \frac{1-x}{1+x}. \tag{9.8–14}$$

Making $n \to \infty$ we see that the circles of each sequence tend to horicycles C_0 and C'_0, passing through $z = 1$ and the points x and x' respectively, where x' depends on x according to

$$\frac{1-x'}{1+x'} = \alpha \frac{1-x}{1+x}. \tag{9.8–15}$$

If r and r' denote the Euclidean radii of C_0 and C'_0, then $x = 1-2r$, $x' = 1-2r'$ and by (9.8–15)

$$r' = \frac{\alpha r}{1-(1-\alpha)r}. \tag{9.8–16}$$

Now let $f(z)$ be a holomorphic function in the interior of the unit circle, such that $|f(z)| < 1$. Suppose that there is a sequence of numbers z_1, $z_2, \ldots$, such that

$$\lim_{n\to\infty} z_n = 1, \quad \lim_{n\to\infty} f(z_n) = 1 \tag{9.8–17}$$

and

$$\lim_{n\to\infty} \frac{1-|f(z_n)|}{1-|z_n|} = \alpha, \tag{9.8–18}$$

where α is finite. It follows from (9.8–10) that

$$\alpha \geqq \frac{1-|f(0)|}{1+|f(0)|} > 0. \tag{9.8–19}$$

Introducing the numbers $u_n = 1-|z_n|$, $u'_n = 1-|f(z_n)|$, $n = 1, 2, \ldots$, we may construct the circles C_n, C'_n considered above. They tend to horicycles C_0 and C'_0, with $z = 1$ as an ideal point. We might also construct circles C_n, C'_n with equal hyperbolic radii about the points z_n, $f(z_n)$. These are obtained from those constructed before by a rotation around the origin. But since by (9.8–17) the angles of rotation tend to zero as $n \to \infty$, the new circles tend to the same horicycles C_0 and C'_0.

Let z lie inside C_0. From a certain index upwards it lies within C_n. Hence $f(z_n)$ lies within C'_n and, therefore, within C'_0. Thus we have Julia's theorem:

If there exist sequences $z_1, z_2, \ldots$ and $f(z_1), f(z_2), \ldots$ tending simultaneously to $+1$ under the condition (9.8–18) *and if $f(z)$ is inside a hori-*

cycle C_0 *touching* $|z| = 1$ *at* $z = 1$, *then* $f(z)$ *is inside a horicycle* C_0' *which is the image of* C_0 *under the tranformation*

$$\frac{1-z'}{1+z'} = \alpha \frac{1-z}{1+z}, \qquad \alpha > 0. \tag{9.8–20}$$

In addition we may assert that *if* z *is on the boundary of* C_0, *then* $f(z)$ *is inside or on the boundary of* C_0' *and in the latter case* $f(z)$ *is an automorphism of the interior of the unit circle.*

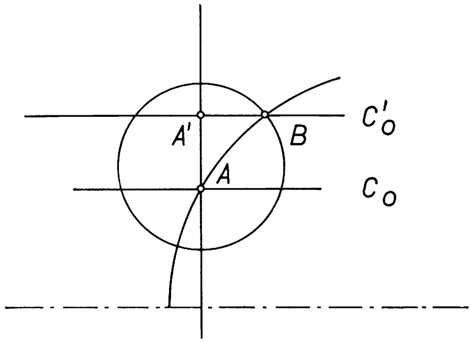


Fig. 9.8–2. The shortest distants between two horicycles.

The last part of this theorem deserves still a proof, which is not trivial. To this end we observe that two horicycles having the same ideal point intersect equal hyperbolic segments on the diameters. This is at once clear by considering the Poincaré model for horicycles having their ideal point at $z = \infty$, (fig. 9.8–2). It is also seen that this segment represents the shortest distance between two points on the horicycles. In fact, let C_0' be inside C_0, A on C_0 and A' on C_0' such that AA' is a common diameter. If any other hyperbolic line through A intersects C_0' in B then the hyperbolic circle about A through B has A' in its interior. Hence dist $(A, B) >$ dist (A, A').

Assume now that z_0 is a point on C_0 such that $f(z_0)$ is on C_0', (fig. 9.8–3). Let C_0^* denote a horicycle within C_0 and $C_0^{*\prime}$, its image under the transformation (9.8–20). Through z_0 passes a diameter of C_0 which cuts C_0^* in z_0^*. By the first part of Julia's theorem $f(z_0^*)$ is on or within $C_0^{*\prime}$. Now dist (z, z_0^*) is equal to the shortest distance between $f(z_0)$ and the points of C_0^*. Hence

$$\text{dist}\,(z_0, z_0^*) \leqq \text{dist}\,(f(z_0), f(z_0^*)).$$

But from (9.8–8) follows

$$\text{dist}\,(z_0, z_0^*) \geqq \text{dist}\,(f(z_0), f(z_0^*)).$$

Hence equality occurs and $f(z)$ is an automorphism. This entails that for every z on C_0 the point $f(z)$ is also on C_0'.

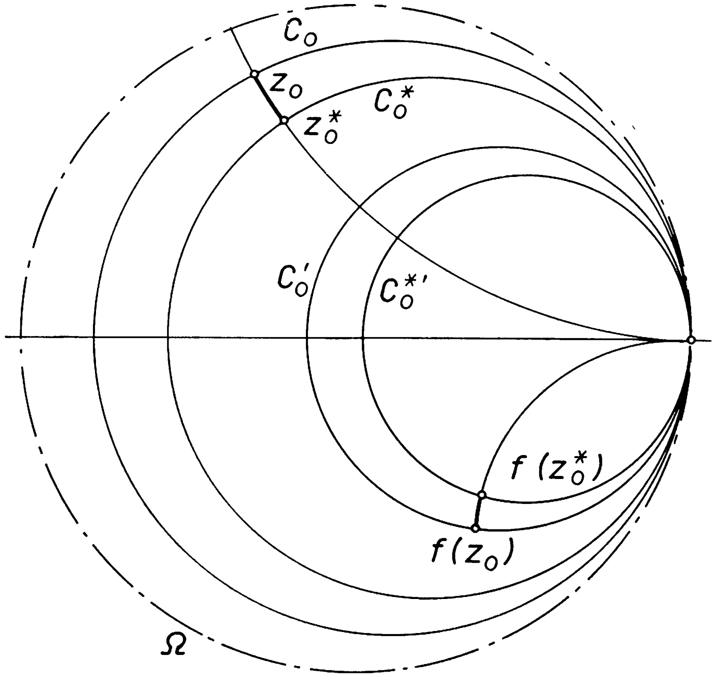


Fig. 9.8–3. Proof of the last part of Julia's theorem.

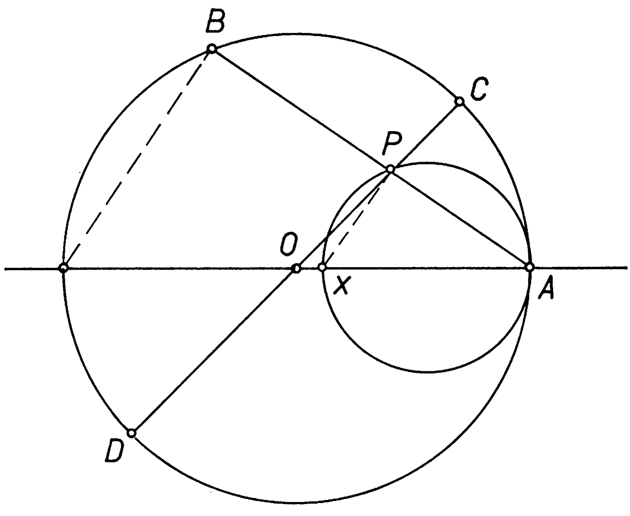


Fig. 9.8–4. Geometric proof of (9.8–21).

By two rotations we can extend the theorem to the case that z_n and $f(z_n)$, $n = 1, 2, \ldots$, tend simultaneously to two arbitrary points on the circumference of the unit circle.

From Julia's theorem an interesting inequality follows, which can be derived by geometric arguments. In fig. (9.8–4) we see that

$$\frac{AP}{BP} = \frac{1-x}{1+x}.$$

If P, P' denote the points $z, f(z)$ then according to Julia's theorem

$$\frac{AP'}{P'B} \leqq \frac{1-x'}{1+x'}.$$

In view of (9.8–20) we have

$$\frac{AP'}{P'B} \leqq \alpha \frac{AP}{PB}.$$

But

$$\frac{AP}{PB} = \frac{AP^2}{AP \cdot PB} = \frac{AP^2}{CP \cdot PD} = \frac{|1-z|^2}{1-|z|^2}$$

and a similar expression for, $AP'/P'B$ in terms of $f(z)$. Thus we obtain the inequality

$$\frac{|1-f(z)|^2}{1-|f(z)|^2} \leqq \alpha \frac{|1-z|^2}{1-|z|^2}. \tag{9.8–21}$$

Finally we wish to mention that Julia has also established a similar theorem for hypercycles.

9.8.4 – A CONVERSE OF JULIA'S THEOREM

The following theorem is a converse of Julia's theorem

Let $f(z)$ be holomorphic and of modulus < 1 throughout the region $|z| < 1$. If the inequality (9.8–21) *is satisfied at every point of this region for a certain positive number α, then it is possible to find a sequence z_n tending to* 1 *such that $f(z_n)$ also tends to* 1 *as $n \to \infty$ and that*

$$\frac{1-|f(z_n)|}{1-|z_n|}$$

has a limit not exceeding α.

Let us take z on the real axis at x and construct a horicycle C_0 including x passing through the ideal point $z = 1$ with radius r. The inequality (9.8–21) expresses that $f(x)$ is on or in a horicycle C_0' with ideal point at $z = 1$ and of Euclidean radius r' given by (9.8–16):

$$r' = \frac{\alpha r}{1-(1-\alpha)r}.$$

It follows that

$$|1-f(x)| \leqq 2r' = \frac{2\alpha r}{1-(1-\alpha)r}.$$

Since $1-x = 2r$ we have

$$\frac{1-|f(x)|}{1-x} \leqq \frac{|1-f(x)|}{1-x} \leqq \frac{\alpha}{1-(1-\alpha)r}.$$

Making $x \to 1$ (which amounts to $r \to 0$) we get

$$\limsup_{x\to 1} \frac{1-|f(x)|}{1-x} \leqq \limsup_{x\to 1} \frac{|1-f(x)|}{1-x} \leqq \alpha.$$

Let now x_n denote a sequence tending to 1. By the previous result the upper limit of

$$\frac{1-|f(x_n)|}{1-x_n}$$

will not exceed α, hence $|1-f(x_n)|$ tends to zero and $f(x_n)$ to 1. Again, if

$$\limsup_{n\to\infty} \frac{1-|f(x_n)|}{1-x_n} = \beta$$

then $\beta \leqq \alpha$. It is possible to find a subsequence x_{n_i} such that $f(x_{n_i})$ tends to 1 and

$$\frac{1-|f(x_{n_i})|}{1-x_{n_i}}$$

to β. This proves the theorem.

9.9 – The theorem of Bloch

9.9.1 – Statement of bloch's theorem

Let Φ denote the family of all functions $f(z)$ holomorphic in the open unit disc $|z| < 1$ and normalized by the condition $f'(0) = 1$. By the *Bloch number* B_f of a function f of this family is meant the least upper bound of the set of positive numbers r such that there exists a subregion $\mathfrak{R}$ of $|z| < 1$ that f maps *univalently* onto an open disc of radius r. The *Bloch constant* is defined as

$$B = \inf B_f, \tag{9.9–1}$$

where the greatest lower bound, indicated by "inf", is taken with respect to the functions of the family Φ.

A famous theorem discovered by A. Bloch states

The constant B is positive.

The remarkable feature of this theorem is that despite the vastness of the family Φ there is an absolute positive constant B and an open disc of radius $\geqq B$ which is the one-to-one image of a subregion of $|z| < 1$ under a mapping $w = f(z)$. The exact value of B is unknown.

An interesting and important application of Bloch's theorem is an "elementary" proof of Picard's theorem, already stated in section 6.11.1. The original proof of this theorem was based on the elliptic modular function and we shall discuss it in paragraph 14.4. Proofs avoiding the use of the modular function are considered as elementary (which is not synonymous with easy!). The shortest elementary proof of Picard's theorem is due to E. Landau, who observed that Bloch's theorem is not needed in its sharpest form, but that a weaker statement suffices.

To this end we introduce the *Landau number* L_f of a function of Φ, being the least upper bound of the set of positive numbers r such that the image of $|z| < 1$ contains an open disc of radius r. The *Landau constant* is the number

$$L = \inf L_f. \tag{9.9–2}$$

It is clear that $B_f \leqq L_f$ and so $B \leqq L$. We shall prove that $L \geqq 1/16$.

9.9.2 – The Bloch–Landau theorem

We start with the following lemma.

Let $\varphi(z)$ be holomorphic in the disc $|z| < R$, $R > 0$. We assume that $\varphi(0) = 0$, $|\varphi'(0)| = p > 0$ and $|\varphi'(z)| < M$ throughout the disc. If c is a number not taken by $\varphi(z)$ for $|z| < R$, then

$$|c| \geqq \frac{Rp^2}{4M}. \tag{9.9–3}$$

From Darboux's inequality (2.4–17) applied to

$$\varphi(z) = \int_0^z \varphi'(\zeta)d\zeta,$$

the path of integration being a rectilinear segment connecting 0 and z, $|z| < R$, we deduce

$$|\varphi(z)| \leqq M|z| < MR. \tag{9.9–4}$$

By assumption $c \neq 0$. Hence $1-\varphi(z)/c$ is holomorphic in $|z| < R$ with no zero. In view of the monodromy theorem of section 9.1.4 we may infer that there is a function

$$\psi(z) = 1 - \frac{\varphi'(0)}{2c}z + \ldots,$$

holomorphic in $|z| < R$ such that

$$\psi^2(z) = 1 - \frac{\varphi'(0)}{c} z + \ldots = 1 - \frac{\varphi(z)}{c}.$$

From (9.9–4) we deduce

$$|\psi^2(z)| < 1 + \frac{MR}{c}, \qquad |z| < R. \tag{9.9–5}$$

If $0 < r < R$ we have by Parseval's theorem (2.18–5)

$$1 + \frac{|\varphi'(0)|^2}{4c^2} r^2 \leqq \frac{1}{2\pi} \int_0^{2\pi} |\psi(re^{i\theta})|^2 d\theta < 1 + \frac{MR}{c},$$

whence

$$|c| > \frac{r^2 |\varphi'(0)|^2}{4MR} = \frac{r^2 p^2}{4MR}.$$

Since we can take r as near R as we please, we finally have

$$|c| \geqq \frac{Rp^2}{4M},$$

as asserted.

Now we can prove the *Bloch-Landau* theorem which states

If $g(z)$ is holomorphic in the disc $|z| < 1$ and if $g'(0) = 1$, then the image of the disc as given by $g(z)$ covers a circular disc of radius $1/16$.

We set

$$M(s) = \max_{|z| \leqq s} |g'(z)|, \qquad 0 \leqq s \leqq r < 1, \qquad r > 0.$$

It is clear that the function

$$\mu(s) = sM(r-s)$$

is continuous, takes the value 0 at $s = 0$ and the value r at $s = r$. Hence there is a least value of s, say $2R$, for which $\mu(2R) = r$; we have $0 < 2R \leqq r$.

Now we take a number a with

$$|a| = r - 2R, |g'(a)| = M(r-2R) = r/2R.$$

The function

$$\varphi(z) = g(z+a) - g(a) \tag{9.9–6}$$

is holomorphic for $|z| < 2R$, because

$$|z+a| \leqq |z| + |a| < 2R + (r-2R) = r.$$

It takes the value 0 at $z = 0$ and

$$|\varphi'(0)| = |g'(a)| = r/2R > 0.$$

If now $|z| < R$, then

$$|z+a| \leq |z|+|a| < R+(r-2R) = r-R$$

and

$$|\varphi'(z)| = |g'(z+a)| \leq M(r-R).$$

By the choice of R we find that

$$\mu(R) = RM(r-R) < r,$$

whence

$$|\varphi'(z)| < \frac{r}{R}, \qquad |z| < R.$$

On applying the lemma we find that if c is a number not taken by $\varphi(z)$ for $|z| < R$

$$|c| \geq \frac{R(r/2R)^2}{4r/R} = \frac{r}{16}.$$

This expresses the fact that the image of $|z| < R$ as given by $\varphi(z)$ covers an open disc with radius $r/16$. The disc $|z| < 1$ certainly includes the disc $|z-a| < R$, since

$$|z| \leq |z-a|+|a| < R+(r-2R) < r < 1.$$

From (9.9–6) we have

$$g(z) = \varphi(z-a)+g(a)$$

and it follows that the image of $|z| < 1$ as given by $g(z)$ covers a disc about $g(a)$ of radius $r/16$. This is true for all $r < 1$. Hence we may let $r \to 1$ and conclude that the image of $|z| < 1$ covers a disc of radius $1/16$.

It is not claimed that the number $1/16$ is the best possible. There does exist a sharp $1/16$-theorem. It will be stated and proved in section 14.5.10.

9.9.3 – Picard's theorem

Following Landau we may base a proof of Picard's theorem on the lemma

Let $\mathfrak{R}$ be a simply connected region in the open plane including the point $z = 0$. Let $f(z)$ be holomorphic throughout $\mathfrak{R}$ and omit the values 0 and 1 for all z in $\mathfrak{R}$. Then there exists a function $g(z)$ holomorphic in $\mathfrak{R}$ such that

$$f(z) = -\exp\left(\pi i \cosh\left(2g(z)\right)\right); \qquad (9.9\text{–}7)$$

$g(0)$ depends only on $f(0)$ and the image of $\mathfrak{R}$ as given by $g(z)$ does not cover any open disc of radius 1.

On applying the monodromy theorem of section 9.1.4 we can determine successively functions $h(z)$, $u(z)$, $v(z)$ and $g(z)$ all holomorphic in $\mathfrak{R}$ which satisfy the conditions listed below. In each case it is necessary to make a definite choice among the various possibilities, but it suffices to make this choice at $z = 0$; the function in question is then uniquely determined throughout $\mathfrak{R}$.

Since $f(z)$ has no zero we can find a function $h(z)$ such that

$$f(z) = \exp\bigl(2\pi i h(z)\bigr).$$

Since $f(z) \neq 1$ implies $h(z) \neq 0$, we can choose $u(z)$ such that

$$h(z) = u^2(z).$$

The condition $f(z) \neq 1$ implies also $h(z) \neq 1$ and, consequently, there exists a function $v(z)$ so that

$$h(z) = 1+v^2(z).$$

It is clear that $u(z) \neq v(z)$, for $u^2(z)-v^2(z) = 1$. Hence there exists a function $g(z)$ such that

$$u(z)-v(z) = \exp g(z).$$

As a consequence

$$u(z)+v(z) = \frac{1}{u(z)-v(z)} = \exp\bigl(-g(z)\bigr),$$

whence

$$u(z) = \cosh g(z)$$

and

$$\cosh 2g(z) = 2\cosh^2 g(z)-1 = 2u^2(z)-1.$$

It follows that

$$2\pi i h(z) = 2\pi i u^2(z) = \pi i \cosh\bigl(2g(z)\bigr)+\pi i$$

and finally

$$f(z) = -\exp\bigl(\pi i \cosh(2g(z))\bigr),$$

as asserted. It is clear that $g(0)$ depends only on $f(0)$.

In order to prove the last assertions of the theorem we construct a set of points not taken as images of the points of $\mathfrak{R}$ under the mapping as given by $g(z)$ with the property that each open disc of radius unity covers at least one of the points of this set. We contend that these are the points represented by

$$b = \pm \log\,(\sqrt{m}+\sqrt{m-1})+\tfrac{1}{2}n\pi i, \tag{9.9–8}$$

where m and n are integers, $m \geqq 1$. These points form the vertices of a rectangular net. The height of each rectangle is $\frac{1}{2}\pi < \sqrt{3}$; the width is

$$\log(\sqrt{m+1}+\sqrt{m})-\log(\sqrt{m}+\sqrt{m-1}) = \begin{cases} = \log(\sqrt{2}+1) < 1, & \text{if } m = 1, \\ < \log\sqrt{\dfrac{m+1}{m-1}} \leqq \log\sqrt{3} < 1, & \text{if } m > 1. \end{cases}$$

As a consequence to every point a there is a point b with

$$|\mathrm{Re}\, b - \mathrm{Re}\, a| < \tfrac{1}{2}, \qquad |\mathrm{Im}\, b - \mathrm{Im}\, a| < \tfrac{1}{2}\sqrt{3}$$

i.e.,

$$|b-a| < \sqrt{\tfrac{1}{4}+\tfrac{3}{4}} = 1.$$

It remains to be proved that $g(z)$ does not take any value b as long as z is in $\mathfrak{R}$. Suppose this were not true. Then we could find a z_0 in $\mathfrak{R}$ such that $g(z_0) = b$ and

$$\cosh(2g(z_0)) = \tfrac{1}{2}((\sqrt{m}+\sqrt{m-1})^2+(\sqrt{m}-\sqrt{m-1})^2 = 2m-1,$$

so that

$$f(z_0) = -\exp(2m-1)\pi i = 1.$$

This is contradictory to the hypothesis.

It is now an easy matter to prove *Picard's theorem*

An integral function which omits two different values is a constant.

This theorem is sharp as the function $\exp z$ which omits only the value 0 shows.

Without loss of generality we may assume that $f(z)$ omits the values 0 and 1. For if $f(z)$ omits the values a and b, $a \neq b$, then

$$\frac{f(z)-a}{b-a}$$

is again an integral function and is constant if and only if $f(z)$ is constant.

We apply the lemma to $f(z)$, where $\mathfrak{R}$ is now the open plane. Suppose $f(z)$ is not constant. Then the same can be asserted about the auxilary function $g(z)$ of the lemma which is also an integral function. It is possible to find a number a with $g'(a) \neq 0$ and the function

$$\tfrac{1}{16}g\left(\frac{16z}{g'(a)}+a\right) = \tfrac{1}{16}g(a)+z+\ldots$$

is an integral function whose values do not cover any open disc of radius 1/16. This contradicts, however, the covering theorem of the previous section. We conclude that $f(z)$ is a constant, as asserted.

9.9.4 – Landau's theorem

The lemma of the previous section leads to astonishing results concerning the influence of the first two terms of a power series on the properties of the functions defined by the series. Landau discovered the following remarkable theorem

Let

$$f(z) = a_0+a_1 z+ \ldots, a_1 \neq 0, \tag{9.9–9}$$

be holomorphic in the region $|z| < R$. *Then there exists a positive number* $R(a_0, a_1)$ *depending only on* $a_0 = f(0)$ *and* $a_1 = f'(0)$ *such that if* $R > R(a_0, a_1)$, *the function takes at least one of the values* 0 *or* 1 *within the disc* $|z| < R$.

Suppose that $f(z)$ omits the values 0 and 1. The lemma of the previous section asserts that there is an auxiliary function $g(z)$ having the properties stated in the lemma. Differentiating we find

$$f'(z) = -2\pi i f(z) g'(z) \sinh(2g(z))$$

and, taking $z = 0$,

$$a_1 = -2\pi i a_0 g'(0) \sinh(2g(0)). \tag{9.9–10}$$

By assumption $a_0 \neq 0$. Since also $a_1 \neq 0$ we see that $g(0) \neq 0$, $g'(0) \neq 0$. It follows from the lemma that $g(0)$ depends only on $a_0 = f(0)$ and, consequently, $g'(0)$ depends only on a_0 and a_1.

If $0 < r < R$ the function

$$\frac{g(rz)}{rg'(0)} = c_0+z+ \ldots$$

is holomorphic in the disc $|z| < 1$ and does not cover any disc of radius $1/r|g'(0)|$. On the other hand it follows from the covering theorem of section 9.9.2 that it covers a disc of radius $1/16$. Hence

$$\frac{1}{r|g'(0)|} > \tfrac{1}{16}$$

so that

$$r < R(a_0, a_1) = \frac{16}{|g'(0)|}.$$

Since this is true for all $r < R$ we may infer that

$$R \leqq R(a_0, a_1)$$

and this result is equivalent to the statement of Landau's theorem.

Landau's theorem implies Picard's theorem.

Let $f(z)$ denote a non-constant integral function. Then there is a point in the finite plane such that $f(z_0) \neq 0$, $f'(z_0) \neq 0$. If we set $a_0 = f(z_0)$ $a_1 = f'(z_0)$ then it follows from Landau's theorem that in every region $|z-z_0| < R$, $R > R(a_0, a_1)$, the function $f(z)$ takes at least one of the values 0 or 1.

9.9.5 – Schottky's theorem

Another application of the lemma of section 9.9.3 is *Schottky's theorem.* *Let*

$$f(z) = a_0 + a_1 z + \dots \tag{9.9–11}$$

be holomorphic throughout $|z| < 1$ *and omit the values* 0 *and* 1. *Let* ϑ *be a number between* 0 *and* 1. *Then a number* $\varphi(a_0, \vartheta)$ *exists depending only on* a_0 *and* ϑ *such that for* $|z| \leqq \vartheta$

$$|f(z)| \leqq \varphi(a_0, \vartheta). \tag{9.9–12}$$

Let $g(z)$ be the auxiliary function of the lemma of section 9.9.3. Take r such that $\vartheta < r < 1$. It is clear that the function

$$\frac{g(a+(1-\vartheta/r)z)}{(1-\vartheta/r)g'(a)} = c_0 + z + \dots, \qquad |a| < \vartheta$$

is holomorphic in $|z| < r$, provided that $g'(a) \neq 0$, for

$$|a+(1-\vartheta/r)z| \leqq |a| + (1-\vartheta/r)|z| < \vartheta + r - \vartheta = r.$$

It does not cover a disc with radius $1/(1-\vartheta/r)|g'(a)|$ and it follows from the Bloch-Landau theorem that

$$\frac{1}{(1-\vartheta/r)|g'(a)|} > \tfrac{1}{16},$$

or

$$|g'(a)| < \frac{16}{1-\vartheta/r}.$$

This is also true if $g'(a) = 0$. If $|z| \leqq \vartheta$ then

$$|g(z)-g(0)| = \left| \int_0^z g'(\zeta) d\zeta \right| \leqq \frac{16\vartheta}{1-\vartheta/r} < \frac{16}{1-\vartheta/r}$$

and since this is true for all r as near 1 as we please, it follows that

$$|g(z)-g(0)| \leqq \frac{16}{1-\vartheta},$$

whence

$$|g(z)| \leqq |g(0)| + \frac{16}{1-\vartheta}$$

and so

$$|f(z)| \leqq \exp\left(\pi \exp\left(|g(0)|\right) + \frac{16}{1-\vartheta}\right) = \varphi(a_0, \vartheta),$$

since $g(0)$ depends only on a_0.

Schottky's theorem implies Landau's theorem.

For assume that $f(z)$ is holomorphic in $|z| < R$ and omits the values 0 and 1. Then $f(Rz) = a_0 + a_1 Rz + \ldots, a_1 \neq 0$, is holomorphic in $|z| < 1$ and

$$|a_1|R \leqq 2 \max_{|z|=\frac{1}{2}} |f(Rz)| \leqq \varphi(a_0, \tfrac{1}{2}),$$

whence

$$R \leqq \frac{1}{|a_1|} \varphi(a_0, \tfrac{1}{2}),$$

as stated in Landau's theorem.

9.9.6 – Ahlfors's extension of Schwarz's lemma

In order to obtain a positive lower bound for Bloch's constant L. V. Ahlfors formulated a theorem which may be considered an extension of Schwarz's lemma.

Any real, non-negative and continuous function $\lambda(z)$, defined throughout a region $\mathfrak{R}$ gives rise to a metric in $\mathfrak{R}$ if we define the length of a curve $C : z = z(t),\ 0 \leqq t \leqq 1$, by means of

$$l = \int_C \lambda(\zeta)|d\zeta| = \int_0^1 \lambda(z(t))|z'(t)|dt, \qquad (9.9\text{–}13)$$

It reduces to the ordinary Euclidean metric if $\lambda(z) = 1$ identically.

Suppose that a region $\mathfrak{R}_z$ has been mapped conformally onto a region $\mathfrak{R}_w$ by means of the function $w(z)$. If we are given a metric $\mu(w)$ in $\mathfrak{R}_w$, then the length of the curve $w(z(t))$ corresponding to the curve C in $\mathfrak{R}_z$ is given by

$$\int_0^1 \mu(w(z(t)))|w'(z(t))||z'(t)|dt = \int_0^1 \lambda(z(t))|z'(t)|dt,$$

with

$$\lambda(z) = \mu(w(z))|w'(z)|. \qquad (9.9\text{–}14)$$

In section 9.6.2 we introduced the metric

$$\lambda(z) = \frac{2}{1+Kz\bar{z}}, \qquad |z| < 1. \qquad (9.9\text{–}15)$$

It is easily verified that it satisfies the condition

$$\triangle \log \lambda(z) = -K\lambda^2(z). \qquad (9.9\text{–}16)$$

In fact, with reference to (9.2–12) we have

$$\frac{\partial}{\partial z}\log \lambda(z) = -\frac{K\bar{z}}{1+Kz\bar{z}},$$

$$\triangle \log \lambda(z) = -4\frac{\partial}{\partial \bar{z}}\frac{K\bar{z}}{1+Kz\bar{z}} = -\frac{4K}{(1+Kz\bar{z})^2} = -K\lambda^2(z).$$

A metric $\lambda(z)$ is called *regular* at a point $z = a$ if in some neighbourhood of this point $\lambda(z)$ has derivatives of at least the second order which are continuous in this neighbourhood.

The expression

$$\boxed{K(z) = -\frac{\triangle \log \lambda(z)}{\lambda^2(z)},} \tag{9.9–17}$$

evaluated at a point where $\lambda(z)$ is regular and different from zero is called the *Gaussian curvature* of the metric at this point. Thus the metric (9.9–15) has the constant Gaussian curvature K throughout the disc $|z| < 1$.

Let now $\lambda(z)$ be a metric within the open disc $|z| < 1$. Suppose that $\lambda(a) \neq 0$ at a point $z = a$ of the disc. A metric $\lambda_a(z)$, which is regular at $z = a$ and satisfies the conditions:

(i) $\lambda_a(z) \leqq \lambda(z)$ in some neighbourhood of $z = a$;

(ii) $\lambda_a(a) = \lambda(a)$;

(iii) in some neighbourhood of $z = a$ we have

$$\triangle \log \lambda_a(z) \geqq \lambda_a^2(z), \tag{9.9–18}$$

will be said to *support* the metric λ at $z = a$.

The third condition expresses the fact, that the Gaussian curvature of the metric at the point $z = a$ does not exceed the Gaussian curvature of the hyperbolic metric, of Gaussian curvature -1.

Ahlfors's theorem states

Let there be given a non negative continuous function $\lambda(z)$ in the disc $|z| < 1$. Suppose that for each $z = a$ in the disc for which $\lambda(a) \neq 0$ there exists a supporting metric λ_a. Then

$$\boxed{\lambda(z) \leqq \frac{2}{1-|z|^2},} \qquad |z| < 1. \tag{9.9–19}$$

We set

$$u(z) = \log \lambda(z), v(z) = \log \frac{2R}{R^2-|z|^2}, \qquad |z| < R < 1.$$

It is clear that $u(z)$ is continuous at all points of $|z| < 1$ which are not zeros of $\lambda(z)$. We contend that throughout the disc $|z| < R$ the inequality

$$u(z) \leqq v(z) \tag{9.9–20}$$

holds.

Suppose that not everywhere in the disc $|z| < R$ the relation (9.9–20) holds. Then at some point $u(z)-v(z)$ is positive. If z tends to a point of $|z| = R$ then $v(z) \to +\infty$ and $u(z)-v(z) \to -\infty$. The same is true if z tends to a zero of $\lambda(z)$. Hence the difference $u(z)-v(z)$ has a positive maximum inside the disc $|z| < R$. Let this maximum be taken at $z = a$. This point is not a zero of $\lambda(z)$

By assumption there is a regular metric $\lambda_a(z)$ supporting $\lambda(z)$ at $z = a$. If $u_a(z) = \log \lambda_a(z)$, we evidently have $u_a(a)-v(a) = u(a)-v(a) > 0$ and also $u_a(z)-v(z) > 0$ in some neighbourhood of $z = a$, by continuity. By condition (i) we have in a sufficiently small neighbourhood of $z = a$ the inequality $u_a(z)-v(z) \leqq u(z)-v(z)$. This means that the function

$$\varphi(z) = u_a(z)-v(z) \tag{9.9–21}$$

has a positive maximum at $z = a$ with respect to a neighbourhood $|z-a| < r_0$. We shall prove that this is imposssible.

The relations (9.9–18) and

$$\Delta u_a \geqq e^{2u_a} \tag{9.9–22}$$

are equivalent. Proceeding as in the proof of (9.9–16) we also have

$$\Delta v = e^{2v}. \tag{9.9–23}$$

Hence

$$\Delta\varphi = e^{2u_a}-e^{2v} > 0, \qquad |z-a| < r_0.$$

Now we employ the expression (9.2–15) and we find, if $r < r_0$,

$$0 < \iint\limits_{|z-a|<r} \Delta\varphi\, dx\, dy = \int_0^{2\pi}\int_0^r \left(\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\varphi}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2\varphi}{\partial\theta^2}\right)\rho\, d\rho\, d\theta.$$

Along the circumference $|z-a| = r$ the function φ is periodic, whence

$$\int_0^{2\pi} \frac{\partial^2\varphi}{\partial\theta^2}\, d\theta = 0$$

and it follows that

$$0 < \int_0^{2\pi}\int_0^r \frac{\partial}{\partial\rho}\left(\rho\frac{\partial\varphi}{\partial\rho}\right) d\rho\, d\theta = \int_0^{2\pi} r\frac{\partial\varphi}{\partial r}\, d\theta = r\int_0^{2\pi}\frac{\partial\varphi}{\partial r}\, d\theta,$$

or

$$0 < \frac{\partial}{\partial r}\int_0^{2\pi} \varphi(a+re^{i\theta})d\theta.$$

Integrating from $r = 0$ to $r = r_0$ yields

$$\int_0^{2\pi} \varphi(a+r_0 e^{i\theta})d\theta - \int_0^{2\pi} \varphi(a)d\theta > 0,$$

whence

$$\varphi(a) < \frac{1}{2\pi}\int_0^{2\pi} \varphi(a+r_0 e^{i\theta})d\theta. \tag{9.9–24}$$

Since $\varphi(z)$ has a maximum at $z = a$ we have

$$\varphi(a) \geqq \varphi(a+r_0 e^{i\theta})$$

and

$$\varphi(a) \geqq \frac{1}{2\pi}\int_0^{2\pi} \varphi(a+r_0 e^{i\theta})d\theta,$$

in contradiction with (9.9–24).

We conclude that

$$\lambda(z) \leqq \frac{2R}{R^2-|z|^2}, \qquad |z| < R$$

and since we may take R as near 1 as we please, we get the desired inequality (9.9–19).

Let, as in Schwarz's lemma, $f(z)$ denote a holomorphic function in $|z| < 1$ and $|f(z)| < 1$. If in the circle $|w| < 1$ we are given a hyperbolic metric it induces by $w = f(z)$ a metric

$$\lambda(z) = \frac{2|f'(z)|}{1-|f(z)|^2}$$

in accordance with (9.9–14). Using (9.2–13) and (9.9–15) with $K = -1$ we readily find that

$$\triangle \log \lambda(z) = \lambda^2(z)$$

at every point where $f'(z) \neq 0$. Hence this metric supports itself and as a consequence of Ahlfors's inequality (9.9–19) we have

$$\frac{|f'(z)|}{1-|f(z)|^2} \leqq \frac{1}{1-|z|^2}, \qquad |z| < 1,$$

in accordance with (9.8–7).

9.9.7 – Proof of Bloch's theorem

We start with a function $f(z)$ holomorphic throughout the unit disc $|z| < 1$ and normalized by the condition $f'(0) = 1$. Let z_0 be a point of the disc with $f'(z_0) \neq 0$. Then there is number r such that the disc $|w-w_0| < r$, $w_0 = f(z_0)$, corresponds univalently to a region $\mathfrak{R}_{z_0}(r)$ in the disc $|z| < 1$ under the mapping $w = f(z)$. Suppose that on the boundary of this region there is no point of the circumference $|z| < 1$, and that there is not a point at which $f'(z)$ vanishes neither. About every point of the circumference $|w-w_0| = r$ we can assign a small open disc which corresponds univalently to a certain region in $|z| < 1$ which overlaps the region $\mathfrak{R}_{z_0}(r)$. From the Heine-Borel theorem (section 1.2.4) follows that the circumference $|w-w_0| = r$ can be covered by a finite number of such discs and that we can extend $|w-w_0| < r$ to a larger open disc which is still in a one-to-one correspondence with a subregion in $|z| < 1$. The least upper bound of the numbers r having the property that the mapping $w = f(z)$ causes a one-to-one correspondence between the disc $|w-w_0| < r$ and a subregion $\mathfrak{R}_{z_0}(r)$ of the disc $|z| < 1$ will be denoted by $\rho(z_0)$. Thus $\rho(z_0)$ is defined at every point at which $f'(z_0) \neq 0$. It is clear that $\rho(z) \leqq B_f$ and that $\rho(z)$ is bounded if we assume that B_f has a finite value.

Let R_{z_0} be the region corresponding to $|w-w_0| < \rho(z_0)$. From the above considerations follows that on the boundary of $\mathfrak{R}_{z_0}$ there is either a point of the circumference $|z| = 1$ or a point at which $f'(z)$ vanishes.

We take a point $z = a$ in $|z| < 1$, where $f'(a) \neq 0$. If z_0 is in $\mathfrak{R}_a$, then $f'(z_0) \neq 0$ (section 9.3.2). The radius $\rho(z_0)$ of the disc which corresponds univalently to $\mathfrak{R}_{z_0}$ is not less than the shortest distance of $f(z_0)$ to a point of the circumference $|w-f(a)| = \rho(a)$ and does not exceed the distance of $f(z_0)$ to the remotest point on this circumference.

Hence

$$|\rho(z)-\rho(a)| \leqq |f(z)-f(a)|, \; f'(a) \neq 0, \qquad (9.9\text{–}25)$$

if z is in some neighbourhood of $z = a$.

If, however, $f'(a) = 0$ then in some neighbourhood of $z = a$ we have $f'(z) \neq 0$, $z \neq a$. If z_0 is sufficiently near a the disc about $f(z_0)$ with radius $\rho(z_0)$ corresponds univalently to a subregion in $|z| < 1$ which has no boundary points in common with $|z| = 1$. But then a zero of $f'(z)$ must be on the boundary of $\mathfrak{R}_{z_0}$ and this can only be at $z = a$. Hence $|w-f(z_0)| = \rho(z_0)$ passes through $f(a)$, and so

$$\rho(z) = |f(z)-f(a)|, \quad f'(a) = 0. \qquad (9.9\text{–}26)$$

The relations (9.9–25) and (9.9–26) express the fact that $\rho(z)$ *is continuous throughout the disc* $|z| < 1$, provided that we take $\rho(z_0) = 0$ if $f'(z_0) = 0$. This function will be used to construct a suitable metric in $|z| < 1$.

Let again $z = a$ denote a point of $|z| < 1$ at which $f'(a) \neq 0$. By b

we denote a point on the circumference $|w-f(a)| = \rho(a)$ which is not the image of a point at which $f'(z)$ does not vanish. The function $w-b$ does not take the value 0 in the disc $|w-f(a)| < \rho(a)$ and we can, therefore, define a single-valued branch $\sqrt{w-b}$ of the square root of $w-b$ throughout the open disc.

The hyperbolic metric with Gaussian curvature -1 in a w_1-plane is by means of the mapping

$$w_1 = \frac{1}{A}\sqrt{w-b}$$

transformed into

$$\frac{A}{(A^2-|w-b|)\sqrt{w-b}}$$

which has a meaning throughout $|w-f(a)| < \rho(a)$, provided that A is sufficiently large, say $> \sqrt{2B_f}$. The mapping as given by $w = f(z)$ transforms this metric into

$$\lambda_a(z) = \frac{A|f'(z)|}{(A^2-\rho_a(z))\sqrt{\rho_a(z)}}, \tag{9.9–27}$$

with

$$\rho_a(z) = |f(z)-b|. \tag{9.9–28}$$

This metric is regular and since it is derived from an hyperbolic metric with Gaussian curvature -1, the relation (9.9–18) is true (in this case we even have equality).

Replacing $\rho_a(z)$ by $\rho(z)$ in (9.9–27) we obtain a more general metric

$$\lambda(z) = \frac{A|f'(z)|}{(A^2-\rho(z))\sqrt{\rho(z)}} \tag{9.9–29}$$

which has a meaning at all points where $f'(z) \neq 0$ (for then $\rho(z) \neq 0$).

We contend that this metric is continuous throughout $|z| < 1$. This is clear at every point $z = a$ where $f'(a) \neq 0$. It follows for a point $z = a$ where $f'(a) = 0$, provided we can prove that $\lambda(z)$ has a limit as $z \to a$. For then we can take $\lambda(a)$ as being equal to this limit.

Assuming that $f'(a) = 0$, we have for all z in some neighbourhood of $z = a$, taking into account (9.9–26),

$$\lambda(z) = \frac{A|f'(z)-f'(a)|}{(A^2-\rho(z))\sqrt{|f(z)-f(a)|}} = \frac{A\dfrac{|f'(z)-f'(a)|}{|z-a|}}{(A^2-\rho(z))\sqrt{\dfrac{|f(z)-f(a)|}{|z-a|^2}}}$$

and this tends to $(1/A)\sqrt{\tfrac{1}{2}|f''(a)|}$, as $z \to a$. Hence if we define

$$\lambda(z) = \frac{1}{A}\sqrt{\tfrac{1}{2}|f''(z)|} \tag{9.9–30}$$

at the points where $f'(z) = 0$, then $\lambda(z)$ is continuous throughout $|z| < 1$.

Next we wish to show that at all points where $\lambda(z) \neq 0$ this metric has a supporting metric. Consider first the case that $f'(a) = 0$. Then we may take the number b occurring in (9.9–28) as $f(a)$ and it follows that $\lambda_a(z) \leqq \lambda(z)$ in some neighbourhood of $z = a$. Let $f'(a) \neq 0$. The function

$$(A^2 - t)t^{\frac{1}{2}} \tag{9.9–31}$$

is increasing in $0 \leqq t \leqq \frac{1}{3}A^2$. Now

$$0 \leqq \rho(z) \leqq B_f < \tfrac{1}{3}A^2$$

if we take $A > \sqrt{3B_f}$. If z is in a sufficiently small neighbourhood of $z = a$ then $\rho_a(z) < \rho(a) + \varepsilon < \frac{1}{3}A^2$. On the other hand, if z is in $\mathfrak{R}_a$, then a circle of radius $\rho(z)$ about $f(z)$ cannot include the point b and it follows that $\rho_a(z) \geqq \rho(z)$. In particular $\rho_a(a) = |f(a) - b| = \rho(a)$. Hence $\lambda_a(z) \leqq \lambda(z)$ if z is sufficiently near a and the conditions (i), (ii) and (iii) listed in the previous sections are fulfilled. It follows from Ahlfors's theorem that

$$\frac{A|f'(z)|}{(A^2 - \rho(z))\sqrt{\rho(z)}} \leqq \frac{2}{1 - |z|^2}, \qquad |z| < 1.$$

Taking $z = 0$ and observing that $f'(0) = 1$, by assumption, then we find

$$A \leqq (A^2 - \rho(0))\sqrt{\rho(0)} \leqq 2(A^2 - B_f)\sqrt{B_f},$$

taking again into account the fact that the function (9.9–31) is increasing in the interval $0 \leqq t \leqq B_f$. By letting $A \to \sqrt{3B_f}$ we obtain

$$\sqrt{3B_f} \leqq 4B_f\sqrt{B_f},$$

whence $B_f \geqq \frac{1}{4}\sqrt{3}$ and so

$$B \geqq \tfrac{1}{4}\sqrt{3}.$$

This concludes the proof of Bloch's theorem.

As M. Heins has pointed out the actual value of B is $> \frac{1}{4}\sqrt{3}$.

9.9.8 – A lower bound for Landau's constant

Ahlfors's method also yields a lower estimate of Landau's constant which is much better than that obtained in section 9.9.2. We start again with a function $f(z)$, holomorphic throughout $|z| < 1$ and normalized by the condition $f'(0) = 1$. Without loss of generality we may assume that the Landau number of L_f of f is finite.

If $|z_0| < 1$ we define $\rho(z_0)$ as the radius of the largest circle about $f(z_0)$ which is still contained in the image of $|z| < 1$ as given by the mapping $w = f(z)$. It is clear that $\rho(z_0) \leqq L_f$ and that (9.9–25) holds for this function $\rho(z)$. As a consequence $\rho(z)$ is continuous throughou $|z| < 1$.

The hyperbolic metric with Gaussian curvature -1 in a w_1-plane is by means of the mapping

$$w_1 = \frac{1-w}{1+w}$$

transformed into

$$\frac{4}{|1+w|^2\left(1-\left|\frac{1-w}{1+w}\right|^2\right)} = \frac{4}{|1+w|^2-|1-w|^2}$$

$$= \frac{4}{(1+w)(1+\overline{w})-(1-w)(1-\overline{w})} = \frac{2}{w+\overline{w}} = \frac{1}{\operatorname{Re} w}.$$

Let b denote a point on the circumference $|w-f(a)| = \rho(a)$ which is also a point of the boundary of the image of $|z| < 1$ as given by $f(z)$. Since $f(z)-b$ has no zero in $|z| < 1$ we can define a single valued branch of

$$\log \frac{A}{f(z)-b}, \qquad A > 0, \qquad |z| < 1,$$

and it gives rise to a regular metric

$$\lambda_a(z) = \frac{|f'(z)|}{\rho_a(z) \log \dfrac{A}{\rho_a(z)}}, \tag{9.9–32}$$

where $\rho_a(z) = |f(z)-b|$. At the points where $f'(z) \neq 0$ the condition (9.9–18) is fulfilled (with equality).

It is easy to see that the metric

$$\lambda(z) = \frac{|f'(z)|}{\rho(z) \log \dfrac{A}{\rho(z)}} \tag{9.9–33}$$

is continuous throughout $|z| < 1$ and is supported at every point $z = a$ with $f'(a) \neq 0$ by $\lambda_a(z)$, provided A is sufficiently large. For the function

$$t \log \frac{A}{t} \tag{9.9–34}$$

is increasing in the interval $0 \leqq t \leqq A/\mathrm{e}$ and we may take $A > \mathrm{e}L_f$. Then in a sufficiently small neighbourhood of $z = a$ also $\rho(z) < A$.

According to Ahlfors's theorem we have

$$\frac{|f'(z)|}{\rho(z)\log\dfrac{A}{\rho(z)}} \leqq \frac{2}{1-|z|^2},$$

whence, by taking $z = 0$,

$$1 \leqq 2\rho(0)\log\frac{A}{\rho(0)} \leqq 2L_f \log\frac{A}{L_f}.$$

By letting $A \to eL_f$ we find $1 \leqq 2L_f$, and so

$$L \geqq \tfrac{1}{2}$$

This is a formidable improvement of the poor estimate $L \geqq \frac{1}{16}$.

9.9.9 – Bloch functions

It is our aim to prove that there exist functions in the class Φ with $B_f = B$. These functions will be called *Bloch functions.*

We need two lemmas. First

The Bloch constant B is the greatest lower bound of all Bloch numbers B_F, where F belongs to Φ and is regular on the closed disc $|z| \leqq 1$.

It is clear that $B_F \geqq B$ and, consequently, $\inf B_F \geqq B$. There exists a function f in Φ such that $B_f < B+\varepsilon$, where ε is a positive number.

If $0 < \vartheta < 1$, then the function

$$F(z) = \frac{f(\vartheta z)}{\vartheta}$$

belongs to Φ and is regular for $|z| \leqq 1$. Now $f(\vartheta z)$ covers univalently a disc of radius $\leqq B_f$ and so $F(z)$ covers univalently a disc of radius $\leqq B_f/\vartheta$, whence

$$B_F \leqq \frac{B_f}{\vartheta} < \frac{B+\varepsilon}{\vartheta}.$$

By making $\varepsilon \to 0$ and $\vartheta \to 1$, we even have

$$B_F \leqq B$$

and, therefore, $\inf B_F \leqq B$ if F runs through the set of functions of Φ which are regular throughout $|z| \leqq 1$.

Secondly

The Bloch constant B is the greatest lower bound of all numbers B_f, where f belongs to Φ and satisfies the condition

$$|f'(z)| \leqq \frac{1}{1-|z|^2}. \tag{9.9–35}$$

From the first lemma follows that there is a function $F(z)$ in the family Φ which is regular on the disc $|z| \leqq 1$ and whose Bloch number satisfies

$$B_F < B+\varepsilon,$$

ε being a given positive number.

The function

$$(1-|z|^2)|F'(z)|$$

is continuous throughout the closed disc $|z| \leqq 1$ and takes a maximum at $z = z_0$. Since the function is 1 at $z = 0$ this maximum is $\geqq 1$ and because the function vanishes on $|z| = 1$ the point z_0 is in the interior of the disc.

The transformation

$$w = \frac{z+z_0}{1+z\bar{z}_0}, \tag{9.9–36}$$

being a transformation of the type (9.5–7), transforms $|z| \leqq 1$ univalently into $|w| \leqq 1$.

The function

$$f(z) = \frac{F(w)}{(1-|z_0|^2)F'(z_0)}, \tag{9.9–37}$$

where w is given by (9.9–36) is holomorphic in $|z| < 1$ and belongs to the family Φ. In fact, from

$$f'(z) = \frac{F'(w)}{(1-|z_0|^2)F'(z_0)} \frac{dw}{dz} \tag{9.9–38}$$

and

$$\frac{dw}{dz} = \frac{1-|z_0|^2}{(1-z\bar{z}_0)^2}$$

follows that $f'(0) = 1$.

Since

$$(1-|z_0|^2)|F'(z_0)| \geqq 1$$

we have

$$|f(z)| \leqq |F(w)|, |z| < 1$$

and, as a consequence,

$$B_f \leqq B_F < B+\varepsilon$$

It remains to be proved that (9.9–35) holds. To this end we observe that

$$\frac{|dw|}{1-|w|^2} = \frac{|dz|}{1-|z|^2}, \qquad |z| < 1,$$

if w and z are related by (9.9–36). This is a consequence of the invariance of the hyperbolic line element (section 9.5.7). We may write this as

$$\left|\frac{dw}{dz}\right| = \frac{1-|w|^2}{1-|z|^2}, \qquad |z| < 1.$$

Now we find from (9.9–38), assuming $|z| < 1$,

$$|f'(z)| = \frac{(1-|w|^2)|F'(w)|}{(1-|z_0|^2)|F'(z_0)|} \cdot \frac{1}{1-|z|^2} \leqq \frac{1}{1-|z|^2}.$$

This completes the proof of the theorem.

The existence of Bloch functions can readily be shown now. We shall prove

In the family Φ there is a function which does not cover univalently any disc of radius exceeding the Bloch constant.

The following proof is due to R. M. Robinson.

For every positive n choose a function $f_n(z)$ whose Bloch number is less than $B+1/n$ and which satisfies the inequality

$$|f_n'(z)| \leqq \frac{1}{1-|z|^2}, \qquad |z| < 1.$$

If $z = re^{i\theta}$, $0 < r < 1$, we have

$$|f_n(z)-f_n(0)| = \left|\int_0^r f_n'(\rho e^{i\theta})e^{i\theta}\,d\rho\right|$$

$$\leqq \int_0^r |f_n'(\rho e^{i\theta})|d\rho \leqq \int_0^r \frac{d\rho}{1-\rho^2} = \tfrac{1}{2}\log\frac{1+r}{1-r}.$$

Thus the sequence $f_n(z)$ is uniformly bounded in the disc $|z| \leqq r < 1$. Analyzing the proof of Vitali's theorem of section 2.22.1 we see that we used only the fact that the sequence (2.22–1) is uniformly bounded on every closed disc included in $\mathfrak{R}$. Hence the corollary stated in section 2.22.2 is already valid if the sequence (2.22–18) is uniformly bounded on every closed and bounded set included in $\mathfrak{R}$. Applied to the above situation we infer that a subsequence $f_{n_1}, f_{n_2}, \ldots$ can be chosen which converges uniformly on every disc $0 \leqq |z| \leqq r < 1$, and hence converges to a holomorphic function $f(z)$ in $|z| < 1$.

Suppose that in $|z| < 1$ the function $f(z)$ covers univalently the circle $|w-w_0| > B+4\varepsilon$, $\varepsilon > 0$. There is a region $\mathfrak{R}_4$ in $|z| < 1$ on which $f(z)$ is univalent and maps $\mathfrak{R}_4$ onto the open disc $|w-w_0| < B+4\varepsilon$. There are regions $\mathfrak{R}_2$ and $\mathfrak{R}_3$ within $\mathfrak{R}_4$, which $f(z)$ maps on $|w-w_0| < B+2\varepsilon$ and $|w-w_0| < B+3\varepsilon$ respectively. The boundary of this latter region is a simple closed analytic arc C, being the image of $|w-w_0| = B+3\varepsilon$ as given by the inverse of $f(z)$. This arc encloses $\mathfrak{R}_2$.

Let z_0 denote a point in $\mathfrak{R}_2$ and $b = f(z_0)$. The function $b-f(z)$ does not vanish on C, for $f(z)$ is univalent in $\mathfrak{R}_4$. Hence there is a positive number m such that for z on C

$$|f(z)-b| \geqq m > 0.$$

Since the sequence $f_{n_1}, f_{n_2}, \ldots$, is uniformly convergent on C we can find a number n_0 such that for $n_k > n_0$

$$|f_{n_k}(z)-b-(f(z)-b)| = |f_{n_k}(z)-f(z)| < m,$$

if z is on C. We then have

$$|f_{n_k}(z)-f(z)| < |f(z)-b|,$$

with $|g_{n_k}(z)| < |f(z)|$. By Rouché's theorem (section 3.10.2) we conclude that $f_{n_k}(z)-b$, $n_k > n_0$ has the same number of zeros inside C as $f(z)-b$. It follows that $f_{n_k}(z)$ is univalent on $\mathfrak{R}_2$.

Since $f_{n_k}(z) \to f(z)$ uniformly in $\mathfrak{R}_2$ the functions $f_{n_k}(z)$ cover univalently the open disc $|w-w_0| < B+\varepsilon$, provided n_k is sufficiently large. But this leads to a contradiction for the functions with $1/n < \varepsilon$. Hence $f(z)$ does not cover univalently a disc with radius $> B$, i.e., the Bloch number of f is equal to the Bloch constant. This concludes the proof of the theorem.

CHAPTER 10

CONFORMAL MAPPING OF SIMPLY CONNECTED REGIONS

10.1 – The principle of symmetry

10.1.1 – GENERAL THEOREMS ON CONFORMAL MAPPING

The theory of conformal mapping is a very important part of the theory of functions. It enables us to illustrate properties of functions by geometric facts and thus the insight in the character of the functions is very often increased. First we shall state some general theorems which are basic for the theory.

In section 3.12.2 we stated that the image of a region under a holomorphic (and therefore also conformal) mapping is again a region. If the mapping is slightly restricted, we may state.

Let w be a univalent function holomorphic throughout a simply connected region $\mathfrak{R}_z$. Then the image $\mathfrak{R}_w$ is also simply connected.

If $\mathfrak{R}_z$ is the extended plane or the open plane, then $\mathfrak{R}_w$ is again the extended or the open plane (section 9.3.2). We assume, therefore, that $\mathfrak{R}_w$ (and hence also $\mathfrak{R}_z$) admits at least one other external point b beyond $w = \infty$. The function $w-b$ does not vanish on $\mathfrak{R}_w$ and, consequently, $w(z)-b$ has no zeros in $\mathfrak{R}_z$. It follows that $w'(z)/(w(z)-b)$ is holomorphic throughout $\mathfrak{R}_z$. Let C_w denote an arbitrary cycle in $\mathfrak{R}_w$ and C_z the corresponding cycle in $\mathfrak{R}_z$ (since $w(z)$ is univalent the function is invertible). In view of (2.3–26), Cauchy's integral theorem, and the fact that $C_z \sim 0$ in $\mathfrak{R}_z$ we have

$$\int_{C_w} \frac{dw}{w-b} = \int_{C_z} \frac{w'(\zeta)}{w(\zeta)-b}\, d\zeta = 0. \tag{10.1–1}$$

In view of (2.3–30) this means that $\Omega_{C_w}(b) = 0$, i.e., $C_w \sim 0$ in $\mathfrak{R}_w$. This proves the assertion.

If w admits a (simple) pole then the theorem is still valid, for we can eliminate the pole by combining the mapping with a suitable homeomorphism of the extended plane.

A famous theorem of Riemann, to be considered in section 10.5.2, states that all simply connected regions, with at least two boundary points, are univalently (and hence conformally) equivalent. This is the reason that without restricting the problem we can confine ourselves to the study of images of the interior of a circular disc. The situation in the case of a multiply connected region is not so simple. However, for applications in the field of applied mathematics the mapping properties of simply connected regions are also extremely useful.

The following theorem is of fundamental importance

Let $w(z)$ be a function holomorphic in the region $|z| < 1$ and regular at each point of its boundary C_z. Let C_w denote the curve into which w transforms C_z and assume that there is no zero of w on C_z. Then, denoting by N the number of zeros of w in the interior of the disc, we have

$$N = \Omega_{C_w}(0). \tag{10.1–2}$$

The number of zeros inside C_z is necessarily finite. Since the winding number of the circumference with respect to an inner point is unity, the truth of the statement follows from (3.10–4).

The theorem remains true if there are isolated points on C_z where w is only continuous. For then we can indent C_z by small arcs and if we contract them to the respective points the winding number varies continuously, i.e., remains constant, for it is an integer.

A more general version of the above theorem states:

Let $w(z)$ be a function continuous on the closed disc $|z| \leqq 1$ and holomorphic in the interior, C_w the curve into which the function w transforms the circumference C_z of the disc in the z-plane and finally assume that C_w does not pass through the origin. Then (10.1–2) *holds, if N is the number of zeros of w in the interior of the disc.*

Let C_n, $n = 1, 2, \ldots$, be a sequence of contours given by the functions $z_n(t)$ respectively, $0 \leqq t \leqq 1$, these functions being continuous. Let C denote a contour given by $z(t)$ on the same interval such that

$$z_n(t) \to z(t),$$

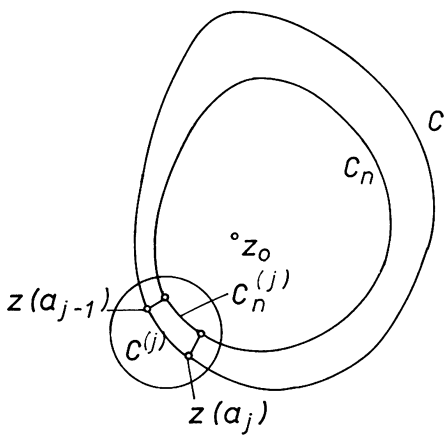


Fig. 10.1–1. Proof of (10.1–3)

uniformly as $n \to \infty$. Then for every point z_0 not lying on C or C_n we have

$$\Omega_{C_n}(z_0) = \Omega_C(z_0), \tag{10.1–3}$$

provided that n is sufficiently large.

In order to prove this we divide the interval $0 \leqq t \leqq 1$ into m equal parts by the points $0 = a_0 < a_1 < \ldots < a_m = 1$ and denote by $C^{(j)}$ the arc of the curve C on the interval $a_{j-1} \leqq t \leqq a_j$, $j = 1, \ldots, m$. If m is sufficiently large then every $C^{(j)}$ is contained in a circle which does not contain z_0.

Similarly, for every $n = 1, 2, \ldots, j = 1, 2, \ldots, m$ we denote by $C_n^{(j)}$ the arc of the curve C_n on the same interval (fig. 10.1–1). Connecting $z_n(a_{j-1})$ with $z(a_{j-1})$ and $z_n(a_j)$ with $z(a_j)$ we may complete $C^{(j)}$ and $C_n^{(j)}$ to a cycle whose winding number with respect to z_0 is zero (for this cycle is contained in an open circular disc not containing z_0, this disc being simply connected), provided that n is sufficiently large. Taking the sum of all these cycles we obtain the chain $C - C_n$ whose winding number with respect to z_0 is also zero.

Since

$$\Omega_{C-C_n}(z_0) = \Omega_C(z_0) - \Omega_{C_n}(z_0)$$

the truth of (10.1–3) follows.

Now we are sufficiently prepared to prove the theorem. Let r_n, $n = 1, 2, \ldots$, be an arbitrary increasing sequence of numbers tending to 1 and let N_n denote the number of zeros of $w(z)$ in the disc $|z| < r_n$. We denote by $C_w(r_n)$ the image of $|z| = r_n$ under the mapping w. Notice that $w(z)$ cannot have infinitely many zeros, for these would have an accumulation point, necessarily on the boundary $|z| = 1$, and since the mapping is continuous the curve C_w would pass through $w = 0$. From a certain value of n upwards all zeros of $w(z)$ are within $|z| = r_n$ and according to the previous theorem

$$\Omega_{C_{w(r_n)}}(0) = N. \tag{10.1–4}$$

By introducing an appropriate parameter we may represent C_z as $z = z(t)$, $0 \leqq t \leqq 1$, and the circumferences $|z| = r_n$ as $z = z_n(t)$ on the same interval. If δ is a given positive number we have evidently $|z(t) - z_n(t)| < \delta$ uniformly with respect to t, provided that n is sufficiently large. On the other hand $w(z)$ is uniformly continuous on $|z| = 1$ (section 1.2.3), i.e., if $\varepsilon > 0$ is given then we can find a number δ such that $|w(z') - w(z'')| < \varepsilon$, provided that $|z' - z''| < \delta$. Hence, from a certain index upwards we have

$$|w(z(t)) - w(z_n(t))| < \varepsilon,$$

uniformly with respect to t. Hence from a certain index upwards

$$\Omega_{C_{w(r_n)}}(0) = \Omega_{C_w}(0)$$

and from (10.1–4) we obtain the desired result.

In practical situations $\mathfrak{R}_z$ is often the upper half of the z-plane. By an elementary transformation e.g. (9.5–28), this can be transformed univalently into a circular disc and we may apply the above result. In this case we therefore have

Let $w(z)$ be a function continuous on the closed half plane $\operatorname{Im} z \geqq 0$ *and holomorphic in the open half plane, C_w the curve into which the function w transforms the real axis percorsed from the left to the right and finally assume that C_w does not pass through the point w_0. Then*

$$N = \Omega_{C_w}(w_0)$$

is the number of zeros of $w - w_0$ in the half plane.

In this theorem it is tacitly understood, of course, that C_w is a closed curve.

Another proof of this theorem may be given as follows. Since $w(z)$ remains finite as $z \to \infty$ along the real axis, a large half circle $|z| = R$, $\operatorname{Im} z \geqq 0$ corresponds to a small arc connecting two points near $w(\infty)$ on C_w. If R is sufficiently large $\Omega_{C_w}(w_0)$ remains unchanged if now C_w denotes the indented curve. The previous theorem may be applied now.

The following example may serve as an illustration. Consider the mapping

$$w = \left(\frac{1+iz}{1-iz}\right)^2.$$

It is easily verified that if z varies from $-\infty$ to $+\infty$ then w percorses the circumference of a unit circle twice. Hence $\Omega_{C_w}(\frac{1}{4}) = 2$, in accordance with the fact that $w(z) = \frac{1}{4}$ has two solutions in the upper half plane, viz. $z = \frac{1}{3}i$ and $z = 3i$.

In certain circumstances, however, continuity on the boundary already implies regularity (with exception of isolated points). This is a famous discovery due to H. A. Schwarz which will occupy us in the subsequent sections. It is the so-called symmetry principle.

10.1.2 – A BOUNDARY VALUE THEOREM

In order to obtain the symmetry principle in a rather general form we must base our discussions on a theorem about harmonic functions which take prescribed values on the circumference of a circle.

In section 2.15.2 we obtained the following result: If $f(z)$ is holomorphic throughout a region which contains a circular disc of radius R around the origin, then (2.15–13) holds, that is to say

$$f(z) = iv(0) + \frac{1}{2\pi}\int_0^{2\pi} u(R, \varphi)\frac{\zeta+z}{\zeta-z}\,d\varphi, \qquad (10.1\text{–}5)$$

with $f = u+iv$, $\zeta = \text{Re}^{i\varphi}$ and z inside the disc. Taking $f(z) = 1$, identically, we find

$$1 = \frac{1}{2\pi}\int_0^{2\pi} \frac{\zeta+z}{\zeta-z}\, d\varphi. \tag{10.1–6}$$

A consequence of (10.1–5) is Poisson's formula (2.15–17) which we shall write in the form

$$\text{Re} f(z) = \frac{1}{2\pi}\int_0^{2\pi} u(R, \varphi)\, \frac{\mu}{d^2}\, d\varphi, \tag{10.1–7}$$

where μ is the power of the point z with respect to the circumference $|z| = R$ with reversed sign and d is the distance between z_0 and ζ.

Thus we see that the values of the harmonic function $\text{Re} f(z)$ are uniquely determined by its values on the circumference. The question arises whether there is a function harmonic throughout $|z| < R$ and having prescribed values on the boundary $|z| = R$. This question can be answered in the affirmative under the assumption that the boundary values constitute a piecewise continuous and bounded function.

Let $G(\varphi)$ denote this function. We can summarize the answer in the following theorem

If $G(\varphi)$ is a piecewise continuous and bounded real function on the circumference $|z| = R$, then there exists a function $g(z)$ holomorphic throughout $|z| < R$ such that the real part of this function tends to $G(\varphi)$ if z tends to a point $z = \text{Re}^{i\varphi}$, where $G(\varphi)$ is continuous.

From the lemma proved in section 2.8.1 we deduce that the function

$$g(z) = \frac{1}{2\pi}\int_0^{2\pi} G(\varphi)\, \frac{\zeta+z}{\zeta-z}\, d\varphi \tag{10.1–8}$$

is holomorphic within the region $|z| < R$, for

$$\frac{\zeta+z}{\zeta-z} = \frac{2\zeta}{\zeta-z} - 1$$

and so, neglecting an additive constant, the integral (10.1–8) takes the form (2.8–1). We have achieved our purpose when we prove that

$$\text{Re}\, g(z) = \frac{1}{2\pi}\int_0^{2\pi} G(\varphi)\, \frac{\mu}{d^2}\, d\varphi \tag{10.1–9}$$

is a solution of the boundary value problem, i.e.,

$$\lim_{z\to \text{Re}^{i\varphi_0}} \text{Re}\, g(z) = G(\varphi_0), \tag{10.1–10}$$

provided that $G(\varphi)$ is continuous at $\varphi = \varphi_0$. Since (10.1–6) is equivalent to

$$1 = \frac{1}{2\pi}\int_0^{2\pi} \frac{\mu}{d^2}\, d\varphi,$$

we also have

$$\operatorname{Re} g(z) - G(\varphi_0) = \frac{1}{2\pi}\int_0^{2\pi} (G(\varphi) - G(\varphi_0))\, \frac{\mu}{d^2}\, d\varphi,$$

and, consequently, we may assume, without restricting the generality of the problem, that $G(\varphi_0) = 0$.

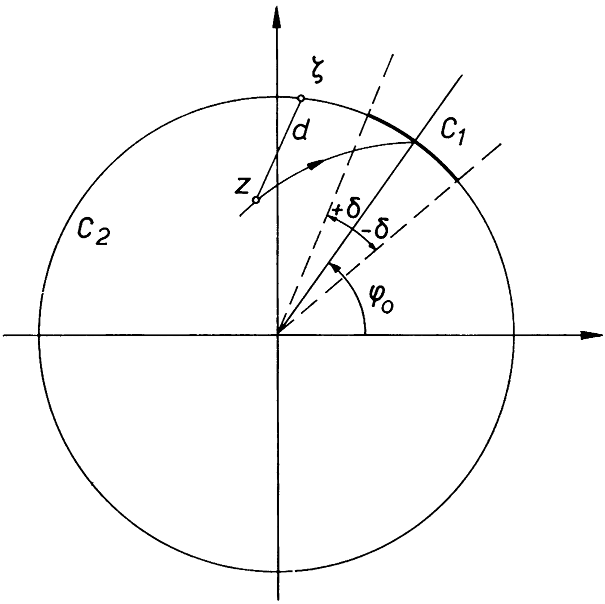


Fig. 10.1–2. Proof of the principle of symmetry

Let C_1 denote a circular arc on $|z| = R$ determined by $\varphi_0 - \delta \leqq \varphi \leqq \varphi_0 + \delta$, δ being a small positive number (e.g. $< \frac{1}{4}\pi$); by C_2 we denote the complementary arc on the circumference, (fig. 10.1–2). If δ is sufficiently small we evidently have, ε being an arbitrary positive number,

$$|G(\varphi)| < \tfrac{1}{2}\varepsilon,$$

provided that $|\varphi - \varphi_0| < \delta$. Let M be an upper bound of $|G(\varphi)|$ on C_2. Then

$$|\operatorname{Re} g(z)| \leqq \frac{\varepsilon}{4\pi}\int_{C_1} \frac{\mu}{d^2}\, d\varphi + \frac{M}{2\pi}\int_{C_2} \frac{\mu}{d^2}\, d\varphi$$

$$< \frac{\varepsilon}{4\pi}\int_0^{2\pi} \frac{\mu}{d^2}\, d\varphi + \frac{M}{2\pi}\int_{C_2} \frac{\mu}{d^2}\, d\varphi = \tfrac{1}{2}\varepsilon + \frac{M}{2\pi}\int_{C_2} \frac{\mu}{d^2}\, d\varphi.$$

Now we observe that μ tends to zero as $z \to Re^{i\varphi_0}$, while d^2 remains above a certain positive value if ζ is on C_2. Hence for all z sufficiently near $Re^{i\varphi_0}$

$$\frac{1}{2\pi}\int_{C_2} \frac{\mu}{d^2}\, d\varphi < \frac{\varepsilon}{2M}$$

and for these values of z we have $|\text{Re}\, g(z)| < \varepsilon$. This concludes the proof of the theorem.

10.1.3 – The symmetry principle

It is our aim to obtain a far reaching generalization of the elementary symmetry principle which has been discussed in section 9.4.2. As in section 9.2.1 we denote by $\mathfrak{A}_{\bar{z}}$ the set obtained by reflecting an open set $\mathfrak{A}_z$ in the real axis. In the section mentioned we proved that if $f(z)$ is holomorphic throughout $\mathfrak{A}_z$ then $g(z) = \overline{f(\bar{z})}$ is holomorphic throughout $\mathfrak{A}_{\bar{z}}$.

We suppose first that $\mathfrak{A}_{\bar{z}}$ is a region $\mathfrak{R}_z$ such that $\mathfrak{R}_z = \mathfrak{R}_{\bar{z}}$, in which case $\mathfrak{R}_z$ is said to be *symmetric* with respect to the real axis. Let $f(z)$ be holomorphic throughout $\mathfrak{R}_z$ and *real on the real axis*. Then $f(z)-\overline{f(\bar{z})}$ is holomorphic throughout $\mathfrak{R}_z$ and vanishes on the real axis. By the identity principle (section 2.11.2) this function is identically zero in $\mathfrak{R}_z$, i.e.,

$$f(z) = \overline{f(\bar{z})}. \tag{10.1–11}$$

Examples are provided by the elementary functions. Thus, for instance,

$$\log z = \log r + i\theta = \overline{\log r - i\theta} = \overline{\log \bar{z}};$$

$\log z$ is defined throughout the principal region $|z|+z \neq 0$ and real on the positive real axis which belongs to the region.

The meaning of (10.1–11) is that $f(z)$ takes conjugate values at conjugate points. Functions obeying the condition (10.1–11) are called *symmetric*.

Continuing our considerations we only require that $\mathfrak{R}_z$ meets the real axis. Then the join $\mathfrak{R}_z + \mathfrak{R}_{\bar{z}}$ is a symmetric region and the intersection of $\mathfrak{R}_z$ and $\mathfrak{R}_{\bar{z}}$ is a union of symmetric regions. If $f(z)$ is holomorphic in $\mathfrak{R}_z$ and real on the real axis then $f(z) = \overline{f(\bar{z})}$ in each component of the intersection of $\mathfrak{R}_z$ and $\mathfrak{R}_{\bar{z}}$. Now we can define a function $g(z)$ on $\mathfrak{R}_z + \mathfrak{R}_{\bar{z}}$ which is equal to $f(\bar{z})$ in $\mathfrak{R}_z$ and equal to $\overline{f(\bar{z})}$ on $\mathfrak{R}_{\bar{z}}$, for these functions coincide in the intersection of these regions. Thus we see that $f(z)$ has a symmetric holomorphic extension to $\mathfrak{R}_z + \mathfrak{R}_{\bar{z}}$. Very often there is no fear of confusion if we designate this extension also by $f(z)$. Summing up we have the following elementary version of *Schwarz's symmetry principle* (or *reflection principle*).

If $f(z)$ is holomorphic throughout a region $\mathfrak{R}_z$ which meets the real axis and if $f(z)$ takes real values on the real axis, then $f(z)$ has a symmetric extension throughout the region $\mathfrak{R}_z + \mathfrak{R}_{\bar{z}}$.

A simple example may illustrate this theorem. The function

$$w = \sqrt{1-z^2} \tag{10.1–12}$$

can be defined as single valued throughout the region $\operatorname{Im} z > 0$ (section 9.1.4) such that it takes the value 1 at $z = 0$. It is still regular at all points of the segment $-1 < x < 1$ and takes real values there. Hence it may be extended to a symmetric function in the z-plane cut along the half rays $x < -1$ and $x > 1$. At points of these half rays the function takes different values. If z tends to $\sqrt{2}$ from above, the limit is i, whereas the limit is $-i$ as z tends to $\sqrt{2}$ from below.

In many cases the theorem is applied under somewhat different circumstances. Suppose that $\mathfrak{R}_z$ is a region in the upper half plane and that its boundary contains an open subset $\mathfrak{a}$ of the real axis. Let

$$f(z) = u(z) + iv(z) \tag{10.1–13}$$

be holomorphic throughout $\mathfrak{R}_z$. If $\mathfrak{R}_{\bar{z}}$ denotes the region obtained by reflecting $\mathfrak{R}_z$ in the real axis, then the union $\mathfrak{R}_z + \mathfrak{a} + \mathfrak{R}_{\bar{z}}$ is a symmetric region and the question arises whether $f(z)$ admits a holomorphic extension in this larger region.

We shall prove that it is sufficient to assume that $v(z)$ *tends to zero as z approaches a point of* $\mathfrak{a}$. The existence of a limit of $u(z)$ is not a part of this hypothesis. In numerous practical cases it is known that $f(z)$ is real and continuous on $\mathfrak{a}$. Then the above condition is certainly fulfilled.

Without loss of generality we may assume that $z = 0$ is an interior point of $\mathfrak{a}$. We consider the semicircular disc $|z| \leqq R$, $\operatorname{Im} z \geqq 0$, whose interior and circular part of the boundary (with exception of the end points) belong to $\mathfrak{R}_z$, while the segment $-R \leqq z \leqq R$ is part of $\mathfrak{a}$.

The function

$$G(\varphi) = \begin{cases} 0, & \text{if } \varphi = 0 \text{ or } \varphi = \pi, \\ v(R, \varphi), & \text{if } 0 < \varphi < \pi, \\ -v(R, -\varphi), & \text{if } -\pi < \varphi < 0, \end{cases} \tag{10.1–14}$$

is clearly continuous and bounded along the circumference $|z| = R$ and takes opposite values at points of the circumference which are symmetric with respect to the real axis. As a consequence the function

$$g(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\varphi) \frac{\zeta + z}{\zeta - z} \, d\varphi \tag{10.1–15}$$

is holomorphic throughout $|z| < R$ and from the theorem of the previous sections it follows that its real part takes precisely the values $G(\varphi)$ on the boundary.

Next we write

$$\begin{aligned} g(z) &= \frac{1}{2\pi}\int_0^\pi G(\varphi)\frac{\zeta+z}{\zeta-z}\,d\varphi+\frac{1}{2\pi}\int_{-\pi}^0 G(\varphi)\frac{\zeta+z}{\zeta-z}\,d\varphi \\ &= \frac{1}{2\pi}\int_0^\pi v(R,\varphi)\frac{\zeta+z}{\zeta-z}\,d\varphi-\frac{1}{2\pi}\int_{-\pi}^0 v(R,-\varphi)\frac{\zeta+z}{\zeta-z}\,d\varphi \\ &= \frac{1}{2\pi}\int_0^\pi v(R,\varphi)\frac{\zeta+z}{\zeta-z}\,d\varphi+\frac{1}{2\pi}\int_{\pi}^0 v(R,\varphi)\frac{\bar\zeta+z}{\bar\zeta-z}\,d\varphi \\ &= \frac{1}{2\pi}\int_0^\pi v(R,\varphi)\left(\frac{\zeta+z}{\zeta-z}-\frac{\bar\zeta+z}{\bar\zeta-z}\right)d\varphi. \end{aligned}$$

The expression between brackets inside the last integral is the difference of two conjugate complex numbers if z is real; hence $g(z)$ is purely imaginary on the real axis. It follows that the function

$$\operatorname{Im}\big(f(z)-ig(z)\big) = v(z)-\operatorname{Im} ig(z) \tag{10.1–16}$$

tends to zero as z tends to a point of $\mathfrak{a}$ between $-R$ and R, these end points included. It is clear that this function vanishes on the upper half of the circle $|z| = R$. Since it is harmonic in the interior of the semicircle the principle of extreme values stated in section 2.14.1 implies that this function is zero. A function with vanishing imaginary part is a real constant c (section 3.12.2). Hence the difference of $f(z)$ and $ig(z)$ is equal to c if Re $z > 0$. The function

$$w(z) = ig(z)+c$$

is defined throughout $|z| < R$ and holomorphic. Hence this function is symmetric.

The function $f(z)$ in $\mathfrak{R}_z$ gives rise to a function $\overline{f(\bar z)}$ in $\mathfrak{R}_{\bar z}$. At a point of $\mathfrak{a}$ we carry out the construction as described above and we find a function $w(z)$ which coincides with $f(z)$ in the upper part of the disc and with $\overline{f(\bar z)}$ in the lower part. Now we identify $f(z)$ and $w(z)$ throughout the disc. It is clear that in overlapping discs the functions are the same in the intersection. Thus we have extended the given function $f(z)$ to a function holomorphic throughout $\mathfrak{R}_z+\mathfrak{a}+\mathfrak{R}_{\bar z}$ which takes real values on $\mathfrak{a}$ and is also regular at every point of $\mathfrak{a}$. Summing, up:

If $f(z)$ is holomorphic in a region $\mathfrak{R}_z$ in the upper half plane, whose boundary contains an open subset of the real axis, and if Im $f(z)$ *tends to zero as z approaches a point of* $\mathfrak{a}$, *then $f(z)$ admits of a holomorphic symmetric extension.*

This theorem has obvious generalizations. We may assume that $\mathfrak{a}$ is an open arc of a circumference and that the values of $w(z)$ approach a point on an arbitrary circumference (or a straight line) when z tends to a point of $\mathfrak{a}$. This case can be reduced to the former case by means of a suitable homeomorphism of the whole plane. The extension satisfies a symmetry relation with respect to the circles just as in the case of linear fractional transformations. Hence

If $\mathfrak{a}$ is an open arc of a circumference belonging to the boundary of a region $\mathfrak{R}$, if $w(z)$ is holomorphic in $\mathfrak{R}$ and if its values tend to the points of a circumference (which may be a straight line) then $w(z)$ is regular at each point of $\mathfrak{a}$ and can be extended beyond $\mathfrak{a}$.

In this form the symmetry principle has many applications. It should be noticed that $\mathfrak{R}$ is supposed to be inside or outside the circumference from which $\mathfrak{a}$ is a part.

10.1.4 – ANALYTIC ARCS

It is possible to generalize the symmetry principle to the case that the boundary contains a particular type of Jordan curves, viz. an analytical arc. Consider a function $z(t)$ holomorphic throughout a region $\mathfrak{R}_t$ containing the segment $0 \leqq t \leqq 1$. The set of points corresponding to those of the segment is called an *analytical arc*. An *analytical Jordan arc* is one with $z(t_1) \neq z(t_2)$ for $t_1 \neq t_2$. As usual it is assumed that $z'(t) \neq 0$ for all t in the interval.

By $\mathfrak{R}(\rho)$ we denote a subregion of $\mathfrak{R}_t$ consisting of all points whose distance from the segment $0 \leqq t \leqq 1$ is $< \rho$, where ρ does not exceed the (minimum) distance of the boundary of $\mathfrak{R}_t$ to the segment. It is clear that $\mathfrak{R}(\rho)$ is symmetric with respect to the real axis in the t-plane.

Analytical arcs have some simple geometric properties which make them useful for applications in various problems.

An analytical arc meets a straight line only in a finite number of points.

A line may be given by an equation

$$Ax+By+C = 0, \tag{10.1–17}$$

where A, B and C are real numbers. The function

$$A(z(t)+\overline{z(\bar{t})})+\frac{B}{i}(z(t)-\overline{z(\bar{t})})+2C \tag{10.1–18}$$

is holomorphic throughout $\mathfrak{R}(\rho)$ and has only a finite number of zeros in $\mathfrak{R}(\rho)$. The real zeros correspond to the points of intersection of the arc and the line.

By a similar argument we may prove

An analytic arc sends only a finite number of tangents through a given point.

Without restricting the generality we may assume that this point is at the origin. A tangent at a point t passes through the origin if there is a number λ such that

$$x(t) = \lambda x'(t), \qquad y(t) = \lambda y'(t),$$

or

$$x(t)y'(t) - x'(t)y(t) = 0.$$

Now we observe that the function

$$(z(t)+\overline{z(\bar{t})}(z'(t)-\overline{z'(\bar{t})}) - (z'(t)+\overline{z'(\bar{t})})(z(t)-\overline{z(\bar{t})})$$

has only a finite number of zeros in $\mathfrak{R}(\rho)$ and that the real zeros correspond to the points at which the tangents pass through the origin.

Suppose now that C is an analytical Jordan arc. Then we contend that $z(t)$ is univalent in $\mathfrak{R}(\rho)$ if ρ is sufficiently small. Let t_0 be a point of the segment. Since $z'(t_0) \neq 0$ there is a neighbourhood of t_0 in which $z(t)$ is univalent. Assume that we could find pairs $t_{1n} \neq t_{2n}$, $n = 1, 2, \ldots$ tending to the segment such that $z(t_{1n}) = z(t_{2n})$. These sequences contain convergent subsequences t_{1n_k}, t_{2n_k} with limits t_1, t_2 and by continuity $z(t_1) = z(t_2)$. By hypothesis $t_1 = t_2 = t_0$, say. For sufficient large k the points t_{1n_k}, t_{2n_k} are in a given neighbourhood of t_0 and thus we arrive at a contradiction.

If t and $\bar{t}$ are conjugate points in $\mathfrak{R}(\rho)$ the corresponding points $z(t)$ and $z(\bar{t})$ are called *symmetric with respect to the analytical Jordan arc C*, (fig. 10.1–3).

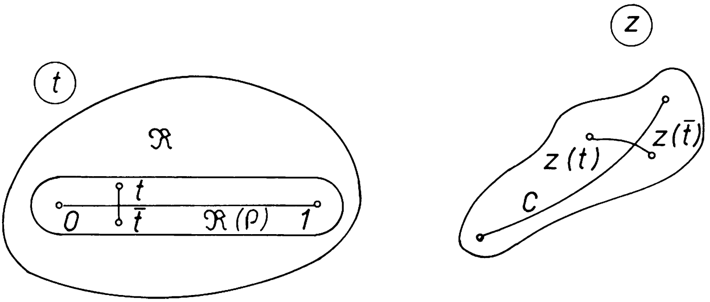


Fig. 10.1–3. Symmetry with respect to an analytic arc

It is necessary to show that this notion depends only on the point set C and not on the particular choice of the parameter. Let $z_1(t_1)$ and $z_2(t_2)$ be parametric representations of the same arc C, $(0 \leqq t_1 \leqq 1, 0 \leqq t_2 \leqq 1)$, and choose corresponding regions $\mathfrak{R}_1$, $\mathfrak{R}_2$ as above, their images being

$\mathfrak{S}_1$ and $\mathfrak{S}_2$ respectively. Let $\mathfrak{D}$ be a component of the intersections of $\mathfrak{S}_1$ and $\mathfrak{S}_2$, containing C and $\mathfrak{D}_1$ and $\mathfrak{D}_2$ its inverse images in the t_1-plane and t_2-plane respectively, (fig. 10.1–4). The mappings $z = z_1(t_1)$ and $z = z_2(t_2)$ induce a one-to-one holomorphic correspondence between $\mathfrak{D}_1$ and $\mathfrak{D}_2$, and by the first theorem of the previous section conjugate values of t_1 correspond to conjugate values of t_2. The points $z(t)$ for which Im t has the same sign are called points on the same side of C. Since two points on the same side can be joined by an arc on which no point is symmetric to itself, this notion is also independent of the parameter.

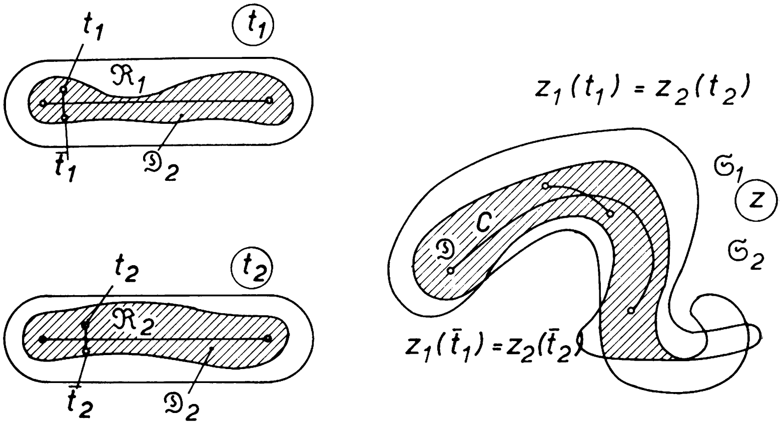


Fig. 10.1–4. Independence of symmetry of the parameter

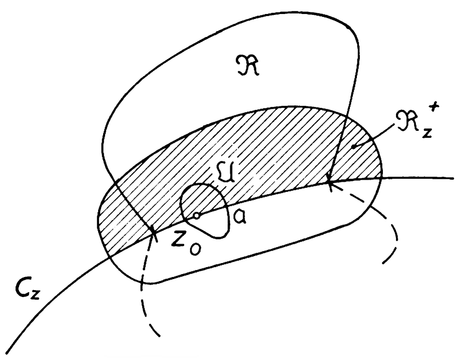


Fig. 10.1–5. Symmetry principle for analytic arcs

Now we are sufficiently prepared to state the symmetry principle for analytic Jordan arcs, (fig. 10.1–5). Consider an analytic Jordan arc C_z and a region $\mathfrak{R}_z$ in which symmetry with respect to C_z is defined. Let

$\mathfrak{R}_z^+$ denote a subregion consisting of all points on the same side of C_z. Let further $\mathfrak{R}$ denote a region and $\mathfrak{a}$ a subset of C_z subject to the condition: every z_0 on $\mathfrak{a}$ has a neighbourhood $\mathfrak{U}$ such that the intersection of $\mathfrak{U}$ and C_z is in $\mathfrak{a}$ and the intersections of $\mathfrak{U}$ and $\mathfrak{R}$ coincides with the intersection of $\mathfrak{U}$ and $\mathfrak{R}_z^+$. Now we contend:

If $f(z)$ is holomorphic in $\mathfrak{R}$ and if the limits of $f(z)$ as z approaches points of $\mathfrak{a}$ all lie on an analytical Jordan arc C_w in the w-plane, then $f(z)$ has a holomorphic extension to a region which contains $\mathfrak{R}+\mathfrak{a}$.

If C_z is given by $z = z(t)$ and C_w by $w = w(\tau)$, then $w(\tau) = f(z(t))$ defines a relation between τ and t which is a holomorphic function in $\mathfrak{R}_t^+$. If t tends to real values, then τ also tends to real values. Hence the conditions of the second theorem of the previous section are satisfied and we may conclude that τ is extensible beyond the real t-axis. As a consequence $f(z)$ can be extended beyond the arc C_z.

10.2 – Examples of conformal mapping

10.2.1 – Introduction

In the subsequent sections we shall list some illustrative examples of the images of the interior of the unit circle as effected by functions which are meromorphic in the interior. Our examples are, however, of a restricted type, since they are continuous on the circumference. The case that they take an infinite value on the circumference is included, provided the function is chordally continuous there. This infinity can then be removed by combining the mapping with a suitable Möbius transformation.

Instead of using a circular disc we can also start with a half plane, for these two domains are univalently equivalent. We consider this fact in

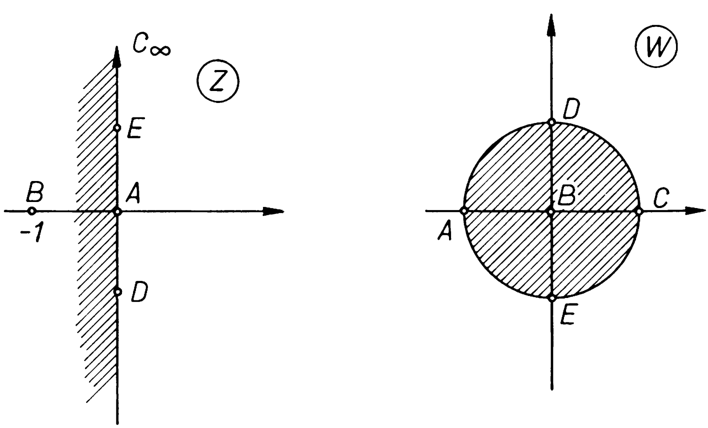


Fig. 10.2–1. The mapping (10.2–1)

more detail by directing our attention to the mapping

$$w = \frac{z+1}{z-1}. \tag{10.2–1}$$

The inverse mapping is

$$z = \frac{w+1}{w-1}. \tag{10.2–2}$$

Hence the mapping is involutory. We contend that the half plane Re $z < 0$ corresponds to the region $|w| < 1$, (fig. 10.2–1). In fact

$$w\bar{w}-1 = \frac{z+1}{z-1}\frac{\bar{z}+1}{\bar{z}-1}-1 = \frac{2(z+\bar{z})}{|z-1|^2} < 0,$$

if Re $z = \frac{1}{2}(z+\bar{z}) < 0$. Of course Re $w < 0$ is also transformed into $|z| < 1$. This can be verified directly as follows

$$w+\bar{w} = \frac{z+1}{z-1}+\frac{\bar{z}+1}{\bar{z}-1} = 2\,\frac{z\bar{z}-1}{|z-1|^2} < 0,$$

if and only if $z\bar{z} = |z|^2 < 1$. Finally we observe that Re $w > 0$ corresponds to the exterior of the unit circle.

Combining the transformations

$$w_1 = \frac{z+1}{z-1}$$

and

$$w = -iw_1,$$

the latter denoting a rotation through a right angle to the right, we obtain

$$\boxed{w = i\,\frac{1+z}{1-z}, \qquad z = \frac{w-i}{w+i},} \tag{10.2–3}$$

which maps the interior of $|z| = 1$ onto the upper half of the w-plane. This is the mapping (9.4–4).

Another involutory mapping is

$$w = \frac{1}{z}, \tag{10.2–4}$$

which *transforms the interior of the unit circle onto the exterior of the unit circle.*

The mapping (10.2–1) has still another remarkable property: *it maps the upper half plane onto the lower half plane.* This follows from

$$w-\bar{w} = \frac{z+1}{z-1}-\frac{\bar{z}+1}{\bar{z}-1} = 2\,\frac{\bar{z}-z}{|z-1|^2}.$$

The involutory character is now geometrically evident.

The Möbius transformations can often serve to simplify certain given situations. For instance, the region bounded by two intersecting circles or by two circles having internal contact, (fig. 10.2–2), can be transformed into an angular region or a strip respectively by means of a Möbius transformation whereby a point common to the two circles corresponds to the point ∞. As we shall see further on these regions can easily be mapped onto a half plane or, which amounts to the same, onto the interior of a circle.

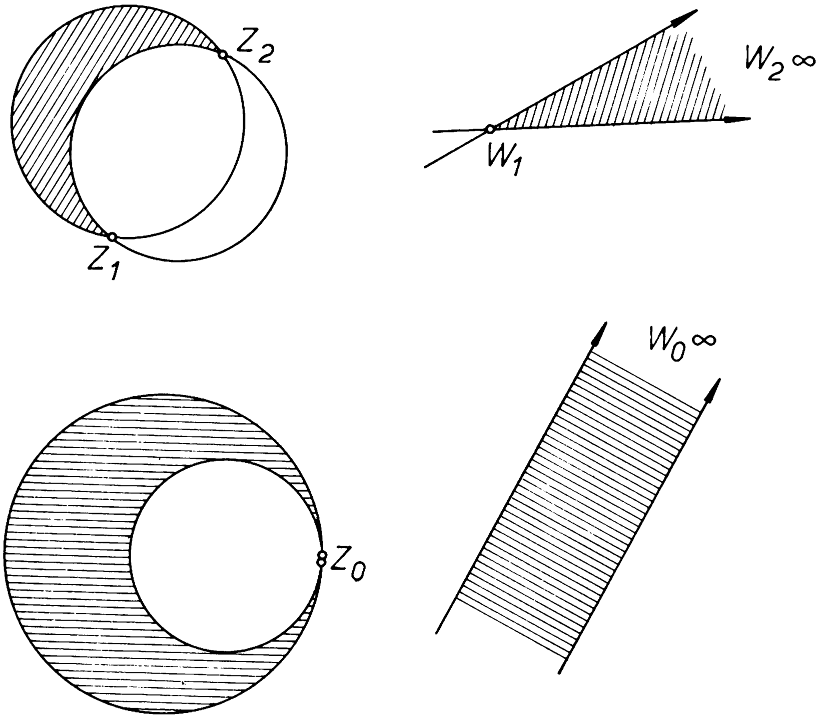


Fig. 10.2–2. The mapping of a lunular region onto an angular region or onto a strip

10.2.2 – The angular region

An arbitrary angular region whose angular width we denote by $\alpha\pi$, $0 < \alpha \leqq 2$ and whose vertex we suppose to be at $w = 0$ can be transformed into a half plane by means of

$$w = z^\alpha, \tag{10.2–5}$$

the function on the right being a branch which is positive for $z > 0$. Indeed, on introducing polar coordinates, this relation takes the form

$$w = r^\alpha e^{i\alpha\theta}$$

and the image of the half ray θ = constant in the z-plane corresponds to the half ray $\varphi = \alpha\theta$ = constant in the w-plane. Hence the region $\operatorname{Im} z > 0$ corresponds to an angular region of width $\alpha\pi$. If we agree that (10.2–5) represents the branch of the power function which takes real values if z is real and positive, then the positive real axis in the z-plane corresponds to the positive real axis in the w-plane and to the negative real axis in the z-plane corresponds the half ray $\varphi = \alpha\pi$. The case $\alpha = 1$ corresponds to the identity mapping. If $\alpha = 2$ the image of the upper half of the z-plane is the w-plane slit along the positive real axis.

If we suppose that in (10.2–5) α is negative $-2 \leqq \alpha < 0$, we consider

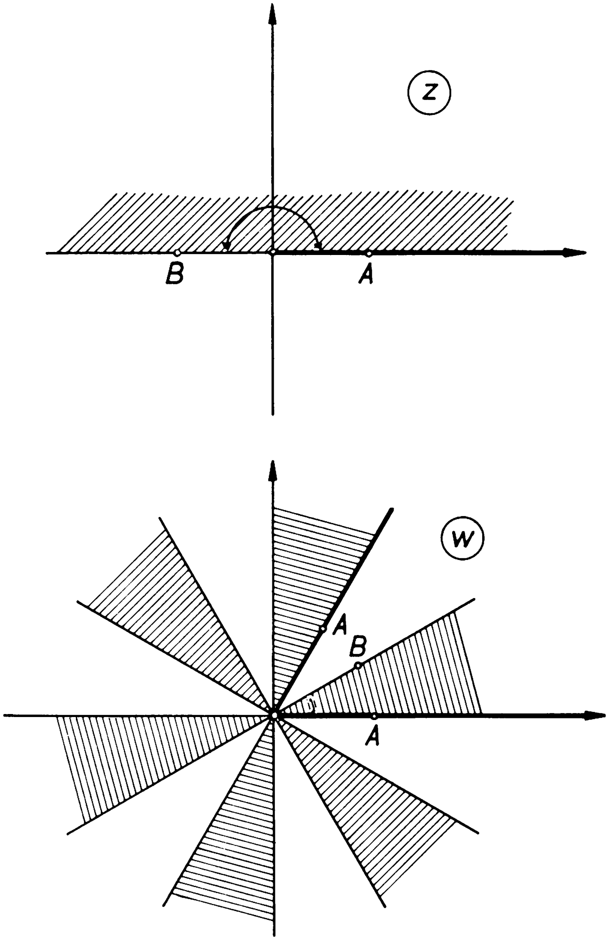


Fig. 10.2–3. The mapping $w = z^{\frac{1}{3}}$

the combination of the mappings

$$z_1 = \frac{1}{z}, \qquad w = z_1^{-\alpha}.$$

The first mapping interchanges the upper and the lower half of the z-plane. Hence, on account of Schwarz's symmetry principle the image of the upper half of the z-plane is mapped onto a region obtained from that found above by reflecting with respect to the real axis in the w-plane.

Of particular interest is the case that $\alpha = 1/n$, where n is a natural number. We obtain an angular region with width π/n. In fig. 10.2–3 this region is shaded. The inverse function

$$z = w^n \tag{10.2–6}$$

is regular everywhere in the angular region and also on the boundary. Since the bounding leg which is not the positive real axis is transformed into the negative real axis we can extend the function (10.2–6) beyond this leg by means of Schwarz's symmetry principle. Thus we obtain a congruent region, unshaded in the figure, which corresponds to the lower z-plane. The union of these regions and the common boundary correspond to the whole z-plane slit along the positive real axis.

Repeating the process of reflection we obtain n shaded and unshaded regions which correspond to the upper and lower half plane respectively by means of (10.2–6) and which cover the w-plane without gaps or overlappings. Thus we have verified that w^n is a single-valued function defined throughout the whole w-plane.

The pattern plotted in fig. 10.2–3 illustrates a remarkable property of symmetry of the functions w^n, viz.,

$$w^n = (\eta w)^n \tag{10.2–7}$$

with $\eta = \exp(2\pi i/n)$. In fact, multiplying by η^m means a rotation through an angle equal to a multiple of $2\pi/n$ and this leaves the pattern in the w-plane as a whole invariant.

In this example we encounter the important notion of *fundamental domain* of a group related to a function. It is clear that the rotation

$$w = \eta w, \qquad \eta = \exp(2\pi i/n) \tag{10.2–8}$$

generates a cyclic group of order n. Two points in the w-plane are said to be "congruent" with respect to the group if they correspond under a transformation of the group. Now a fundamental domain may be characterized as follows:

i) No two points of the domain are congruent;

ii) Every point of the w-plane is congruent to precisely one point of the domain, (see also section 13.2.1).

It should be noticed that a fundamental domain is not uniquely determined by the group, for from a given domain we may delete a certain part and add a part congruent with this.

It is easy to verify that the domain $0 \leqq \arg w < 2\pi/n$ ($w = 0$ included) is an example of a fundamental domain of the cyclic group generated by (10.2–8).

The function w^n, invariant under the transformations of this group, is said to be *automorphic* with respect to this group. The automorphic functions constitute a remarkable class of functions; in subsequent sections we shall encounter other interesting examples and chapter 13 is devoted to the general theory of these functions.

10.2.3 – The exterior of an ellipse

We start with the problem of representing the z-plane, cut along the

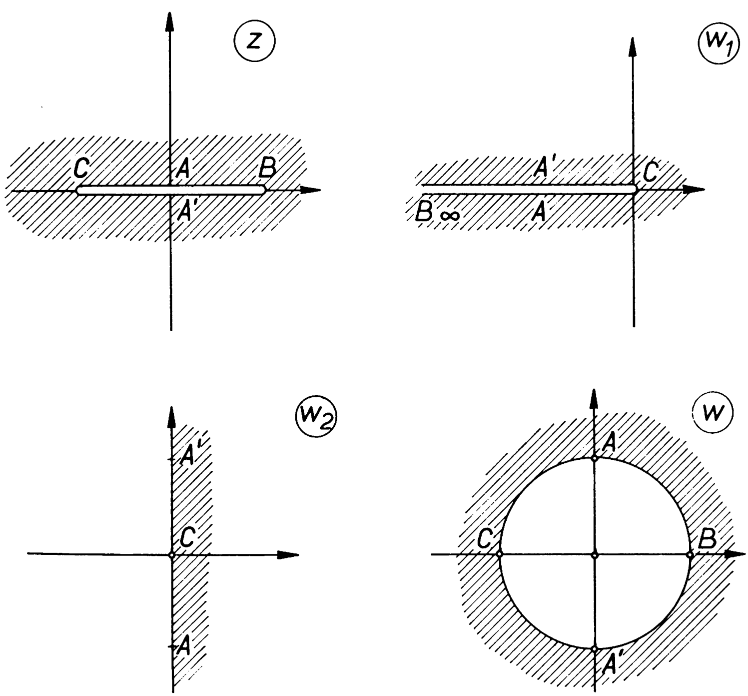


Fig. 10.2–4. The mapping (10.2–9)

segment $-1 \leqq z \leqq 1$ onto the exterior $|w| > 1$ of the unit circle, so that the points $z = \infty$ and $w = \infty$ correspond. In view of the results obtained in section 10.2.1 we see that the transformation

$$w_1 = \frac{z+1}{z-1}, \qquad z = \frac{w_1+1}{w_1-1}$$

transforms the cut z-plane into the w_1-plane cut along the negative real axis. The transformation $w_2 = \sqrt{w_1}$ transforms this into the half plane $\operatorname{Re} w_2 > 0$ and this half plane corresponds to the exterior of the unit circle by means of

$$w = \frac{w_2+1}{w_2-1}.$$

We eliminate the auxiliary variables by writing

$$w + \frac{1}{w} = 2\frac{w_2^2+1}{w_2^2-1} = 2\frac{w_1+1}{w_1-1} = 2z,$$

thus the relation

$$z = \frac{1}{2}\left(w + \frac{1}{w}\right) \tag{10.2–9}$$

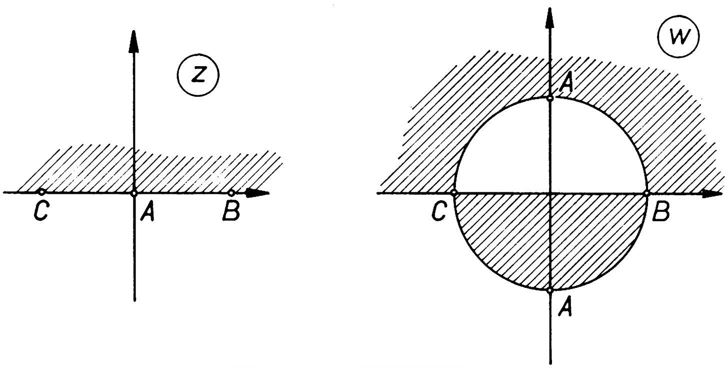


Fig. 10.2–5. The pattern associated with the mapping (10.2–9)

effects the desired mapping, (fig. 10.2–4). It is clear that it also maps the slit plane onto the interior of the unit circle.

The figure in the z-plane is symmetric with respect to the half lines $z < -1$ and $z > +1$ which correspond to $w < -1$ and $w > +1$ in the w-plane. Hence, in view of the symmetry principle, the upper half of the z-plane is mapped onto the upper half of the w-plane outside the unit circle (fig. 10.2–4), the segment $-1 < z < 1$ corresponding to the semicircular arc. As a consequence the interior of the upper half of the semicircle in the w-plane corresponds to the lower half of the z-plane.

Repeating the process of reflection we obtain the pattern of fig. 10.2–5.

The function (10.2–9) is automorphic with respect to a group of order two generated by the transformation $w \to 1/w$.
A fundamental domain is evidently the union of the region $|w| < 1$ and the arc $|w| = 1, 0 \leqq \arg w < \pi$.

Next we interchange the symbols z and w and we consider the relation

$$w = \frac{1}{2}\left(z + \frac{1}{z}\right). \qquad (10.2\text{–}10)$$

Introducing polar coordinates in the z-plane: $z = re^{i\theta}$ we have

$$w = u + iv = \frac{1}{2}\left(r + \frac{1}{r}\right)\cos\theta + \tfrac{1}{2}i\left(r - \frac{1}{r}\right)\sin\theta \qquad (10.2\text{–}11)$$

and it appears that the circle $|z| = r > 1$ corresponds to an ellipse in the w-plane with semi axes

$$a = \frac{1}{2}\left(r + \frac{1}{r}\right), \qquad b = \frac{1}{2}\left(r - \frac{1}{r}\right). \qquad (10.2\text{–}12)$$

If r increases a and b do so too. Hence the exterior of the circle $|z| = r > 1$ (or the interior of the circle $|z| = 1/r < 1$) corresponds to the exterior of the ellipse (10.2–11), (fig. 10.2–6). If r tends to 1 the ellipse shrinks into the segment $-1 \leqq w \leqq 1$.

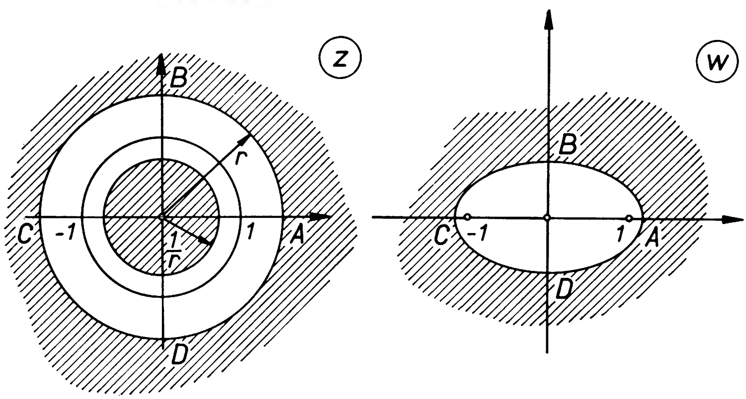


Fig. 10.2–6. The mapping (10.2–10)

The representation of the interior of the ellipse cannot be obtained in this simple way. We shall solve this problem in section 10.2.12, using a Jacobian elliptic function.

Another application of the mapping (10.2–10) which deserves mention is the following. We modify it slightly by writing

$$w_1 = z + \frac{1}{z} - 2.$$

It is clear that the interior of the unit circle of the z-plane corresponds to the w_1-plane cut along the segment $-4 \leqq w_1 \leqq 0$. This is transformed into the w-plane cut along the half ray $w \leqq -\frac{1}{4}$ by means of $w = 1/w_1$, (fig. 10.2–7). Eliminating w_1 we obtain the function

$$\boxed{w = \frac{z}{(1-z)^2},} \tag{10.2–13}$$

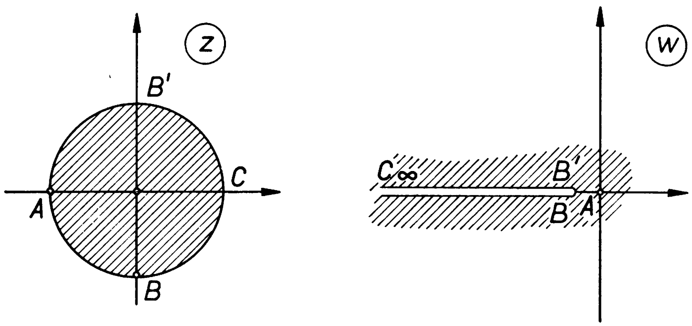


Fig. 10.2–7. The mapping as given by Koebe's function (10.2–13)

known as *Koebe's function*, which plays an important part in the theory of univalent functions to be dealt with in the next chapter. It has the remarkable expansion

$$w = z + 2z^2 + 3z^3 + \ldots = \sum_{\nu=1}^{\infty} \nu z^\nu \tag{10.2–14}$$

all coefficients being equal to the exponent of z.

10.2.4 – The dihedral function

A natural generalization of (10.2–9) is the *dihedral function*

$$z = \frac{1}{2}\left(w^n + \frac{1}{w^n}\right), \tag{10.2–15}$$

where n is an integer $\geqq 2$. This function is automorphic with respect to the cyclic group of rotations of order n but also with respect to the extended group which is obtained by adjoining the transformation $w \to 1/w$. This group can be generated by the transformations

$$\mathsf{A} : w \to \eta w, \qquad \eta = \exp(2\pi i/n)$$
$$\mathsf{B} : w \to 1/w,$$

with the defining relations

$$\mathsf{A}^n = \mathsf{E}, \quad \mathsf{B}^2 = \mathsf{E}, \quad \mathsf{AB} = \mathsf{BA}^{-1}. \tag{10.2–16}$$

The transformation as given by (10.2–15) is the product of

$$z = \frac{1}{2}\left(w_1 + \frac{1}{w_1}\right)$$

and

$$w_1 = w^n,$$

and it is easily seen that the region $\text{Im}\, z > 0$ is transformed into the region $0 < \arg w < \pi/n$ outside the circle $|w| = 1$. On applying the symmetry principle we obtain a pattern of shaded and unshaded regions (fig. 10.2–8) which correspond to the upper and lower half of the z-plane respectively. A fundamental domain consists of an angular region containing a shaded and an unshaded part together with a bounding half ray.

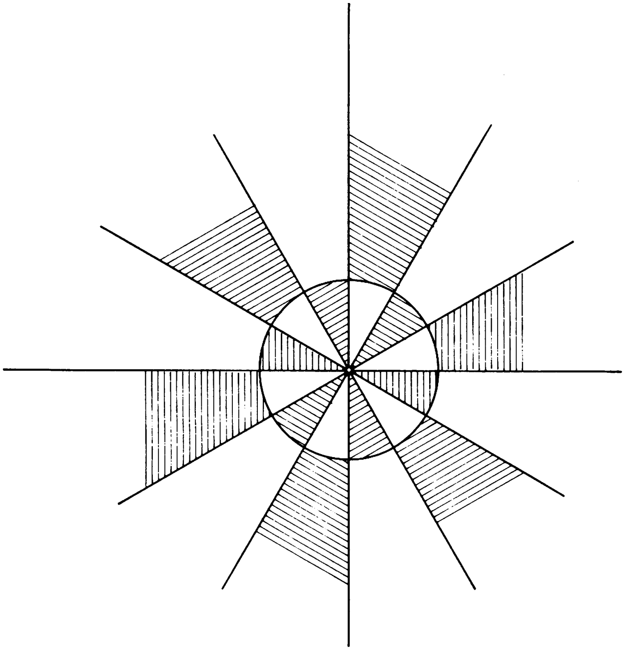


Fig. 10.2–8. The pattern associated with the dihedral function with $n = 6$

The case $n = 2$ deserves mention. Then the dihedral group introduced above has the defining relations

$$\mathsf{A}^2 = \mathsf{E}, \quad \mathsf{B}^2 = \mathsf{E}, \quad \mathsf{AB} = \mathsf{BA}.$$

Writing $\mathsf{C} = \mathsf{AB}$, we evidently also have $\mathsf{C}^2 = \mathsf{ABAB} = \mathsf{AABB} = \mathsf{E}$, $\mathsf{BC} = \mathsf{BBA} = \mathsf{A} = \mathsf{ABB} = \mathsf{CB}$, $\mathsf{CA} = \mathsf{BAA} = \mathsf{B} = \mathsf{AAB} = \mathsf{AC}$. The group possesses four elements and is known as the *four group* of F. Klein.

10.2.5 – THE SEMICIRCLE AND THE SLIT CIRCLE

In section 10.2.3 we found that the relation (10.2–9) whose inverse is

$$w = z + \sqrt{z^2 - 1} \qquad (10.2\text{–}17)$$

maps a half z-plane onto a semicircle in the w-plane, more precisely,

$$z_1 = \frac{1}{2}\left(w + \frac{1}{w}\right)$$

maps the lower half of the z_1-plane onto the semicircle $|w| < 1$, Im $w > 0$. The transformation

$$z = \frac{z_1 + 1}{z_1 - 1}$$

maps the lower half of the z_1-plane onto the upper half of the z-plane. Eliminating z_1 we find that

$$z = \left(\frac{w+1}{w-1}\right)^2 \qquad (10.2\text{–}18)$$

and its inverse

$$w = \frac{\sqrt{z} - 1}{\sqrt{z} + 1}, \qquad (10.2\text{–}19)$$

where $\sqrt{z}$ is such that it takes positive values for positive values of z, maps the upper half of the z-plane onto a semicircle in the w-plane (fig. 10.2–9).

It is easy to see that

$$w = \left(\frac{\sqrt{z} - 1}{\sqrt{z} + 1}\right)^2 \qquad (10.2\text{–}20)$$

maps the region Im $z > 0$ onto the interior of the unit circle in the w-plane cut along the radius $0 \leqq w < 1$, (fig. 10.2–10). It is instructive to consider the correspondence between the boundaries in more detail. If z increases from 0 to 1 then w decreases from 1 to 0 and if z increases from 1 to ∞ then w increases from 0 to 1. If $z = -x$, $x > 0$ then $|w| = 1$, hence the negative real axis in the z-plane corresponds to the circular part of the boundary

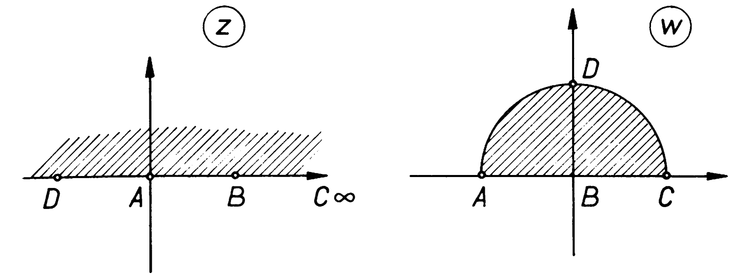


Fig. 10.2–9. The mapping (10.2–19)

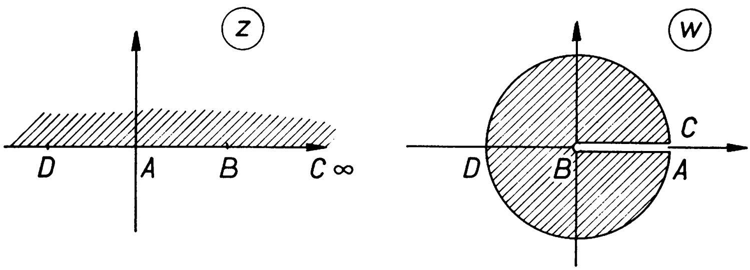


Fig. 10.2–10. The mapping (10.2–20)

10.2.6 – The infinite strip

The mapping functions so far employed (except for (10.2–5) if α is not rational) are all algebraic. A greater variety of possibilities is obtained by considering the elementary transcendental functions too. We start with the mapping (10.2–5) and observe that it does not differ essentially from

$$w = \frac{z^\alpha - 1}{\alpha}, \qquad \alpha > 0. \tag{10.2–21}$$

Now $z = 0$ corresponds to $w = -1/\alpha$. If $\alpha \to 0$ the angular region tends to a strip, for the angular width tends to zero and the vertex to $-\infty$. Since

$$\frac{z^\alpha - 1}{\alpha} = \frac{\exp(\alpha \log z) - 1}{\alpha} \to \log z$$

as $\alpha \to 0$, it is natural to consider the mapping

$$w = \log z, \tag{10.2–22}$$

where the logarithm may denote the principal branch. If $z = re^{i\theta}$, $0 \leqq \theta \leqq \pi$, then $w = \log r + i\theta$. Hence a half ray $\theta =$ constant corresponds to a horizontal line at a distance θ above the real axis in the w-plane. Thus the region $\text{Im } z > 0$ corresponds to the infinite strip $0 < \text{Im } w < \pi$.

The inverse function

$$z = \exp w \tag{10.2–23}$$

may be extended beyond the line $\text{Im } w = \pi$ and thus we obtain an unshaded strip corresponding to $\text{Im } z < 0$. Repeating the process of reflection on both sides of the strip we find a pattern of shaded and unshaded

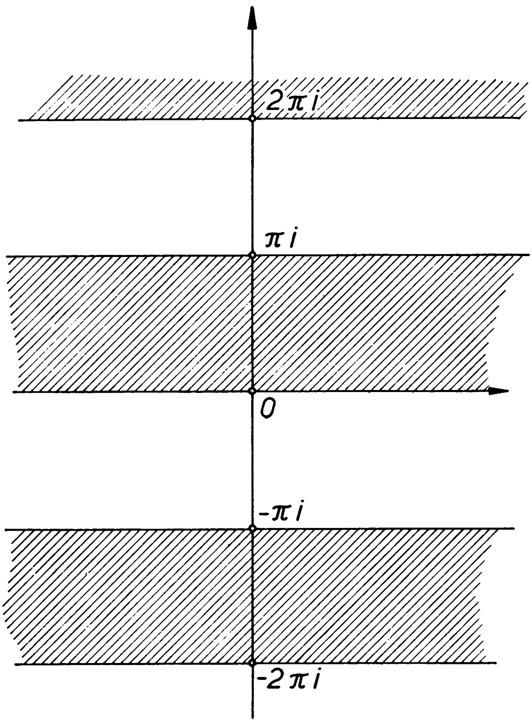


Fig. 10.2–11. The pattern associated with the logarithmic function

strips (fig. 10.2–11) covering the whole w-plane without gaps or overlappings. This is in accordance with the fact that the exponential function is defined throughout the whole plane. The transformation

$$w \to w + 2\pi i$$

generates a cyclic group of translations in the vertical direction, the order of this group being infinite. This illustrates the fact that the exponential function is automorphic with respect to this group, i.e., this function is periodic. The domain $0 \leq \text{Im}\, w < 2\pi$ is a fundamental domain.

10.2.7 – The semi infinite strip

The hyperbolic and circular functions may be obtained from the exponential function by simple algebraic operations. Consider first

$$w_1 = \exp w, \tag{10.2-24}$$

the mapping of the strip $0 < \text{Im}\, w < \pi$ onto the upper half of the w_1-

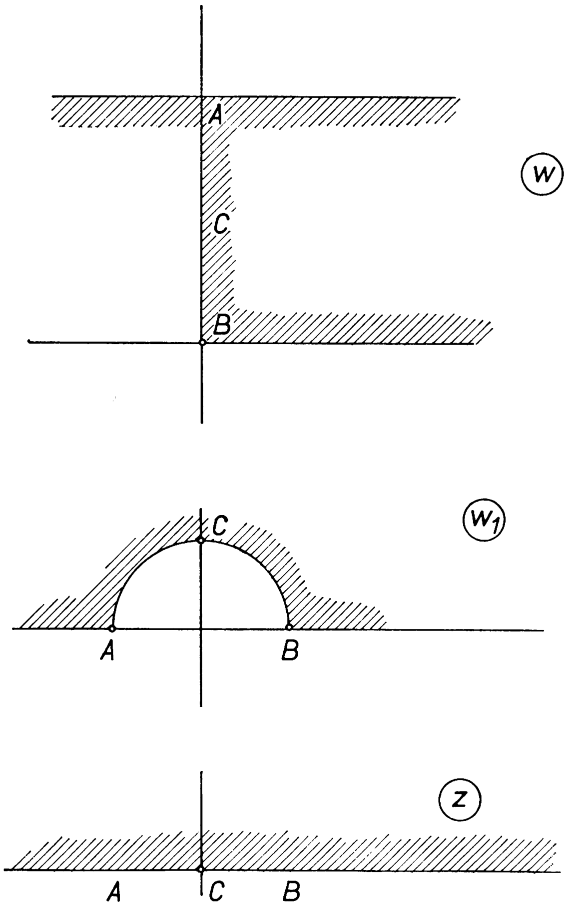


Fig. 10.2–12. The mapping as given by (10.2–26)

plane. If we write $w = u+iv$ then

$$w_1 = e^u \cos v + ie^u \sin v.$$

Take first $v = \pi$ and let u decrease from $+\infty$ to 0; then w_1 moves from $-\infty$ to -1. If now u remains equal to 0 and v decreases from π to 0 the corresponding point w_1 describes the semicircle $\operatorname{Im} w_1 \geqq 0$, $|w_1| = 1$ from the left to the right. If, finally, $v = 0$ and u increases from 0 to ∞ then w_1 increases from 1 to ∞. The line $u = 0$ is a line of symmetry of the infinite strip in the w-plane and the half of the unit circle in the w_1-plane is a line of symmetry in the upper half of the w_1-plane. It follows from Schwarz's symmetry principle that the half strip $\operatorname{Re} w > 0$, $0 < \operatorname{Im} w < \pi$ corresponds to the upper half of the w_1 plane outside the unit circle, (fig. 10.2–12). In section 10.2.3, we found that this region is mapped onto

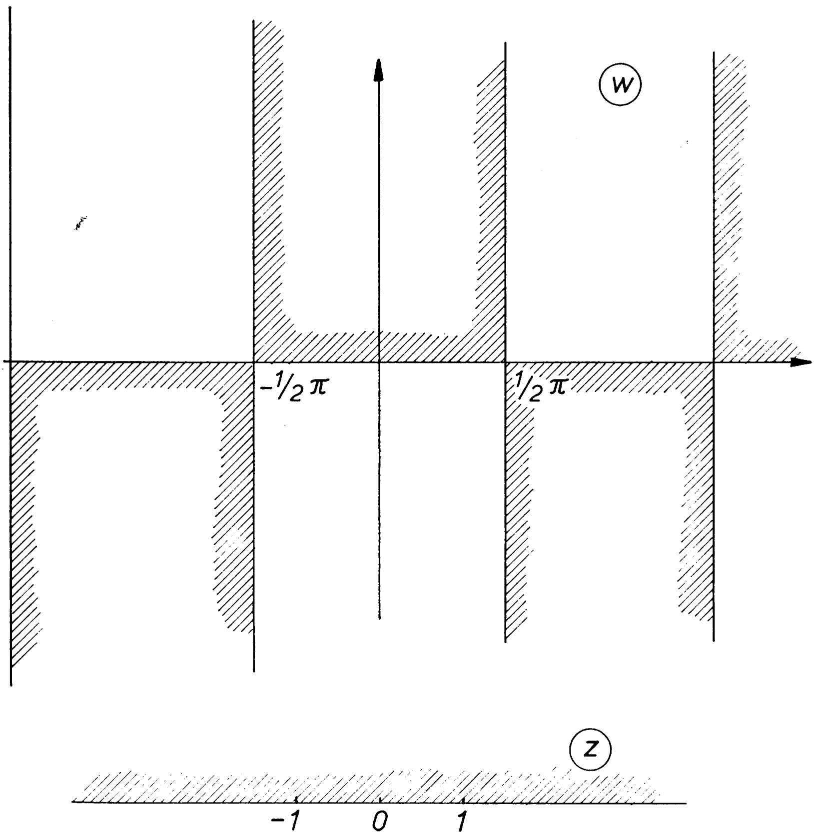


Fig. 10.2–13. The pattern associated with (10.2–28)

the upper half of the z-plane by means of

$$z = \frac{1}{2}\left(w_1 + \frac{1}{w_1}\right) \tag{10.2–25}$$

and it follows by eliminating w_1 from (10.2–24) and (10.2–25) that

$$z = \cosh w \tag{10.2–26}$$

maps the half strip described above onto the region $\operatorname{Im} z > 0$. Replacing w by w/i means a rotation in the w-plane to the right through a right angle.

Hence the function

$$z = \cos w \tag{10.2–27}$$

maps the half strip $0 < \operatorname{Re} w < \pi$, $\operatorname{Im} w < 0$, onto the upper half of the z-plane. Replacing w by $\frac{1}{2}\pi - w$ means a reflection in the origin followed by a shift through the distance $\frac{1}{2}\pi$ to the right. As a consequence *the function*

$$z = \sin w \tag{10.2–28}$$

maps the strip $-\frac{1}{2}\pi < \operatorname{Re} w < \frac{1}{2}\pi$, $\operatorname{Im} w > 0$, *onto the upper half of the z-plane.* Applying the symmetry principle we obtain a pattern of shaded and unshaded half strips, (fig. 10.2–13) which is invariant under the transformation

$$w \to \pi - w. \tag{10.2–29}$$

Hence the sine function is automorphic with respect to this group of transformations generated by (10.2–29).

The inverse of the function (10.2–28) may be represented as

$$w = \int_0^z \frac{dt}{\sqrt{1-t^2}}. \tag{10.2–30}$$

We remove the ambiguity if we agree that the argument of the denominator is equal to $\frac{1}{2}(\theta_1+\theta_2-\pi)$, where $\theta_1 = \arg(z+1)$, $\theta_2 = \arg(z-1)$, $0 \leqq \theta_1, \theta_2 \leqq \pi$. The integral does not depend on the path of integration. If z moves from -1 to $+1$ along the real axis then w describes the horizontal segment $-\frac{1}{2}\pi \leqq w \leqq \frac{1}{2}\pi$, for then $\theta_1 = 0$, $\theta_2 = \pi$, whence $\theta = 0$. If $z > 1$ we may write, since $\theta = -\frac{1}{2}\pi$,

$$w = \int_0^1 \frac{dt}{\sqrt{1-t^2}} + i\int_1^z \frac{dt}{\sqrt{t^2-1}},$$

the argument of the second integrand being $\frac{1}{2}\pi$ if $z > 1$, so w describes the half line Re $w = \frac{1}{2}\pi$ from below to above. If $z < -1$ we may write

$$w = -i\int_{-1}^{z} \frac{dt}{\sqrt{t^2-1}} + \int_{-1}^{0} \frac{dt}{\sqrt{1-t^2}}$$
$$= i\int_{z}^{-1} \frac{dt}{\sqrt{t^2-1}} - \int_{0}^{1} \frac{dt}{\sqrt{1-t^2}}$$

for $\theta = \frac{1}{2}\pi$, i.e., the argument of the first integrand is $-\frac{1}{2}\pi$ if $z < -1$.

10.2.8 – The mapping of an infinite strip onto a circle

It is easily verified that the relation

$$w_1 = e^{2iw} = e^{-2v}\cos 2u + ie^{-2v}\sin 2u \qquad (10.2\text{–}31)$$

effects the mapping of the infinite strip $-\frac{1}{4}\pi <$ Re $w < \frac{1}{4}\pi$ onto the right

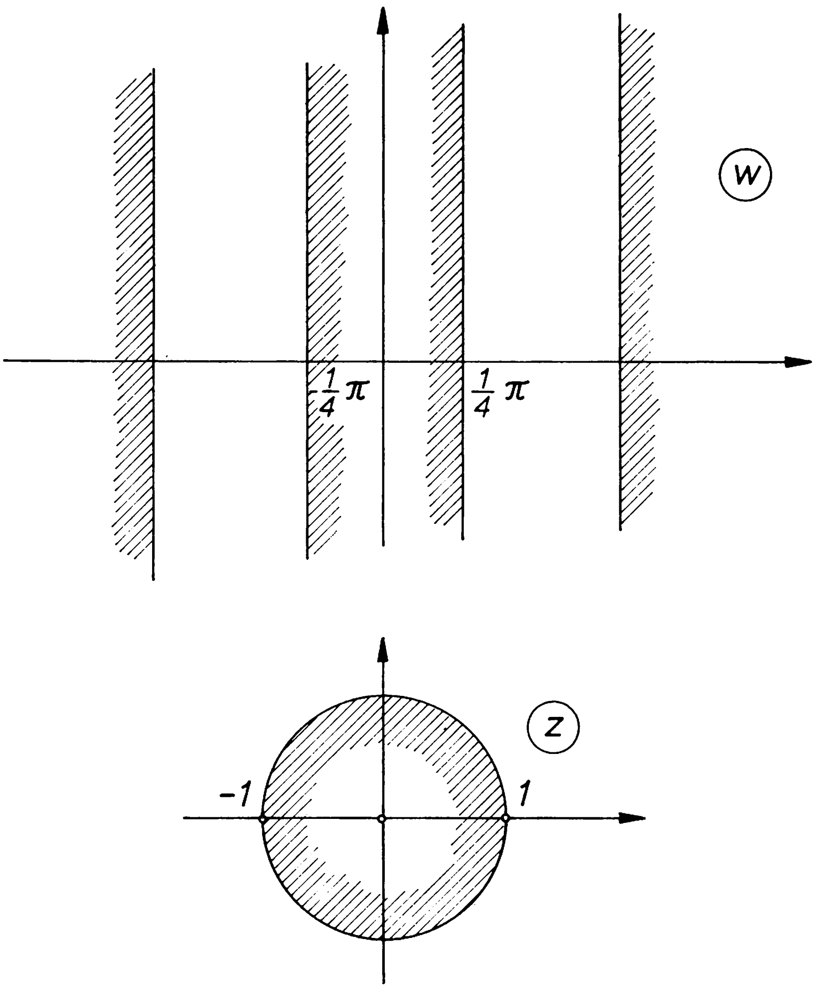


Fig. 10.2–14. The mapping as given by (10.2–33)

half plane Re $w_1 > 0$. By an argument employed in section 10.2.1 we deduce that

$$z = \frac{1}{i}\frac{w_1-1}{w_1+1} \tag{10.2–32}$$

maps Re $w_1 > 0$ onto the interior of the unit circle. Eliminating w_1 from (10.2–31) and (10.2–32) we get

$$z = \tan w \tag{10.2–33}$$

representing the mapping of $-\frac{1}{4}\pi <$ Re $w < \frac{1}{4}\pi$ *onto the region* $|z| < 1$. Applying the symmetry principle we see that $\frac{1}{4}\pi <$ Re $w < \frac{3}{4}\pi$ corresponds to the exterior $|z| > 1$ of the unit circle and proceeding in the usual way we obtain a pattern of shaded and unshaded strips, (fig. 10.2–14). This pattern is invariant under the group of transformations generated by

$$w \to w+\pi,$$

illustrating the fact that the tangent is periodic with period π.

The inverse function is

$$w = \int_0^z \frac{dt}{1+t^2} \tag{10.2–34}$$

defined throughout the disc $|z| \leqq 1$ except at $z = \pm i$. If z moves on the right half of the circumference we have

$$\begin{aligned} w &= \int_0^1 \frac{dt}{1+t^2} + i\int_0^\varphi \frac{e^{i\theta}}{1+e^{2i\theta}}\,d\theta = \tfrac{1}{4}\pi + \tfrac{1}{2}i\int_0^\varphi \frac{d\theta}{\cos\theta} \\ &= \tfrac{1}{4}\pi - \tfrac{1}{2}i \log\tan\left(\tfrac{1}{4}\pi - \tfrac{1}{2}\varphi\right). \end{aligned}$$

Thus we see that w describes the line Re $w = \frac{1}{4}\pi$. A similar expression is obtained if z describes the left half of the circumference.

If $z = iy$, then the integral (10.2–34) becomes

$$w = i\int_0^y \frac{dt}{1-t^2}$$

i.e., Re $w = 0$. Hence the right half of the unit circle is transformed onto the strip $0 <$ Re $w < \frac{1}{4}\pi$. Applying the symmetry principle to the right half of the circumference we see that (10.2.34) can be extended beyond this arc to the right half of the z-plane. A similar argument holds for the left half of the circumference. We conclude that the integral can be defined as a single valued function in the whole z-plane slit along two vertical half lines issuing from $+i$ and $-i$ respectively. This result is in accordance with that obtained in section 1.12.3.

10.2.9 – The parabola

We consider the mapping

$$w = z^2 \tag{10.2–35}$$

from another point of view than in section 10.2.2. Let us write $w = u+iv$, $z = x+iy$. Then

$$u = x^2-y^2, \qquad v = 2xy. \tag{10.2–36}$$

The image of the line $x = 1$ has the equation

$$v^2 = 4(1-u) \tag{10.2–37}$$

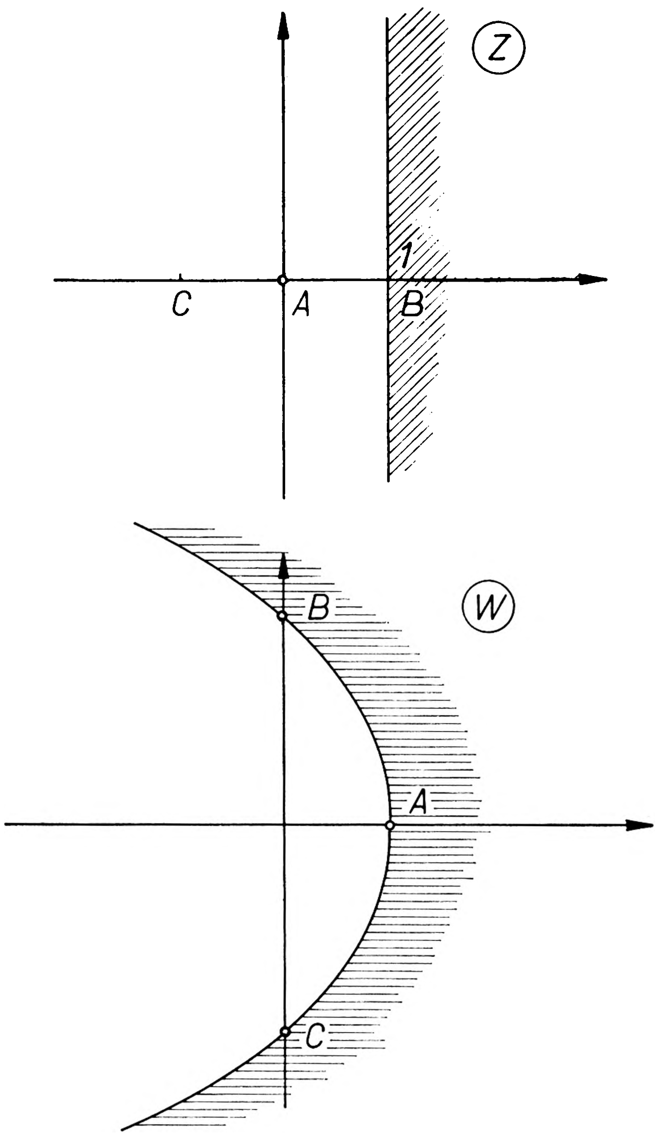


Fig. 10.2–15. The mapping of the region Re $z > 1$ as given by (10.2–35)

and is, therefore, a parabola whose focus is at the origin and whose vertex is the point $w = 1$, (fig. 10.2–15). If $x > 1$ then

$$v^2 = 4x^2y^2 > 4y^2 = 4(x^2-u) > 4(1-u)$$

and thus we see that the exterior of the parabola corresponds to $\text{Re}\, z > 1$. It follows that

$$w = (iz+1)^2 \tag{10.2–38}$$

maps the exterior of the parabola onto the upper half of the z-plane. To $z = 0, 1, -1$ correspond $w = 1, 2i, -2i$ respectively.

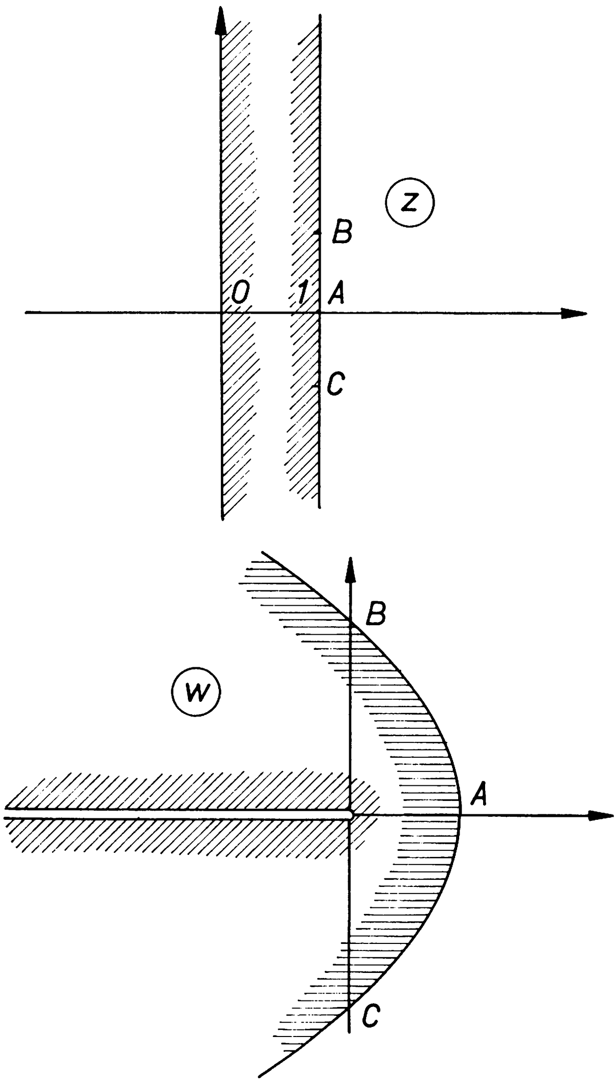


Fig. 10.2–16. The mapping of a vertical strip onto the interior of a parabolic slit along the negative real axis

For the interior of the parabola the problem is not quite so simple. Eliminating y from the two relations (10.2–36) yields

$$v^2 = 4x^2(x^2 - u) \qquad (10.2\text{–}39)$$

which represents a parabola which is in the interior of (10.2–37) if $0 < x < 1$. To $x = 0$ corresponds the negative real axis in the w-plane counted doubly. Hence the interior of the parabola slit along the negative real axis corresponds to the interior of the strip $0 < \operatorname{Re} z < 1$, (fig. 10.2–16). Next we consider the transformations

$$w_1 = \sqrt{w}, \qquad z = \exp \pi w_1,$$

which combine to

$$z = \exp \pi\sqrt{w}. \qquad (10.2\text{–}40)$$

The branch of the root is such that $\sqrt{1} = 1$. It is clear that (10.2–40) transforms the interior of the slit parabola into the upper half of the z-plane.

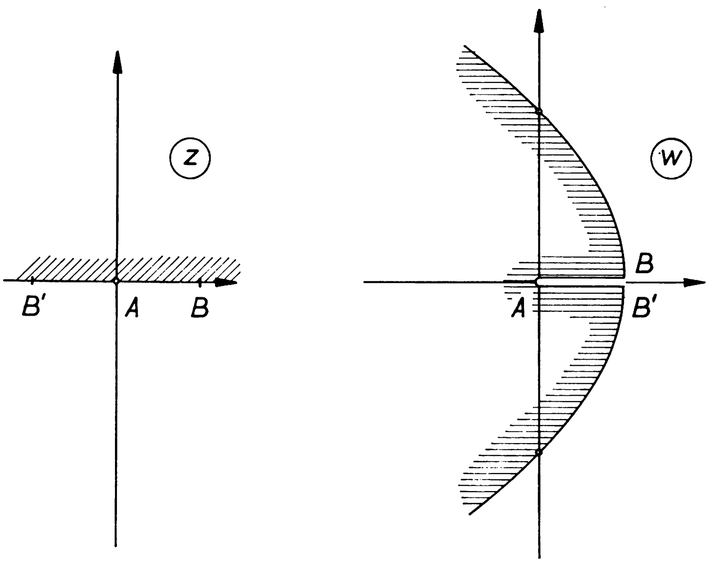


Fig. 10.2–17. The mapping as given by (10.2–41)

Starting again with the mapping (10.2–35) it is easily seen that the upper half of the interior of the parabola corresponds to the upper half of the strip $0 < \operatorname{Re} z < 1$, for the segment $0 < w < 1$ corresponds to $y = 0$, $0 < x < 1$. The left side of the half strip corresponds to the negative real axis in the w-plane. Reflecting with respect to this negative real axis we obtain a parabola slit from 0 to 1, (fig. 10.2–17), which corresponds to the semi strip $-1 < \operatorname{Re} z < 1$. It is now clear that the relation

$$z = \sin \tfrac{1}{2}\pi\sqrt{w} \tag{10.2–41}$$

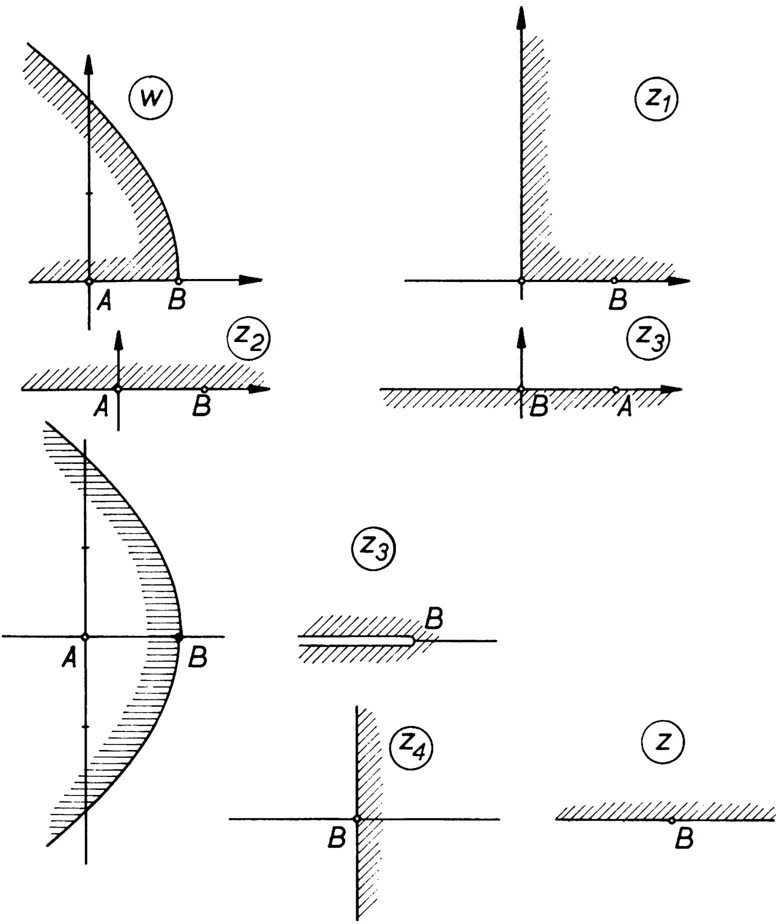


Fig. 10.2–18. The mapping as given by (10.2–43).

maps the parabolic region slit from the focus to the vertex onto the region $\operatorname{Im} z > 0$. In particular the line $v = 0$, $u < 0$ corresponds to

$$z = \tfrac{1}{2}i(e^{\frac{1}{2}\pi\sqrt{u}} - e^{-\frac{1}{2}\pi\sqrt{u}}) = i \sinh \tfrac{1}{2}\pi\sqrt{u},$$

that is to say the positive part of the imaginary axis. Writing

$$z_1 = \sin \tfrac{1}{2}\pi\sqrt{w} \tag{10.2–42}$$

it follows that the upper half of the parabolic region is transformed onto the first quadrant of the z_1-plane and this is carried into the upper half of the z_2-plane by means of $z_2 = z_1^2$. The relation

$$z_3 = 1-z_2 = 1-z_1^2 = \cos^2 \tfrac{1}{2}\pi\sqrt{w}$$

effects the mapping onto the lower half of the z_3-plane such that to the half ray $w < 1$ corresponds to $z_3 > 0$. This is at once clear if $0 \leqq w < 1$ and it follows from $z_3 = \cosh^2 \frac{1}{2}\pi\sqrt{w}$ if $w > 0$. On applying the symmetry principle we find that the whole interior of the parabola corresponds to the z-plane slit along the negative real axis. Hence $z_1 = \cos \frac{1}{2}\pi\sqrt{w}$ transforms this into the right half z_4-plane and

$$z = i \cos \tfrac{1}{2}\pi\sqrt{w} \tag{10.2–43}$$

gives the solution *of the problem of the mapping of the interior of the parabola* (10.2–37) *onto the upper half of the z-plane*, (fig. 10.2–18).

10.2–10 – The rectangle

A new aspect of the theory of conformal mapping is shown if we leave the field of elementary functions and turn our attention to the elliptic functions. We start with Legendre's integral

$$w = \int_0^z \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}, \qquad 0 < k < 1. \tag{10.2–44}$$

As usual z varies throughout $\operatorname{Im} z \geqq 0$. The integral is regular at each point of the real axis, except at $z = \pm 1$, $z = \pm 1/k$, where it is still continuous. We agree to take that branch of the square root which takes the value 1 at $z = 0$. This can be specified as follows. Let $\theta_1, \theta_2, \theta_3, \theta_4$ denote the principal arguments of $z+1/k$, $z+1$, $z-1$ and $z-1/k$ respectively, therefore these angles are between 0 and π (these values included). The argument of the denominator in (10.2–44) is chosen as

$$\theta = \tfrac{1}{2}(\theta_1 - \pi + \theta_2 - \pi + \theta_3 + \theta_4),$$

i.e., the argument of the integrand is $-\theta$. It follows that the argument of the integrand is 0 for z between -1 and $+1$, for then $\theta_1 = \theta_2 = 0$, $\theta_3 = \theta_4 = \pi$. The argument is $\frac{1}{2}\pi$ for z between 1 and $1/k$ and for z between $-1/k$ and -1. Finally the argument is π for $z > 1/k$ and $z < -1/k$.

If z moves from 0 to 1 the integral is a steadily increasing function and takes the value

$$\mathsf{K} = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} \tag{10.2–45}$$

(in accordance with (5.14–4). If z is between 1 and $1/k$ we may write

$$w = \mathrm{K}+i\int_1^z \frac{dt}{\sqrt{(t^2-1)(1-k^2t^2)}}. \tag{10.2–46}$$

It must be confessed that the integrand is not regular at $z = 1$, but the integral is continuous there and we may avoid the singularity by a small indentation. If we let its radius tend to zero we obtain (10.2–46). If $z \to 1/k$ the second integral becomes

$$\int_1^{1/k} \frac{dt}{\sqrt{(t^2-1)(1-k^2t^2)}}.$$

This integral can be thrown into a more elegant form by making the substitution

$$t = \frac{1}{\sqrt{(1-k'^2t'^2)}}$$

where

$$k' = \sqrt{1-k^2}.$$

Omitting afterwards the primes at the variable t' we find

$$\mathrm{K}' = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k'^2t^2)}}. \tag{10.2–47}$$

Hence to $z = 1/k$ corresponds $w = \mathrm{K}+i\mathrm{K}'$. Thereupon we let z vary between $1/k$ and ∞ and we get

$$w = \mathrm{K}+i\mathrm{K}'-\int_{1/k}^z \frac{dt}{\sqrt{(t^2-1)(k^2t^2-1)}}.$$

If we perform the substitution $t' = 1/kt$ the last integral becomes (omitting the primes)

$$\int_1^{1/kz} \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}$$

and making $z \to \infty$ it becomes $-\mathrm{K}$. Thus to $z = \infty$ corresponds $i\mathrm{K}'$. The other half of the real axis may be treated in a similar way and thus it appears that the real axis, percorsed from the left to the right, corresponds to the perimeter of a rectangle with vertices K, $\mathrm{K}+i\mathrm{K}'$, $-\mathrm{K}+i\mathrm{K}'$, $-\mathrm{K}$, percorsed in the counter clockwise sense. If w_0 is a point inside the rectangle the winding number is exactly $+1$, for the mapping of the real axis onto the perimeter is one-to-one. Hence, in view of the last theorem of section 10.1.1:

The function (10.2–44) *maps the upper half of the z-plane univalently onto the interior of a rectangle with sides* $2\mathrm{K}$ *and* K'.

As a consequence the function is invertible within the rectangle and on the sides. If $-1 < z < 1$ its inverse is the Jacobian sine (section 5.14.1 and 5.14.7), for the substitution $t = \sin\varphi$ leads to

$$w = \int_0^{\arcsin z} \frac{d\varphi}{\sqrt{1-k^2\sin^2\varphi}}.$$

On account of the identity principle (section 2.11.2) we, therefore, have

$$z = \operatorname{sn} w, \tag{10.2–48}$$

the modulus of this function being equal to k.

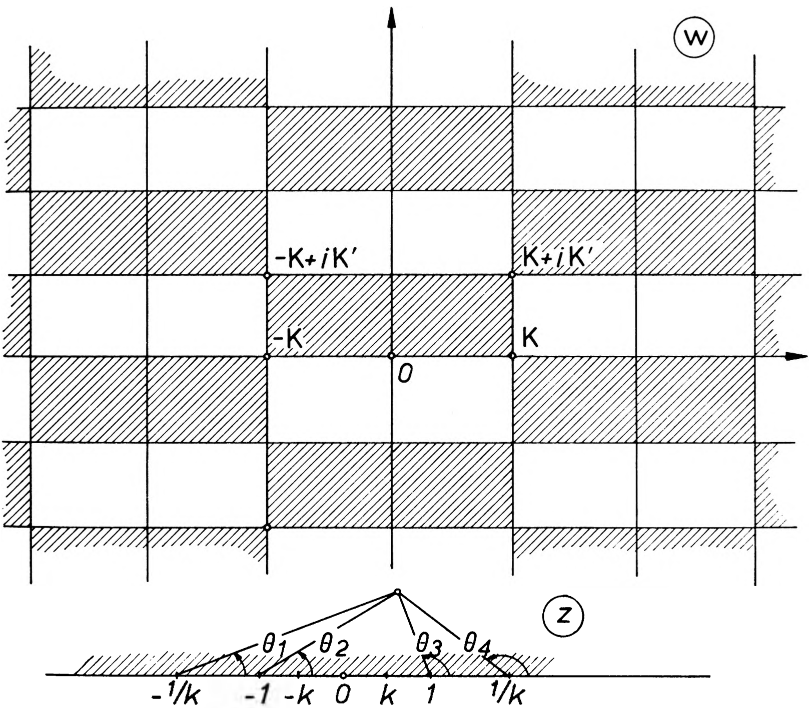


Fig. 10.2–19. The pattern associated with the mapping (10.2–44)

Applying the symmetry principle it is easily seen that this function is extensible throughout the whole w-plane and is everywhere holomorphic except at isolated poles, viz. at iK', mod $(2K, 2iK')$. The w-plane can be covered smoothly by shaded and unshaded rectangles of width $2K$ and height K', (fig. 10.2–19), corresponding to the upper and the lower half of the z-plane respectively. This illustrates the double periodicity of the Jacobian sine. Otherwise stated: the Jacobian sine is automorphic with respect to the group of translations generated by

$$w \to w+4\mathsf{K}, \qquad w \to w+2i\mathsf{K}'. \tag{10.2–49}$$

A fundamental domain is a period parallelogram containing two shaded and two unshaded rectangles.

Let us now assume that a rectangle is given, with the vertices a, $a+ib$, $-a+ib$ and $-a$, where a and b are positive. First we evaluate

$$\tau = i\frac{b}{a}, \qquad q = \exp \pi i\tau = \exp\left(-\pi\frac{b}{a}\right)$$

and with the theta functions belonging to this value of τ we find from (5.15–6) and (5.15–9)

$$k = \frac{\vartheta_2^2(0)}{\vartheta_3^2(0)}, \qquad \mathsf{K} = \tfrac{1}{2}\pi\vartheta_3^2(0), \tag{10.2–50}$$

or, which amounts to the same,

$$\mathsf{K} = \int_0^{\frac{1}{2}\pi} \frac{d\varphi}{\sqrt{1-k^2\sin^2\varphi}}.$$

Now the desired mapping is

$$z = \operatorname{sn}\frac{\mathsf{K}w}{a}, \tag{10.2–51}$$

k being the modulus of the Jacobian function.

For later applications we need to know the image of the line segment connecting the midpoints of the vertical sides of the rectangle mapped by (10.2–48), i.e., the segment connecting the points $\mathsf{K}+i\mathsf{K}'$ and $-\mathsf{K}+i\mathsf{K}'$. This segment is the set of points $u+\frac{1}{2}i\mathsf{K}'$, where u is real and between $-\mathsf{K}$ and $+\mathsf{K}$. The image is

$$z = \operatorname{sn}(u+\tfrac{1}{2}i\mathsf{K}')$$

and we contend that this image is a semicircle.

The proof of this assertion requires some computation. In the addition theorems (5.16–7) we put $a = b$ and we find

$$\operatorname{sn} 2a = \frac{2\operatorname{sn} a \operatorname{cn} a \operatorname{dn} a}{1-k^2\operatorname{sn}^4 a},$$

$$\operatorname{cn} 2a = \frac{\operatorname{cn}^2 a-\operatorname{sn}^2 a\operatorname{dn}^2 a}{1-k^2\operatorname{sn}^4 a},$$

$$\operatorname{dn} 2a = \frac{\operatorname{dn}^2 a-k^2\operatorname{sn}^2 a\operatorname{cn}^2 a}{1-k^2\operatorname{sn}^4 a}.$$

These equations are remarkable since it is possible to solve $\mathrm{sn}^2 a$, $\mathrm{cn}^2 a$ and $\mathrm{dn}^2 a$ rationally. We need only an expression for $\mathrm{sn}^2 a$. Some computation yields, taking into account (5.14–8) and (5.14–10),

$$\mathrm{sn}^2\, a = \frac{1-\mathrm{cn}\, 2\, a}{1+\mathrm{dn}\, 2\, a}.$$

In this equation we put $a = u+\frac{1}{2}i\mathrm{K}'$. Using (5.16–11) we easily find

$$\mathrm{sn}^2(u+\tfrac{1}{2}i\mathrm{K}') = \frac{1}{k}\,\frac{k\,\mathrm{sn}\, 2u+i\,\mathrm{dn}\, 2u}{\mathrm{sn}\, 2u-i\,\mathrm{cn}\, 2u}.$$

Referring again to (5.14–8) and (5.14–10) we see that the modulus of the numerator and that of the denominator in the second fraction are both unity provided u is real. Hence

$$|\mathrm{sn}\,(u+\tfrac{1}{2}i\mathrm{K}')\,| = \frac{1}{\sqrt{k}}, \tag{10.2–52}$$

that is to say: the image of the segment connecting the midpoints of the vertical sides is the arc of a circle with radius $1/\sqrt{k}$ above the real axis, (fig. 10.2–20).

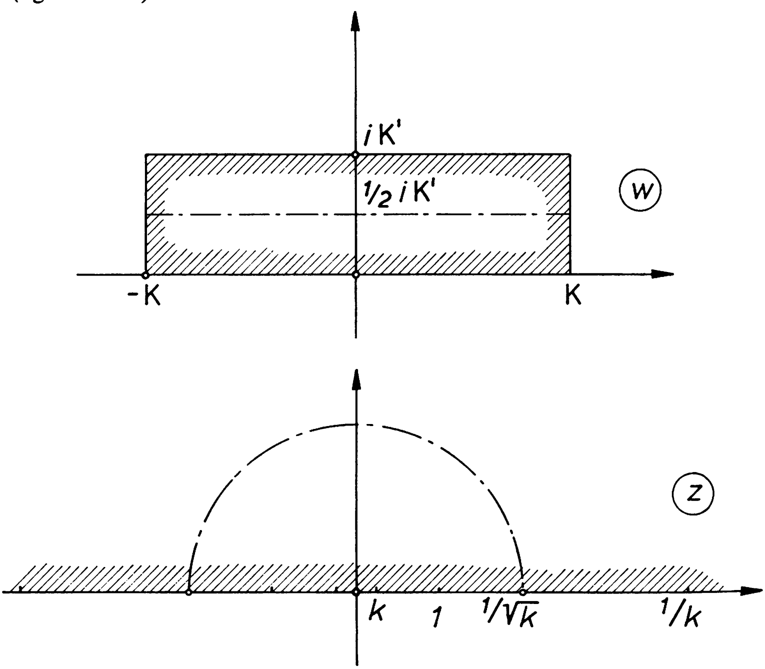


Fig. 10.2–20. The image of the midline of a rectangle which is mapped onto the upper half plane by means of (10.2–48)

10.2.11 – THE RECTANGLE (continued)

An alternative approach to the problem of mapping a rectangle is possible by means of Weierstrass's elliptic integral

$$w = \int_z^\infty \frac{dt}{\sqrt{4t^3 - g_2 t - g_3}}, \tag{10.2–53}$$

where g_2 and g_3 are real and such that the polynomial in the denominator has real roots e_1, e_2, e_3, ($e_3 < e_2 < e_1$). Again z varies through $\text{Im}\, z \geqq 0$ and it is easily verified that the integral is continuous at $z = \infty$. The path of integration may be any curve connecting z with a point on the real axis, combined with a half ray on the real axis issuing from this point. As we shall see presently it is of no importance whether this half ray tends to the right or to the left. It is necessary to select a branch of the square root occurring within the sign of integration. Let θ_1, θ_2 and θ_3 denote the arguments between 0 and π (these values included) of $z - e_3$, $z - e_2$ and $z - e_1$ respectively. We take the argument of the dominator as $\theta = \frac{1}{2}(\theta_1 + \theta_2 + \theta_3)$, i.e., that of the integrand as $-\theta$.

The further discussion will be facilitated by the following considerations. The integrand is holomorphic in the region $\text{Im}\, z > 0$ and regular at each point of the real axis, except at e_1, e_2, e_3 and ∞. At these points, however the integral is continuous. Let C denote a contour consisting of a semicircle $z = |R|$, $\text{Im}\, z > 0$, completed by the segment $-R \leqq z \leqq R$. Along this contour the integral

$$\int_C W(t)dt,$$

where

$$W(z) = \frac{1}{\sqrt{4(z-e_1)(z-e_2)(z-e_3)}}$$

is zero (possible singularities can be avoided by small indentations). On the other hand the integral tends to zero like $R^{-\frac{1}{2}}$ as $R \to \infty$. Hence the integral taken along the whole real axis is zero.

If z increases from e_1 to ∞ the argument of the denominator is 0 and

$$W(z) = \frac{1}{\sqrt{4(z-e_1)(z-e_2)(z-e_3)}}$$

where the square root is positive. Hence w decreases from

$$\omega = \int_{e_1}^\infty \frac{dt}{\sqrt{4t^3 - g_2 t - g_3}}$$

(compare (5.13–9)) to zero along a linear path.

If z is between $-\infty$ and e_3, the argument of the integrand is $\frac{3}{2}\pi$. Hence

$$w = \int_z^\infty W(t)dt = i\int_z^\infty \frac{dt}{\sqrt{4(e_1-t)(e_2-t)(e_3-t)}}$$
$$= -i\int_{-z}^{-\infty} \frac{dt}{\sqrt{4(t+e_1)(t+e_2)(t+e_3)}}$$
$$= -i\int_{-z}^{\infty} \frac{dt}{\sqrt{4(t+e_1)(t+e_2)(t+e_3)}},$$

since

$$\int_{-\infty}^{\infty} \frac{dt}{\sqrt{4(t+e_1)(t+e_2)(t+e_3)}} = 0.$$

If z tends to e_3 then w tends to

$$-i\int_{-e_3}^{\infty} \frac{dt}{\sqrt{4t^3-g_2t+g_3}} = -i\frac{\omega'}{i} = -\omega'.$$

Next we observe that

$$0 = \int_{-\infty}^{\infty} W(t)dt = \int_{-\infty}^{e_3} W(t)dt + \int_{e_3}^{e_2} W(t)dt + \int_{e_2}^{e_1} W(t)dt + \int_{e_1}^{\infty} W(t)dt.$$

We already know that

$$\omega = \int_{e_1}^{\infty} W(t)dt.$$

On the other hand

$$\int_{-\infty}^{e_3} W(t)dt = -i\int_{\infty}^{-e_3} \frac{dt}{\sqrt{4t^3-g_2t+g_3}} = i\int_{-e_3}^{\infty} \frac{dt}{\sqrt{4t^3-g_2t+g_3}} = \omega'.$$

Hence

$$\int_{e_3}^{e_2} W(t)dt + \int_{e_2}^{e_1} W(t)dt = -\omega-\omega'.$$

The first integral on the left is real, the second is purely imaginary. It follows that

$$\int_{e_2}^{e_3} W(t)dt = \omega, \qquad \int_{e_1}^{e_2} W(t)dt = \omega'. \tag{10.2–54}$$

Taking into account the fact that the argument of the first integral is $-\pi = \pi \pmod{2\pi}$ and that of the second integral is $-\frac{1}{2}\pi = \frac{3}{2}\pi \pmod{2\pi}$

we have the following expressions for the half periods ω and ω'

$$\boxed{\omega = \int_{e_3}^{e_2} \frac{dt}{\sqrt{4t^3-g_2t-g_3}},} \tag{10.2–55}$$

and

$$\boxed{\frac{\omega'}{i} = \int_{-e_1}^{-e_2} \frac{dt}{\sqrt{4t^3-g_2t+g_3}}.} \tag{10.2–56}$$

If z is between e_3 and e_2, then

$$w = \int_z^{\infty} W(t)dt = \int_z^{e_2} W(t)dt + \int_{e_2}^{e_1} W(t)dt + \int_{e_1}^{\infty} W(t)dt$$

$$= \int_z^{e_2} W(t)dt - \omega' + \omega$$

varies between $-\omega-\omega'+\omega = -\omega'$ and $-\omega'+\omega$ on a rectilinear path. Finally, if z is between e_2 and e_1 then

$$w = \int_z^{e_1} W(t)dt + \int_{e_1}^{\infty} W(t)dt = \int_z^{e_1} W(t)dt + \omega$$

varies between $-\omega'+\omega$ and ω on a rectilinear path.

By means of the same arguments as employed in the previous section we may deduce that the upper half of the z-plane corresponds univalently to the interior of the rectangle with vertices 0, $-\omega'$, $-\omega'+\omega$, ω (fig. 10.2–21).

The function (10.2−53) is invertible, its inverse is denoted by

$$z = \wp(w;\, g_2, g_3). \tag{10.2–57}$$

This function is defined throughout the closed rectangle mentioned above ($w = 0$ being a pole) and coincides with the Weierstrass pe function on the real axis. Hence it is identical with the Weierstrass pe function (compare section 5.13.1).

Assume that $\omega > 0$ and $\omega'/i > 0$ are given. Then $\tau = \omega'/\omega$ is known. From (5.10–2) and $e_1+e_2+e_3 = 0$ we deduce

$$2e_1+e_3 = \frac{\pi^2}{4\omega^2}\, \vartheta_4^4(0),$$

whilst

$$e_1-e_3 = \frac{\pi^2}{4\omega^2}\, \vartheta_3^4(0).$$

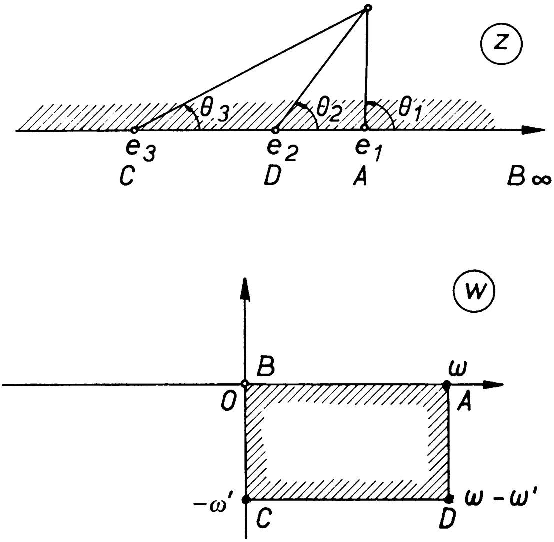


Fig. 10.2–21. The mapping (10.2–53).

Hence

$$e_1 = \frac{1}{3}\frac{\pi^2}{4\omega^2}(\vartheta_3^4(0)+\vartheta_4^4(0)).$$

Similarly

$$e_2 = \frac{1}{3}\frac{\pi^2}{4\omega^2}(\vartheta_2^4(0)-\vartheta_4^4(0))$$

and

$$e_3 = -\frac{1}{3}\frac{\pi^2}{4\omega^2}(\vartheta_2^4(0)+\vartheta_3^4(0)),$$

all theta functions belonging to the above value of τ.

An interesting particular case is presented by a *square*. Then $i\omega = \omega'$ and from (5.8–8) it follows

$$\tau = i, \qquad q = e^{-\pi}. \tag{10.2–58}$$

From Jacobi's imaginary transformation (section 5.10–3) we deduce $\vartheta_2(0) = \vartheta_4(0)$, whence $e_2 = 0$, $e_1 = -e_3$. We take ω such that $e_1 = 1$. The pe function is $\wp(w; 4, 0)$ now and from (10.2–55) follows

$$\omega = \int_{-1}^{0}\frac{dt}{2\sqrt{t(t^2-1)}} = \int_0^1 \frac{dt}{2\sqrt{t(1-t^2)}} = \frac{1}{4}\int_0^1 \frac{dt}{t^{\frac{3}{4}}(1-t)^{\frac{1}{2}}} = \tfrac{1}{4}B(\tfrac{1}{4}, \tfrac{1}{2}).$$

An easy computation yields, (sections 4.6.13 and 4.6.14),

$$\omega = \frac{1}{4\sqrt{2\pi}}\,\Gamma^2(\tfrac{1}{4}). \tag{10.2–59}$$

Summing up we have:

The function

$$z = \wp(\omega; 4, 0) \tag{10.2–60}$$

maps a square with side length (10.2–59) *onto the upper half of the z-plane.*

By a geometric reasoning based on the symmetry principle we may obtain other interesting results. We start again with (10.2–60), the inverse mapping being

$$w = \int_z^{\infty} \frac{dt}{\sqrt{4t(t^2-1)}}. \tag{10.2–61}$$

Now we observe that along the imaginary axis in the upper half of the z-plane the argument of the integrand has a constant value, viz. $-\frac{3}{4}\pi$. Hence w describes a straight line as z moves along this axis. Since $w(\infty) = 0$ and $w(0) = w(e_2) = \omega-\omega'$ the point w describes the diagonal joining the points 0 and $\omega-\omega'$, (fig. 10.2–22). It is now easily seen that the triangle with vertices ω, 0, $-\omega'$ $(= -i\omega)$ is mapped onto the first quadrant in the z-plane. It follows that

A triangle with angles $\frac{1}{4}\pi$, $\frac{1}{2}\pi$, $\frac{1}{4}\pi$ can be mapped univalently onto the upper half of the z-plane by means of the function

$$z = \wp^2(w; 4, 0). \tag{10.2–62}$$

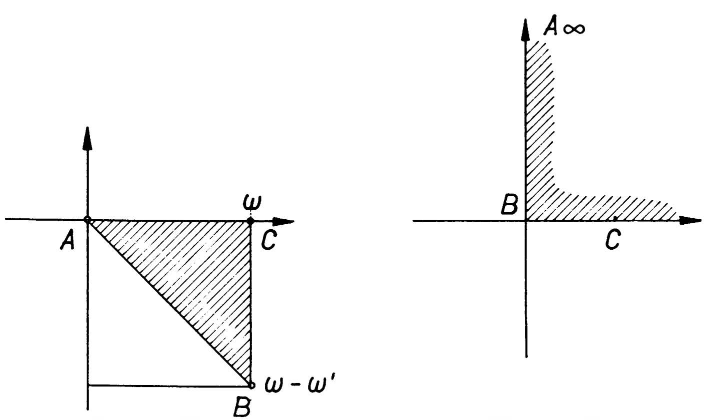


Fig. 10.2–22. The mapping of a triangle onto a quadrant by means of (10.2–60)

The image of the diagonal connecting ω and $-\omega'$, which is also a line of symmetry for the square, is readily found if we apply the formula (5.5–5). Observing that this diagonal is characterized by

$$w = (1-t)\omega - it\omega = \omega - t(\omega + i\omega), \qquad 0 \leqq t \leqq 1,$$

(fig. 10.2–23) remembering the fact that the pe function is even, we have

$$\wp(w) = 1 + \frac{2}{\wp(t(\omega+i\omega))-1} = \frac{\wp(t(\omega+i\omega))+1}{\wp(t(\omega+i\omega))-1}.$$

The line $w = t(w+i\omega)$, however, is obtained by reflecting the diagonal from 0 to $\omega - i\omega$ in the real axis and on this diagonal the pe function takes purely imaginary values. Hence $\wp(t(\omega+i\omega))$ is purely imaginary and it follows that $|\wp(w)| = 1$, if w describes the diagonal from ω to $-i\omega$.

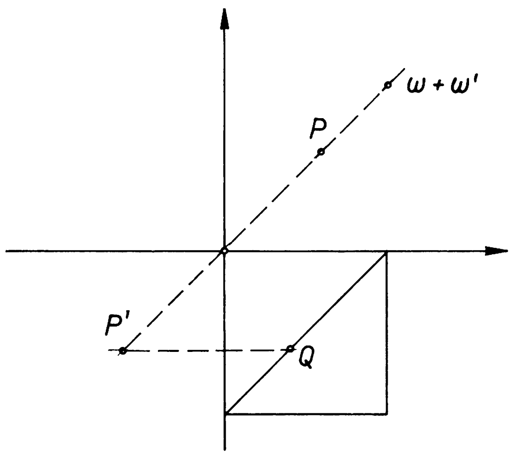


Fig. 10.2–23. $P = t(\omega+\omega')$, $P' = -t(\omega+\omega')$, $Q = \omega - t(\omega+\omega')$

The image is a semicircle in the z-plane. The interior of the triangle ω, 0, $-i\omega$ is transformed into the upper half of the z-plane outside the semicircle and it is clear that $1/\wp(w; 4, 0)$ maps this triangle onto the lower half of the unit circle. Taking the square root we find a correspondence between the triangle and a quadrant of a circle. Applying again Schwarz's symmetry principle several times we find

The relation

$$z^2 = \frac{1}{\wp(w; 4, 0)} \qquad (10.2\text{–}63)$$

defines a univalent mapping of a square with diagonal length

$$2\omega = \frac{1}{2\sqrt{2\pi}} \Gamma^2(\tfrac{1}{4}) \tag{10.2–64}$$

onto the interior of the unit circle in the w-plane.

The vertices of the square are ω, $i\omega$, $-\omega$, $-i\omega$, (fig. 10.2–24).

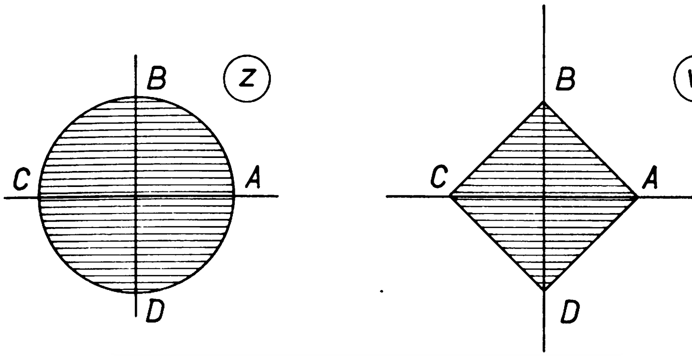


Fig. 10.2–24. The mapping of a square onto a circular disc by means of (10.2–63)

In all these examples the figures in the w-plane are specially situated. But by the movement $w \to c_1 w + c_2$, $|c_1| = 1$, we can carry them into any other situation.

10.2.12 – The interior of an ellipse

In the w-plane we consider an ellipse with semi axis a and b. We can represent the ellipse by means of the equation

$$\frac{u^2}{a^2} + \frac{v^2}{b^2} = 1. \tag{10.2–65}$$

For the sake of convenience we assume that the foci are situated at the ponts ± 1, i.e.,

$$a^2 - b^2 = 1, \qquad a > 1. \tag{10.2–66}$$

We contend that the mapping

$$w = \sin z_1 \tag{10.2–67}$$

transforms the upper half of the interior of the ellipse into a rectangle in the z_1-plane. If $z_1 = x_1 + iy_1$ then

$$\begin{aligned} u &= \sin x_1 \cosh y_1, \\ v &= \cos x_1 \sinh y_1 \end{aligned} \tag{10.2–68}$$

and if $y_1 = h$ is a fixed number, then these equations represent the ellipse

(10.2–65) in parameter form, provided h satisfies the relations

$$\cosh h = a, \qquad \sinh h = b,$$

that is to say

$$\tanh h = \frac{b}{a},$$

or

$$h = \tfrac{1}{2} \log \frac{a+b}{a-b} = \log(a+b),$$

taking into account (10.2–66).

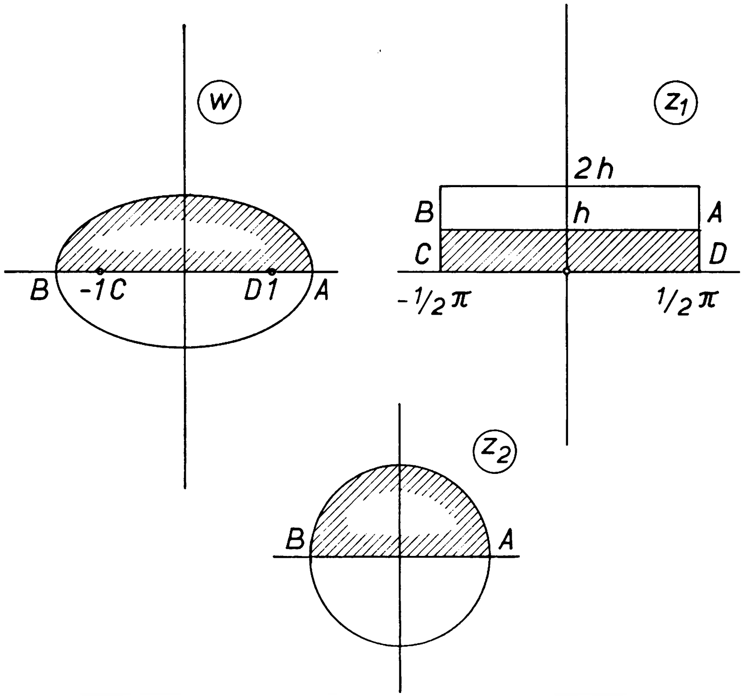


Fig. 10.2–25. The mapping of the interior of an ellipse onto a circular disc

From (10.2–68) we easily deduce that the segment $-1 \leqq w \leqq 1$ corresponds to the segment $-\frac{1}{2}\pi \leqq x_1 \leqq \frac{1}{2}\pi$, $y_1 = 0$, the segment $1 \leqq w \leqq a$ to $x_1 = \frac{1}{2}\pi$, $0 \leqq y_1 \leqq h$, the segment $-a \leqq w \leqq -1$ to $x_1 = -\frac{1}{2}\pi$, $0 \leqq y_1 \leqq h$ and the elliptical arc to the segment $y_1 = h$, $-\frac{1}{2}\pi \leqq x_1 \leqq \frac{1}{2}\pi$, (fig. 10.2–25).

It is necessary to employ a trick now. We do not map this rectangle, for then afterwards the symmetry principle does not yield the whole interior of the ellipse (but that of a slit region as may be seen afterwards). Instead we map a rectangle of the same width, its height being however, doubled, i.e. $2h$. As in section 10.2.10 we define

$$\tau = \frac{2hi}{\pi/2} = \frac{4hi}{\pi} = \frac{4i}{\pi}\log(a+b), \qquad q = \frac{1}{(a+b)^4}. \tag{10.2–69}$$

Then k and K of the corresponding Jacobian sine can be evaluated and the desired mapping is

$$z_2 = \operatorname{sn}\frac{2\mathsf{K}}{\pi}z_1. \tag{10.2–70}$$

In section 10.2.10 we proved that the segment connecting the points $\frac{1}{2}\pi+ih$ and $-\frac{1}{2}\pi+ih$ is transformed into the arc of a semicircle with radius $1/\sqrt{k}$. It is easily verified that the w-plane is mapped onto the z_2-plane in such a way that the rectilinear part of the semi-ellipse corresponds to the horizontal diameter of the circle, whilst the elliptical arc is carried into the circular arc. Applying the symmetry principle we find that the elliptical region is transformed into the interior of the circle with radius $1/\sqrt{k}$, and that finally this is changed into the unit circle by means of $z = z_2\sqrt{k}$. Thus

The elliptical region bounded by the ellipse

$$\frac{u^2}{a^2}+\frac{v^2}{b^2} = 1, \qquad a^2-b^2 = 1,$$

in the w-plane, is mapped onto the region $|z| < 1$ *in the z-plane by means of the transformation*

$$z = \sqrt{k}\operatorname{sn}\left(\frac{2\mathsf{K}}{\pi}\arcsin w\right), \tag{10.2–71}$$

where the modulus and the quarter period are determined by (10.2–69).

10.2.13 – The annulus

The method used for the mapping of the interior of an ellipse can also be employed to map an annulus bounded by two concentric circles. This is, however, not a simply connected region and therefore we cannot expect that an image is the interior of a circle. As we shall see it is possible to map the annulus onto the interior of a unit circle slit along a segment, hence also a doubly connected region.

In the w-plane we consider an annulus bounded by two circles around the origin with radii R_1, R_2, $(R_1 < R_2)$. By means of

$$z_1 = \log\frac{w}{\sqrt{R_1R_2}} \tag{10.2–72}$$

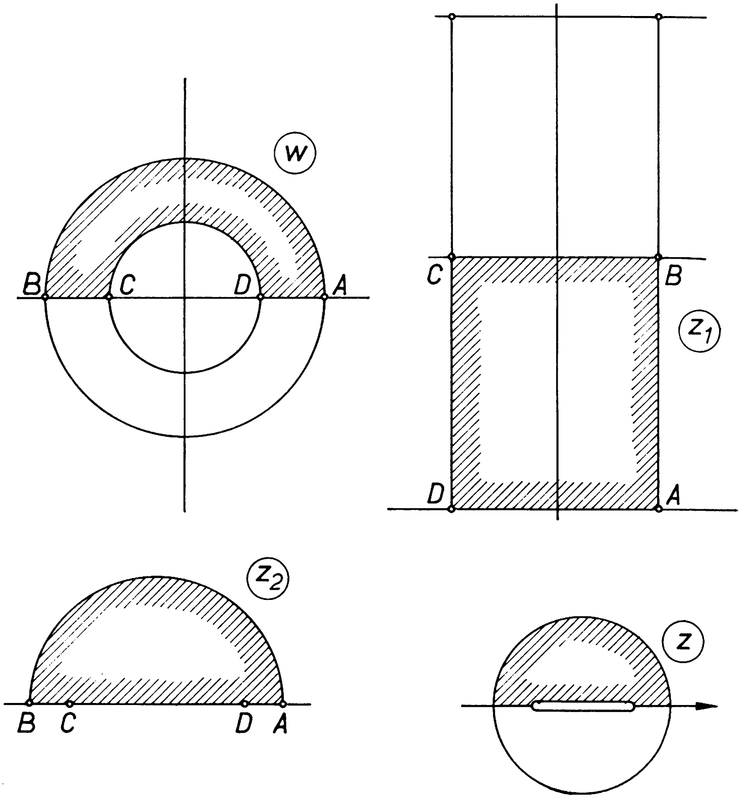


Fig. 10.2–26. The mapping of an annulus onto a slitted disc

the upper half of the annulus is transformed into a rectangle in the z_1-plane with height π and vertices at $\frac{1}{2}\log(R_2/R_1)$ and $-\frac{1}{2}\log(R_2/R_1)$, (fig. 10.2–26). The semicircles correspond to the left and the right vertical sides respectively. Next we determine τ from

$$\tau = \frac{4\pi i}{\log(R_2/R_1)} \tag{10.2–73}$$

in order to map the rectangle with double height. This rectangle is mapped onto the upper half of the z_2 plane by means of

$$z_2 = \operatorname{sn} \frac{2K}{\log(R_2/R_1)} z_1 \tag{10.2–74}$$

and we see that the upper half of the annulus corresponds to the interior of a semicircle with radius $1/\sqrt{k}$ as in the previous section. The rectilinear parts of the boundary of the semiannulus correspond to the segments between 1 and $1/\sqrt{k}$, and $-1/\sqrt{k}$ and -1 respectively. The largest arc corresponds with the semicircle in the z_2-plane and the smallest arc with the segment between 1 and $+1$. Applying the symmetry principle we find that the whole annulus corresponds to the circle with radius $1/\sqrt{k}$ slit from -1 to $+1$. Performing the final transformation $z = z_2\sqrt{k}$ we have

An annulus bounded by the circles $|w| = R_1$, $w = |R_2|$, $R_1 < R_2$ in the w-plane can be mapped by means of

$$z = \sqrt{k}\,\mathrm{sn}\left(\frac{2\mathrm{K}}{\log (R_2/R_1)} \log \frac{w}{\sqrt{R_2/R_1}}\right) \tag{10.2–75}$$

onto the interior of the unit circle in the z-plane cut along a horizontal segment $-\sqrt{k} \leqq x \leqq \sqrt{k}$. Here k is determined by (10.2–73).

It should be noticed that the length of the slit is determined by k, i.e., by the ratio of R_2 and R_1. It can be proved that it is not possible to modify the mapping in such a way that the slit takes an arbitrary length. Hence two annuli are then and only then conformally equivalent if the ratio of their radii is the same.

If we had taken

$$\tau = \frac{2\pi i}{\log (R_2/R_1)} \tag{10.2–76}$$

a simple reasoning would have led to the result

An annulus bounded by the circles $|w| = R_1$, $|w| = R_2$, $R_1 < R_2$, in the w-plane can be mapped by means of

$$z = \mathrm{sn}\left(\frac{2\mathrm{K}}{\log R_2/R_1} \log \frac{w}{\sqrt{R_1 R_2}}\right) \tag{10.2–77}$$

onto the whole z-plane slit along the segments $1 \leqq x \leqq 1/k$ and $-1/k \leqq x \leqq -1$, where k is determined by (10.2–76) *now*, (fig. 10.2–27).

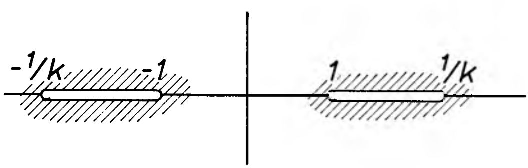


Fig. 10.2–27. The image of an annulus as given by (10.2–77)

Finally we mention that *the above annulus can be mapped by means of*

$$z = \wp\left(\log \frac{w}{R_1}\right), \tag{10.2–78}$$

where the pe-function is determined by $\omega = \log(R_2/R_1)$ *and* $\omega' = \pi i$, *onto the whole z-plane slit along the half line* $x \leqq e_3$ *and the segment* $e_2 \leqq x \leqq e_1$, (fig. 10.2–28), where e_1, e_2 and e_3 have the usual meaning. We leave the details to the reader.

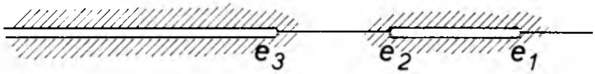


Fig. 10.2–28. The image of an annulus as given by (10.2–78)

We conclude this section by establishing the following theorem: *There does not exist a function* $f(z)$ *which provides a bicontinuous map of the annulus* $r_1 \leqq |z| \leqq 1$ *onto the annulus* $r_2 \leqq |z| \leqq 1$, $r_1 \neq r_2$, *in such a way that* $f(z)$ *is univalent on* $r_1 < |z| < 1$.

Bicontinuous means that $f(z)$ is continuous and has a continuous inverse. It is understood, of course, that $f(z)$ is holomorphic in the interior of the annular region.

If such a function exists we can extend it, by symmetry to a region $0 < |z|$. The extension is called $f(z)$ again. It has no singularity at $z = 0$ and, evidently, $f(0) = 0$. Similarly $f(z)$ has a pole at $z = \infty$. Hence $f(z)$ is an automorphism of the extended plane and, therefore, has the form

$$\frac{az+b}{cz+d}.$$

Since $f(0) = 0$, $f(\infty) = \infty$, $|f(z)| = 1$ for $|z| = 1$ it is easily seen that $f(z) = e^{i\theta}z$. It follows that $|f(z)| = r_1$ for $|z| = r_1$, in contradiction with the fact that by assumption $|f(z)| = r_2 \neq r_1$.

10.3 – The mapping of a polygon

10.3.1 – Jordan's theorem for a polygon

Generally speaking the problem of finding a function which transforms a given simply connected region into the interior of the unit circle (or the upper half of the complex plane) is very difficult, for in advance we do not know sufficiently about the special properties of the function. In most cases we cannot find a useful analytic expression and we are obliged to use numerical methods. In the case of a polygon, however, a systematic theory exists and we can write down an explicit formula solving the mapping problem.

First we wish to prove a theorem about closed polygons which is an elementary counterpart of Jordan's theorem, referred to in section 1.2.5. It is not our aim to expose all mathematical details which are necessary if we would attain satisfactory rigour, but we prefer to throw light on the main points which are important for the adaptation of geometric reasoning to obtain a solid understanding of basic function theoretic aspects.

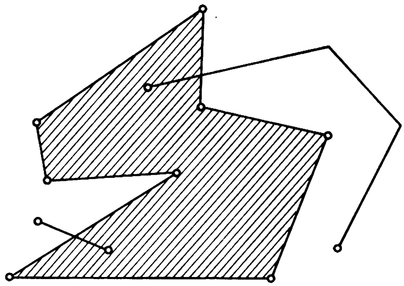


Fig. 10.3–1. Jordan's theorem for a polygon

A *closed polygon* consists of n finite points $p_1, \ldots, p_n$, the *vertices*, and the connecting line segments p_1p_2, $p_2p_3, \ldots$, p_np_1, the *sides*. If we omit the last segment the polygon is called a *polygonal arc* connecting the points p_1 and p_n. A closed polygon is called *simple* if all vertices are different, no vertex is on a side and no two sides intersect. It is our intention to prove:

A closed polygon divides the (open) plane into two parts such that 1) *each point which does not lie on the polygon belongs to precisely one part,* 2) *it is impossible to connect a point of one part with a point of the other part by a polygonal arc which does not meet the given polygon*; 3) *the end points of a segment which cuts one side of the polygon belong to different parts*, (fig. 10.3–1).

Let p denote a point not on the polygon and consider a half line issuing from p. We define the number of intersections of this half ray in the following way. If the ray crosses a side between the vertices it is counted as an intersection. If the ray passes through a vertex p_k it is counted as an intersection if the adjacent sides are on different sides of the ray. If, finally, one or more successive sides are on the ray, then it counts as an intersection, provided the adjacent sides are on different sides of the ray.

Let now L_1 and L_2 be two half rays issuing from p. We assert that the sum of the numbers of intersections of these rays with the polygon is always even. In order to prove this we divide the sides of the polygon in parts, such that each part meets only one of the half rays L_1 or L_2. The end points of these parts (that are the original vertices and possible newly introduced points) are denoted in order by $q_1, \ldots, q_m$. To each point q_i we relate a number κ_i having the value 0 or 1 according as q_i is inside or outside the angle formed by L_1 and L_2. If q_i is on one of the legs of the angle we put $q_i = q_{i-1}$ (it is understood that the enumeration is such that q_1 is not on a leg). The difference $\kappa_{i+1} - \kappa_i$ is always zero, except when between q_i and q_{i+1} or at q_i or along the side $q_{i-1}q_i$ an intersection occurs. Then $\kappa_{i+1} - \kappa_i = \pm 1$. The number of terms in the identity

$$(\kappa_2 - \kappa_1) + (\kappa_3 - \kappa_2) + \ldots + (\kappa_n - \kappa_{n-1}) + (\kappa_1 - \kappa_n) = 0$$

equal to 1 is the same as the number of those equal to -1. Hence the number of intersections is even.

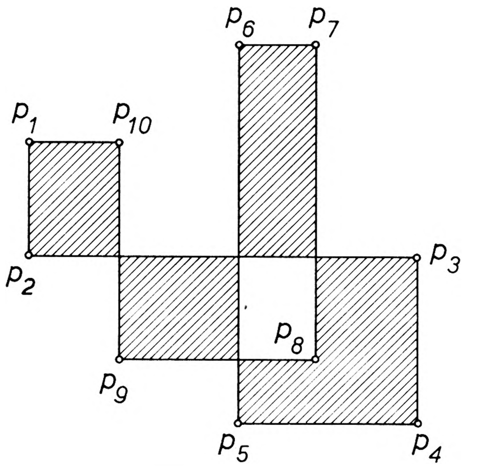


Fig. 10.3–2. The interior and the exterior points of a polygon (schematic)

Now we define: If the number of intersections of the half ray issuing from p (not on the polygon) is odd, then we call p a point of the first kind. If, however, this number is even, then p is called a point of the second kind. As pointed out above this distinction is independent of the choice of the half ray. Thus we proved 1).

Let now p and q denote two points not on the polygon such that the segment pq does not meet the polygon. The half ray from p through q can be divided into the segment pq and its extension issuing from q. It is clear that both half rays have the same number of intersections, i.e., p and q are of the same kind. This proves 2). By the same argument we see that the end points of a segment pq which crosses exactly one side of the polygon are of different kind. Thus we proved 3). It is common practice to call the points of the first kind the *internal* points of the polygon and those of the second kind the *external* points. In fig. 10.3–2 the set of internal points of the polygon is shaded. In addition we have the lemma:

If p and q are end points of a segment which crosses just one side of a closed polygon L between its vertices, then the winding numbers of L with respect to p and q differ by unity, i.e.,

$$\Omega_L(p) - \Omega_L(q) = \pm 1. \tag{10.3–1}$$

Let $p_k p_{k+1}$ be the side of the polygon crossed by the segment. We assume that L is percorsed in a certain sense, e.g., in the sense $p_1, \ldots, p_n$. There is a square Q with vertices p_k, p_{k+1}, p_0, q_0 such that if Q is percorsed in the appropriate sense the orientation induced by $p_k p_{k+1}$ is opposite to that induced by L, (fig. 10.3.–3). Hence in the chain $L_0 = L + Q$ this

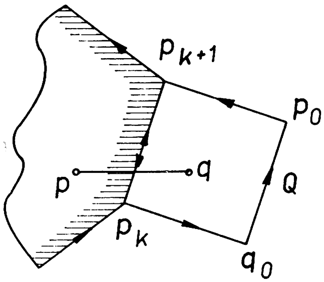


Fig. 10.3–3. Proof of (10.3–1)

segment is missing (for it is percorsed in two opposite ways). Since the winding number depends continuously on the reference point we may assume that at least one of the points p or q, let us say q, is inside Q. Then the other is outside and we have

$$\Omega_L(q) = \Omega_{L_0-Q}(q) = \Omega_{L_0}(q) - \Omega_Q(q) = \Omega_{L_0}(q) + \varepsilon, \quad \varepsilon = \pm 1,$$
$$\Omega_L(p) = \Omega_{L_0-Q}(p) = \Omega_{L_0}(p) - \Omega_Q(p) = \Omega_{L_0}(p).$$

Since the segment pq does not meet L_0 we have $\Omega_{L_0}(p) = \Omega_{L_0}(q)$. Hence (10.3–1) is true.

A consequence is

The winding number of a closed polygon with respect to an internal point is odd, that with respect to an external point is even.

This is a direct consequence of the above lemma and of the fact that the winding number depends continuously on the reference point.

10.3.2 – The mapping of the interior of a polygon

We are sufficiently prepared now to establish the following basic theorem:

Let $w(z)$ be holomorphic throughout $|z| < 1$ and univalent and continuous on $|z| = 1$. Assume that the image of this circumference is a closed polygon. Then $w(z)$ is univalent throughout the disc.

Since $w(z)$ is univalent on $|z| = 1$ the image is a simple closed polygon. Let w_0 denote a point on a side between two vertices corresponding to z_0 on the circumference and C_0 a sufficiently small circle around w_0 (such that it intersects the polygon in two points only on the side through w_0), (fig. 10.3–4); this side divides C_0 into two parts.

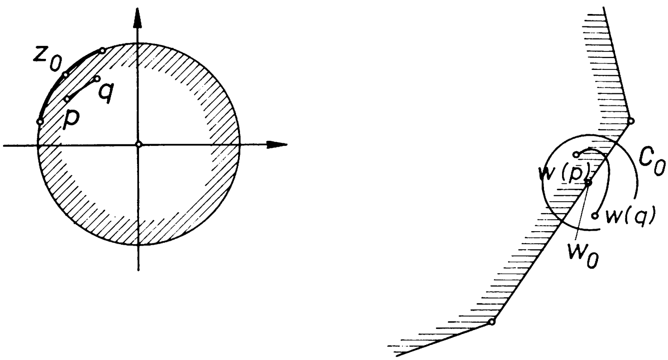


Fig. 10.3–4. Univalence of the mapping function

If $\delta > 0$ is chosen suitably the image $w(z)$ of a point z with $|z-z_0| < \delta$ is inside C_0. All these images belong to the same part, for if $|p-z_0| < \delta$, $|q-z_0| < \delta$, $|p| < 1$ and $|q| < 1$ and $w(p)$ and $w(q)$ are in different parts, then the image of a curve connecting p and q which lies in $|z| < 1$ is a curve connecting $w(p)$ and $w(q)$, which lies inside C_0, This curve, however, meets the common boundary of the parts, in contradiction with the fact that the mapping is univalent on $|z| = 1$. Hence the points of one part are not images of points inside $|z| = 1$, and, consequently, the winding number of the polygon with respect to these points is zero, as follows from the second theorem of section 10.1.1. In view of the lemma of the previous section the winding number with respect to any point of the other part is ± 1. Since the image of $|z| < 1$ is connected, the polygon has the same winding number with respect to every point of this image, this number being either $+1$ or -1. If the circumference $|z| = 1$ is percorsed in the counter clockwise sence this number cannot be -1, for $w(z)$ has no poles inside

$|z| = 1$. Referring again to section 10.1.1 we may conclude that the function $w(z) - w_1$ where $w_1 = w(z_1)$, $|z_1| < 1$, has exactly one zero, i.e., the function $w(z)$ takes the value w_1 just once.

It is now clear that the polygon L, the image of $|z| = 1$, divides the plane in exactly two regions. The region of the internal points is characterized by the fact, that the winding number of L with respect to each of its points is unity. They are the images of the internal points of $|z| \leq 1$. The region of external points is the set of all points with respect to which the winding number of L is 0.

Applying a linear fractional transformation we may replace the disc $|z| \leq 1$ by the closed half plane $\operatorname{Im} z \geqq 0$ (with the point ∞ added); the circumference of the circle corresponds to the real axis. Thus we have

If a function $w(z)$ is holomorphic in the half plane $\operatorname{Im} z > 0$ *and continuous and univalent on the real axis, and if it transforms this axis into a closed polygon, then this function transforms the open half plane univalently into a region bounded by the polygon and the closed half plane into the closure of the polygon.*

10.3.3 – The Schwarz-Christoffel formula

It is to be expected that an appropriate generalization of the formulas discussed in the sections 10.2.10 and 10.2.11 will provide us with a formula which effects the mapping of the upper half of the z-plane onto a region bounded by a polygon. Let $a_1, \ldots, a_{n-1}$ be finite real numbers such that

$$a_1 < a_2 < \ldots < a_{n-1}, n > 1.$$

We consider the integral

$$\boxed{w = \int_{z_0}^{z} \frac{dt}{(t-a_1)^{\lambda_1} \ldots (t-a_{n-1})^{\lambda_{n-1}}},} \quad \operatorname{Im} z \geqq 0, \operatorname{Im} z_0 \geqq 0 \tag{10.3–2}$$

where $\lambda_1, \ldots, \lambda_{n-1}$ are real numbers such that the integral is continuous at $a_1, \ldots, a_{n-1}$. This requires $\lambda_1 > -1, \ldots, \lambda_{n-1} > -1$. To investigate the behaviour at $z = \infty$ we perform the substitution $t = 1/t'$ and after omitting the primes we find

$$w = -\int_{1/z_0}^{1/z} \frac{dt}{t^{2-(\lambda_1+,,,+\lambda_{n-1})}(1-a_1 t)^{\lambda_1} \ldots (1-a_{n-1} t)^{\lambda_{n-1}}}.$$

Hence the integral is also continuous at $z = \infty$ if $\lambda_n = 2 - (\lambda_1 + \ldots + \lambda_{n-1}) > -1$. On geometrical grounds it will be necessary to suppose $\lambda_1 < 1, \ldots, \lambda_n < 1$ as we shall see presently. Hence to (10.3–2) are associated n real numbers $\lambda_1, \ldots, \lambda_n$, such that

$$-1 < \lambda_1 < 1, \ldots, -1 < \lambda_n < 1, \tag{10.3–3}$$

and

$$\lambda_1 + \ldots + \lambda_n = 2. \tag{10.3–4}$$

Without further comment the integral (10.3–2) is not uniquely defined; we must make an agreement about the argument of the integrand. We assume $\operatorname{Im} z \geqq 0$ and we take $\theta_k = \arg(z - a_k)$, $k = 1, \ldots, n-1$, such that $0 \leqq \theta_k \leqq \pi$. The number

$$\theta = \lambda_1\theta_1 + \ldots + \lambda_{n-1}\theta_{n-1}$$

will be taken as the argument of the denominator of the integrand, i.e., $-\theta$ is the argument of the integrand itself. The integrand is holomorphic throughout $\operatorname{Im} z > 0$ and continuous at every point of the real axis: z_0 is an arbitrary point of $\operatorname{Im} z \geqq 0$ and the path of integration can be taken at will. For the sake of convenience we denote the integrand in (10.3–2) by $W(t)$. Let x' denote a point between a_{k-1} and a_k, ($a_{k-1} = -\infty$ if $k = 1$), and x'' a point between a_k and a_{k+1}, ($a_{k+1} = \infty$ if $k = n-1$). Let the argument of $W(x')$ be φ. Then the argument of $W(x'')$ is $\varphi + \lambda_k\pi$. Hence

$$w(a_k) - w(x') = e^{i\varphi} \int_{x'}^{a_k} |W(t)|\,dt.$$

Thus we see that $w(z)$ describes a linear segment as long as x' is between a_{k-1} and a_k. Further

$$w(x'') - w(a_k) = e^{i(\varphi + \lambda_k\pi)} \int_{a_k}^{x''} |W(t)|\,dt$$

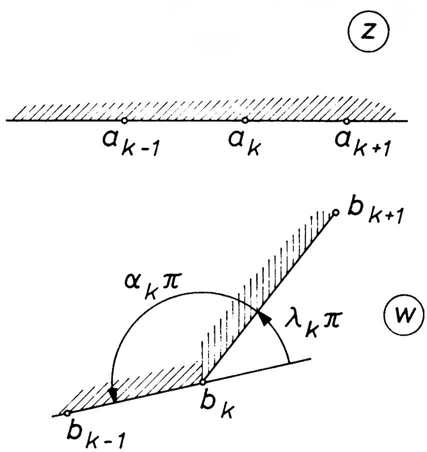


Fig. 10.3–5. Behaviour of the integral (10.3–2) at a point a_k

and $w(z)$ also describes a linear segment as long as x'' describes the segment $a_k a_{k+1}$. It is clear that the angle between the vectors $w(a_k) - w(a_{k-1})$ and $w(a_{k+1}) - w(a_k)$ is $\lambda_k \pi$, (fig. 10.3–5). We can also say that the argument of the integral increases by $\lambda_k \pi$ if z passes through the point a_k.

We introduce the numbers

$$b_1 = w(a_1), \ldots, b_{n-1} = w(a_{n-1}), b_n = w(\infty).$$

The preceding discussion leads to the result that the real axis is transformed into a closed polygon. If z moves from $-\infty$ to $+\infty$ the argument of the integral increases by $(\lambda_1 + \ldots + \lambda_{n-1})\pi = (2 - \lambda_n)\pi = -\lambda_n \pi$ (mod 2π). Hence the angle between $b_{n-1} b_n$ and $b_n b_1$ is $\lambda_n \pi$. These angles are called the *exterior angles* of the polygon. By the *interior angles* we understand the numbers

$$\alpha_1 \pi = (1 - \lambda_1)\pi, \ldots, \alpha_n \pi = (1 - \lambda_n)\pi \tag{10.3–5}$$

and, evidently,

$$\alpha_1 + \ldots + \alpha_n = n - 2. \tag{10.3–6}$$

Now we make the additional assumption that this polygon has no multiple points. Then we may apply the second theorem of the previous section. Hence

The integral (10.3–2) *transforms the region* $\operatorname{Im} z > 0$ *into a region bounded by a polygon with angles* (10.3–5), *provided the exponents in the denominator are subject to the conditions* (10.3–3) *and* (10.3–4) *and the integral is univalent for real values of* z.

The formula (10.3–2) is called the *Schwarz-Christoffel formula.* It should be noticed that it is, in general, rather difficult to tell from this formula whether it transforms the real axis into a polygon without multiple points. In the case of a triangle there is no difficulty.

The length of a side $b_k\, b_{k+1}$, may be evaluated from

$$l_k = |w(a_{k+1}) - w(a_k)| = \left| \int_{a_k}^{a_{k+1}} W(t)dt \right|. \tag{10.3–7}$$

It may be happen that $\lambda_n = 0$. Then the angle at b_n is π and the image is an $(n-1)$-gon.

Given an n-gon with vertices $b_1, \ldots, b_n$ the numbers $\lambda_1, \ldots, \lambda_n$ are known and we can write down the integral (10.3–7) with undeterminate values of $a_1, \ldots, a_{n-1}$. Since also $c_1 w + c_2$ is a mapping onto a polygon we can take e.g. $a_1 = 0$, $a_2 = 1$ and evaluate c_1 and c_2 such that b_1 corresponds to a_1 and b_2 to a_2. There remain $n-3$ values of a_i which correspond to b_i, $(i = 3, \ldots, n-1)$. The evaluation of these constants is an extremely difficult problem and a general solution is not known. In the particular case of a rectangle we are able to solve this problem by means of the Jacobian thetafunctions as we have seen in the examples of section 10.2.10 and 10.2.11.

In particular applications it is sometimes convenient to change the sign of one or more monomials in the denominator of $W(t)$. This means only a change in the argument. Thus, for instance, $(a_k - z)^{\lambda_k}$ contributes $\lambda_k(\theta_k + \pi)$ or $\lambda_k(\theta_k - \pi)$ to θ.

By a simple transformation we can also derive a formula which effects the mapping of the interior of a circle onto a polygonal region. The linear fractional transformation (10.2–3), now written as

$$z = i\,\frac{1+z'}{1-z'}, \tag{10.3–8}$$

maps the interior of $|z'| = 1$ onto the region $\operatorname{Im} z > 0$. Performing the substitution we have, if a'_k corresponds to a_k,

$$t - a_k = i\left(\frac{1+t'}{1-t'} - \frac{1+a'_k}{1-a'_k}\right) = \frac{2i}{1-a'_k}\,\frac{t'-a'_k}{1-t'}$$

and

$$dt = \frac{2i\,dt'}{(1-t')^2}.$$

Taking into account (10.3–4), omitting the multiplicative and additive constants and dropping the primes, we arrive at

$$\boxed{w = \int_0^z \frac{dt}{(t-a_1)^{\lambda_1} \ldots (t-a_n)^{\lambda_n}},} \qquad \lambda_1 + \ldots + \lambda_n = 2, \tag{10.3–9}$$

where now $a_1, \ldots, a_n$ are on the unit circle. Thus, formally, we have an expression similar to (10.3–2).

10.3.4 – Illustrative examples

i) A simple application of the use of (10.3–9) is the construction of a function $w(z)$ which maps the upper half plane $\operatorname{Im} z > 0$ onto the interior of a triangle. In this case we are sure that the boundary is a simple polygon. Let the interior angles of the triangle be $\alpha_1\pi$, $\alpha_2\pi$, $\alpha_3\pi$. Then $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and, if we take $\lambda_1 = 1-\alpha_1$, $\lambda_2 = 1-\alpha_2$, $\lambda_3 = 1-\alpha_3$, the conditions (10.3–3) and (10.3–4) are satisfied. We may take $a_1 = 0$, $a_2 = 1$ and the desired formula will be

$$w(z) = \int_0^z \frac{dt}{t^{\lambda_1}(1-t)^{\lambda_2}} = \int_0^z t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1}\,dt, \tag{10.3–10}$$

where we have written $1-t$ rather than $t-1$. In this case $w(0)$ and $w(1)$ are real. In particular

$$w(1) = \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1}\,dt = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

$$= \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(1-\alpha_3)} = \frac{1}{\pi}\,\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\sin\pi\alpha_3\,.$$

Hence the length of the side opposite the vertex b_3 is

$$l_3 = \frac{1}{\pi}\,\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\sin\pi\alpha_3\,.$$

It is easy to see that the length of the side opposite the vertex b_1 is

$$l_1 = \int_1^\infty t^{\alpha_1-1}(t-1)^{\alpha_2-1}\,dt = \int_0^1 t^{1-\alpha_1}(1-t)^{\alpha_2-1}\frac{dt}{t^{1+\alpha_2}}$$

$$= \int_0^1 t^{-(\alpha_1+\alpha_2)}(1-t)^{\alpha_2-1}\,dt = \int_0^1 t^{\alpha_3-1}(1-t)^{\alpha_2-1}\,dt,$$

or

$$l_1 = \frac{1}{\pi}\,\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\sin\pi\alpha_1\,.$$

In almost the same way we finally have

$$l_2 = \frac{1}{\pi}\,\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\sin\pi\alpha_2\,.$$

Notice that these results are in accordance with the sine rule

$$\frac{l_1}{\sin\pi\alpha_1} = \frac{l_2}{\sin\pi\alpha_2} = \frac{l_3}{\sin\pi\alpha_3}$$

of elementary trigonometry.

ii) The examples of the sections 10.3.3 and 10.3.4 are particular cases of the Schwarz-Christoffel formula. In both cases we have $\lambda_1=\lambda_2=\lambda_3 = \lambda_4 = \frac{1}{2}$. In contrast to the Legendre case the Weierstrass integral transforms $z = \infty$ into a vertex of the rectangle.

iii) An interesting application of (10.3–9) is the formula

$$w = \int_0^z \frac{dt}{(1-t^n)^{2/n}}\,. \tag{10.3–11}$$

We have $a_1 = 1$, $a_2 = \eta, \ldots, a_n = \eta^{n-1}$, with $\eta = \exp 2\pi i/n$ and $\lambda_1 = \ldots = \lambda_n = 2/n$. The image of the circle $|z| = 1$ is a polygon with angles $(1-2/n)\pi$. The length of a side $b_{k-1}b_k$ is

$$|w(a_k)-w(a_{k-1})| = \left| \int_{a_{k-1}}^{a_k} \frac{dt}{(1-t^n)^{2/n}} \right| .$$

If we perform the substitution $t' = \eta t$ and drop the primes the integral becomes

$$\frac{1}{\eta} \int_{a_{k-1}\eta}^{a_k\eta} \frac{dt}{(1-t^n)^{2/n}} = \frac{1}{\eta} \int_{a_k}^{a_{k+1}} \frac{dt}{(1-t^n)^{2/n}} ,$$

whence

$$|w(a_{k-1})-w(a_k)| = |w(a_k)-w(a_{k-1})|.$$

Hence the polygon is regular.

The radius of the circumscribed circle of this polygon is

$$R = \int_0^1 \frac{dt}{(1-t^n)^{2/n}} = \frac{1}{n} \int_0^1 t^{1/n-1}(1-t)^{-2/n}\, dt.$$

In view of (4.6–13) we have

$$R = \frac{1}{n} \frac{\Gamma\left(\frac{1}{n}\right)\Gamma\left(1-\frac{2}{n}\right)}{\Gamma\left(1-\frac{1}{n}\right)} = \frac{1}{n} \frac{\Gamma\left(1-\frac{2}{n}\right)}{\Gamma^2\left(1-\frac{1}{n}\right)} \frac{\pi}{\sin\frac{\pi}{n}} .$$

The length of a side is $l = 2R \sin \pi/n$, hence

$$l = \frac{2\pi}{n} \frac{\Gamma\left(1-\frac{2}{n}\right)}{\Gamma^2\left(1-\frac{1}{n}\right)} = \frac{2\pi}{n} \frac{\Gamma^2\left(1-\frac{2}{n}\right)}{\Gamma^2\left(1-\frac{1}{n}\right)} \frac{1}{\Gamma\left(1-\frac{2}{n}\right)} .$$

Applying (4.6–26) we find

$$\Gamma\left(\tfrac{1}{2}-\frac{1}{n}\right)\Gamma\left(1-\frac{1}{n}\right) = 2^{2/n}\Gamma\left(1-\frac{2}{n}\right)\sqrt{\pi}$$

and after some computation we find

$$l = \frac{1}{n} 2^{1-4/n} \frac{\Gamma^2\left(\frac{1}{2}-\frac{1}{n}\right)}{\Gamma\left(1-\frac{2}{n}\right)} . \qquad (10.3\text{–}12)$$

Summing up:

The relation (10.3–11) *effects a univalent mapping of the unit circle onto a regular polygon whose sides have the length* (10.3–12).

We list some particular cases. If $n = 3$ we have an equilateral triangle whose sides have the length

$$l = \frac{1}{3\sqrt[3]{2}} \frac{\Gamma^2(\frac{1}{6})}{\Gamma(\frac{1}{3})}.$$

This result can be simplified. From (4.6–26) follows

$$\Gamma(\tfrac{1}{6})\Gamma(\tfrac{2}{3}) = 2^{\frac{2}{3}}\Gamma(\tfrac{1}{3})\sqrt{\pi}$$

and, according to (4.6–13),

$$\Gamma(\tfrac{1}{3})\Gamma(\tfrac{2}{3}) = \frac{2\pi}{\sqrt{3}},$$

whence

$$\Gamma(\tfrac{1}{6}) = 2^{-\frac{1}{3}}\sqrt{3}\,\frac{\Gamma^2(\frac{1}{3})}{\sqrt{\pi}} \tag{10.3–13}$$

and thus the length of the side turns out to be

$$l = \frac{1}{2\pi}\Gamma^3(\tfrac{1}{3}). \tag{10.3–14}$$

In the case $n = 4$ we have

$$l = \frac{1}{4\sqrt{\pi}}\Gamma^2(\tfrac{1}{4}). \tag{10.3–15}$$

This is in accordance with (10.2–64).

10.3.5 – The mapping of the exterior of a polygon

The problem of representing the interior of the unit circle in the z-plane upon the exterior of a simple polygon in the w-plane can be solved by a formula closely resembling the Schwarz-Christoffel formula (10.3–2). We assume that the numbers $\lambda_1, \ldots, \lambda_n$ are the same as in section 10.3.3 and satisfy the same relations (10.3–3) and (10.3–4). Now we consider the integral

$$\boxed{w = \int_{z_0}^{z} (t-a_1)^{\lambda_1} \ldots (t-a_n)^{\lambda_n} \frac{dt}{t^2},} \qquad z_0 \neq 0, \tag{10.3–16}$$

where $a_1, \ldots, a_n$ are on $|z| = 1$. It is evident that the function (10.3–16) has a simple pole at $z = 0$. In order to prove that $|z| = 1$ is transformed into a polygon it is more convenient to apply the substitution (10.3–8) and by omitting the primes and irrelevant constants we find

$$w = \int_{z_0}^{z} \frac{(t-a_1)^{\lambda_1} \ldots (t-a_n)^{\lambda_n}}{(1+t^2)^2}\, dt, \tag{10.3–17}$$

$a_1 < a_2 < \ldots < a_n$. If $a_n = \infty$ the corresponding factor is missing. We agree that now the argument of the integrand is the number *0*, defined in (10.3–5).

Reasoning as in section 10.3.3 we see that (10.3–16) maps the real axis onto a closed polygon L and, consequently, the relation (10.3–17) maps the circumference $|z| = 1$ on L. We must make the additional assumption that this polygon is simple. It should be noticed that if z percorses the real axis and passes through a point a_k, the argument of the integrand in (10.3–17) *decreases* by $\lambda_k \pi$, (fig. 10.3–6).

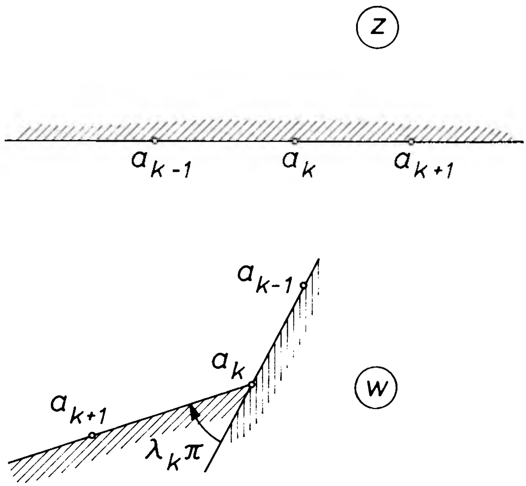


Fig. 10.3–6. Behaviour of the function (10.3–16) at a point a_k

Our further discussion is based on (10.3–16). Let z_1 denote a point inside $|z| = 1$ and $w_1 = w(z_1)$ its image. Since z_1 can be connected with $z = 0$ by a continuous arc which does not meet the circumference we can join w_1 with ∞ by means of a continuous curve which does not meet L, since $w(z)$ is univalent on $|z| = 1$. Hence $\Omega_L(w_1) = 0$. By a reasoning similar to that employed in section 10.1.1 we make sure the fact that the argument principle of section 3.10.1 is applicable and we find that the sum of the numbers of zeros and poles of $w(z) - w_1$ inside $|z| = 1$ is zero. Since $w(z)$ has a simple pole it follows that $w(z) - w_1$ has a simple zero. Notice that (10.3–17) is finite at $z = \infty$, by virtue of (10.3–4). Summing up

The relation (10.3–16) *effects a univalent mapping of the interior of the unit circle onto the exterior of a polygon L, provided this polygon is simple, and the closed disc onto the closure of this exterior region.*

A similar theorem holds for (10.3–17). We need only replace the disc $|z| \leqq 1$ by $\operatorname{Im} z \geqq 0$.

i) It is clear that the function

$$w = \int_{z_0}^{z} \sqrt{1-t^4}\,\frac{dt}{t^2}, \qquad z_0 \neq 0 \tag{10.3–18}$$

maps the interior of the unit circle onto the exterior of a square.

ii) The integral

$$w = 2\int_0^z \frac{\sqrt{(t^2-c^2)(t^2-1/c^2)}}{(t^2+1)^2}\,dt, \qquad c \text{ real}, \tag{10.3–19}$$

represents the mapping of the exterior of a rectangle onto the upper half of the z-plane. The vertices correspond to $\pm c$, $\pm 1/c$ on the real axis. We assume $0 < c < 1$.

In this case it is also possible to express the mapping function in terms of Jacobian functions. We follow a method essentially due to W.G. Bickley.

First we introduce a variable χ and a constant α by means of

$$z = \tan \tfrac{1}{2}\chi, \qquad c = \tan \tfrac{1}{2}\alpha. \tag{10.3–20}$$

By an elementary, somewhat troublesome computation, it can be verified that the integral (10.3–18) can be cast into the form

$$w = \tfrac{1}{2} \csc \alpha \int_0^{\chi} \sqrt{2(\cos 2\theta - \cos 2\alpha)}\,d\theta. \tag{10.3–21}$$

Indeed, replacing t by $\tan \frac{1}{2}\theta$, the square of the integrand of (10.3–19) takes the form

$$(\tan^2 \tfrac{1}{2}\theta - \tan^2 \tfrac{1}{2}\alpha)\left(\tan^2 \tfrac{1}{2}\theta - \frac{1}{\tan^2 \frac{1}{2}\alpha}\right)\frac{1}{(1+\tan^2 \frac{1}{2}\theta)^4}$$

$$= (\tan^2 \tfrac{1}{2}\theta - \tan^2 \tfrac{1}{2}\alpha)(\tan^2 \tfrac{1}{2}\theta \tan^2 \tfrac{1}{2}\alpha - 1)\frac{\cos^8 \frac{1}{2}\theta}{\tan^2 \frac{1}{2}\alpha}.$$

First we have

$$\tan \tfrac{1}{2}\theta - \tan \tfrac{1}{2}\alpha = \tan \tfrac{1}{2}(\theta-\alpha)(1+\tan \tfrac{1}{2}\theta \tan \tfrac{1}{2}\alpha),$$

whence

$$\tan^2 \tfrac{1}{2}\theta - \tan^2 \tfrac{1}{2}\alpha = \tan \tfrac{1}{2}(\theta-\alpha) \tan \tfrac{1}{2}(\theta+\alpha)(1-\tan^2 \tfrac{1}{2}\theta \tan^2 \tfrac{1}{2}\alpha).$$

Next

$$\tan \tfrac{1}{2}(\theta-\alpha) \tan \tfrac{1}{2}(\theta+\alpha) = -\frac{\cos\theta-\cos\alpha}{\cos\theta+\cos\alpha} = -\frac{1}{2}\frac{\cos 2\theta-\cos 2\alpha}{(\cos\theta+\cos\alpha)^2}.$$

From

$$1-\tan\tfrac{1}{2}\theta \tan\tfrac{1}{2}\alpha = \frac{\cos\tfrac{1}{2}(\theta+\alpha)}{\cos\tfrac{1}{2}\theta\cos\tfrac{1}{2}\alpha}$$

follows

$$1-\tan^2\tfrac{1}{2}\theta\tan^2\tfrac{1}{2}\alpha = \frac{\cos\theta+\cos\alpha}{2\cos^2\tfrac{1}{2}\theta\cos^2\tfrac{1}{2}\alpha}.$$

Thus the expression we started with takes the form

$$\frac{\cos 2\theta-\cos 2\alpha}{2(\cos\theta+\cos\alpha)^2}\,\frac{(\cos\theta+\cos\alpha)^2\cos^8\tfrac{1}{2}\theta}{4\cos^4\tfrac{1}{2}\theta\cos^4\tfrac{1}{2}\alpha\tan^2\tfrac{1}{2}\alpha}$$

$$= \frac{\cos 2\theta-\cos 2\alpha}{8\sin^2\tfrac{1}{2}\alpha\cos^2\tfrac{1}{2}\alpha}\cos^4\tfrac{1}{2}\theta = \frac{\cos 2\theta-\cos 2\alpha}{2\sin^2\alpha}\cos^4\tfrac{1}{2}\theta$$

and the integrand of (10.3–19) becomes

$$\frac{1}{2\sin\alpha}\sqrt{2(\cos 2\theta-\cos 2\alpha)}\cos^2\tfrac{1}{2}\theta.$$

Because

$$dt = \frac{1}{2\cos^2\tfrac{1}{2}\theta}\,d\theta,$$

the truth of (10.3–21) follows.

Next we put

$$\operatorname{dn} t = \cos\theta, \qquad \operatorname{dn} w_1 = \cos\chi, \qquad k\operatorname{sn} w_1 = \sin\chi, \tag{10.3–22}$$

where the modulus k of the elliptic functions is determined by

$$k = \sin\alpha. \tag{10.3–23}$$

Performing the substitution we find

$$w = k\int_0^{w_1}\operatorname{cn}^2 t\,dt = \frac{1}{k}\int_0^{w_1}(\operatorname{dn}^2 t-k'^2)dt$$

or, in view of (5.17–2) and (5.17–14),

$$w = \frac{1}{k}\left(\operatorname{zn} w_1+\left(\frac{\mathsf{E}}{\mathsf{K}}-k'^2\right)w_1\right). \tag{10.3–24}$$

The desired mapping function is obtained by eliminating χ and w_1.

The elimination of χ does not give rise to difficulties, but provides us with useful information. From (10.3–22) and (10.3–20) it follows that

$$\frac{2z}{1+z^2} = k \text{ sn } w_1$$

or

$$\left(\frac{1+z}{1-z}\right)^2 = \frac{1+k \text{ sn } w_1}{1-k \text{ sn } w_1}. \tag{10.3–25}$$

From (10.2–18) and (10.2–1) it follows that this relation defines a univalent correspondence between a rectangle in the w_1-plane and a semicircle in the z-plane, the radius of this circle being unity. To $w_1 = 0$ corresponds $z = 0$ and to $w_1 = \mathsf{K}+i\mathsf{K}'$ corresponds $z = 1$. Hence the upper side of the rectangle is transformed into the circular arc.

An alternative expression for the relation between z and w_1 can be obtained by starting from the first equation (10.3–22). We have

$$z^2 = \frac{1-\text{dn } w_1}{1+\text{dn } w_1} = \frac{(1-\text{dn } w_1)^2}{1-\text{dn}^2 w_1} = \frac{(1-\text{dn } w_1)^2}{k^2 \text{ sn}^2 w_1},$$

or

$$z = \frac{1-\text{dn } w_1}{k \text{ sn } w_1}. \tag{10.3–26}$$

By the symmetry principle we find that (10.3–26) or the equivalent relation (10.3–25) maps the upper half of the z-plane onto a rectangle with vertices $\pm\mathsf{K}$, $\pm\mathsf{K}+2i\mathsf{K}'$. As a consequence (10.3–24) *maps the exterior of a rectangle in the w-plane onto the interior of a rectangle in the w_1-plane.*

Turning back to (10.3.23) we see that to $w_1 = \mathsf{K}$ corresponds

$$z = \frac{1-k'}{k} = \frac{1-\cos\alpha}{\sin\alpha} = \tan\tfrac{1}{2}\alpha = c. \tag{10.3–27}$$

To $w_1 = \mathsf{K}+2i\mathsf{K}'$ corresponds

$$z = \frac{1+k'}{k} = \frac{1-k'^2}{k(1-k')} = \frac{k}{1-k'} = 1/c,$$

since $\text{dn }(\mathsf{K}+2i\mathsf{K}') = -\text{dn } \mathsf{K}$, in accordance with (5.14–30) and $\text{sn}(\mathsf{K}+2i\mathsf{K}') = 1$.

Now we wish to consider a rectangle in the w-plane with vertices $\pm a \pm bi$, $a > 0$, $b > 0$, and try to find a mapping function such that the exterior corresponds to a half z-plane, whereby $a-bi$ and $-a-bi$

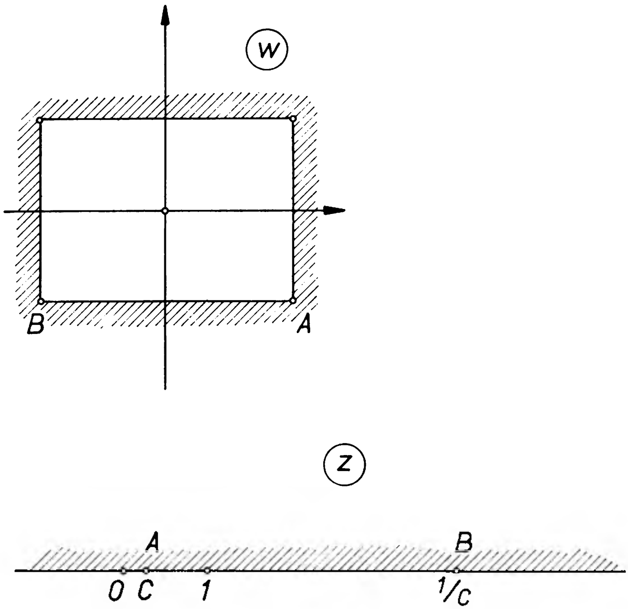


Fig. 10.3–7. The mapping of the exterior of a rectangle

correspond to c and $1/c$ respectively, (fig. 10.3–7). We shall prove that the desired mapping function is of the form

$$Aw+B,$$

where w is the function of z found by eliminating w_1 from (10.3–24) and (10.3–26). Since

$$\operatorname{zn} \mathsf{K} = 0, \qquad \operatorname{zn} (\mathsf{K}+2i\mathsf{K}') = -\frac{\pi i}{\mathsf{K}},$$

by (5.17–15), we have to determine A and B such that

$$a-bi = \frac{A}{k}(\mathsf{E}-k'^2\mathsf{K})+B,$$

$$-a-bi = \frac{A}{k}(\mathsf{E}-k'^2\mathsf{K})+B+\frac{A}{k}\left(-\frac{\pi i}{\mathsf{K}}+2i\frac{\mathsf{EK}'-k'^2\mathsf{KK}'}{\mathsf{K}}\right)$$

$$= \frac{A}{k}(\mathsf{E}-k'^2\mathsf{K})+B-\frac{2Ai}{k}(\mathsf{E}'-k^2\mathsf{K}'), \qquad (10.3\text{–}28)$$

where we have used Legendre's relation (5.17–8). It follows that

$$2a = \frac{2Ai}{k}(\mathsf{E}'-k^2\mathsf{K}'),$$

or

$$A = -\frac{iak}{\mathsf{E}'-k^2\mathsf{K}'}. \tag{10.3–29}$$

Hence A is purely imaginary and, therefore,

$$B = a. \tag{10.3–30}$$

From the first of the equations (10.3–28) it also follows

$$A = \frac{-ibk}{\mathsf{E}-k'^2\mathsf{K}} \tag{10.3–31}$$

and thus k is determined by

$$\frac{b}{a} = \frac{\mathsf{E}-k'^2\mathsf{K}}{\mathsf{E}'-k^2\mathsf{K}'}. \tag{10.3–32}$$

By the aid of numerical tables k can be evaluated if b/a is known. The constant c is determined by (10.3–27).

The special case of a square deserves some extra consideration. Now we have $a = b$, $k = k' = \frac{1}{2}\sqrt{2}$, whence

$$c = \sqrt{2}-1, \qquad 1/c = \sqrt{2}+1.$$

The quarter period K of the associated elliptic functions is

$$\mathsf{K} = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-\frac{1}{2}t^2)}} = \sqrt{2}\int_0^1 \frac{dt}{\sqrt{(1-t^2)(2-t^2)}}.$$

Performing the substitution $x^2 = 2-t^2$, we find

$$\mathsf{K} = \int_0^1 \frac{dx}{\sqrt{1-x^4}} = \frac{\sqrt{2}}{4}\int_0^1 x^{-\frac{3}{4}}(1-x)^{-\frac{1}{2}}\,dx = \frac{\sqrt{2}}{4}\,\frac{\Gamma(\frac{1}{4})\Gamma(\frac{1}{2})}{\Gamma(\frac{3}{4})}$$

whence, by (10.2–59)

$$\mathsf{K} = \frac{1}{4\sqrt{\pi}}\Gamma^2(\tfrac{1}{4}). \tag{10.3–33}$$

From Legendre's relation we get

$$\mathsf{E}-\tfrac{1}{2}\mathsf{K} = \frac{\pi}{4\mathsf{K}},$$

and since K is known, we can evaluate A if a is given.

10.3.6 – Degenerate cases

The scope of the Schwarz-Christoffel formula is wider than is implied by the theorems of section 10.3.3. Various generalizations depend on a particular choice of the exponents λ_k; we then obtain a transformation of the half plane (or the unit disc) into generalized polygonal regions the boundaries of which may consist not only of segments, but also of half lines. Also the mapping onto the boundary needs not be one-to-one either. It should be pointed out that the use of the Schwarz-Christoffel formula in this degenerate cases requires special justification and very often it is easier to verify the result directly. If we retain the condition (10.3–4) we can interpret the figures as degenerate polygons.

We wish to discuss some typical examples which may serve as an illustration.

First we consider again the mapping

$$w = z^\alpha, \qquad 0 < \alpha \leqq 2$$

which we shall write in the form

$$w = \alpha \int_0^z \frac{dt}{t^{1-\alpha}}. \tag{10.3–34}$$

In this case we satisfy (10.3–4) by taking $\lambda_1 = 1-\alpha$, $\lambda_2 = 1+\alpha$. The angular region which is the image of $\operatorname{Im} z > 0$ may be considered as a 2-gon with internal angles $\alpha\pi$ at $w = 0$ and $-\alpha\pi$ at $w = \infty$ (compare section 9.2.8). In accordance with (10.3–6) the sum of the internal angles is zero.

A similar example is

$$w = \log z,$$

written as

$$w = \int_1^z \frac{dt}{t}. \tag{10.3–35}$$

The values of λ are $\lambda_1 = 1$, $\lambda_2 = 1$ and the internal angles (at the vertices coinciding with $w = \infty$) are both zero. This situation can be pictured in a more illustrative form on applying a linear fractional transformation which transforms the point at infinity into a finite point. The infinite parallel strip which is the image of $\operatorname{Im} z > 0$ under (10.3–4) becomes the region between two circles with internal contact, (fig. 10.3–8).

A somewhat more complicated example is the following: the function

$$w_1 = \frac{1}{2}\left(\sqrt{z} + \frac{1}{\sqrt{z}}\right) \tag{10.3–36}$$

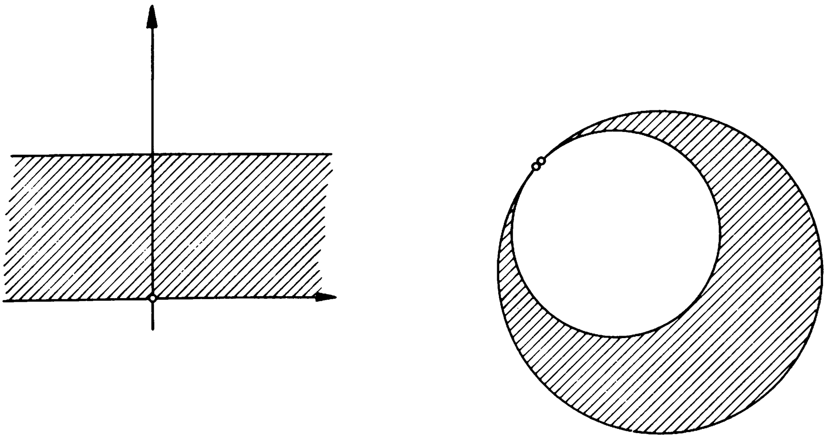


Fig. 10.3–8. Linear equivalence between an infinite strip and a region bounded by two circles with interval contact

is defined throughout $\operatorname{Im} z > 0$. We agree that the square root is such that $\sqrt{1} = 1$. This may be written as an integral

$$1 + \frac{1}{4}\int_1^z \frac{t-1}{t^{\frac{3}{2}}}\, dt. \tag{10.3–37}$$

The transformation $z_1 = \sqrt{z}$ maps the upper half of the z-plane onto the first quadrant in the z_1-plane. Let $z_1 = re^{i\theta}$. Then

$$w_1 = \tfrac{1}{2}\left(r + \frac{1}{r}\right)\cos\theta + \tfrac{1}{2}i\left(r - \frac{1}{r}\right)\sin\theta.$$

If $\theta = 0$ then w_1 decreases from $+\infty$ to 1 if r moves from 0 to 1 and increases from 1 to $+\infty$ if r moves from 1 to ∞. If, however, $\theta = \frac{1}{2}\pi$, then w_1 moves from ∞i to $-\infty i$ along the imaginary axis. Thus we see that the image is the half plane $\operatorname{Re} w_1 > 0$ slit along the half ray $w_1 \geqq 1$, (fig. 10.3–9). The integral (10.3–16) suggests the values $\lambda_1 = \lambda_2 = \frac{3}{2}$, $\lambda_3 = -1$. Hence the figure is a degenerate triangle with a vertex at $w_1 = 1$, corresponding to $z = 1$ and a doubly counted vertex at $w_1 = \infty$ (corresponding to $z = 0$ and $z = \infty$). The internal angles are $-\frac{1}{2}\pi$, $-\frac{1}{2}\pi$, 2π. By means of

$$w = \frac{w_1 - 1}{w_1 + 1} \tag{10.3–38}$$

the half w-plane is transformed into the unit circle slit along the radius $w \leqq 1$. Inserting (10.3–36) into (10.3–38) we obtain (10.2–20).

Now let us consider the mapping defined by means of

$$w = \log z - z. \tag{10.3–39}$$

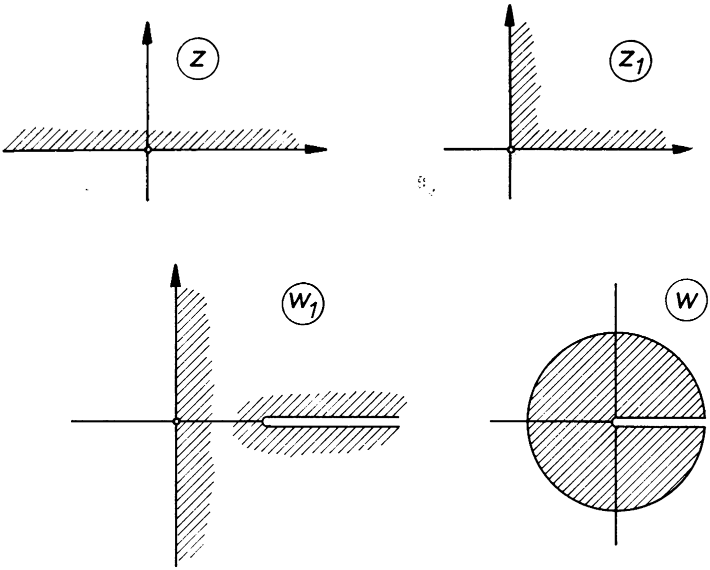


Fig. 10.3–9. The mapping as provided by the combination of (10.3–36) and (10.3–38)

We put $z = re^{i\theta}$ and find

$$u = \log r - r\cos\theta, \qquad v = \theta - r\sin\theta.$$

To the positive real axis in the z-plane corresponds $v = 0$, $u = \log r - r$. This function u increases from $-\infty$ to -1 as r increases from 0 to 1 and decreases from -1 to $-\infty$ as r increases from 1 upwards. The negative real axis corresponds to the line $v = \pi$ described from ∞ to $-\infty$ if z increases from $-\infty$ to 0. Thus the image of the upper half of the z-plane is the half plane $\operatorname{Im} w < \pi$ slit along the half ray $u \leqq -1$, (fig. 10.3–10). Expressed as a Schwarz-Christoffel integral the function (10.3–18) appears as

$$w = -1 - \int_1^z \frac{t-1}{t}\, dt.$$

If we take $\lambda_1 = -1$, $\lambda_2 = 1$, $\lambda_3 = 2$ we can interpret the image described above as a triangle with the angles 2π, 0, $-\pi$, the last two at coinciding vertices at infinity. Transforming this point to a finite point we find a figure as pictured in fig. 10.3–11.

A similar function is

$$w = \log z - z^2. \tag{10.3–40}$$

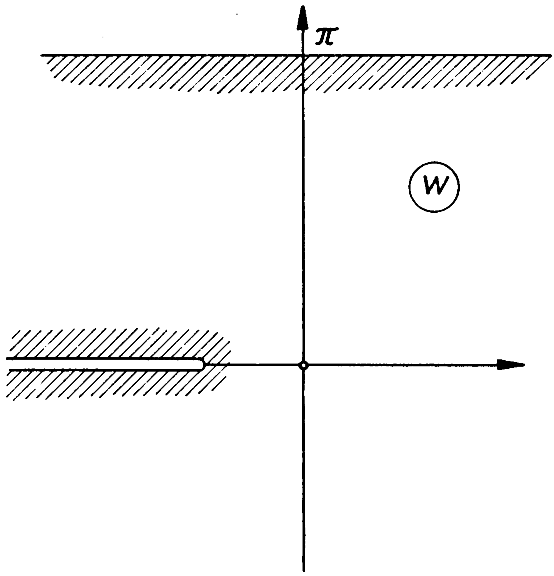


Fig. 10.3–10. The image of the upper half of z-plane as given by (10.3–39)

A discussion, quite the same as above, reveals that the image of the upper half of the z-plane is the whole w-plane slit along two half rays, viz. $u \leqq -1$, $v = 0$ and $u \leqq -1$, $v = \pi$, (fig. 10.3–12). Written as a Schwarz-Christoffel integral (10.3–40) assumes the form

$$-1+\int_1^z \frac{1-2t^2}{t}\,dt. \qquad (10.3\text{–}41)$$

If we take $\lambda_1 = -1$, $\lambda_2 = -1$, $\lambda_3 = 1$, $\lambda_4 = 3$ we can interpret the image as a quadrangle with internal angles 2π, 2π, 0, -2π, the latter two angles

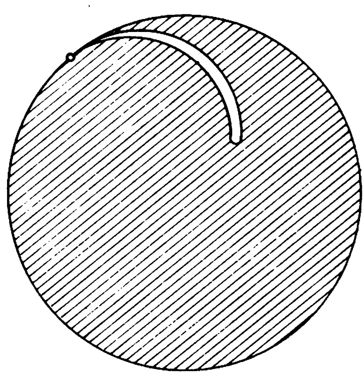


Fig. 10.3–11. A linear equivalent of the region pictured in Fig. 10.3–10

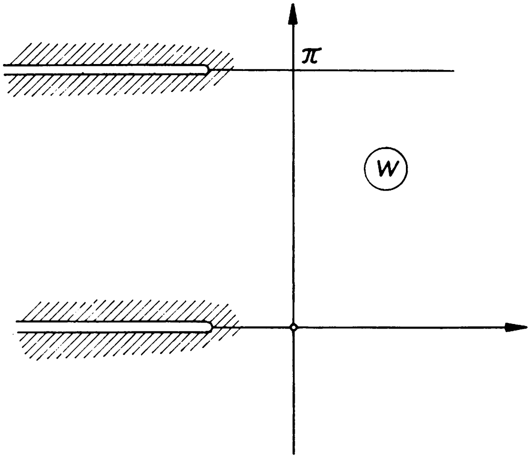


Fig. 10.3–12. The image of the upper half of the z-plane as given by (10.3–40)

corresponding to a doubly counted vertex at infinity. Bringing this to a finite point we obtain the figure pictured in fig. 10.3–13.

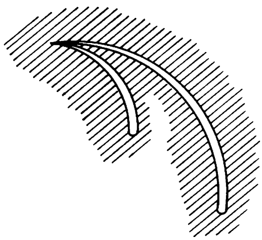


Fig. 10.3–13. A linear equivalent of Fig. 10.3–12

10.4 – Functions related to the mapping of a square

10.4.1 – The Gaussian lemniscate functions

We obtain an interesting pair of functions if we study more closely the mapping of a square onto a circular disc by means of the formula

$$w = \int_0^z \frac{dt}{\sqrt{1-t^4}}. \tag{10.4–1}$$

It is clear that the points corresponding to $z = 1$ and $z = -1$ are real and those corresponding to $z = i$ and $-i$ are purely imaginary.

The inverse function $z(w)$ is defined throughout the square and is denoted by

$$z = \operatorname{sl} w. \tag{10.4–2}$$

It is called the *lemniscate sine*. A complementary function, the *lemniscate cosine*, is defined as the inverse of

$$w = \int_z^1 \frac{dt}{\sqrt{1-t^4}} \tag{10.4–3}$$

and denoted by

$$z = \operatorname{cl} w. \tag{10.4–4}$$

The length of the diagonal of the square in the w-plane shall be denoted by $\tilde{\pi}$. It is equal to

$$\tilde{\pi} = 2\int_0^1 \frac{dt}{\sqrt{1-t^4}} = \frac{1}{2}\int_0^1 t^{-\frac{3}{4}}(1-t)^{-\frac{1}{2}}\,dt = \tfrac{1}{2}B(\tfrac{1}{4}, \tfrac{1}{2}),$$

whence, by (10.2–59) and (10.3–30)

$$\tilde{\pi} = \frac{1}{2\sqrt{2\pi}}\Gamma^2(\tfrac{1}{4}) = \mathsf{K}\sqrt{2}, \tag{10.4–5}$$

where K is the real quarter period of the Jacobian elliptic functions with modulus $\frac{1}{2}\sqrt{2}$.

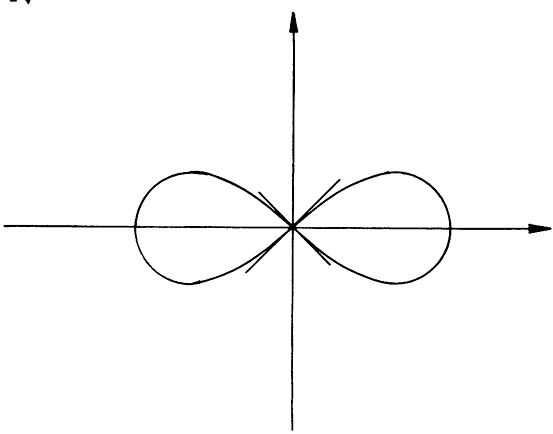


Fig. 10.4–1. The lemniscate

The integral (10.4–1) appears in the problem of rectifying the arc of the lemniscate. This curve can be represented in polar coordinates as

$$r^2 = \cos 2\theta. \tag{10.4–6}$$

The curve consists of two loops and has a double point at the origin, (Fig. 10.4–1). The arc length is calculated from

$$\left(\frac{ds}{dr}\right)^2 = 1+\left(r\,\frac{d\theta}{dr}\right)^2 = \frac{1}{1-r^4}\,.$$

Hence the perimeter of the half of a loop is

$$\tilde{\pi} = \int_0^1 \frac{dr}{\sqrt{1-r^4}}\,.$$

It is clear that the complementary lemniscate functions are related as follows

$$\mathrm{cl}\; w = \mathrm{sl}(\tfrac{1}{2}\tilde{\pi}-w), \quad \mathrm{sl}\; w = \mathrm{cl}(\tfrac{1}{2}\tilde{\pi}-w). \tag{10.4–7}$$

10.4.2 – Lemniscate functions as doubly periodic functions

If we reflect the square which is mapped onto the interior of the unit circle in the z-plane with respect to one of its sides we obtain another square which corresponds univalently to the exterior of the unit circle. In a figure the first square is shaded. Repeating this process indefinitely we obtain a regular pattern of shaded and unshaded squares, covering the whole w-plane without gaps or overlappings, (fig. 10.4–2). In two

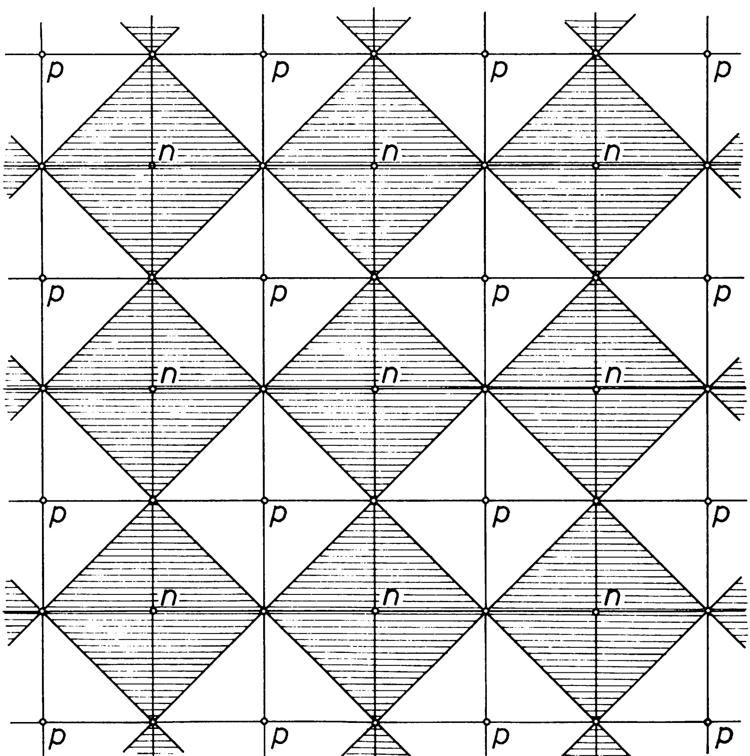


Fig. 10.4–2. The pattern associated with the lemniscate sine (n = zero, p = pole)

shaded squares we can find *homologous* points, i.e., points which correspond to the same point inside the unit circle. A similar remark holds for the unshaded squares. Each point of the w-plane belongs to a lattice of homologous points.

It is now clear that the lemniscate functions can be extended beyond the boundary of the original square by applying the symmetry principle repeatedly. It follows that these extended functions (also denoted by sl and cl) are defined throughout the w-plane and take the same values at homologous points. Since the pattern remains unchanged if we shift it horizontally or vertically through a distance $2\tilde{\pi}$, it follows that the lemniscate functions are doubly periodic with periods $2\tilde{\pi}$ and $2\tilde{\pi}i$.

The zeros of the lemniscate sine are at the centres of the shaded squares, whereas the centres of the unshaded squares are the poles of this function. It is clear that the order of a pole is unity. The pattern of the zeros and the poles of the lemniscate cosine is obtained from that of the lemniscate sine by the transformation $w \to \frac{1}{2}\tilde{\pi} - w$.

The investigation of the pattern of the zeros and the poles of the lemniscate functions reveals that the pattern of the zeros of the lemniscate sine is similar to that of the zeros of the first theta function ϑ_1, (fig. 5.8–1) and that of the poles is similar to that of the zeros of the third theta function ϑ_3. The quotient ϑ_1/ϑ_3 is doubly periodic and we may deduce from Liouville's theorem (section 5.2.2) that

$$\operatorname{sl} w = c_1 \frac{\vartheta_1(\lambda w)}{\vartheta_3(\lambda w)}, \tag{10.4–8}$$

where c_1 and λ are constants. Since the periods of the theta quotient are 1 and τ, we find that $\lambda = 1/\tilde{\pi}$ (the distance between two consecutive zeros being $\tilde{\pi}$) and $\tau = i$, for the network of zeros and poles consists of squares with horizontal and vertical sides. Taking into account (5.8–27) and (5.15–8), we find

$$1 = \operatorname{sl} \tfrac{1}{2}\tilde{\pi} = c_1 \frac{\vartheta_1(\frac{1}{2})}{\vartheta_3(\frac{1}{2})} = c_1 \frac{\vartheta_2(0)}{\vartheta_4(0)} = c_1 \sqrt{\frac{k}{k'}}.$$

But since $\tau = i$ Jacobi's imaginary transformation (5.10–1) is the identity now and $k = k' = \frac{1}{2}\sqrt{2}$. Hence $c_1 = 1$. Proceeding along the same lines, and taking into account the fact that $\operatorname{cl} 0 = 1$, we finally have,

$$\boxed{\operatorname{sl} w = \frac{\vartheta_1(w/K\sqrt{2})}{\vartheta_3(w/K\sqrt{2})}, \qquad \operatorname{cl} w = \frac{\vartheta_2(w/K\sqrt{2})}{\vartheta_4(w/K\sqrt{2})},} \tag{10.4–9}$$

where we have replaced $\tilde{\pi}$ by its value (10.4–5).

It is not difficult to relate the lemniscate functions to the Jacobian elliptic functions with modulus $k = \frac{1}{2}\sqrt{2}$. From (5.15–10) we deduce at once

$$\boxed{\operatorname{sl} w = \frac{1}{\sqrt{2}} \frac{\operatorname{sn} w\sqrt{2}}{\operatorname{dn} w\sqrt{2}}, \qquad \operatorname{cl} w = \operatorname{cn} w\sqrt{2}.} \tag{10.4–10}$$

Finally from (5.14–16), (5.14–18) and (5.13–12) we find that

$$z^2 = \operatorname{sl}^2 w = \frac{1}{\wp(w; 4, 0)}, \tag{10.4–11}$$

this result being in accordance with (10.2–63).

In a wider sense all elliptic functions related to $\tau = i$ are called lemniscate functions.

10.5 – Riemann's theorem

10.5.1 – Preliminary lemmas

We are going to prove one of the most important theorems in the theory of functions of a complex variable: the fact that any simply connected region whose boundary is not empty and which does not reduce to a single point can be mapped conformally onto the interior of the unit circle. As a consequence two regions of this kind are always conformally equivalent, that is to say there always exists a univalent and meromorphic function defined in one of these regions which provides a mapping of this region onto the other region. For the sake of brevity we shall call a mapping provided by a univalent and meromorphic function a conformal mapping.

This theorem, although formulated by Riemann, was not completely proved by him. The first complete proof we owe to Koebe and Carathéodory. Since then many mathematicians have simplified the original proof considerably.

It is worth noting that the situation in the case of multiply connected regions is not quite so simple. For instance, not every two doubly connected regions are conformally equivalent, as we illustrated in an example exhibited in section 10.2.13.

The theorem is not valid for a region which coincides with the extended plane or has only one boundary point. Without loss of generality we may assume that in this latter case the boundary point is at infinity. A mapping of either region of the kind under consideration onto the interior of the unit circle is provided by a function $f(z)$ holomorphic throughout the z-plane and bounded. Hence, by Liouville's theorem this function reduces to a constant, and is, therefore, not univalent.

Before stating and proving Riemann's theorem we mention some lemmas which facilitate the main proof.

First we have

Every simply connected region $\mathfrak{R}$ *whose boundary is not empty and which does not reduce to a single point can be transformed conformally into a bounded region.*

This lemma is obvious when the complement of the region with respect to the extended plane contains at least one interior point. If this point is at infinity, then $\mathfrak{R}$ is already bounded. If this point is at $z = a$, then we perform the mapping which interchanges the interior and the exterior of a sufficiently small circle around $z = a$. This mapping is provided by a fractional linear transformation. Therefore it is sufficient that $\mathfrak{R}$ can be transformed into a region having this property. Without loss of generality we may assume that $z = \infty$ does not belong to $\mathfrak{R}$. For if $z = \infty$ is in $\mathfrak{R}$ we can find a (finite) point $z = a$ not in $\mathfrak{R}$, by hypothesis, and the function $1/(z-a)$ provides a mapping of $\mathfrak{R}$ onto a region not containing the point at infinity. From now on we assume that $\mathfrak{R}$ is situated in the open z-plane.

By hypothesis there is a finite point $z = a$ not in $\mathfrak{R}$. The function $z-a$ is holomorphic throughout $\mathfrak{R}$ and has no zero in $\mathfrak{R}$. Hence (section 9.1.4) we can define a single-valued branch $g(z)$ of $\log(z-a)$. The function $g(z)$ is evidently univalent in $\mathfrak{R}$. Now let z_0 denote a point of $\mathfrak{R}$. From the first theorem of section 3.12.1 follows that there exists a circle around $g(z_0)$ such that the image of a suitable chosen circle around z_0 is inside this circle. We translate this circle about $g(z_0)$ through the distance 2π upwards and we assert that the circle thus obtained has no point in common with the image of $\mathfrak{R}$ as provided by g. For in the contrary case we could find two points z_1 and z_2 in $\mathfrak{R}$ such that $g(z_1) = g(z_2)+2\pi i$, whence $\exp g(z_1) = z_1-a = \exp g(z_2) = z_2-a$, i.e., $z_1 = z_2$. This is a contradiction. It follows that the function $1/(g(z)-g(z_0)-2\pi i)$ is holomorphic and univalent throughout $\mathfrak{R}$ and, moreover, bounded. Thus we obtain the desired mapping.

Secondly we have

If $\mathfrak{R}$ *is a simply connected region containing the point* $z = 0$ *and contained in the open disc* $|z| < 1$, *but not coinciding with this disc, then there exists a holomorphic and univalent function* $f(z)$ *in* $\mathfrak{R}$ *such that*

$$|f(z)| < 1, \qquad f(0) = 0, \qquad f'(0) > 1.$$

Let z_0 denote a point in the disc $|z| < 1$, but not in $\mathfrak{R}$. An automorphism (9.5–7) brings z_0 into the origin and transforms $\mathfrak{R}$ into a region $\varphi_1(\mathfrak{R})$, φ_1 denoting this automorphism, which does not contain $z = 0$. Hence we can define $\varphi_2(z)$ as a single-valued branch of $\sqrt{z}$ in $\varphi_1(\mathfrak{R})$. Finally we apply an automorphism φ_3 of the type (9.5–7) which brings

$\varphi_2(\varphi_1(0))$ into the origin. The function $\varphi(z) = \varphi_3(\varphi_2(\varphi_1(z)))$ is obviously holomorphic in $\mathfrak{R}$ and univalent. In particular $\varphi(0) = \varphi_3(\varphi_2(\varphi_1(0))) = 0$ and $\varphi'(0) \neq 0$. If we take

$$f(z) = \frac{|\varphi'(0)|}{\varphi'(0)} \varphi(z)$$

then, evidently, $f'(0) > 0$.

The inverse of the function φ is the function $\psi(z) = \psi_1(\psi_2(\psi_3(z)))$ where ψ_1, ψ_2, ψ_3 are the inverses of $\varphi_1, \varphi_2, \varphi_3$ respectively. It is clear that this function is defined throughout $|z| < 1$. Moreover $|\psi(z)| < 1$ as $|z| < 1$, whilst $\psi(0) = 0$. Since $\psi_2(z) = z^2$, the function ψ is not a linear fractional transformation. By virtue of Schwarz's lemma we have $|\psi'(0)| < 1$, and it follows that

$$f'(0) = \frac{1}{\psi'(0)} > 1.$$

This concludes the proof of the second lemma.

Finally we need the lemma

Let $f_n(z)$, $n = 1, 2, \ldots$, be holomorphic and univalent in an open set $\mathfrak{A}$ and suppose that the sequence of this function is uniformly convergent in every bounded and closed subset of $\mathfrak{A}$. If

$$\lim_{n\to\infty} f_n'(0) \neq 0,$$

then

$$f(z) = \lim_{n\to\infty} f_n(z)$$

is holomorphic and univalent throughout $\mathfrak{A}$.

By virtue of Weierstrass's theorem of section 2.20.1 $f_n(z)$ tends to a holomorphic limiting function $f(z)$ as $n \to \infty$ and $f_n'(z)$ tends to $f'(z)$. Hence $f'(0) \neq 0$, that is to say, $f(z)$ is not constant. If $f(z)$ is not univalent a value b exists such that $f(z_1) = f(z_2) = b$, $z_1 \neq z_2$ in $\mathfrak{A}$. Hence the function $f(z) - b$ has at least two zeros inside $\mathfrak{A}$. By Hurwitz's theorem (section 3.11.1) there are infinitely many functions $f_n(z) - b$ which admit a zero inside an arbitrarily small circle around z_1, i.e., there is a subsequence tending to $f(z) - b$ consisting of functions with at least one zero in the considered circle. By the same argument there are among these functions infinitely many which also admit a zero inside an arbitrarily small circle about z_2. Hence among the given functions there are infinitely many which take the value b at different points. Thus we obtain a contradiction.

10.5.2 – Statement and proof of Riemann's theorem

Now we are sufficiently prepared to establish the following fundamental theorem:

Given a simply connected region $\mathfrak{R}$ which is not the whole z-plane and a point z_0 in $\mathfrak{R}$, there exists a unique holomorphic and univalent function $w(z)$ in $\mathfrak{R}$, normalized by the conditions

$$w(z_0) = 0, \qquad w'(z_0) > 0 \tag{10.5–1}$$

which provides a conformal mapping of $\mathfrak{R}$ onto the open disc $|w| < 1$.

In view of the results obtained in the previous section we may assume that $\mathfrak{R}$ is bounded. After a suitable translation we may suppose that $z_0 = 0$ and, applying a dilation with centre O, if necessary, that $\mathfrak{R}$ is within the unit circle around the origin.

Now we consider the family Ψ of univalent and holomorphic functions $f(z)$ on $\mathfrak{R}$ with the additional property that $f(0) = 0$, $f'(0) > 0$, $|f(z)| < 1$. Let m denote the least upper bound of all numbers $f'(0)$. We do not claim at this stage that m is finite. We can extract a sequence $f_1(z)$, $f_2(z), \ldots$ from this family such that $\lim_{n\to\infty} f_n'(0) = m$. The functions of this sequence are bounded and by virtue of the theorem of section 2.22.2 we can select a subsequence which converges uniformly in any closed set contained in $\mathfrak{R}$. This sequence tends to a holomorphic function $w(z)$ and it is clear that $w'(0) = m > 0$. Hence m is finite. Moreover $w(0) = 0$, $|w(z)| < 1$. From the last lemma of section 10.5.12 it follows that $w(z)$ is univalent.

It remains to show that $w(z)$ takes every value w with $|w| < 1$ as z is in $\mathfrak{R}$. According to the second lemma of the previous section we can construct a function $\varphi(w)$ on $w(\mathfrak{R})$, such that $|\varphi(w)| < 1$, $\varphi(0) = 0$, $\varphi'(0) > 1$ provided $w(\mathfrak{R})$ does not coincide with the open disc $|w| < 1$. The function $f(z) = \varphi(w(z))$ belongs to the family Ψ, as may be verified at once, whereas

$$f'(0) = \varphi'(0), \qquad w'(0) > m.$$

This contradicts the definition of m and, consequently, $w(z)$ fills out the whole disc $|w| < 1$, if z runs through $\mathfrak{R}$.

The uniqueness is obvious, for if w_1 and w_2 are functions with the property stated in the theorem then $\chi(w) = w_1(\check{w}_2(w))$ is univalent, where $\check{w}_2(w)$ is the inverse of $w_2(z)$, whilst $\chi(0) = 0$, $|\chi(w)| \leqq 1$. It follows from Schwarz's lemma that $|\chi(w)| \leqq |w|$. Since $\chi(w)$ is invertible, we also have $|w| \leqq |\chi(w)|$, whence $|\chi(w)| = |w|$. From $\chi'(0) = w_1'(0)/w_2'(0) > 0$ follows $\chi(w) = w$, i.e., χ the identity mapping and, therefore, $w_1(z) = w_2(z)$. This concludes the proof of Riemann's mapping theorem.

10.5.3 – THE MAPPING OF THE BOUNDARY

The mapping theorem of Riemann states that it is always possible to establish a one-to-one conformal mapping between two simply connected regions $\mathfrak{R}_z$ and $\mathfrak{R}_w$ whose boundaries contains at least two points. However, it yields no information regarding the points of the boundaries of these regions. In particular the mapping theorem does not say that the mapping is continuous in the closure of $\mathfrak{R}_z$ and that it establishes a one-to-one correspondence between the closures of $\mathfrak{R}_z$ and $\mathfrak{R}_w$. That this cannot be expected follows from the fact that the concept of a region with more than one boundary point is very general and that quite peculiar cases may occur.

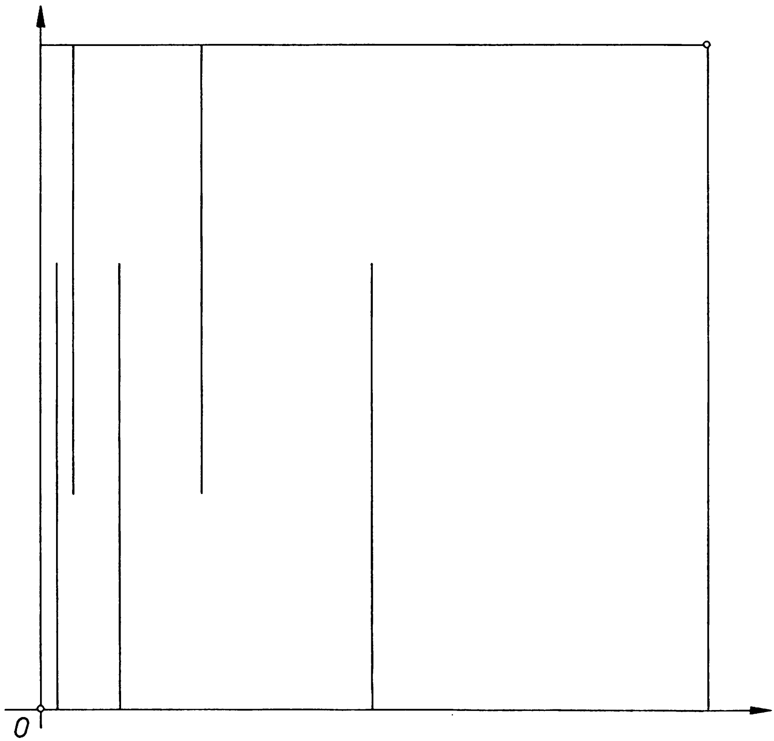


Fig. 10.5–1. Inaccessible boundary points

To illustrate what might happen consider the region indicated in fig. 10.5–1. The region $\mathfrak{R}$ consists of the square $0 < x < 1$, $0 < y < 1$, from which the vertical segments

$$x = \frac{1}{2^{2k-1}}, \quad 0 < y < \tfrac{2}{3}, \qquad x = \frac{1}{2^{2k}}, \quad \tfrac{1}{3} < y < 1,$$

$k = 1, 2, \ldots$, have been removed. The remaining set represents a region whose boundary consists of the perimeter of the square and all indicated segments. Obviously no point of the set $x = 0$, $0 \leqq y \leqq 1$ can be connected with an interior point by a Jordan arc which is entirely in $\mathfrak{R}$. For if we accept the possibility of such an arc we must take for the ordinates of its points a continuous function $y(t)$ which tends to $y(t_0) = b$, $0 \leqq b \leqq 1$, as $t \to t_0$ which is impossible, for it must take infinitely many times values greater than $\frac{2}{3}$ and values less than $\frac{1}{3}$.

Points of the boundary which can be connected by a Jordan arc with interior points, with only one point in common with the boundary, are called *accessible boundary points*. Points without this property are called *inaccessible* boundary points of the region $\mathfrak{R}$. In our example the points $x = 0$, $0 \leqq y \leqq 1$, are inaccessible (from the interior) and all other boundary points are accessible. It is clear that the conformal mapping of the interior of the unit circle onto $\mathfrak{R}$ cannot be continuous at all points of the circumference.

In order to obtain continuity it becomes necessary to make restrictive assumptions which exclude such phenomena as described above.

Let $f(z)$ denote a univalent conformal mapping of the open disc $|z| < 1$ onto a region $\mathfrak{R}_w$ in the w-plane (compared with the previous section we have interchanged the roles of the z- and the w-plane). We consider a sequence $z_1, z_2, \ldots$, of points in the region $\mathfrak{R}_z$ described by $|z| < 1$ converging to a point z_∞ on $|z| = 1$. As regards the corresponding points $w_1, w_2, \ldots$, we can only assert that the accumulation points of this sequence belong to the boundary of $\mathfrak{R}_w$. In fact, such an accumulation point cannot be an exterior point of $\mathfrak{R}_w$, for all points w_n belong to $\mathfrak{R}_w$ and an exterior point possesses a neighbourhood which has no points in common with $\mathfrak{R}_w$. Neither can it be an interior point, for a suitably chosen neighbourhood of an interior point is mapped by the inverse function $\check{f}(w)$ onto an open set within $\mathfrak{R}_z$. This region would contain infinitely many points z_n in contradictions with the assumption that they converge to a point on the circumference.

In particular cases the sequence $w_1, w_2, \ldots$, corresponding to $z_1, z_2, \ldots$, may have one accumulation point w_∞. If, however, we consider another sequence $z'_1, z'_2, \ldots$, in $\mathfrak{R}_z$ which also converges to z_∞, there is no reason to believe that also the sequence $w'_1, w'_2, \ldots$, of the corresponding points in $\mathfrak{R}_w$ converges to w_∞. It may happen that it does not converge, or converge to a point w'_∞ different from w_∞.

Next we consider two points z_∞ and w_∞ having the property: *to every sequence* $z_1, z_2, \ldots$ *which converges to* z_∞ *corresponds a sequence*

$w_1, w_2, \ldots$ *which converges to* w_∞. In this case we shall say that *the boundary point* w_∞ *corresponds sequentially to the boundary point* z_∞ by the mapping $w = f(z)$.

Suppose that this situation is realized for every point of the circumference $|z| = 1$. We may extend the mapping function $f(z)$ defined in $|z| < 1$ to a function defined on the closed disc $|z| \leqq 1$ if we assign to every point z_∞ on $|z| = 1$ the point w_∞ as described above. This function shall also be denoted by $f(z)$. Now we assert:

If every boundary point of $\mathfrak{R}_w$ *corresponds sequentially to a point of* $|z| = 1$ *then the extended mapping function* $f(z)$ *is continuous throughout the closed disc.*

Suppose that $f(z)$ is not continuous at a certain point z_∞ of $|z| = 1$. Then we can find a sequence $z_1, z_2, \ldots$, tending to z_∞, of points in $|z| \leqq 1$ such that $|f(z_n)-f(z_\infty)| \geqq \alpha$, where α is a positive number. There are infinitely many points z_n on $|z| = 1$, for, if not, then we could omit them and obtain a sequence in $|z| < 1$ tending to z_∞ and hence $f(z_n)$ tends to $w_\infty = f(z_\infty)$ according to the definition of w_∞. As a consequence we may assume that all points of the sequence z_n are on $|z| = 1$. By hypothesis $f(z)$ is defined at every point of $|z| = 1$. To every point z_n on $|z| = 1$ we can take a point z_n' in $|z| < 1$ such that $|z_n - z_n'| < 1/n$ and $|f(z_n)-f(z_n')| \leqq \frac{1}{2}\alpha$. Then we have $|f(z_n')-f(z_\infty)| \geqq \frac{1}{2}\alpha > 0$ and this is impossible, for the sequence $z_1', z_2', \ldots$ consists of points in $|z| < 1$ tending to z_∞. This concludes the proof of the theorem.

If z describes the circumference $|z| = 1$ then $f(z)$ describes a continuous curve C_w consisting of boundary points of $\mathfrak{R}_w$. We shall prove that C_w *exhausts the whole boundary* of $\mathfrak{R}_w$. Assume that w' is a boundary point of $\mathfrak{R}_w$ which does not belong to C_w. It is always possible to find a sequence $w_1', w_2', \ldots$ consisting of inner points of $\mathfrak{R}_w$ converging to w'. The corresponding points $z_1', z_2', \ldots$ are in $|z| < 1$ and the accumulation points of this sequence are on $|z| = 1$. Consider one of the accumulation points z_∞ and extract from the above sequence a subsequence $z_1, z_2, \ldots$ tending to z_∞. The corresponding points $w_1, w_2, \ldots$ constitute a sequence tending to the point $w_\infty = f(z_\infty)$ of the curve C_w, but also to $w' \neq w_\infty$. This is impossible. Hence

If by means of $w = f(z)$ *to every point of the circumference* $|z| = 1$ *corresponds a boundary point in the above described manner, then the boundary of* $\mathfrak{R}_w$ *is a continuous curve* C_w.

The curve C_w may have multiple points, i.e., to different points of $|z| = 1$ may correspond the same point of C_w. The last example of section 10.2.5 provides an illustration.

The general problem of the boundary behaviour of the Riemann mapping function and its inverse is not easy and it has received consider-able attention in a series of investigations since the beginning of this century. Special mention should be made of the important pioneer work of Carathéodory on the general boundary problem which he could solve by introducing the fundamental concept of a "prime end".

It is not our intention to pursue the general problem, for we need only a rather special case. In the next section we shall formulate a criterion which enables us to conclude that the situation as described in this section is realized. The faultless verification of this criterion in concrete examples often requires detailed topological considerations but in most cases we can accept its truth without further proof.

10.5.4 – Normal boundary points

It will turn out that the considerations of section 10.5.3 may be applied if the boundary points of $\mathfrak{R}_w$ are of a particular kind. By a *cut* into the interior of $\mathfrak{R}_w$ is meant a Jordan curve joining an interior point of $\mathfrak{R}_w$ to a point w_∞ of the boundary, and such that any infinite sequence of points of the curve either has at least one accumulation point within $\mathfrak{R}_w$ or else has w_∞ as its only accumulation point. A boundary point w_∞ of $\mathfrak{R}_w$ is called *normal* if every sequence $w_1, w_2, \ldots$ of points in $\mathfrak{R}_w$ tending to w_∞ contains a subsequence of points lying on a cut ending in w_∞. All boundary points of the region plotted in fig. 10.5–1 except those on the side $x = 0$ are normal.

In order to be able to decide whether a given boundary point is normal the following lemma may render service

A sequence $w_1, w_2, \ldots$ of points in a region $\mathfrak{R}_w$ which converges to a boundary point w_∞ lies on a cut C if to every $\varepsilon > 0$ we can assign a number n_0 such that all points w_n of the sequence with $n \geqq n_0$ can be joined two by two by means of a polygonal arc included in the intersection of $\mathfrak{R}_w$ and the disc $|w - w_\infty| < \varepsilon$.

According to the assumption to every number ε_k of a sequence $\varepsilon_1, \varepsilon_2, \ldots$ of positive numbers tending to zero we can assign a number n_k such that all points w_n with $n \geqq n_k$ can be joined by a polygonal arc within the intersection of the region $\mathfrak{R}_w$ and the disc $|w - w_\infty| < \varepsilon_k$. First we connect $w_1, \ldots, w_{n_1}$ within $\mathfrak{R}_w$ in this order. Then $w_{n_1}, w_{n_1+1}, \ldots, w_{n_2}$ within the intersection of $\mathfrak{R}_w$ and the disc $|w - w_\infty| < \varepsilon_1$, etc. Thus we obtain a half open polygon in $\mathfrak{R}_w$ which admits the only accumulation point w_∞ on the boundary of $\mathfrak{R}_w$. In general this polygon has multiple points. If we shift a little the vertices we can obtain a polygon L with only isolated self-intersections not coinciding with the points w_n.

Starting from w_1 we percorse the polygon in a direction which we shall call positive. We meet the double points which we denote in succession

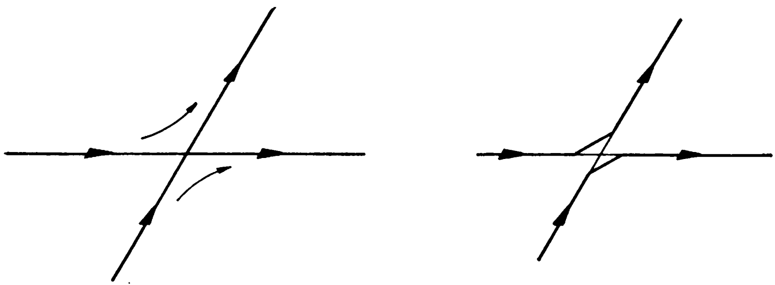


Fig. 10.5–2. The elimination of a contact point

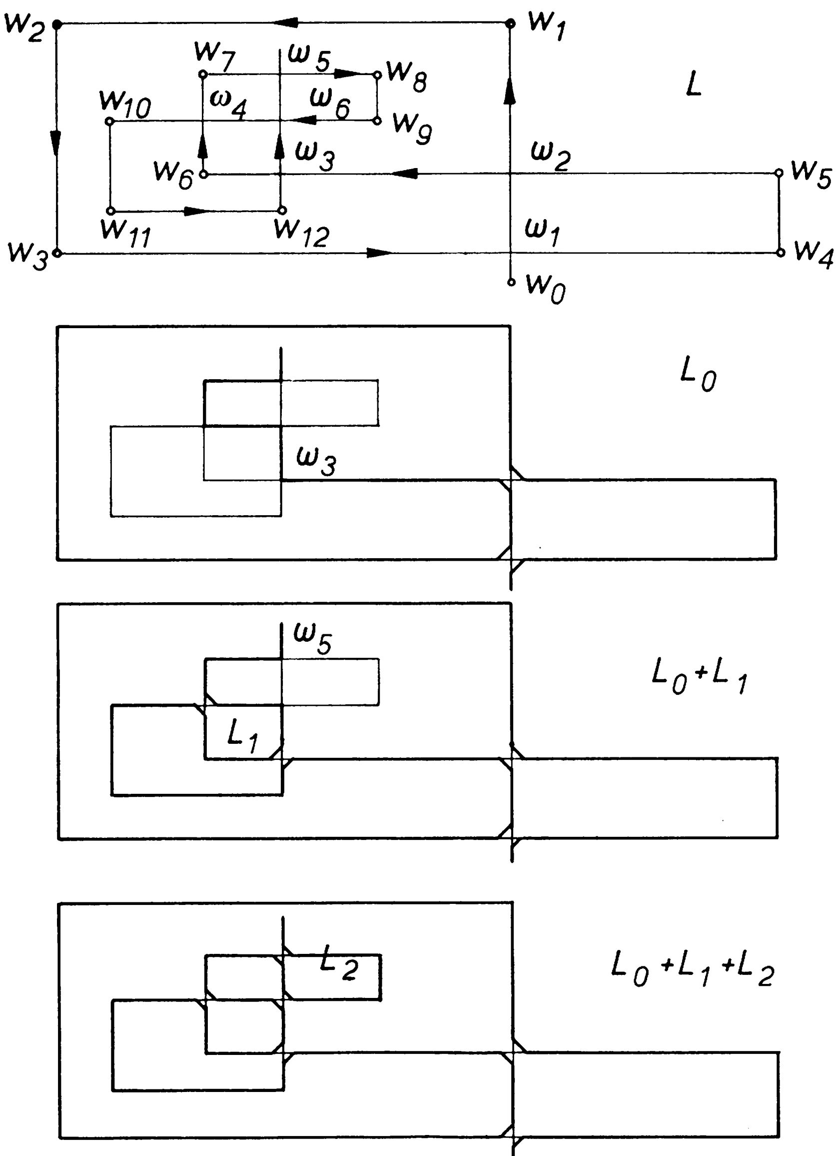


Fig. 10.5–3. The successive eliminations of the crossing points (schematic)

by $\omega_1, \omega_2, \ldots$. First we reach ω_1. We do not traverse the crossing but move along the polygon to the right or to the left in the positive direction. Then we reach a second crossing ω_{m_1} and proceed as before. Continuing we obtain a polygon L_0 which has the only accumulation point w_∞ on the boundary of $\mathfrak{R}_w$. The polygon L_0 has no crossing points, but may have contact points. We remove them by means of two short segments inside a neighbourhood of the contact point which does not contain points of the original sequence, (fig. 10.5–2).

Next we percorse L again starting from w_1. Let ω_r denote a crossing point of L which was not a contact point of L_0. We can find a part L_1 of L beginning and ending at ω_r which bounds a simply connected region. This polygon contacts with L_0 in ω_r and possibly in other points of L_0. By removing small segments issuing from ω_r we can connect L_1 to L_0 by means of two other small segments, (fig. 10.5–3), and the remaining contact points are removed as above in the case of L_0. Thus we have extended L_0 to a polygonal arc L_0+L_1. This being done we move further along L until we reach again a crossing point which was not a contact point of L_0+L_1 and proceed as at ω_r. Thus we construct a polygon $C = L_0+L_1+\ldots$ without double points, containing all points $w_1, w_2, \ldots$ and having the only accumulation point w_∞ on the boundary of $\mathfrak{R}_w$.

10.5.5 – The central theorem

The question of the behaviour of a mapping function at the boundary can be answered completely if we make the assumptions stated in the theorem

If $f(z)$ maps the interior of the disc $|z| < 1$ univalently and conformally onto a region $\mathfrak{R}_w$ which has only normal boundary points, then $f(z)$ is continuous throughout the closed disc $|z| \leq 1$.

By the same arguments as used for the proof of Riemann's theorem we may assume that $\mathfrak{R}_w$ is bounded and even included in the open disc $|w| < 1$. Let $z_0, z_1, \ldots$ denote a sequence of points having a point z_∞ on $|z| = 1$ as the only accumulation point. It is our aim to show that the corresponding sequence $w_1, w_2, \ldots$ tends to a boundary point w_∞ of $\mathfrak{R}_w$ and that w_∞ is uniquely determined by z_∞, i.e., it does not depend on the sequence tending to z_∞. Suppose that the sequence of the points w_n has at least two accumulation points w' and w''. Then we can find two subsequences $w_1', w_2', \ldots$ and $w_1'', w_2'', \ldots$ tending to w' and w'' respectively and having the additional property that the points of the first sequence are on a cut C_w' and those of the second sequence on a cut C_w'', corresponding to two curves C_z' and C_z'' in $|z| < 1$. The curve C_z'

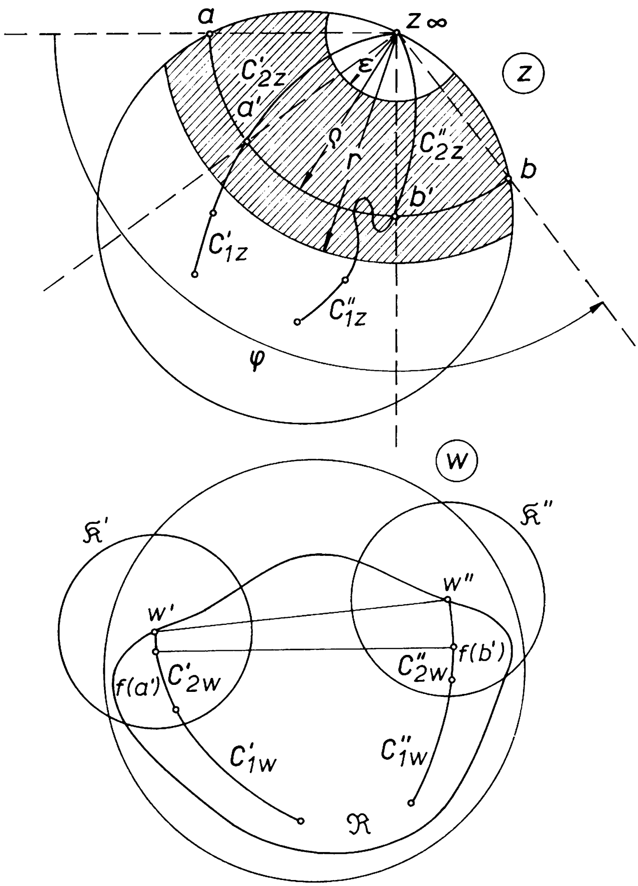


Fig. 10.5–4. Proof of the central theorem

contains infinitely many points z_n' corresponding to w_n' and C_z'' contains infinitely many points z_n'' corresponding to w_n''. Both have z_∞ as accumulation point.

Let d denote the positive number $|w' - w''|$. About w' and w'' we describe two discs $\mathfrak{K}'$ and $\mathfrak{K}''$ having a radius equal to $\frac{1}{3}d$. On C_w' we determine a closed initial part C_{1w}' such that the remainder C_{2w}' on C_w' is in $\mathfrak{K}'$. Similarly we define C_{1w}'' and C_{2w}''. The originals C_{1z}', C_{1z}'' of C_{1w}', C_{1w}'', respectively, are closed sets in $|z| < 1$ and they do not contain z_∞. Hence we can separate z_∞ from these sets by means of a circular arc C_ρ about z_∞ of radius ρ within $|z| < 1$, (fig. 10.5–4) with end points a and b on the circumference $|z| = 1$. It is clear that C_{2z}' and C_{2z}'' meet C_ρ

at points a', b' corresponding to points on C'_{2w} and C''_{2w}; we may assume that we do not meet a point of C_ζ any more if we move along C'_{2z}, C''_{2z} from a', b' resp. to z_∞. It follows that

$$\tfrac{1}{3}d \leqq |f(b')-f(a')| = \left|\int_{a'}^{b'} f'(\zeta)d\zeta\right|,$$

where ζ is on C_ρ. Put $\zeta = z_\infty + \rho e^{i\theta}$. Since the angle φ' between the half rays issuing from z_∞ and passing through a' and b' does not exceed the angle φ between the half rays issuing from z_∞ and passing through a and b, we have

$$\left|\int_{a'}^{b'} f(\zeta)\,d\zeta\right| = \left|i\rho\int_0^{\varphi'} e^{i\theta} f'(\zeta)d\theta\right| \leqq \rho\int_0^{\varphi'} |f'(\zeta)|d\theta \leqq \rho\int_0^{\varphi} |f'(\zeta)|\,d\theta.$$

Hence

$$\tfrac{1}{3}d \leqq \rho\int_0^{\varphi} |f'(\zeta)|d\theta. \qquad (10.5\text{–}2)$$

This inequality is valid if ρ remains under a certain number r.

Now we apply *Schwarz's inequality*, stating

If $f(t)$ and $g(t)$ are complex continuous functions of a real variable t varying throughout an interval $a \leqq t \leqq b$ and if we put

$$(f, g) = \int_a^b f(t)\bar{g}(t)dt, \qquad (10.5\text{–}3)$$

then

$$\boxed{|(f, g)|^2 \leqq (f, f)(g, g)} \qquad (10.5\text{–}4)$$

Notice that $(f, f) \geqq 0$ and $(f, g) = \overline{(g, f)}$. Further $(f_1 + f_2, g) = (f_1, g) + (f_2, g)$, $(\beta f, g) = \beta(f, g)$ if β is a constant.

The inequality (10.5–4) is trivial if $(g, g) = 0$. Hence we assume that $(g, g) > 0$, i.e., g does not vanish identically. We have

$$0 \leqq (f-\beta g, f-\beta g) = (f, f) - \bar{\beta}(f, g) - \beta\overline{(f, g)} + \beta\bar{\beta}(g, g).$$

Taking $\beta = (f, g)/(g, g)$ we obtain the desired result.

If we replace in Schwarz's inequality f by $f'(\zeta)$, where t is replaced by θ, and g by 1, we find from (10.5–2)

$$\frac{d^2}{9\rho} \leqq \rho\left(\int_0^{\varphi} |f'(\zeta)|d\theta\right)^2 \leqq \rho\int_0^{\varphi} |f'(\zeta)|^2 d\theta \int_0^{\varphi} d\theta$$
$$< 2\pi\rho\int_0^{\varphi} |f'(\zeta)|^2\, d\theta.$$

Integrating from $\rho = \varepsilon$ to $\rho = r$, where $0 < \varepsilon < r$, we get

$$\frac{d^2}{9}\log\frac{r}{\varepsilon} \leqq 2\pi\int_\varepsilon^r \int_0^{\varphi} |f'(\zeta)|^2 \rho\, d\rho\, d\theta. \qquad (10.5\text{–}5)$$

If ρ varies from ε to r the arc C_ρ describes a region $\mathfrak{B}_z$ whose image $\mathfrak{B}_w$ is in $\mathfrak{R}_w$ and hence in $|w| < 1$. Therefore its area is $\leqq \pi$ and we obtain the inequality

$$\iint_{\mathfrak{B}_w} du\,dv \leqq \pi.$$

But

$$\iint_{\mathfrak{B}_w} du\,dv = \iint_{\mathfrak{B}_z} \frac{\partial(u, v)}{\partial(x, y)}\,dx\,dy = \iint_{\mathfrak{B}_z} (u_x^2 + u_y^2)dx\,dy$$

where we have applied the Cauchy-Riemann equations (1.3–7). Combining this with (10.5–5) we find

$$\log \frac{r}{\varepsilon} \leqq \frac{18\pi^2}{d^2},$$

or

$$\log \varepsilon \geqq \log r - \frac{18\pi^2}{d^2}.$$

By making $\varepsilon \to 0$ we find a contradiction. Our conclusion must be that the sequence $w_1 = f(z_1)$, $w_2 = f(z_2), \ldots$, tends to a point w_∞ on the boundary of $\mathfrak{R}_w$.

The uniqueness is readily proved. Let $z_1', z_2', \ldots$, and $z_1'', z_2'', \ldots$, denote two sequences tending to z_∞ on $|z| = 1$. Their corresponding sequences are $w_1', w_2', \ldots$, and $w_1'', w_2'', \ldots$. We form the combined sequence $z_1', z_1'', z_2', z_2'', \ldots$. The corresponding sequence $w_1', w_1'', w_2', w_2'', \ldots$, converges to a unique point w_∞ and so do the above subsequences. We are now in the situation described in section 10.5.3 and the truth of the theorem has been established.

10.5.6 – Boundaries containing analytic arcs

A particular case which often occurs in practical application is the following: Assume that $\mathfrak{R}_z^*$ is a simply connected region in the z-plane which is mapped by $w = f(z)$ onto the open disc $|w| < 1$ in the w-plane and that C_z is an analytic arc (section 10.1.4) on the boundary of $\mathfrak{R}_z^*$. In addition we assume that C_z can be imbedded in a region $\mathfrak{R}_z$ in which symmetry with respect to C_z is defined. Let $\mathfrak{a}$ denote a subset of C_z with the property described at the end of section 10.1.4.

We consider a sequence $z_0, z_1, z_2, \ldots$ of points of $\mathfrak{R}_z$ which tends to $\mathfrak{a}$, i.e., all accumulation points of the sequence are on $\mathfrak{a}$. Then we contend that the images $f(z_0), f(z_1), f(z_2), \ldots$, tend to the circumference $|w| = 1$. If this assertion were not true we could find a subsequence $f(z_{n_0}), f(z_{n_1}), f(z_{n_2}), \ldots$, which converges to a point w_∞ belonging to $|w| < 1$. Since the mapping is one-to-one and bicontinuous the corresponding sequence $z_{n_0}, z_{n_1}, z_{n_2}, \ldots$ would converge to the point $z_\infty = \check{f}(w_\infty)$ in $\mathfrak{R}_z$, contrary to the assumption.

In view of the theorem of section 10.1.4 we may conclude that $f(z)$ *remains regular at every point of* $\mathfrak{a}$ *and that the correspondence as given by* $f(z)$ between $\mathfrak{a}$ and points on $|w| = 1$ is one-to-one. For, assume that $f(z)$ would take the same value at two points of $\mathfrak{a}$. Then $f(z)$ would have to map disjoint neighbourhoods onto open sets which have a point on $|w| = 1$ in common but no common points inside the unit circle.

As a particular instance we mention the following theorem

Let $\mathfrak{R}_w$ *denote the interior of a polygon whose boundary consists of circular arcs with vertices* $b_1, \ldots, b_n$. *If* $f(z)$ *maps* $|z| < 1$ *onto* $\mathfrak{R}_w$, *then* $f(z)$ *is still continuous on* $|z| = 1$ *and there is a one-to-one correspondence between the circumference* $|z| = 1$ *and the boundary of* $\mathfrak{R}_w$. *At every point of* $|z| = 1$ *which does not correspond to a vertex the mapping function is regular and so is its inverse.*

This latter theorem which is of great practical value can be proved in a simple direct way, only with the aid of the Riemann mapping theorem of section 10.5.2. For the sake of convenience we make the additional assumption that *the region is either in the interior or in the exterior of a circle to which a side belongs.* In our practical applications this condition is always fulfilled, (fig. 10.5–5).

There is a fractional linear transformation of the w-plane which maps any side $\mathfrak{a}$ onto the line segment between -1 and $+1$ and maps $\mathfrak{R}_w$ onto a region $\mathfrak{R}_1$ in the upper half of the w-plane. Let $\overline{\mathfrak{R}}_1$ denote the region symmetric with $\mathfrak{R}_1$ with respect to the real axis. Since $\mathfrak{R}_1$ is simply connected its complement with respect to the extended plane is connected (section 9.1.3). It is easy to see that the complement of the region $\mathfrak{R}^*$ consisting of $\mathfrak{R}_1$, $\overline{\mathfrak{R}}_1$ and the segment between -1 and $+1$ is also connected. Hence $\mathfrak{R}^*$ is simply connected and we can map it univalently onto the interior of a unit circle in a z_1-plane by means of a function $z_1 = \varphi(w)$, such that $\varphi(0) = 0$ $\varphi'(0) > 0$. Since this mapping function is unique and since the same mapping is accomplished by the function $\overline{\varphi(\overline{w})}$ it follows that $\varphi(w) = \overline{\varphi(\overline{w})}$ and hence $\varphi(w)$ is real for real w. Thus φ provides a univalent mapping of $\mathfrak{R}_1$ onto the upper half of the unit circle. In addition this function is univalent and continuous on the region $\mathfrak{R}_1$ extended by the segment between -1 and $+1$, as is its inverse. A transformation of the type (10.2–18), viz.,

$$z_2 = \left(\frac{z_1+1}{z_1-1}\right)^2$$

transforms the upper half of the circle in the z_1-plane onto the upper half

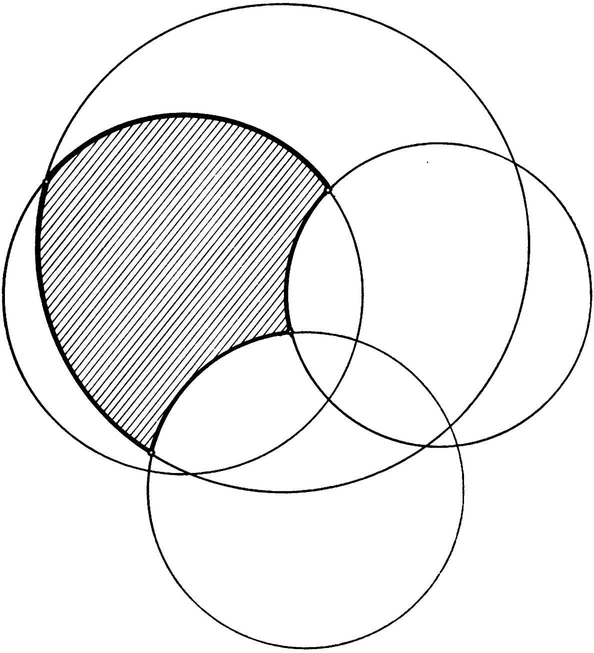


Fig. 10.5–5. Region bounded by circular arcs

of the z_2-plane, such that the segment between -1 and $+1$ corresponds to the positive real axis. Finally we can map the upper half of the z_2-plane onto the interior of the unit disc in the z-plane by means of a suitable fractional linear transformation. Combining all these auxiliary transformations and adjusting constants we can find a function $g(w)$ mapping $\mathfrak{R}_w$ onto the unit disc $|z| < 1$ such that for given w_0 we have $g(w_0) = 0$, $g'(w_0) > 0$. This function is unique and it is continuous and univalent on the side $\mathfrak{a}$ and so is its inverse. Since this side is arbitrarily chosen the same statement holds for the whole boundary of $\mathfrak{R}$ with the possible exception of the vertices.

We investigate the behaviour of $g(w)$ at the vertices. Let two sides meet at a vertex b and let $\alpha\pi \neq 0$ be the interior angle of the two arcs. Our assumption about $\mathfrak{R}_w$ implies $\alpha \leqq 1$ and $\alpha = 1$ is impossible, unless the two arcs belong to the same circle. But then the continuity of $g(w)$ at $w = b$ is obvious. By a suitable linear fractional transformation of the w-plane we map b onto the origin, one side onto part of the positive axis and the other into part of the ray $w = re^{i\alpha}$. Then $\mathfrak{R}_w$ is carried into a part of the angular region $0 < \arg w < \alpha\pi$. Now we apply the transformation

$t^\alpha = w$, (section 10.2.2). Thus we have mapped $\mathfrak{R}_w$ onto a part of the upper t-plane such that a segment of the real t-axis, including $t = 0$ is on the boundary. We can proceed as before and conclude that $g(w)$ is still continuous at b.

If $\alpha = 0$, a linear fractional transformation is used to map into a strip with the image of one side on the real axis and the image of the other on the line $\operatorname{Im} w = \pi$, whilst b corresponds to the point $-\infty$ of the real axis. We apply the mapping $w = \log t$, (section 10.2.6), and proceed as before. This concludes the proof of the theorem.

CHAPTER 11

UNIVALENT FUNCTIONS

11.1 – Preliminary lemmas

11.1.1 – AREA ENCLOSED BY A CONTOUR

We consider a function $f(z)$ which is regular at every point of the circumference $|z| = r$ and univalent. Then the image of this circle as given by the function is a closed simple curve, a contour, (section 1.2.5). The equation of this curve is

$$w = Re^{i\Phi}$$

and by an elementary formula, the area enclosed by it is

$$A_r = \tfrac{1}{2}\int_{C_r} R^2\, d\Phi, \tag{11.1–1}$$

the integration being performed along C_r, the image of $|z| = r$, percorsed in the counter clockwise sense. This may also be written as

$$A_r = \tfrac{1}{2}\int_0^{2\pi} R^2 \frac{\partial\Phi}{\partial\theta}\, d\theta. \tag{11.1–2}$$

From (9.2–16) we deduce

$$R^2 \frac{\partial\Phi}{\partial\theta} = rR\frac{\partial R}{\partial r} = \tfrac{1}{2}r\frac{\partial R^2}{\partial r}.$$

Hence

$$A_r = \tfrac{1}{4}r\int_0^{2\pi} \frac{\partial R^2}{\partial r}\, d\theta. \tag{11.1–3}$$

11.1.2 – THE INTERIOR AREA THEOREM

We wish to apply the formula (11.1–3) to the following problem. Assume that $f(z)$ is holomorphic and univalent throughout the region $|z| < 1$. What can be said about the area of the image of this region as given by $f(z)$?

The Taylor expansion is

$$f(z) = \sum_{\nu=0}^{\infty} a_\nu z^\nu, \tag{11.1–4}$$

valid throughout the interior of the unit circle. Since nothing is claimed about $f(z)$ on the boundary, it is natural to consider first the image of $|z| = r$, with $0 < r < 1$. The area enclosed by this curve can be evaluated, for we have

$$R^2 = f(re^{i\theta})\overline{f(re^{i\theta})} = \sum_{\nu=0}^{\infty} |a_\nu|^2 r^{2\nu} + (\exp ik\theta),$$

where $(\exp ik\theta)$ collects all terms involving a factor of the type $\exp ik\theta$, with $k \neq 0$. It follows that

$$r\frac{\partial R^2}{\partial r} = \sum_{\nu=0}^{\infty} 2\nu|a_\nu|^2 r^{2\nu} + (\exp ik\theta).$$

Hence

$$A_r = \pi \sum_{\nu=0}^{\infty} \nu|a_\nu|^2 r^{2\nu}. \tag{11.1–5}$$

Now two cases are possible. First we assume that A_r is bounded for $0 < r < 1$. Let M denote an upper bound. Then certainly

$$\pi \sum_{\nu=0}^{n} \nu|a_\nu|^2 r^{2\nu} < M, \tag{11.1–6}$$

where n is a fixed, but arbitrary integer. Since all terms in the series (11.1–5) are not negative the expression on the left of (11.1–6) does not decrease with r and has a limit as $r \to 1$. Hence

$$\pi \sum_{\nu=0}^{n} \nu|a_\nu|^2 \leqq M.$$

If $n \to \infty$ we obtain

$$A = \lim_{r\to 1} A_r = \pi \sum_{\nu=0}^{\infty} \nu|a_\nu|^2. \tag{11.1–7}$$

This number is called the *interior area* of the image. Secondly we assume that A_r is not bounded. Then the series on the right of (11.1–6) is divergent and the interior area has an infinite value.

11.1.3 – The exterior area theorem

A very interesting and important result is obtained if we consider the function

$$g(z) = \frac{1}{z} + \sum_{\nu=0}^{\infty} b_\nu z^\nu, \tag{11.1–8}$$

supposed to be holomorphic and univalent throughout the region

$0 < |z| < 1$. It is distinguished from the preceding function by the addition of a simple pole (with residue 1). Writing again $g(z) = Re^{i\Phi}$, we now have

$$R^2 = \frac{1}{r^2} + \sum_{\nu=0}^{\infty} |b_\nu|^2 r^{2\nu} + (\exp ik\theta)$$

whence

$$r\frac{\partial R^2}{\partial r} = -\frac{2}{r^2} + \sum_{\nu=0}^{\infty} 2\nu|b_\nu|^2 r^{2\nu} + (\exp ik\theta), \qquad 0 < r < 1.$$

It follows that

$$A_r = -\frac{\pi}{r^2} + \pi \sum_{\nu=0}^{\infty} \nu|b_\nu|^2 r^{2\nu}, \tag{11.1–9}$$

where A_r denotes again the integral (11.1–3).

In this case, however, A_r is negative, i.e., $-A_r$ is the area of the region complementary to the image of the interior of $|z| = r$. This is a consequence of the fact that C_r is percorsed in the clockwise sense. This can be verified by the following argument. Let z_0 denote a point in the region $|z| < 1$ outside $|z| = r$. Since $g(z)$ is univalent the function $g(z) - g(z_0)$ has no zero inside or on $|z| = r$. But it has a simple pole inside this circumference and so we have (section 3.8.2)

$$\frac{1}{2\pi i}\int_{|\zeta|=r} \frac{g'(\zeta)}{g(\zeta)-g(z_0)}\, d\zeta = -1.$$

This means that the winding number (section 3.10.1) of C_r with respect to the origin is -1.

When proceeding we deduce from (11.1–9) that

$$\sum_{\nu=0}^{\infty} \nu|b_\nu|^2 r^{2\nu+2} < 1. \tag{11.1–10}$$

As in the previous section we see that it is allowed to let $r \to 1$ and thus we obtain *Gronwall's theorem*

If $g(z)$ *is holomorphic and univalent in the region* $0 < |z| < 1$ *and has the expansion*

$$g(z) = \frac{1}{z} + \sum_{\nu=0}^{\infty} b_\nu z^\nu,$$

then

$$\boxed{\sum_{\nu=0}^{\infty} \nu|b_\nu|^2 \leq 1.} \tag{11.1–11}$$

It should be noticed that this theorem does not state anything about the coefficient b_0. This is clear, for changing b_0 means a translation of the image of $|z| < 1$ and the areas considered above do not alter.

We wish to derive an important consequence from (11.1–11) namely

$$\boxed{|b_1| \leqq 1} \tag{11.1–12}$$

Equality occurs only if $b_1 = \eta$, $|\eta| = 1$, and $b_2 = b_3 = \ldots = 0$ Then $g(z)$ is the function

$$g(z) = \frac{1}{z} + \eta z$$

and thus it appears that the inequality (11.1–12) is sharp.

11.1.4 – Prawitz's lemma

Gronwall's theorem is an interesting illustration of the fact that simple geometric considerations give rise to function theoretic results of importance. There is another geometric approach which is based on a very simple idea due to Prawitz. This is embodied in the following lemma

Let C denote a piecewise analytic simple closed curve, surrounding the origin, given by the equation

$$z = \mathrm{Re}^{i\Phi}.$$

Let $F(R)$ be a non-negative monotonous function. Then

$$\int_C F(R)d\Phi \geqq 0, \tag{11.1–13}$$

the integration being performed along C in the counter clockwise sense.

Consider an angular element issuing from the origin with width $|\Delta\Phi|$ and inside this element a half ray which does not touch C. This is possible, for C sends only a finite number of tangents through O, (section 10.1.4). The number of intersections with C is odd. Denote the distances of these points to O successively by $R_1, \ldots, R_{2k+1}$. The contribution of this angular element to the Riemann sum approximating the integral (11.1–13) is

$$\begin{aligned}
&(F(R_1)-F(R_2)+\ldots-F(R_{2k})+F(R_{2k+1}))|\Delta\Phi| \\
&= \sum_{\kappa=1}^{k}(F(R_{2\kappa-1})-F(R_{2\kappa}))+F(R_{2k+1}))||\Delta\Phi| \\
&= (F(R_1)+\sum_{\kappa=1}^{k}(-F(R_{2\kappa})+F(R_{2\kappa+1})))|\Delta\Phi|.
\end{aligned} \tag{11.1–14}$$

We have taken the +sign at the points where C intersects the half ray from the left to the right and the −sign in the contrary case. It is now clear that the sum is $\geqq F(R_{2k+1})$ if $F(R)$ is non-increasing and $\geqq F(R_1)$ if $F(R)$ is non-decreasing. In either case the contribution to the Riemann sum is not negative. The truth of the statement follows easily.

Next we assume that $f(z)$ is holomorphic throughout $|z| < 1$ and univalent. We put

$$f(re^{i\theta}) = Re^{i\Phi}, \qquad 0 < r < 1. \tag{11.1–15}$$

If $C = C_r$ is the image of $|z| = r$ as given by $f(z)$, we have

$$\int_0^{2\pi} F(R)\frac{\partial \Phi}{\partial \theta}\,d\theta \geqq 0. \tag{11.1–16}$$

Two particular cases are of importance. First we take

$$F(R) = R^\lambda, \tag{11.1–17}$$

where λ is real and different from zero. Then, on account of the Cauchy-Riemann equations (9.2–16)

$$0 \leqq \int_0^{2\pi} F(R)\frac{r}{R}\frac{\partial R}{\partial r}\,d\theta = r\int_0^{2\pi} R^{\lambda-1}\frac{\partial R}{\partial r}\,d\theta$$

or

$$\frac{1}{\lambda}\frac{d}{dr}\int_0^{2\pi} R^\lambda d\theta \geqq 0. \tag{11.1–18}$$

This result implies $A_r \geqq 0$, where A_r is defined in (11.1–3), if we take $\lambda = 2$.

Secondly we take

$$F(R) = R_0^\lambda - R^\lambda, \tag{11.1–19}$$

where R_0 is the radius of a circle around O which includes C_r in its interior. In view of (11.1–16) we have

$$\int_0^{2\pi} R_0^\lambda \frac{\partial \Phi}{\partial \theta}\,d\theta \geqq \int_0^{2\pi} R^\lambda \frac{\partial \Phi}{\partial \theta}\,d\theta. \tag{11.1–20}$$

But

$$\int_0^{2\pi} \frac{\partial \Phi}{\partial \theta}\,d\theta = \Phi|_{\theta=0}^{\theta=2\pi} = 2\pi$$

and performing the operations which lead to (11.1–18) we have

$$\frac{r}{\lambda}\frac{d}{dr}\int_0^{2\pi} R^\lambda d\theta \leqq 2\pi R_0^\lambda. \tag{11.1–21}$$

11.1.5 – A mean value theorem

Let again $f(z)$ be holomorphic and univalent throughout $|z| < 1$. We make the additional assumption that $f(0) = 0$. Let $M(r)$ denote the maximum of $|f(z)|$ if $|z| = r$. Next we take $R_0 = M(r)+\varepsilon$, $\varepsilon > 0$. Assuming that $\lambda > 0$ we find from (11.1–21)

$$\frac{d}{dr}\int_0^{2\pi} |f(re^{i\theta})|^\lambda d\theta \leq 2\pi\lambda \frac{(M(r)+\varepsilon)^\lambda}{r}.$$

Since this is valid for every $\varepsilon > 0$ we may let $\varepsilon \to 0$. Integrating from 0 to r we obtain

If $f(z)$ is univalent and holomorphic within $|z| = 1$ and $f(0) = 0$, then

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^\lambda d\theta \leq \lambda \int_0^r \frac{M(\rho)^\lambda}{\rho} d\rho, \qquad 0 < r < 1,\ \lambda > 0, \qquad (11.1\text{–}22)$$

where $M(r)$ denotes the maximum of $|f(z)|$ on $|z| = r$,

From this result we may derive an estimate for the coefficients of the Taylor series

$$f(z) = \sum_{\nu=1}^{\infty} c_\nu z^\nu \qquad (11.1\text{–}23)$$

which exhibits close resemblance to Cauchy's inequality (2.18–2). In fact, as in section 2.18.1, we have

$$c_n = \frac{1}{2\pi i}\int_{C_r} \frac{f(\zeta)}{\zeta^{n+1}} d\zeta, \qquad n \geq 1,$$

where C_r is a circumference about the origin with radius $r < 1$. It follows that

$$|c_n| \leq \frac{1}{2\pi r^n}\int_0^{2\pi} |f(re^{i\theta})| d\theta$$

and taking $\lambda = 1$ in (11.1–22) we obtain

$$|c_n| r^n \leq \int_0^r \frac{M(\rho)}{\rho} d\rho, \qquad (11.1\text{–}24)$$

the desired formula.

11.1.6 – Faber's functions

In subsequent parts of this chapter we will study functions which are univalent and holomorphic throughout the interior of the unit circle. In most cases the statements about these functions obtain their simplest form if we make some not essential restrictions, namely that they have a zero (and hence the only zero) at the origin and that the value of the derivative there is equal to one. Every univalent function $f(z)$ can be reduced to a function of this kind if we replace it by

$$\frac{f(z)-f(0)}{f'(0)}. \tag{11.1–25}$$

Univalent functions, holomorphic for $|z| < 1$ and satisfying the above conditions will be referred to as *univalent normalized functions*. Their expansions in series are of the type

$$f(z) = z+a_2z^2+\ldots = z+\sum_{\nu=2}^{\infty} a_\nu z^\nu. \tag{11.1–26}$$

Every such function gives rise to a class of associated funtions of the same kind, introduced by Faber. In many considerations about univalent functions they present themselves as very useful.

In order to define them we observe that in

$$f(z^k) = z^k(1+\sum_{\nu=2}^{\infty} a_\nu z^{(\nu-1)k}) \tag{11.1–27}$$

$k = 1, 2, \ldots$ the series between brackets represents a function which is holomorphic for $|z| < 1$ and has no zero. According to section 9.1.4 we can find a holomorphic function

$$h(z) = 1+\frac{a_2}{k}z^k+\ldots,$$

such that

$$h^k(z) = 1+a_2z^k+\ldots.$$

The function

$$f_k(z) = zh(z) = z+\frac{a_2}{k}z^{k+1}+\ldots \tag{11.1–28}$$

is holomorphic for $|z| < 1$ and satisfies the equation

$$f(z^k) = (f_k(z))^k. \tag{11.1–29}$$

We will prove that this function is univalent. In fact, it vanishes at $z = 0$ only and from

$$f_k(z_1) = f_k(z_2)$$

follows

$$f(z_1^k) = f(z_2^k)$$

whence

$$z_1^k = z_2^k$$

and so

$$z_1 = \eta z_2$$

with $\eta^k = 1$. It remains to prove that $\eta = 1$. If $\eta \neq 1$ we have

$$f_k(z_1) \neq \eta f_k(z_1) = f_k(\eta z_1) = f_k(z_2)$$

and this concludes the proof of the assertion.

11.1.7 – An inequality of Golusin

The inequality (11.1–22) may be generalized for an integral including also the derivative of $f(z)$, namely

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^{\lambda} |f'(re^{i\theta})| d\theta. \tag{11.1–30}$$

It is again understood that $f(z)$ is univalent and normalized, while $\lambda > 0$ and $0 < r < 1$. For our considerations we need Schwarz's inequality (10.5–4).

We consider the function

$$f_k(z) = (f(z^k))^{1/k} = \sum_{\nu=0}^{\infty} a_{\nu,k} z^{\nu k+1}, \qquad k = 1, 2, \ldots, \tag{11.1–31}$$

which is, as we know, univalent and normalized as is $f(z)$. Differentiating we get

$$f_k'(z) = z^{k-1}(f(z^k))^{1/k-1} f'(z^k),$$

whence

$$f'(z) = z^{1/k-1}(f(z))^{1-1/k} f_k'(z^{1/k}).$$

Inserting this into (11.1–30) we find

$$\frac{r^{1/k-1}}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^{\lambda+1-1/k} |f_k'(z^{1/k})| d\theta$$

and from Schwarz's inequality (10.5–4) follows that this integral does not exceed

$$\frac{1}{r} P^{\frac{1}{2}} Q^{\frac{1}{2}}$$

with

$$P = \frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^{2\lambda+2-2/k} d\theta \tag{11.1–32}$$

and

$$Q = \frac{1}{2\pi}\int_0^{2\pi} r^{2/k} |f_k'(z^{1/k})|^2 d\theta. \tag{11.1–33}$$

Now we take k such that $\lambda+1-1/k > 0$. Then we may apply (11.1–22) and we find

$$P \leqq \left(2\lambda+2-\frac{2}{k}\right)\int_0^r \frac{(M(\rho))^{2\lambda+2-2/k}}{\rho}\,d\rho. \qquad (11.1\text{–}34)$$

For Q we find

$$Q = \sum_{\nu=0}^{\infty}(\nu k+1)^2|a_{\nu,k}|^2 r^{2\nu+2/k}$$
$$= \sum_{\nu=0}^{\infty}(\nu k+1)r^{\nu+1/k}(\nu k+1)|a_{\nu,k}|^2 r^{(\nu k+1)/k}.$$

Our next task will be estimating the expression

$$(nk+1)r^{n+1/k}.$$

This may be done as follows. If m is any positive number the function

$$x^m(1-x)$$

takes it maximums on the interval $0 < x < 1$ at $x = m/(m+1)$. Hence this maximum is

$$\frac{m^m}{(m+1)^{m+1}} = \frac{1}{m}\frac{1}{\left(1+\dfrac{1}{m}\right)^{m+1}} < \frac{1}{me}$$

whence

$$mx^m < \frac{1}{e(1-x)}, \qquad 0 < x < 1, \qquad m > 0.$$

If we take

$$m = n+\frac{1}{k}, \qquad x = r,$$

we get

$$(nk+1)r^{n+1/k} < \frac{k}{e(1-r)} \qquad (11.1\text{–}35)$$

and thus

$$Q < \frac{k}{e(1-r)}\sum_{\nu=0}^{\infty}(\nu k+1)|a_{\nu,k}|^2 r^{(\nu k+1)/k}.$$

According to (11.1–5) the sum on the right is equal to the area of the image of the disc $|z| = r^{1/2k}$ divided by π as determined by the function $f_k(z)$. This area does not exceed that of a circle with radius the maximum of $|f_k(z)|$ if $|z| = r^{1/2k}$. But this maximum is also that of $|f(z^k)|^{1/k}$

for $|z^k| = \sqrt{r}$ i.e., $M(\sqrt{r})^{1/k}$. As a consequence

$$Q < \frac{k}{e(1-r)} M(\sqrt{r})^{2/k}. \tag{11.1–36}$$

Combining this with (11.1–31) we finally have

If $f(z)$ is univalent and normalized, then for $0 < r < 1$, $\lambda > 0$,

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^\lambda |f'(re^{i\theta})| d\theta$$
$$\leq \frac{1}{r}\left(\frac{2k(1+\lambda)-2}{e}\right)^{\frac{1}{2}} \left(\frac{M(\sqrt{r})^{2/k}}{1-r}\right)^{\frac{1}{2}} \left(\int_0^r \frac{M(\rho)^{2\lambda+2-2/k}}{\rho} d\rho\right)^{\frac{1}{2}}, \tag{11.1–37}$$

where $M(r)$ denotes the maximum of $|f(z)|$ on $|z| = r$, provided the integer k is such that $\lambda+1-1/k > 0$. This result is due to Golusin.

There is another estimate of the expression on the left of (11.1–35) which deserves mention. We notice that

$$(nk+1)r^n = \frac{nk+1}{n+1} r^n(n+1) \leq \frac{nk+k}{n+1}(r^n + \ldots + r^n)$$
$$< k(1+r+\ldots+r^n) < k(1+r+\ldots) = \frac{k}{1-r}.$$

Hence

$$(nk+1)r^{n+1/k} < \frac{kr^{1/k}}{1-r}$$

and thus, reasoning as above,

$$Q < \frac{kr^{1/k}}{1-r} M(\sqrt{r})^{2/k}, \tag{11.1–38}$$

whence

$$\frac{1}{2\pi}\int_0^{2\pi} |f_k'(z^{1/k})|^2 d\theta < \frac{k}{r^{1/k}(1-r)} M(\sqrt{r})^{2/k}. \tag{11.1–39}$$

In the case $k = 1$ this result is due to Littlewood and Payley.

11.2 – Distortion theorems

11.2.1 – Bieberbach's theorem

A map of the interior of the unit circle as given by a univalent normalized function can be regarded as a distortion of the shape of this region and of all subsets of it, resulting in the shape of the corresponding images.

Any result giving limitations for the amount of this distortion is, therefore, called a *distortion theorem.*

As we shall see a basic theorem for deriving distortion theorem is the following theorem due to Bieberbach

If the function (11.1–26) *is univalent for* $|z| < 1$ *then*

$$\boxed{|a_2| \leqq 2.} \tag{11.2–1}$$

In order to prove this statement we consider an associated Faber function with $k = 2$:

$$f_2(z) = z + \tfrac{1}{2}a_2 z^3 + \dots.$$

Then

$$\frac{1}{f_2(z)} = \frac{1}{z} - \tfrac{1}{2}a_2 z + \dots$$

satisfies the conditions of Gronwall's theorem of section 11.1.3. From (11.1–12) follows

$$|\tfrac{1}{2}a_2| \leqq 1$$

and this is equivalent to (11.2–1).

An alternative proof of Bieberbach's theorem may be given by using Prawitz's method. We write

$$f(z) = Re^{i\Phi}, \qquad z = re^{i\theta}, \qquad 0 < r < 1, \tag{11.2–2}$$

and consider the function

$$\left(\frac{z}{f(z)}\right)^{\frac{1}{2}} = c_0 + c_1 z + \dots \tag{11.2–3}$$

which represents a holomorphic branch of the square root of $z/f(z)$. It is easily verified that this branch exists throughout $|z| < 1$ and is univalent. A simple calculation shows that

$$c_0 = 1, \qquad c_1 = -\tfrac{1}{2}a_2. \tag{11.2–4}$$

Next we observe that

$$R^{-1} = \frac{1}{r}\left|\frac{z}{f(z)}\right| = \frac{1}{r}\sum_{\nu=0}^{\infty} c_\nu z^\nu \sum_{\nu=0}^{\infty} \bar{c}_\nu \bar{z}^\nu$$

$$= \sum_{\nu=0}^{\infty} |c_\nu|^2 r^{2\nu-1} + (\exp ik\theta).$$

Hence

$$\int_0^{2\pi} R^{-1}\, d\theta = 2\pi \sum_{\nu=0}^{\infty} |c_\nu|^2 r^{2\nu-1}$$

and it follows from (11.1–18) that

$$-\frac{1}{r^2}+\sum_{\nu=1}^{\infty}(2\nu-1)|c_\nu|^2 r^{2\nu-2} \leqq 0$$

or

$$\sum_{\nu=1}^{\infty}(2\nu-1)|c_\nu|^2 r^{2\nu} \leqq 1.$$

By making $r \to 1$ we find

$$\sum_{\nu=1}^{\infty}(2\nu-1)|c_\nu|^2 \leqq 1.$$

In particular $|c_1| \leqq 1$ and in view of (11.2–4) we find again $|a_2| \leqq 2$.

The estimate (11.2–1) is the best possible. Equality occurs if $f(z)$ is *Koebe's function*

$$k(z) = \frac{z}{(1-z)^2}\,. \tag{11.2–5}$$

We have investigated the mapping as given by this function in section 10.2.3 and we remarked that the expansion in series is

$$k(z) = z+2z^2+\ldots = \sum_{\nu=1}^{\infty}\nu z^\nu. \tag{11.2–6}$$

This led Bieberbach to the conjecture that for all $n \geqq 2$

$$|a_n| \leqq n, \tag{11.2–7}$$

provided the function is normalized. It is one of the deepest unsolved problems in the theory of univalent functions whether (11.2–7) holds for all $n > 1$. Beyond the case $n = 2$ only the cases $n = 3$ and $n = 4$ have been established. The case $n = 3$ will be considered in section 11.5.3.

11.2.2 – The Koebe–bieberbach theorem

A theorem due to Koebe states that the image of $|z| < 1$ as given by a normalized univalent function $f(z)$ covers a circle whose radius is an universal constant. Bieberbach determined the exact value of the radius.

First we observe that $f(z)$ cannot take every complex value, for there is no one-to-one conformal mapping onto the whole z-plane. Hence there is a constant c such that $f(z) \neq c$ if $|z| < 1$. The function

$$\varphi(z) = \frac{cf(z)}{c-f(z)} = z+\left(a_2+\frac{1}{c}\right)z^2+\ldots \tag{11.2–8}$$

is again univalent and normalized. By Bieberbach's theorem we have

$$\left| a_2 + \frac{1}{c} \right| \leqq 2,$$

or

$$\frac{1}{|c|} = \left| \left(a_2 + \frac{1}{c}\right) - a_2 \right| \leqq \left| a_2 + \frac{1}{c} \right| + |a_2| \leqq 4,$$

whence

$$|c| \geqq \tfrac{1}{4}. \tag{11.2–9}$$

Thus we have the so-called $\frac{1}{4}$-*theorem*

A univalent normalized function takes for $|z| < 1$ *all values* c *with* $c < \frac{1}{4}$.

Otherwise stated

The image of the interior of the unit circle as given by a normalized univalent function covers an open disc about the origin with radius $\frac{1}{4}$.

This result is sharp, for Koebe's function does not take the value $-\frac{1}{4}$ inside $|z| = 1$. Hence the constant $\frac{1}{4}$ cannot be replaced by a larger constant.

An easily proved consequence is the following theorem

If two points are situated on a line through the origin and separated by it and are not covered by the image of the interior of the unit circle, then at least one has a distance from the origin not less than $\frac{1}{2}$.

Suppose that the function $f(z)$ does not take the values a and b with $a \neq b$, $\arg a - \arg b = \pi$. Then, evidently, $a \neq 0$ and the function

$$\frac{af(z)}{a - f(z)}$$

is also univalent and normalized. Since it does not take the value $ab/(a-b)$ we have

$$\frac{|ab|}{|a-b|} \geqq \tfrac{1}{4},$$

or

$$\left| \frac{1}{a} - \frac{1}{b} \right| = \left| \frac{1}{a} \right| + \left| \frac{1}{b} \right| \leqq 4.$$

Without loss of generality we may assume that $|b| \leqq |a|$. Then $2/|a| \leqq 4$ i.e., $|a| \geqq \frac{1}{2}$. This proves the assertion.

11.2.3 – A covering theorem for convex functions

In the case that the image of the unit circle as given by $f(z)$ is convex (section 2.2.1) we can improve the $\frac{1}{4}$-theorem of the previous section. In this case we shall say that the function $f(z)$ is *convex*.

If the normalized univalent function $f(z)$ *maps* $|z| < 1$ *onto a convex region, then this region covers an open disc about the origin with radius* $\frac{1}{2}$.

A very short and elegant proof of this theorem has recently been given by T. H. Mac Gregor. He showed that the theorem is a direct consequence of the $\frac{1}{4}$-theorem.

Let $\mathfrak{R}$ denote the image of $|z| < 1$ as given by the normalized univalent function $f(z)$ and suppose that $\mathfrak{R}$ is convex. Let c be a number not taken by $f(z)$. We introduce the auxiliary functions

$$g(z) = (f(z)-c)^2, \qquad h(z) = \frac{c^2-g(z)}{2c}. \tag{11.2–10}$$

It is clear that $g(z)$ has no zero in $|z| < 1$. Moreover, $g(z)$ is univalent. For, let z_1 and z_2 denote two distinct points of the unit disc. Then

$$\begin{aligned} g(z_1)-g(z_2) &= (f(z_1)-c)^2-(f(z_2)-c)^2 \\ &= (f(z_1)-f(z_2))(f(z_1)+f(z_2)-2c). \end{aligned}$$

Since $f(z)$ is univalent we have $f(z_1) \neq f(z_2)$. On the other hand, $\frac{1}{2}(f(z_1)+ +f(z_2))$ belongs to $\mathfrak{R}$, for $\mathfrak{R}$ is convex and since c does not belong to $\mathfrak{R}$ we see that also $f(z_1)+f(z_2)-2c \neq 0$.

Now $g(z) = (-c+z+\ldots)^2 = c^2-2cz+\ldots$, whence $h(z) = z+\ldots$ and $h(z)$ is again univalent. It does not take the value $\frac{1}{2}c$, and by the $\frac{1}{4}$-theorem we have $\frac{1}{2}|c| \geqq \frac{1}{4}$ or $|c| \geqq \frac{1}{2}$, as asserted.

The function

$$f(z) = \frac{z}{1-z}$$

shows that this result is sharp, for this function maps $|z| < 1$ onto the half plane $\operatorname{Re} w > -\frac{1}{2}$.

11.2.4 – General distortion theorems

In this section we shall develop an interesting further group of inequalities for univalent functions as a direct consequence of Bieberbach's theorem (11.2–1). First we wish to prove the following lemma

If $0 \leqq x < 1$, *then*

$$\left| \frac{f''(x)}{f'(x)} - \frac{2x}{1-x^2} \right| \leqq \frac{4}{1-x^2}, \tag{11.2–11}$$

where $f(z)$ *denotes a normalized univalent function.*

The function

$$\frac{z+x}{1+xz},$$

considered as a function of z is univalent for $|z| < 1$ and its modulus is less than unity, for $(1+xz)(1+x\bar{z})-(z+x)(\bar{z}+x) = (1-x^2)(1-z\bar{z}) > 0$. Hence

$$g(z) = f\left(\frac{z+x}{1+xz}\right) = b_0+b_1 z+b_2 z^2+\ldots \qquad (11.2\text{–}12)$$

is univalent and holomorphic for $|z| < 1$. By elementary computation we have

$$g'(z) = f'\left(\frac{z+x}{1+xz}\right)\frac{1-x^2}{(1+xz)^2},$$

$$g''(z) = f''\left(\frac{z+x}{1+xz}\right)\frac{(1-x^2)^2}{(1+xz)^4} - f'\left(\frac{z+x}{1+xz}\right)\cdot 2x\,\frac{1-x^2}{(1+xz)^3}.$$

Hence

$$b_1 = g'(0) = f'(x)(1-x^2), \qquad (11.2\text{–}13)$$

$$b_2 = \tfrac{1}{2}g''(0) = \tfrac{1}{2}(f''(x)(1-x^2)^2-2xf'(x)(1-x^2)).$$

The function

$$\frac{g(z)-b_0}{b_1} = z+\frac{b_2}{b_1}z^2+\ldots$$

is again univalent and normalized. From Bieberbach's theorem we may infer that

$$\frac{1}{2}\left|\frac{f''(x)}{f'(x)}(1-x^2)-2x\right| = \left|\frac{b_2}{b_1}\right| \leqq 2$$

and (11.2–11) follows at once.

An easily deduced consequence are the inequalities

$$\boxed{\frac{1-r}{(1+r)^3} \leqq |f'(z)| \leqq \frac{1+r}{(1-r)^3},} \qquad |z| \leqq r < 1. \qquad (11.2\text{–}14)$$

Thus we have a limitation of the stretching of the mapping at any point within the unit circle. It is usually referred to as *Koebe's distortion theorem.*

Since $f'(z)$ has no zeros $1/f'(z)$ is holomorphic. In view of the maximum modulus principle it is sufficient to prove (11.2–14) for $|z| = r$.

We may even take $z = r$, for the general case can be reduced to this one by considering the function

$$\frac{f(\eta z)}{\eta},$$

with suitably chosen η such that $|\eta| = 1$.

From the lemma we deduce, assuming $0 \leqq r < 1$,

$$|\log|f'(r)|+\log(1-r^2)| = \left|\operatorname{Re}\int_0^r \frac{f''(x)}{f'(x)}\,dx+\log(1-r^2)\right|$$

$$\leqq \left|\int_0^r \frac{f''(x)}{f'(x)}\,dx-\int_0^r \frac{2x}{1-x^2}\,dx\right| \leqq \int_0^r \frac{4}{1-x^2}\,dx = 2\log\frac{1+r}{1-r}.$$

Hence

$$\log|f'(r)| \leqq -\log(1-r^2)+2\log\frac{1+r}{1-r} = \log\frac{1+r}{(1-r)^3},$$

$$\log|f'(r)| \geqq -\log(1-r^2)-2\log\frac{1+r}{1-r} = \log\frac{1-r}{(1+r)^3},$$

and this concludes the proof of (11.2–14).

Next we shall prove

$$\boxed{\frac{r}{(1+r)^2} \leqq |f(z)| \leqq \frac{r}{(1-r)^2},} \qquad |z| \leqq r < 1. \qquad (11.2\text{–}15)$$

From the previous theorem we deduce

$$|f(r)| \leqq \int_0^r |f'(x)|dx \leqq \int_0^r \frac{1+x}{(1-x)^3}\,dx = \frac{r}{(1-r)^2} \qquad (11.2\text{–}16)$$

and as in the above proof it follows that the inequality on the right is valid for all z with $|z| \leqq r$.

The proof of the inequality on the left is not easy. As we remarked above it suffices to prove it for $z = -r$, $0 < r < 1$. Let w and z be related by Koebe's function

$$w = \frac{z}{(1-z)^2} = k(z). \qquad (11.2\text{–}17)$$

The inverse relation may be denoted by

$$z = \varphi(w). \qquad (11.2\text{–}18)$$

The function φ maps the w-plane slit from $-\frac{1}{4}$ to $-\infty$ onto the region $|z| < 1$. If p is a positive constant < 1, the function

$$\psi(w) = \varphi(pk(z)) \tag{11.2–19}$$

is holomorphic and univalent for $|w| < 1$. In fact $pk(z)$ does not take values $\leqq -\frac{1}{4}p$ and, therefore, exceeds $-\frac{1}{4}$. The derivative of Koebe's function is

$$k'(z) = \frac{1+z}{(1-z)^3},$$

and is positive if $-1 < z \leqq 0$. Hence Koebe's function increases and tends to $-\frac{1}{4}$ as z tends to -1 from the right. Its value for $z = -r$, $0 < r < 1$, is

$$-\frac{r}{(1+r)^2}.$$

Now we take

$$p = \frac{4r}{(1+r)^2}. \tag{11.2–20}$$

It follows that if z tends to -1 from the right, then $\psi(w)$ tends to $\varphi(pk(-1)) = \varphi(-\frac{1}{4}p) = \varphi(k(-r)) = -r$.

Continuing we introduce the function

$$\chi(z) = \frac{f(\psi(z))}{p} = z + \ldots$$

which is univalent and holomorphic for $|z| < 1$. If $0 < r < 1$ then

$$\frac{\chi(rz)}{r}$$

is also univalent and normalized. As a consequence of the one quarter theorem of section 11.2.2 $\chi(rz)$ takes all values rc with $|c| < \frac{1}{4}$. Hence $\chi(z)$ takes all values c with $|c| < \frac{1}{4}r$, if $|z| < r$. No such value can be taken on the circumference $|z| = r$, for the function is univalent. Thus we may infer that

$$|\chi(z) \geqq \frac{|z|}{4}, \qquad 0 < |z| < 1.$$

Let $0 < \rho < 1$. Then

$$|\chi(z)| > \tfrac{1}{4}\rho$$

if $-1 < z < -\rho$. If z tends to -1 from the right, then $x(z)$ tends to $f(-r)/p$. It follows that

$$\frac{|f(-r)|}{p} \geqq \tfrac{1}{4}\rho$$

and by making $\rho \to 1$ we find

$$f(-r) \geqq \tfrac{1}{4}p = \frac{r}{(1+r)^2}.$$

Thus we proved the statement on the left of (11.2–15).

An alternative statement is

$$\boxed{\frac{1}{(1+r)^2} \leqq \left|\frac{f(z)}{z}\right| \leqq \frac{1}{(1-r)^2},} \qquad |z| \leqq r < 1. \qquad (11.2\text{–}21)$$

This is clear for $|z| = r$ and it follows for $|z| < r$ from the maximum modulus principle since $f(z)/z$ and $z/f(z)$ are both holomorphic for $|z| < 1$.

We wish to apply this latter statement to the function

$$\frac{g(z)-b_0}{b_1},$$

where $g(z)$ is defined in (11.2–12) in which x is replaced by r. In view of (11.2–21) we readily find

$$\frac{|b_1|}{(1+r)^2} \leqq \left|\frac{f\left(\dfrac{r+z}{1+rz}\right) - f(r)}{z}\right| \leqq \frac{|b_1|}{(1-r)^2}, \qquad |z| \leqq r < 1.$$

Next we replace z by $-r$ and insert the expression (11.2–13) for b_1. Thus we get

$$|f'(r)|\frac{1-r}{1+r} \leqq \left|\frac{f(r)}{r}\right| \leqq |f'(r)|\frac{1+r}{1-r}$$

or

$$\frac{1-r}{1+r} \leqq \left|\frac{rf'(r)}{f(r)}\right| \leqq \frac{1+r}{1-r},$$

whence, by the usual arguments

$$\boxed{\frac{1-r}{1+r} \leqq \left|\frac{zf'(z)}{f(z)}\right| \leqq \frac{1+r}{1-r},} \qquad |z| \leqq r < 1. \qquad (11.2\text{–}22)$$

All these inequalities are sharp. Equality occurs for Koebe's function.

The inequalities (11.2–14), Koebe's distortion theorem, are a limitation for the absolute value of $f'(z)$ inside the unit circle. Bieberbach has also obtained a limitation for $\arg f'(z)$. Taking into account (11.1–16) we have

$$|\arg f'(z)| = |\mathrm{Im}\log f'(z)| = |\mathrm{Im}(\log f'(z)+\log(1-r^2))|$$
$$\leq |\log f'(r)+\log(1-r^2)| = \left|\int_0^r \left(\frac{f''(x)}{f'(x)} - \frac{2x}{1-x^2}\right) dx\right|$$
$$\leq \int_0^r \left|\frac{f''(x)}{f'(x)} - \frac{2x}{1-x^2}\right| dx \leq \int_0^r \frac{4}{1-x^2}\, dx = 2\log\frac{1+r}{1-r},$$

where the argument is that branch which takes the value 0 at $z = 0$. It follows that

$$\boxed{|\arg f'(z)| \leq 2\log\frac{1+r}{1-r},} \qquad 0 < |z| = r < 1. \qquad (11.2\text{–}23)$$

This is *Bieberbach's rotation theorem.*

In contrast with Koebe's theorem this result is not sharp. In fact, the function

$$f(z) = \frac{1}{2(1+i)}\left(\exp\left((1+i)\log\frac{1+z}{1-z}\right) - 1\right) = z+(1+i)z^2+\ldots$$

yields a counter example, for

$$f'(z) = \frac{\exp(1+i)\log\dfrac{1+z}{1-z}}{1-z^2}$$

and along the positive real axis

$$\mathrm{Im}\log f'(x) = \arg f'(x) = \log\frac{1+x}{1-x}.$$

The sharp form of the rotation theorem has been obtained by Golusin, applying powerful methods to be dealt with in a subsequent part of this chapter.

11.2.5 – Koebe's distortion theorem for general regions

Up to now we considered functions which are univalent throughout a circular region. Koebe has also proved a distortion theorem valid in a general region in the open z-plane. It states

Let $\mathfrak{R}$ denote a region in the open z-plane and $\mathfrak{C}$ a closed set included in $\mathfrak{R}$. Then there is a positive constant M depending only on $\mathfrak{R}$ and $\mathfrak{C}$ such that for every function $f(z)$ univalent and holomorphic in $\mathfrak{R}$ and for arbitrary values z_1, z_2 belonging to $\mathfrak{C}$ the inequalities

$$\frac{1}{M} \leq \left| \frac{f'(z_1)}{f'(z_2)} \right| \leq M \tag{11.2–24}$$

hold.

We may assume that $\mathfrak{C}$ is the closure of a region. In fact, any two points of $\mathfrak{C}$ can be connected by an arc included in $\mathfrak{R}$ and the points belonging to all these arcs constitute a closed set $\mathfrak{C}'$ again included in $\mathfrak{R}$. Any point of this set is the centre of an open disc of radius ρ, say, included in $\mathfrak{R}$. We replace it by an open disc of radius $\frac{1}{2}\rho$. Thus $\mathfrak{C}'$ is covered by open discs and from the Heine-Borel theorem we deduce that $\mathfrak{C}'$ can be covered by a finite number of such discs. Their union is a region whose boundary is still contained in $\mathfrak{R}$. Hence every closed set may be imbedded in the closure of a region which is a subset of $\mathfrak{R}$ and we have only to prove the theorem for sets of this kind.

Let now $\sigma > 0$ denote a number not exceeding the distance of $\mathfrak{C}$ to the exterior of $\mathfrak{R}$. We construct a net of squares whose sides have a length $\frac{1}{2}\sigma$. The number n_0 of the closed squares which have points in common with $\mathfrak{C}$ is finite, for $\mathfrak{C}$ is bounded, and their union is the closure $\mathfrak{D}$ of a region, which closure contains $\mathfrak{C}$ and is included in $\mathfrak{R}$. Next we take two points z_1 and z_2 in $\mathfrak{C}$. We can find a sequence of points $s_1 = z_1, s_2, \ldots. s_n = z_2$ such that two consecutive points of this sequence are situated in adjacent squares and n does not exceed n_0. The distance of two consecutive points is at most $\sigma/\sqrt{2} < \frac{3}{4}\sigma$. The open discs $|z-s_k| < \sigma$, $k = 1, \ldots, n$ are all included in $\mathfrak{R}$. Hence the function

$$\frac{f(s_k+\sigma s)-f(s_k)}{\sigma f'(s_k)} = s + \ldots$$

is holomorphic and univalent in $|s| < 1$ for every k. From (11.2–14) in $|s| < 1$ it follows

$$\frac{|f'(s_k+\sigma s)|}{|f'(s_k)|} \leq \frac{1+|s|}{(1-|s|)^3}. \tag{11.2–25}$$

If we take

$$s = \frac{s_{k+1}-s_k}{\sigma},$$

we obtain, since $|s| < \frac{3}{4}$,

$$\frac{|f'(s_{k+1})|}{|f'(s_k)|} \leq \frac{1+\frac{3}{4}}{(1-\frac{3}{4})^3} = 112.$$

Taking $k = 1, \ldots, n-1$ successively and multiplying corresponding members of the inequalities thus obtained we get

$$\frac{|f'(z_2)|}{|f'(z_1)|} \leqq 112^{n-1} < 112^{n_0} = M, \tag{11.2–26}$$

where M depends only on $\mathfrak{R}$ and on $\mathfrak{C}$, but not on $f(z)$ and the selected points z_1, z_2. Interchanging z_1 and z_2 in the inequality we obtain the desired result.

In the case that $\mathfrak{R}$ in a circular disc we easily obtain an explicit expression for M. Let $\mathfrak{R}$ denote the interior of the unit circle. If z_1 and z_2 are two points of the closed disc $|z| \leqq r < 1$ we have, according to (11.2–14),

$$\frac{1-r}{(1+r)^3} \leqq |f'(z_1)| \leqq \frac{1+r}{(1-r)^3}, \quad \frac{1-r}{(1+r)^3} \leqq |f'(z_2)| \leqq \frac{1+r}{(1-r)^3},$$

and it follows that

$$\left(\frac{1-r}{1+r}\right)^4 \leqq \frac{|f'(z_1)|}{|f'(z_2)|} \leqq \left(\frac{1+r}{1-r}\right)^4, \tag{11.2–27}$$

the desired result.

11.2.6 – A test for normal families

Integrating the inequality (11.2–25), and using (11.2–24) we get,

$$\begin{aligned} |f(s_{k+1})-f(s_k)| &= \sigma \left| \int_0^{(s_{k+1}-s_k)/\sigma} f'(s_k+\sigma s)ds \right| \\ &\leqq \sigma |f'(s_k)| \int_0^{\frac{3}{4}} \frac{1+\rho}{(1-\rho)^3}\, d\rho = 12\sigma |f'(s_k)| < 12M\sigma |f'(z_1)|. \end{aligned}$$

Taking $k = 1, \ldots, n-1$, and adding the corresponding members of the inequalities thus obtained, we easily find

$$|f(z_2)-f(z_1)| < 2(n-1)M\sigma |f'(z_1)| \leqq 12Mn_0\sigma |f'(z_1)|,$$

where n_0 is the number introduced in the previous section. Hence

$$|f(z_2)| \leqq |f(z_1)| + 12Mn_0\sigma |f'(z_1)|.$$

Writing $K = 12Mn_0\sigma$, we have

Let $\mathfrak{R}$ denote a region in the open z-plane and $\mathfrak{C}$ a closed set belonging to $\mathfrak{R}$. Then there is a positive constant K, depending only on $\mathfrak{R}$ and on $\mathfrak{C}$, such that for all functions $f(z)$ which are univalent and holomorphic throughout $\mathfrak{R}$ and for every pair z_1, z_2, belonging to $\mathfrak{C}$ the inequality

$$|f(z_2)| \leqq |f(z_1)| + K|f'(z_1)| \tag{11.2–28}$$

holds.

This result has an interesting consequence. Consider a family of functions $f(z)$ univalent and holomorphic throughout a region $\mathfrak{R}$ in the open z-plane, such that at a given point z_0 of the region the inequalities

$$|f(z_0)| < M_1, \qquad |f'(z_0)| < M_2$$

hold, where M_1 and M_2 do not depend on $f(z)$. Then *these functions are uniformly bounded throughout any closed set included in* $\mathfrak{R}$.

This follows at once from (11.2–28).

Now we make the following remark. Analyzing the proof of Vitali's theorem of section 2.22.1 we see that we used only the fact that the sequence (2.22.1) is uniformly bounded on every closed disc included in $\mathfrak{R}$. Hence the corollary stated in section 2.22.2 is already valid if the sequence (2.22–18) is uniformly bounded on any closed and bounded set included in the region $\mathfrak{R}$. It is customary to express this by saying that the sequence (2.22–18) is uniformly bounded in the interior of $\mathfrak{R}$. As a consequence of this we have

If the family of univalent and holomorphic functions in a region $\mathfrak{R}$ *in the open z-plane is such that the values of the functions as well as those of their derivatives are bounded at a given point of* $\mathfrak{R}$, *then the family is a normal family.*

11.3 – Estimates of coefficients

11.3.1 – A THEOREM OF LITTLEWOOD

As we pointed out earlier (section 11.2.1) Bieberbach conjectured that for a univalent and normalized function with the expansion (11.1–26) for $|z| < 1$ the sharp inequalities $|a_n| \leq n$ hold. Littlewood has established the truth of a weaker assertion stating that Bieberbach's conjecture is true within order of magnitude. He proved

If $f(z)$ *is the univalent function* (11.1–26) *then*

$$|a_n| < ne, \qquad n = 2, 3, \ldots. \tag{11.3–1}$$

We are sufficiently prepared to establish this result in a few lines. Using (11.2–16) we find from (11.1–24)

$$|a_n| r^n \leq \int_0^r \frac{d\rho}{(1-\rho)^2} = \frac{r}{1-r}.$$

Now we may take $r = 1 - 1/n$. Hence

$$|a_n| \leq \frac{n^n}{(n-1)^{n-1}} = n\left(1 + \frac{1}{n-1}\right)^{n-1} < ne,$$

as was stated.

Since then this result has been improved by other authors. A first result in this direction is due to Landau who proved

$$|a_n| \leqq n\left(\tfrac{1}{2}+\frac{1}{\pi}\right)\mathrm{e}.$$

The strongest result known at present is due to Bazilevič and is

$$|a_n| < \tfrac{1}{2}n\mathrm{e}+1{,}51.$$

11.3.2 – Odd functions

Imposing certain conditions upon the functions under consideration it is possible to gain more detailed information. A classical result due to Littlewood and Payley states:

The coefficients in the expansion of an odd univalent and normalized function are bounded

$$|a_n| \leqq A. \tag{11.3–2}$$

Here A is a constant not depending on the functions. The best possible A is not known, but we shall be able to evaluate a number for which (11.3–2) is true.

In order to prove this statement we will apply Golusin's inequality (11.1–37), taking $\lambda = 0$, $k = 4$. First we notice that the coefficients occurring in

$$f(z) = \sum_{\nu=1}^{\infty} a_{2\nu-1}z^{2\nu-1} \tag{11.3–3}$$

are on account of (2.16–7) determined by

$$na_n = \frac{1}{2\pi i}\int_{C_r}\frac{f'(\zeta)}{\zeta^n}\,d\zeta, \qquad n = 1, 3, 5, \ldots,$$

whence

$$na_n \leqq \frac{1}{2\pi r^{n-1}}\int_0^{2\pi}|f'(re^{i\theta})|d\theta$$

and thus

$$n|a_n| \leqq \frac{1}{r^n}\sqrt{\frac{6}{\mathrm{e}}}\left(\frac{M(\sqrt{r})^{\frac{1}{2}}}{1-r}\right)^{\frac{1}{2}}\left(\int_0^r\frac{M(\rho)^{\frac{1}{2}}}{\rho}\,d\rho\right)^{\frac{1}{2}}. \tag{11.3–4}$$

Now we observe that $f(z)$ may be considered as a Faber function with $k = 2$ associated with a univalent and normalized function $g(z)$. In fact

$$f^2(z) = z^2+2a_3z^4+\ldots$$

is univalent in terms of the variable z^2. Hence

$$f(z) = g_2(z) = (g(z^2))^{\frac{1}{2}}$$

with

$$g(z) = z+2a_3z^2+\ldots.$$

If we apply the estimate (11.2–16) to $g(z)$ we readily find

$$M(\rho) \leqq \frac{\rho}{1-\rho^2}.$$

Hence

$$\int_0^r \frac{M(\rho)^{\frac{3}{2}}}{\rho}\,d\rho = \int_0^r \frac{\rho^{\frac{1}{2}}d\rho}{(1-\rho^2)^{\frac{3}{2}}} \leqq r^{\frac{1}{2}}\int_0^r \frac{d\rho}{(1-\rho^2)^{\frac{3}{2}}} = \frac{r^{\frac{3}{2}}}{(1-r^2)^{\frac{1}{2}}}.$$

Again

$$\frac{M(\sqrt{r})^{\frac{1}{2}}}{1-r} \leqq \frac{r^{\frac{1}{4}}/(1-r)^{\frac{1}{2}}}{1-r} = \frac{r^{\frac{1}{4}}}{(1-r)^{\frac{3}{2}}}.$$

The product of the expressions obtained is

$$\frac{r^{\frac{3}{2}}}{(1-r)^2}\cdot\frac{r^{\frac{1}{4}}}{(1+r)^{\frac{1}{2}}} < \frac{r^{\frac{3}{2}}}{(1-r)^2}\,\frac{1}{\sqrt[4]{4}},$$

for

$$\frac{r}{(1+r)^2} < \tfrac{1}{4}, \qquad 0 < r < 1,$$

as follows from $0 < (1-r)^2 = (1+r)^2-4r$. Inserting these results into (11.3–4) we get

$$n|a_n| < \frac{2^{\frac{1}{4}}3^{\frac{1}{2}}e^{-\frac{1}{2}}}{r^{n-\frac{1}{2}}(1-r)}.$$

Take

$$r = \frac{2n-1}{2n+1}.$$

Then

$$\frac{1}{r^{n-\frac{1}{2}}(1-r)} = \tfrac{1}{2}\sqrt{\frac{(2n+1)^{2n+1}}{(2n-1)^{2n-1}}} = n\,\frac{\left(1+\dfrac{1}{2n}\right)^{n+\frac{1}{2}}}{\left(1-\dfrac{1}{2n}\right)^{n-\frac{1}{2}}}.$$

Making $n \to \infty$ we find

$$\frac{\left(1+\dfrac{1}{2n}\right)^{n+\frac{1}{2}}}{\left(1-\dfrac{1}{2n}\right)^{n-\frac{1}{2}}} = \left(1-\frac{1}{4n^2}\right)^{\frac{1}{2}} \left(\frac{\left(1+\dfrac{1}{2n}\right)^{2n}}{\left(1-\dfrac{1}{2n}\right)^{2n}}\right)^{\frac{1}{2}} \to \mathrm{e}.$$

Since

$$\frac{(x+1)^{x+1}}{(x-1)^{x-1}}$$

is increasing for $x > 1$ (its derivative being positive) we find that

$$\tfrac{1}{2}\sqrt{\frac{(2n+1)^{2n+1}}{(2n-1)^{2n-1}}} < ne.$$

Hence (11.3–2) is valid if we take

$$A = 2^{\frac{1}{2}}3^{\frac{1}{2}}\mathrm{e}^{\frac{1}{2}} = 3{,}39\ldots.$$

11.3.3 – Typically real functions

There is an interesting case for which Bieberbach's conjecture is true and can be proved by simple means. A function is called *typically real* if $f(z)$ is holomorphic throughout $|z| < 1$ and is real there if and only if z is real. For these functions (which need not necessarily be univalent) the following theorem holds:

If

$$f(z) = z + a_2 z^2 + \ldots \qquad (11.3\text{–}5)$$

is typically real, then

$$|a_n| \leqq n, \qquad n = 2, 3, \ldots.$$

Note that for real values x of z we have $f(x) = \operatorname{Re} f(x) + i \operatorname{Im} f(x)$, where $\operatorname{Im} f(x)$ is a power series in terms of x which vanishes everywhere. Hence all coefficients of this series are zero and, consequently, the coefficients of $f(z)$ are real. Next we put $z = re^{i\theta}$. Then

$$v(z) = \operatorname{Im} f(z) = \operatorname{Im} r(\cos\theta + i\sin\theta) + \operatorname{Im}\sum_{\nu=2}^{\infty} a_\nu r^\nu(\cos\nu\theta + i\sin\nu\theta)$$

$$= \sum_{\nu=1}^{\infty} a_\nu r^\nu \sin\nu\theta,$$

if we take $a_1 = 1$. Multiplying the first and the last member by $\sin n\theta$, observing that the series thus obtained is uniformly convergent with respect to θ and taking into account the fact that

$$\int_0^{\pi} \sin n\theta \sin m\theta \, d\theta = \begin{cases} 0 & \text{if} \quad n \neq m, \\ \frac{1}{2}\pi & \text{if} \quad n = m, \end{cases}$$

we find that

$$a_n r^n = \frac{2}{\pi} \int_0^{\pi} v(re^{i\theta}) \sin n\theta \, d\theta, \qquad n = 1, 2, \ldots.$$

By induction we may prove that

$$|\sin n\theta| \leqq n \sin \theta, \qquad 0 < \theta < \pi.$$

In fact, the assertion is true for $n = 1$ and if n is an integer for which the assertion holds we have

$$|\sin (n+1)\theta| = |\sin n\theta \cos \theta + \cos n\theta \sin \theta| \leqq |\sin n\theta| + \sin \theta.$$

Thus, since $v(re^{i\theta})$ has constant sign for $0 < \theta < \pi$ (for $f(z)$ is only real if z is real), we find

$$|a_n r^n| \leqq \frac{2n}{\pi} \int_0^{\pi} |v(re^{i\theta}) \sin \theta| \, d\theta.$$

Notice that

$$\int_0^{\pi} |v(re^{i\theta}) \sin \theta| \, d\theta = \tfrac{1}{2}\pi r,$$

whence

$$|a_n r^n| \leqq nr.$$

Making $r \to 1$ we obtain the desired result.

Koebe's function $z/(1-z)^2$ shows that this result is sharp.

Now let $f(z)$ denote a univalent and normalized function which has the expansion (11.3–5) with real coefficients. In this case

$$f(\bar{z}) = \overline{f(z)}.$$

If $f(z)$ is real for some complex $z_0 \neq 0$ we have

$$f(\bar{z}_0) = f(z_0).$$

This is impossible, since $\bar{z}_0 \neq z_0$ and f is univalent. It follows that $f(z)$ is typically real and thus we have

If $f(z)$ is univalent and normalized and if its Taylor expansion about the origin has only real coefficients then Bieberbach's conjecture holds.

11.3.4 – Starlike univalent functions

Another important class of univalent functions for which Bieberbach's conjecture holds is provided by those which map $|z| < 1$ on a starshaped region. A region $\mathfrak{R}$ is said to be *starshaped* with respect to a point O in $\mathfrak{R}$ if for any point P in $\mathfrak{R}$ the straight line segment OP also lies in $\mathfrak{R}$ (see also section 8.1.1).

First we will prove

If $\mathfrak{R}$ is the image of $|z| < 1$ and starshaped with respect to the origin, then so is the domain $\mathfrak{R}_r$, the image of $|z| \leqq r < 1$.

If $f(z)$ is in $\mathfrak{R}$ then $tf(z)$, $0 < t < 1$ is in $\mathfrak{R}$ and the function

$$g(z) = \check{f}(tf(z)),$$

where $\check{f}$ denotes the inverse of f, is holomorphic in $|z| < 1$ and satisfies $|g(z)| < 1$, $g(0) = 0$. It follows from Schwarz's lemma that even $|g(z)| \leqq |z|$. Suppose now that $f(z_1)$ is in $\mathfrak{R}_r$, $|z_1| \leqq r < 1$. Then

$$|\check{f}(tf(z_1))| = |g(z_1)| \leqq |z_1| \leqq r.$$

Hence there is a number z_2 with $|z_2| \leqq r$ such that $f(z_2) = tf(z_1)$ and so $tf(z_1)$ is in $\mathfrak{R}_r$, i.e., $\mathfrak{R}_r$ is also starshaped.

A function mapping $|z|<1$ onto a sharshaped region is called *starlike*.

If $z \neq 0$ we have

$$\log f(z) = \log|f(z)| + i \arg f(z)$$

and, if $z = re^{i\theta}$,

$$\frac{\partial}{\partial\theta} \log f(z) = \frac{izf'(z)}{f(z)} = \frac{\partial}{\partial\theta} \log|f(z)| + i\frac{\partial}{\partial\theta} \arg f(z),$$

whence

$$\frac{\partial}{\partial\theta} \arg f(z) = \operatorname{Re}\frac{zf'(z)}{f(z)}. \tag{11.3–6}$$

It is geometrically clear that for a starlike function arg $f(z)$ does not decrease along $|z| = r$. Hence

$$\operatorname{Re}\frac{zf'(z)}{f(z)} \geqq 0, \qquad |z| = r < 1,$$

and in general

If $f(z)$ is univalent and starlike, then

$$\operatorname{Re}\frac{zf'(z)}{f(z)} > 0, \qquad |z| < 1, \tag{11.3–7}$$

and conversely.

In fact, equality cannot occur, on account of the maximum principle of section 2.14.1.

If we write

$$\frac{zf'(z)}{f(z)} = 1+\psi(z) \tag{11.3–8}$$

then, evidently, $\psi(z)$ maps $|z| < 1$ onto a region included in the half plane $\operatorname{Re} z > -1$.

In order to obtain more information about $f(z)$ from this result we apply the following lemma.

Suppose that

$$\varphi(z) = \sum_{\nu=1}^{\infty} \alpha_\nu z^\nu \tag{11.3–9}$$

is convex and univalent (not necessarily normalized) *and maps* $|z| < 1$ *onto a region* $\mathfrak{R}$. *Let*

$$\psi(z) = \sum_{\nu=1}^{\infty} \beta_\nu z^\nu \tag{11.3–10}$$

denote a function holomorphic throughout $|z| < 1$ *and assume there only values which lie in* $\mathfrak{R}$. *Then*

$$|\beta_n| \leqq |\alpha_1|, \qquad n = 1, 2, \ldots. \tag{11.3–11}$$

Consider the function

$$\chi(z) = \check{\varphi}(\psi(z)) = \frac{\beta_1}{\alpha_1} z + \ldots, \tag{11.3–12}$$

where $\check{\varphi}$ denotes the inverse of φ. It is clear that $\chi(z)$ is holomorphic throughout $|z| < 1$ and satisfies $|\chi(z)| < 1$, $\chi(0) = 0$. By Schwarz's lemma we have $|\chi'(0)| \leqq 1$, whence $|\beta_1| \leqq |\alpha_1|$.

Let ω_k, $1 \leqq k \leqq n$, denote the nth roots of unity. Since $\mathfrak{R}$ is convex, the centre of gravity

$$\frac{1}{n} \sum_{\kappa=1}^{n} \varphi(\omega_\kappa z^{1/n}) = \beta_n z + \beta_{2n} z^2 + \ldots$$

is also included in $\mathfrak{R}$. We may apply the previous result and we find $|\beta_n| \leqq |\alpha_1|$.

A simple corollary is

If $f(z)$ *is univalent, normalized and convex, then*

$$|a_n| \leqq 1. \tag{11.3–13}$$

This is clear if we apply the lemma to $\varphi = \psi = f$. This result is sharp as is shown by the function

$$\frac{z}{1-z} = z + z^2 + \ldots, \tag{11.3–14}$$

which maps $|z| < 1$ onto the half plane $\mathrm{Re}\, z > -\frac{1}{2}$.

If we take the function

$$\varphi(z) = \frac{2z}{1-z} = 2z+2z^2+\dots \qquad (11.3\text{–}15)$$

which maps $|z| < 1$ onto the half plane $\mathrm{Re}\, z > -1$, we easily deduce

If

$$f(z) = 1+a_1z+a_2z^2+\dots \qquad (11.3\text{–}16)$$

is a univalent function with a positive real part then

$$|a_n| \leqq 2, \qquad n = 1, 2, \dots. \qquad (11.3\text{–}17)$$

This result is sharp as is shown by the function

$$1+\varphi(z) = \frac{1+z}{1-z} = 1+2z+2z^2+\dots. \qquad (11.3\text{–}18)$$

An estimate for the modulus of functions of this kind has been obtained in section 2.21.3.

Finally we wish to show that the lemma yields also a second proof for the covering theorem of section 11.2.3.

Let $\mathfrak{R}$ be the image of $|z| < 1$ as given by the convex, univalent and normalized function $f(z)$ and let $w_0 = re^{i\theta}$ be a point of smallest modulus lying outside $\mathfrak{R}$. Replacing, if necessary, $f(z)$ by $-\eta^{-1}f(-\eta z)$, where $|\eta| = 1$, we may suppose that $w_0 = -r$. Then no point w with $\mathrm{Re}\, w \leqq -r$ is in $\mathfrak{R}$. Suppose, contrary to this, that $\mathfrak{R}$ contains a point w_1 such that $\mathrm{Re}\, w_1 \leqq -r$. By the convexity of $\mathfrak{R}$ that part of the line through w_0 and w_1 which lies on the side of w_0 opposite to w_1 lies entirely outside $\mathfrak{R}$. But this part contains points of modulus less than r, which leads to a contradiction.

The function

$$\varphi(z) = \frac{2rz}{1-z} = 2rz+\dots$$

maps $|z| < 1$ onto the half plane $\mathrm{Re}\, z > -r$. Since $f(z)$ assumes all its values in this half plane, the lemma applied to $\psi(z) = f(z)$ gives $2r \geqq 1$. This proves the assertion.

Now we return to the case of the starlike functions. For $\varphi(z)$ of the lemma we take the function (11.3–15) which maps $|z| < 1$ onto the half plane $\mathrm{Re}\, z > -1$. For $\psi(z)$ we take the function defined by (11.3–8). Then the lemma states that $|\beta_n| \leqq 2, n = 1, 2, \dots$, as we have seen above. If $f(z)$ has the expansion (11.3–5) it follows from (11.3–8) that

$$\left(\sum_{\nu=1}^{\infty} a_\nu z^\nu\right)\left(1+\sum_{\nu=1}^{\infty} \beta_\nu z^\nu\right) = \sum_{\nu=1}^{\infty} \nu a_\nu z^\nu.$$

Equating coefficients, we have

$$na_n = a_n + \beta_1 a_{n-1} + \ldots + \beta_{n-1}.$$

Thus

$$(n-1)|a_n| = |\beta_1 a_{n-1} + \ldots + \beta_{n-1}| \leqq 2(|a_{n-1}| + \ldots + 1), n > 1. \tag{11.3–19}$$

By induction we can prove that

$$|a_n| \leqq n$$

for, assuming that

$$|a_m| \leqq m, \qquad m = 1, \ldots, n-1, n > 1,$$

we also have

$$(n-1)|a_n| \leqq n(n-1).$$

Thus we proved

For starlike univalent and normalized functions Bieberbach's conjucture holds.

Koebe's function shows that this result is sharp, for this function is starlike with respect to the origin.

We additionally have

For starlike, univalent, normalized odd functions the inequalities

$$|a_n| \leqq 1 \tag{11.3–20}$$

hold.

In (11.3–19) we replace n by $2n+1$ and notice that inside the brackets on the left n terms occur. Assuming $|a_{2m+1}| \leqq 1$ for $m < n$, we have $2n|a_n| \leqq 2n$, and the assertion has been proved by induction.

11.3.5 – Relation between convex and starlike functions

First we need the following lemma

If $\mathfrak{R}$, the image of $|z| < 1$ as given by a univalent normalized function $f(z)$, is convex, then $\mathfrak{D}_r$, the image of $|z| \leqq r < 1$, is also convex.

Suppose that $f(z_1)$ and $f(z_2)$ are in $\mathfrak{D}_r$. Without loss of generality we may assume that $|z_1| \leqq |z_2| \neq 0$. Then

$$\varphi(z) = tf\left(\frac{z_1}{z_2}z\right) + (1-t)f(z)$$

is in $\mathfrak{R}$ for $|z| < 1$ and $0 < t < 1$. Hence $g(z) = \check{f}(\varphi(z))$ is holomorphic in $|z| < 1$, $\check{f}$ being the inverse of f, while $g(0) = 0$, $|g(z)| \leqq 1$. By Schwarz's lemma we have $|g(z)| \leqq |z|$. In particular, if $z = z_2$

$$|\check{f}(tf(z_1) + (1-t)f(z_2))| \leqq r.$$

But the value

$$tf(z_1)+(1-t)f(z_2)$$

is assumed by $f(z)$ at $z = z_0$. Hence $|z_0| \leqq r$, i.e., $f(z_0)$ is in $\mathfrak{D}_r$. The converse is obviously true: If all $\mathfrak{D}_r$ are convex so is $\mathfrak{R}$.

In particular the curve $f(r)$, $|z| = r$ is convex. The tangent through the point $z = re^{i\theta}$ at this curve makes an angle φ with the positive real axis. Since

$$\arg \frac{d}{d\theta} f(re^{i\theta}) = \arg \mathrm{ir}\, e^{i\theta} f'(re^{i\theta}) = \tfrac{1}{2}\pi+\theta+\arg f'(z)$$

we have

$$\varphi = \theta+\tfrac{1}{2}\pi+\arg f'(z). \tag{11.3–21}$$

If s denotes the arc length, the curvature is

$$\kappa = \frac{d\varphi}{ds} = \frac{d\varphi/d\theta}{ds/d\theta}. \tag{11.3–22}$$

In view of (11.3–6) we have

$$\frac{\partial}{\partial\theta} \arg f'(z) = \mathrm{Re}\, \frac{zf''(z)}{f'(z)}. \tag{11.3–23}$$

From (11.3–21) follows

$$\frac{d\varphi}{d\theta} = 1+\mathrm{Re}\, \frac{zf''(z)}{f'(z)}. \tag{11.3–24}$$

The sign of $d\varphi/d\theta$ is the same as that of κ, since $ds/d\theta$ is positive. Hence the expression on the right of (11.3–24) has the same sign as the curvature and is, therefore, not negative for convex functions. Thus

If $f(z)$ is convex, then

$$1+\mathrm{Re}\, \frac{zf''(z)}{f'(z)} > 0, \qquad |z| < 1, \tag{11.3–25}$$

and conversely.

Equality does not occur on account of the maximum principle. The converse is also true.

Now we consider the univalent and normalized functions $f_1(z)$ and $f_2(z)$. Suppose that $f_2(z)$ is connected with $f_1(z)$ by the equation

$$f_2(z) = zf_1'(z). \tag{11.3–26}$$

In view of (11.3–7) then it follows from

$$\mathrm{Re}\, \frac{zf_2'(z)}{f_2(z)} = 1+\mathrm{Re}\, \frac{zf_1''(z)}{f_1'(z)}$$

that $f_2(z)$ is starlike if and only if $f_1(z)$ is convex.

Let

$$f_1(z) = z+c_2z^2+\ldots = \sum_{\nu=1}^{\infty} c_\nu z^\nu,\ C_0=1.$$

Then

$$f_2(z) = z+2c_2z^2+\ldots = \sum_{\nu=1}^{\infty} \nu c_\nu z^\nu = \sum_{\nu=1}^{\infty} a_\nu z^\nu.$$

For convex functions we proved $|c_n| \leqq 1$, (11.3–13). Hence $|a_n| \leqq n$, in accordance with results obtained before.

11.4 – Löwner's theory

11.4.1 – Introduction

The foregoing results, based on simple geometrical considerations, are limited in scope. Very deep results can be achieved by an interesting theory due to K. Löwner. For many years this method was the only capable of producing results of such depth. Recently two other methods, Schiffer's variational method and Jenkin's theory, have appeared, which are capable of giving comparable results. They, however, are beyond the scope of this book. Löwner's theory has initiated a new development in the field of univalent functions and we desire to give an exposition of it which is justified by its many beautiful applications.

For the sake of convenience we will henceforth denote the class of functions which are holomorphic, univalent and normalized in the disc $|z|<1$ by $\mathscr{S}$. The fundamental idea in Löwner's theory is that for many problems it will be sufficient to consider functions of a subclass $\mathscr{S}'$ which is *dense* in $\mathscr{S}$. This means that an arbitrary function $f(z)$ of $\mathscr{S}$ can be approximated as closely as desired by a sequence of functions $f_n(z)$ of $\mathscr{S}'$. More precisely stated: there is a sequence of functions $f_n(z)$ of $\mathscr{S}'$ such that $f_n(z) \to f(z)$ uniformly on every closed set within $|z| < 1$. It follows from the considerations of section 2.20.1 (Weierstrass's theorem) that then all derivatives at an arbitrary point of $|z| < 1$, and in particular the coefficients of the series expansion of $f_n(z)$ tend to those of $f(z)$ as $n \to \infty$. Thus bounds obtained for the functions of the class $\mathscr{S}'$ are also valid for those of the class $\mathscr{S}$.

As we will see we can obtain approximating functions as the functions which map $|z| < 1$ on so-called *slit-regions*, i.e., regions which are the region $|w| < 1$ slit along an arc inside it and ending on the circumference. A remarkable thing is that these functions are closely related to solutions of a certain type of differential equations, and thus much information about these functions can be obtained by a computational method.

11.4.2 – The kernel of a sequence of regions

First we will introduce an important notion due to Carathéodory. Let an arbitrary sequence

$$\mathfrak{R}_1, \mathfrak{R}_2, \ldots \tag{11.4–1}$$

of simply connected regions, each containing the origin but not the point at infinity, be given. We associate with the sequence a certain set of points $\mathfrak{K}$, called the *kernel* of the sequence. If the sequence is such that no circle with the origin as its centre is covered by all regions $\mathfrak{R}_n$, when n is sufficiently large, then its kernel consists of the origin only. In all other cases the kernel $\mathfrak{K}$ is the largest region having the property that every connected closed set, containing the origin and contained in $\mathfrak{K}$ lies within every region $\mathfrak{R}_n$ for sufficiently large n (or, as we shall also say, in almost all $\mathfrak{R}_n$).

The kernel of a sequence of regions is uniquely determined.

This may be seen by means of the following construction. Suppose that the point z_0 of the complex plane has a neighbourhood which is covered by almost all $\mathfrak{R}_n$ of the sequence. Let $C(z_0)$ denote the largest circle with centre z_0 such that all smaller concentric circles are covered by almost all $\mathfrak{R}_n$. The union of the interior of all these circles $C(z_0)$, if z_0 varies throughout the open plane, is an open set $\mathfrak{A}$, which may be either empty, or does not contain the origin. In this case the kernel is a single point, viz. the origin. If, however, $\mathfrak{A}$ contains the origin, the kernel $\mathfrak{K}$ consists of all points of $\mathfrak{A}$ which can be connected with the origin by a polygonal arc within $\mathfrak{A}$. This proves the assertion.

Let us consider two examples. First we define $\mathfrak{R}_n$ as the region obtained from the z-plane by omitting all points of the real axis with $\operatorname{Re} z \leqq -1/n$. The set $\mathfrak{A}$ of the above construction is obviously the principal region. This does not contain the origin. Hence the kernel is the origin.

Secondly we define $\mathfrak{R}_n$ as the open z-plane cut along an arc of the unit circle joining the points $e^{i\pi/n}$ and $e^{-i\pi/n}$, and $2\pi(1-1/n)$ in length. The set $\mathfrak{A}$ consists of the interior and the exterior of the unit circle; the kernel $\mathfrak{K}$ is the interior of the unit circle.

A sequence (11.4–1) is said to *converge to its kernel* if every subsequence

$$\mathfrak{R}_{n_1}, \mathfrak{R}_{n_2}, \ldots, n_1 < n_2 < \ldots \tag{11.4–2}$$

has the same kernel as the original sequence. It is clear that the sequences of the above examples converge to their kernels respectively.

11.4.3 – Carathéodory's convergence theorem

The following theorem, due to Carathéodory, relates the convergence of a sequence of functions to the convergence of a sequence of regions.

Let

$$f_1(z), \quad f_2(z), \ldots \tag{11.4–3}$$

be a sequence of univalent functions, holomorphic and uniformly bounded in the region $|z| < 1$, *with the additional property that they all vanish at the origin and that their first derivatives are positive there. These functions converge uniformly on every closed set in* $|z| < 1$ *to a holomorphic function* $f(z)$ *or to zero if and only if the sequence* (11–4–1) *of the images of* $|z| < 1$ *in the w-plane consists of uniformly bounded regions converging to the kernel of the sequence. If the limiting function* $f(z)$ *is not identically zero, then it maps the open disc* $|z| < 1$ *conformally onto the kernel* $\mathfrak{K}$ *and the sequence of inverse functions* $\check{f}_1(w), \check{f}_2(w), \ldots$ *converges uniformly on every closed set within the kernel to the inverse function* $\check{f}(w)$ *of* $f(z)$.

The proof of this theorem is rather long. Let us first assume that the sequence $f_n(z)$ is convergent within $|z| < 1$, uniformly on every closed set of $|z| < 1$. Then it follows from the last lemma of section 10.5.1 that the limiting function is either the constant 0, or a univalent holomorphic function which transforms $|z| < 1$ into a region $\mathfrak{K}$ in the w-plane containing the point $w = 0$. We suppose that $f(z)$ is not identically zero.

Now we take a closed and connected set $\mathfrak{C}_w$ in $\mathfrak{K}$ such that it includes $w = 0$. It corresponds by $f(z)$ to a closed set $\mathfrak{C}_z$ within $|z| < 1$. Hence there is a circumference $|z| = r < 1$ which encloses $\mathfrak{C}_z$. Either $\mathfrak{C}_w$ is contained in almost all $\mathfrak{R}_n$, or there is an infinite sequence $n_1 < n_2 < \ldots$ such that for each n_k the function

$$w = f_{n_k}(z) \tag{11.4–4}$$

maps $|z| \leqq r$ onto a set of points which does not contain the whole set $\mathfrak{C}_w$. By this assumption not all points of $\mathfrak{C}_w$ correspond to points inside $|z| = r$ on account of the function (11.4–4), for in the contrary case the winding number of the image of $|z| = r$ with respect to each point of $\mathfrak{C}_w$ would be equal to that with respect to $w = 0$ (for $\mathfrak{C}_w$ is connected) and consequently different from zero. But then $\mathfrak{C}_w$ would contain no other points than those of the image of $|z| \leqq r$ by virtue of (11.4–4). As a consequence $\mathfrak{C}_w$ contains at least one point w_{n_k} which is the image of a point z_{n_k} on $|z| = r$ on account of (11.4–4). The sequence of numbers n_1, $n_2, \ldots$, contains a subsequence $m_1, m_2, \ldots$, such that the sequence of points z_{m_k} converges to a point z_0 with $|z_0| = r$.

The following remark is now of importance. Let $z_n \to z_0$ as $n \to \infty$, where $z_1, z_2, \ldots$, are points of a closed set within $|z| < 1$. Since the sequence (11.4–3) is uniformly convergent on this set we have, if ε is an arbitrary positive number,

$$|f_n(z) - f(z)| < \tfrac{1}{2}\varepsilon$$

for all points z of the set, provided n is sufficiently large. In particular

$$|f_n(z_n)-f(z_n)| < \tfrac{1}{2}\varepsilon.$$

But since $f(z)$ is continuous at $z = z_0$, we also have

$$|f(z_n)-f(z_0)| < \tfrac{1}{2}\varepsilon,$$

provided that n is sufficiently large. It follows that from a certain index upwards

$$|f_n(z_n)-f(z_0)| < \varepsilon$$

and this means that the sequence of the points $w_n = f_n(z_n)$ is convergent.

Now we return to our proof. Is clear that the points

$$w_{m_k} = f_{m_k}(z_{m_k})$$

constitute a convergent sequence, tending to the point $w_0 = f(z_0)$. Hence $\mathfrak{C}_z$ meets $|z| = r$, in contradiction to the construction of this circumference. Thus we see that every closed and connected set containing the point $w = 0$ and contained in $\mathfrak{K}$ is contained in almost all regions $\mathfrak{R}_n$.

Next we assume that $\mathfrak{K}^*$ is a region with the same property, i.e., it contains the origin and every closed and connected set of $\mathfrak{K}^*$ is covered by almost all regions $\mathfrak{R}_n$. Then it follows that $f(z)$ cannot be a constant. In order to prove this, we take an arbitrary point w_0 of $\mathfrak{K}^*$. Consider a region $\mathfrak{R}$ containing the points $w = 0$ and w_0 and such that the closure of $\mathfrak{R}$ is also in $\mathfrak{K}$. Since, by hypothesis, $\mathfrak{R}$ is covered by almost all $\mathfrak{R}_n$, the inverse functions $\check{f}_n(w)$ of $f_n(z)$ are defined throughout $\mathfrak{R}$, provided n is sufficiently large. These functions are uniformly bounded, for their moduli do not exceed unity. Hence they constitute a normal family and there is a subsequence of functions which converge to a function $\varphi(w)$, this convergence being uniform on every closed subset of $\mathfrak{R}$. Applying again the above mentioned lemma of section 10.5.1 we may infer that either $\varphi(w)$ is the constant 0 (for $\varphi(0) = 0$) or the equation $z = \varphi(w)$ gives a mapping of $\mathfrak{R}$ onto a region within $|z| < 1$. In both cases $z_0 = \varphi(w_0)$ is a point in the interior of the unit circle in the z-plane. Thus, if $z_{n_k} = \check{f}_{n_k}(w_0)$, it follows that $z_{n_k} \to z_0$ and since z_0 is within $|z| < 1$, $f_{n_k}(z_{n_k}) \to f(z_0)$ by the strength of the same argument as used above. Observing that $f_{n_k}(r_{n_k}) = w_0$, it follows that $w_0 = f(z_0)$. But w_0 is an arbitrary point of $\mathfrak{K}^*$ and so $f(z)$ cannot be a constant and $|z| < 1$ is mapped by $f(z)$ onto a region $\mathfrak{K}$.

The following consequences are obvious now:

1) $\mathfrak{K}^*$ is a subregion of $\mathfrak{K}$. Hence $\mathfrak{K}$ is the kernel of the sequence of the images of $|z| < 1$ as given by the functions (11.4–3).

2) If the sequence (11.4–3) tends to a function which is identically zero, then there is no neighbourhood of $w = 0$ which is covered by almost all regions $\mathfrak{R}_n$ and in this case the kernel is the point $w = 0$.

3) The function $\varphi(w)$ is the inverse of $f(z)$ if $f(z)$ is not identically equal to zero and consequently all subsequences of the sequence $\check{f}_n(w)$ tend to the same limiting function, i.e., the sequence is convergent, uniformly on every closed subset of $\mathfrak{K}$.
4) Any subsequence of (11.4–1) tends to $f(z)$ and this function maps $|z| < 1$ onto the kernel of the sequence (11.4–1). It follows that this kernel coincides with $\mathfrak{K}$ and so that the sequence (11.4–1) converges to its kernel. This concludes the necessity part of the theorem.

In order to prove the sufficiency part, we start with the assumption that all regions of the sequence (11.4–1) are uniformly bounded and converge to their kernel $\mathfrak{K}$. To each $\mathfrak{R}_n$ corresponds a function $f_n(z)$ with

$$f_n(0) = 0, \qquad f_n'(0) > 0, \qquad n = 1, 2, \ldots, \tag{11.4–5}$$

which maps $|z| < 1$ onto $\mathfrak{R}_n$. The sequence (11.4–3) is uniformly bounded and constitutes, therefore, a normal family. If this series were not convergent it would be possible to find two subsequences

$$f_{n_1}(z), f_{n_2}(z), \ldots,$$

and

$$f_{m_1}(z), f_{m_2}(z), \ldots,$$

converging in $|z| < 1$ to two different holomorphic functions $g(z)$ and $h(z)$ respectively. But both functions map $|z| < 1$ onto the kernel $\mathfrak{K}$ and from (11.4–5) it follows that they vanish at $z = 0$, having positive derivatives there. This entails $g(z) = h(z)$ and this concludes the proof.

11.4.4 – Löwner's approximation theorem

Carathéodory's theorem enables us to give a simple geometrical construction of a set of functions of $\mathcal{S}'$ which is dense in $\mathcal{S}$. First we will show that an arbitrary function $f(z)$ can be approximated (in the sense of section 11.4.1) by a family of functions also belonging to $\mathcal{S}$ which map $|z| < 1$ onto a region bounded by a contour. For these functions we can take the functions $f(\rho z)/\rho$, where ρ is between 0 and 1. It is clear that $f(\rho z)$ maps $|z| < 1$ onto the image of the interior of $|z| = \rho$, which is the boundary, and it is also clear that $f(\rho z)/\rho$ is in $\mathcal{S}$. We have to show that these functions tend to $f(z)$ as $\rho \to 1$ uniformly on any closed disc $|z| \leqq r < 1$. If $f(z)$ has the expansion

$$f(z) = z + \sum_{\nu=2}^{\infty} a_\nu z^\nu \tag{11.4–6}$$

then

$$f(\rho z)/\rho = z + \sum_{\nu=2}^{\infty} a_\nu \rho^{\nu-1} z^\nu$$

and

$$|f(z)-f(\rho z)/\rho| = (1-\rho)|\sum_{\nu=2}^{\infty} a_\nu \sum_{\mu=0}^{\nu-2} \rho^\mu z^\nu| \leqq (1-\rho)\sum_{\nu=2}^{\infty}(\nu-1)|a_\nu||z|^\nu.$$

The radius of convergence of this latter series is the same as that of (11.4–6) (as follows e.g. on applying the Cauchy-Hadamard test of section 1.6.5) and, therefore, the series is bounded throughout $|z| \leqq r < 1$. The truth of the statement follows easily now.

Thus it is sufficient to state an approximation theorem for the functions $f(\rho z)/\rho$, $0 < \rho < 1$.

Next we introduce the so-called *slitregions*. Let L be a simple Jordan arc having one end point on the circumference $|w| = 1$ and lying otherwise in $|w| < 1$. We suppose that L does not pass through $w = 0$. The set $\mathfrak{R}$ consisting of all points of $|w| < 1$ not on L will be a simply connected region. In fact, if P and Q are points of $\mathfrak{R}$ near L we can pass from P to Q along a path near L, which, if necessary, will go round the tip of L and along the other side. Thus any two points of $\mathfrak{R}$ can be joined to some point near L, for instance to the tip of L. Thus $\mathfrak{R}$ is connected. That it is open is evident. The complement of $\mathfrak{R}$ with respect to the closed z-plane consists of $|w| \geqq 1$ together with L and so it is closed and connected (last theorem of section 9.1.1). Hence $\mathfrak{R}$ is simply connected (section 9.1.3).

Now we are in a position to prove an approximation theorem, which will turn out to be of utmost importance. First we observe that if M is sufficiently large the function $f(\rho z)/\rho M$ maps $|z| < 1$ onto a region within $|w| < 1$ bounded by a contour. Further, if L is a slit in $|w| < 1$ as described above then on account of Riemann's mapping theorem there exists a unique function

$$w = \beta(z+a_2 z^2 + \ldots), \qquad \beta > 0, \tag{11.4–7}$$

which maps $|z| < 1$ one-to-one onto $|w| < 1$ except for the slit L.

Let now $\varphi(z)$ denote a function which maps $|z| < 1$ onto a region $\mathfrak{R}$ bounded by a closed contour C within $|w| < 1$. We define a sequence of slits in the following way. Let L_n consist of a straight line segment from $w = 1$ to the nearest point P of C, (fig. 11.4–1) and a part of C which is described from P along the whole of C in the positive sense, except for an arc $P_n P$ of diameter $1/n$. Let $\mathfrak{R}_n$ consist of $|w| < 1$ except for this slit L_n and let

$$f_n(z) = \beta_n(z+a_{2,n} z^2 + \ldots), \qquad \beta_n > 0, \tag{11.4–8}$$

where $a_{2,n}$, etc. also depend on n, of course, map $|z| < 1$ onto $\mathfrak{R}_n$. Now it is easy to see that $\mathfrak{R}$ is the kernel of the sequence $\mathfrak{R}_1, \mathfrak{R}_2, \ldots$, and that this sequence converges to $\mathfrak{R}$. By Carathéodory's theorem the

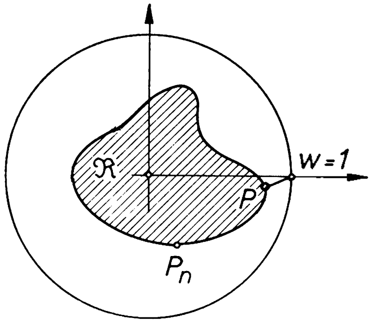


Fig. 11.4–1. Löwner's approximation theorem

functions (11.4–8) tend to $\varphi(z) = f(\rho z)/\rho M$ as $n \to \infty$ and, consequently, $f_n(z)/\beta_n$ to $f(\rho z)/\rho$, uniformly on every closed set within $|z| < 1$.

Our next task is proving that the functions (11.4–7) constitute a family, dense in $\mathscr{S}$. Let $f(z)$ denote an arbitrary function of $\mathscr{S}$ and ρ_n run through an increasing sequence of numbers between 0 and 1, such that $\rho_n \to 1$ as $n \to \infty$. The functions $f(\rho_n z)/\rho_n$ form a sequence which tends to $f(z)$, uniformly on every closed subset of $|z| < 1$. Further we take an increasing sequence of numbers $r_1, r_2, \ldots$, also between 0 and 1, such that $r_n \to 1$ as $n \to \infty$ and, additionally, a decreasing sequence of positive numbers $\varepsilon_1, \varepsilon_2, \ldots$, tending to 0. In view of the results obtained, there is a function $f_1(z)/\beta_1$ in $\mathscr{S}$ of the type described above such that

$$|f_1(z)/\beta_1 - f(\rho_1 z)/\rho_1| < \varepsilon_1,$$

for all $|z| \leqq r$. Again a function $f_2(z)/\beta_2$ in $\mathscr{S}$ such that

$$|f_2(z)/\beta_2 - f(\rho_2 z)/\rho_2| < \varepsilon_2$$

and so on. Thus we obtain a sequence

$$f_1(z)/\beta_1, \qquad f_2(z)/\beta_2, \ldots$$

of functions in $\mathscr{S}$ and it remains to show that they tend to $f(z)$ uniformly for $|z| \leqq r < 1$. Given the number r we can find a number N_1 such that $r_n > r$ for all $n > N_1$. Given $\varepsilon > 0$ we can find a number N_2 such that $\varepsilon_n < \frac{1}{2}\varepsilon$ for all $n > N_2$. Finally there is a number N_3 such that $|f(\rho_n z)/\rho_n - f(z)| < \frac{1}{2}\varepsilon$ for $n > N_3$, all z satisfying $|z| \leqq r < 1$. It follows that for all $n > N = \max(N_1, N_2, N_3)$ and $|z| \leqq r$

$$|f_n(z)/\beta_n - f(z)| \leqq |f_n(z)/\beta_n - f(\rho_n z)/\rho_n| + |f(\rho_n z)/\rho_n - f(z)| < \varepsilon$$

and this proves the assertion.

Calling the functions $f(z)/\beta$, where $f(z)$ is the function (11.4–7), *bounded* *slit mappings* we may state *Löwner's approximation theorem* in the form

The set of bounded slit mappings is dense in $\mathscr{S}$.

In the next section we will study the slit mappings more closely.

11.4.5 – Introduction of a parameter

Let L be a sectionally analytic slit inside $|w| < 1$ given by

$$w = \omega(t), \qquad 0 \leqq t \leqq t_0, \quad t_0 > 0, \tag{11.4–9}$$

where $\omega(t) \neq 0$, $|\omega(t)| < 1$, $(0 \leqq t < t_0)$ and $|\omega(t_0)| = 1$. By $L_{t't''}$ we denote the arc corresponding to the values t which satisfy $t' \leqq t \leqq t''$ and by L_t the arc L_{tt_0}. Let $\mathfrak{R}_t$ consist of $|w| < 1$ except for L_t. As t increases from 0 to t_0 the region $\mathfrak{R}_t$ expands from $\mathfrak{R} = \mathfrak{R}_0$ to $|w| < 1$. In accordance with (11.4–7) we denote by

$$w = g(z, t) = \beta(t)(z + a_2(t)z^2 + \ldots), \qquad \beta(t) > 0 \tag{11.4–10}$$

the function which maps $|z| < 1$ onto $\mathfrak{R}_t$, and by $\check{g}(w, t)$ the inverse of this function.

First we observe that $\beta(t)$ *is continuous*. In fact, $\mathfrak{R}_{t''}$ is a subregion of $\mathfrak{R}_{t'}$, if $t' < t''$ and it is easy to show that the sequence $\mathfrak{R}_{t_1}, \mathfrak{R}_{t_2}, \ldots$, converges to the kernel $\mathfrak{R}_{t^*}$, if $t_1, t_2, \ldots$ tends to t^*. Hence $g(z, t) \to g(z, t^*)$ as t runs through a sequence tending to t^*, uniformly on every closed set within $|z| < 1$. As a consequence $\beta(t) \to \beta(t^*)$.

Now we introduce the function

$$h(z, t', t'') = \check{g}(g(z, t'), t''), \qquad 0 \leqq t' < t'' \leqq t_0. \tag{11.4–11}$$

Its expansion in series is evidently

$$h(z, t', t'') = \frac{\beta(t')}{\beta(t'')} z + \ldots. \tag{11.4–12}$$

The function $g(z, t')$ maps $|z| < 1$ onto $\mathfrak{R}_{t'}$ and $\check{g}(w, t'')$, which is the inverse of $g(z, t'')$, maps $\mathfrak{R}_{t''}$ onto $|z| < 1$. Hence $|h(z, t', t'')| < 1$, while $h(0, t', t'') = 0$. Thus the conditions of Schwarz's lemma are satisfied and we may infer that

$$h'(0, t', t'') = \frac{\beta(t')}{\beta(t'')} < 1, \qquad t' < t'',$$

for it is clear that $h(z, t', t'')$ is not the identity mapping. As a consequence $\beta(t)$ is a strictly increasing function as is

$$\tau = \log \frac{\beta(t)}{\beta(0)}.$$

We may, therefore, take τ for our parameter t, which has been left undetermined so far. With this normalization we have, replacing again τ by t,

$$g(z, t) = \beta e^t(z + a_2 z^2 + \ldots), \qquad 0 \leqq t \leqq t_0 = \log(1/\beta), \tag{11.4–13}$$

$\beta = \beta(0)$ being a positive constant. It follows that

$$h(z, t', t'') = e^{t'-t''}z + \ldots. \tag{11.4–14}$$

In view of the definition (11.4–11) we see intuitively that

$$h(z, t', t'') \to z \tag{11.4–15}$$

as $t'-t'' \to 0$, but the formal proof is not trivial. As we will see this result is of paramount importance for the further development of the theory.

11.4.6 – The continuity property

We recall that $w = g(z, t')$ maps $|z| < 1$ onto $\mathfrak{R}_{t'}$, that is $|w| < 1$ cut along $L_{t'}$. Also that $z = \check{g}(w, t'')$ maps $\mathfrak{R}_{t''}$ onto $|z| < 1$ and so $\mathfrak{R}_{t'}$ corresponds to $|z| < 1$, except for the image of $L_{t't''}$ by $\check{g}(w, t'')$, for $L_{t't''}$ lies in $\mathfrak{R}_{t''}$, except for one point $\omega(t'')$ but not in $\mathfrak{R}_{t'}$. The image of $L_{t't''}$ by $\check{g}(w, t'')$ will be denoted by $S_{t't''}$. Thus $h(z, t', t'')$ maps $|z| < 1$ onto $|z| < 1$ cut along $S_{t't''}$.

If $w = g(z, t)$ then the inverse function $z = \check{g}(w, t)$ is continuous at $\omega(t)$, i.e. as $w \to \omega(t)$ in any manner from $\mathfrak{R}_t$, z approaches a point $\lambda(t)$ such that $|\lambda(t)| = 1$. This point will play an important part in the theory.

We may prove this in quite the same way as the theorem of section 10.5.5. It is clear that the points of $|z| = 1$ are normal boundary points of $|z| < 1$. Let $w_0, w_1, \ldots,$ denote a sequence of points in $\mathfrak{R}_t$ tending to $\omega(t)$. Suppose that the sequence of points $z_0, z_1, \ldots,$ corresponding to these by the mapping $z = \check{g}(w, t)$ has two subsequences tending to z' and z'' respectively. They are on two cuts C'_z and C''_z respectively, corresponding to the curves C'_w and C''_w. We determine closed initial parts C'_1 and C''_1 such that the remainders C'_2 and C''_2 are within two open discs $\mathfrak{K}'$ and $\mathfrak{K}''$ about z' and z'' with radius $\frac{1}{3}d = \frac{1}{3}|z'-z''|$. Now it is not difficult to see how to modify the discussion of section 10.5.5, (fig. 11.4–2), in order to be applicable to the problem under consideration.

The first theorem of section 10.5.5 states that $h(z, t', t'')$ is still continuous at all points of $|z| = 1$. The arc on $|z| = 1$ which corresponds to $S_{t't''}$ will be denoted by $B_{t't''}$, (fig. 11.4–3).

Suppose first that $t'' \to t'$, $(t' < t'')$, while t' remains fixed. Then the arc $L_{t't''}$ shrinks to the fixed point $\omega(t')$. The mapping $\check{g}(w, t')$ is continuous at $\omega(t')$. Hence, given $\varepsilon > 0$, we can find a circumference about

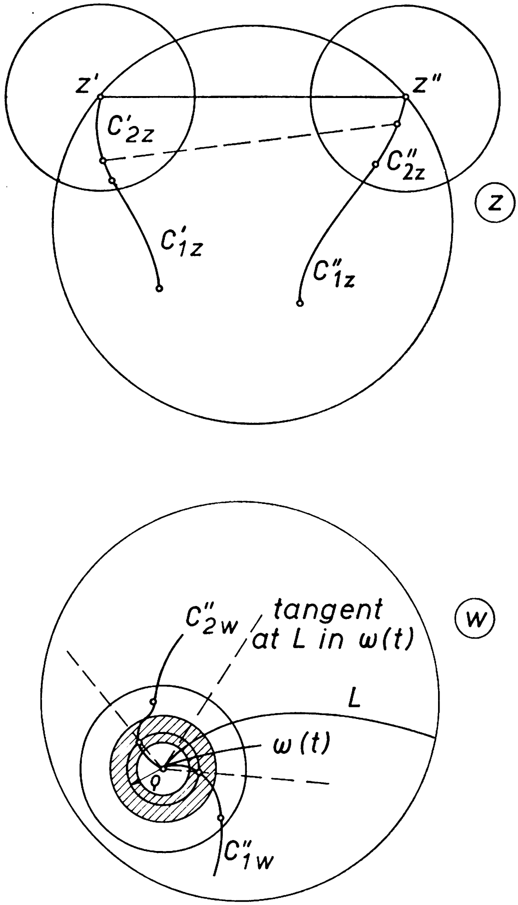


Fig. 11.4–2. Proof of the continuity of the function g at $w = \omega(t)$

$\omega(t')$ with radius δ such that all points of $\mathfrak{R}_{t'}$ within this circle correspond to points z with $|z-\lambda(t')| < \varepsilon$. If t'' is sufficiently near t' the arc $L_{t't''}$ is inside $|w-w(t')| < \delta$. If w tends to a point of this arc then the corresponding point $\check{g}(w, t)$ tends to a point the distance of which from $\lambda(t')$ is not larger than ε. Hence $B_{t't''}$ is contained in the region $|z-\lambda(t')| < \varepsilon$.

Secondly we suppose that $t' \to t''$, while t'' remains fixed. Since $\check{g}(w, t'')$ is continuous at $\omega(t'')$ it is at once clear that the tip of $S_{t't''}$ approaches $\lambda(t'')$ as $t' \to t''$, for then $\omega(t') \to \omega(t'')$.

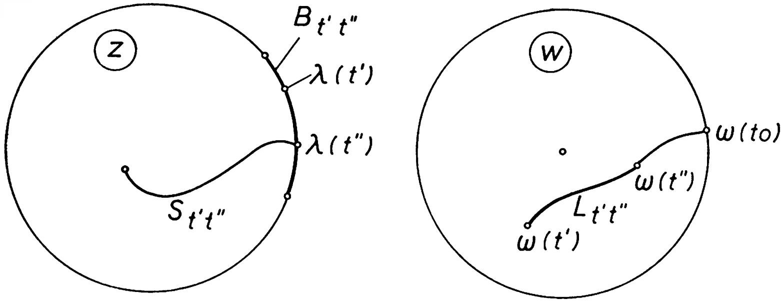


Fig. 11.4–3. The mapping of $|z| < 1$ onto the region $|w| < 1$ slit along the arc L

Now we wish to prove the fundamental result of this section:

As $t'-t'' \to 0$, while either $t = t'$ or $t = t''$ remains fixed, both $S_{t't''}$ and $B_{t't''}$ approach the point $\lambda(t)$. Moreover, $\lambda(t)$ is continuous.

We give the proof for the case that $t = t'$. The proof in the case $t = t''$ is similar.

Take a decreasing sequence of positive numbers $\varepsilon_1, \varepsilon_2, \ldots$ tending to zero and such that $\varepsilon_1 < 1$. In the z-plane on the ray issuing from the origin and passing through $\lambda(t')$ there is a point which is the centre of a circle C_{ε_n} with radius ε_n, $n = 1, 2, \ldots$, which cuts the unit circle orthogonally and is within C_{ε_n}. This circle contains $\lambda(t')$ in its interior. Let $\mathfrak{D}_n$ denote the region inside $|z| < 1/\varepsilon_n$ and outside the circle C_{ε_n}.

Take t'' so near t' that $B_{t't''}$ is inside C_{ε_n}. The function $h(z, t', t'')$ can be continued analytically beyond the arc C outside C_{ε_n} on the unit circle by means of the reflection principle. Hence its values taken at points symmetric with respect to $|z| = 1$ correspond to points also symmetric with respect to the same circumference. Since $e^{t'-t''}$ remains above a certain constant c, taking into account (11.4–14), we find from (11.2–15) that

$$|h(z, t', t'')| \geq \frac{cr}{(1+r)^2}, \qquad |z| = r < 1.$$

It follows from the reflection principle that on $|z| = 1/\varepsilon_n$ we have

$$|h(z, t', t'')| \leq \frac{(1+\varepsilon_n)^2}{c\varepsilon_n} \leq \frac{4}{c\varepsilon_n}.$$

With reference to the maximum modulus principle we infer that $h(r, t', t'')$ is uniformly bounded in $\mathfrak{D}_n$ and the family consisting of all these functions is normal on $\mathfrak{D}_n$.

Now let $t'' \to t'$ through any sequence of values. There is a subsequence such that the corresponding functions

$$h_{11}(z, t'), \qquad h_{12}(z, t'), \ldots \tag{11.4–16}$$

constitute a sequence which converges throughout $\mathfrak{D}_1$, uniformly on every closed subset of $\mathfrak{D}_1$. From a certain index upwards these functions are also defined throughout $\mathfrak{D}_2$ and, therefore, from the sequence (11.4–16) we can find a subsequence

$$h_{21}(z, t'), \qquad h_{22}(z, t'), \ldots \tag{11.4–17}$$

which converges in the same way throughout $\mathfrak{D}_2$. And so on. Now we apply the diagonal method and consider the sequence

$$h_{11}(z, t'), \qquad h_{22}(z, t'), \ldots \tag{11.4–18}$$

From a certain index upwards its members are defined throughout $\mathfrak{D}_n$, where n is fixed, but arbitrary. Hence the sequence is convergent, uniformly on every closed subset of $\mathfrak{D}_n$, and tends, therefore, to a function holomorphic throughout $\mathfrak{D}_n$. It follows that this function has a meaning on every bounded region which does not contain $\lambda(t')$. But the function is locally bounded at $\lambda(t')$ and on account of Riemann's theorem (section 2.8.3) it is still regular there. Thus we obtain a function, holomorphic throughout the whole (open) z-plane, taking the value 0 at $z = 0$ and which has a positive derivative there. From (11.4–14) it follows that this derivative at $z = 0$ is unity and hence the limiting function is the function z. Thus we have proved (11.4–15), the convergence being uniform on every closed set within $|z| < 1$.

We already know that $B_{t't''}$ shrinks to $\lambda(t')$. Now we have the more precise result that $S_{t't''}$ also shrinks to $\lambda(t')$. We conclude the proof by showing that $\lambda(t)$ is continuous.

If z lies on $B_{t't''}$ and s is a corresponding point on $S_{t't''}$ by $h(z, t', t'')$ we have $|z-s| < \varepsilon$ finally. Now $S_{t't''}$ contains $\lambda(t'')$ and $B_{t't''}$ contains $\lambda(t')$. Take $z = \lambda(t')$ and choose t'' so near t' that the diameter of $S_{t't''}$ is less than ε. Then $|s-\lambda(t'')| < \varepsilon$, $|s-z| < \varepsilon$, and so $|\lambda(t'')-\lambda(t')| < 2\varepsilon$. Thus we see that $\lambda(t'') \to \lambda(t)$ as $t'' \to t = t'$. In a similar way we can prove that $\lambda(t') \to \lambda(t)$ as $t' \to t'' = t$.

11.4.7 – Löwner's differential equation

We are now in a position to relate the approximating slit mappings to the solutions of a certain differential equation.

Let L denote the slit $\omega(t)$, $0 \leqq t \leqq t_0$, defined in section 11.4.5 and let $f(z)$, which is the function (11.4–7), map $|z| < 1$ onto $|w| < 1$ except for L. The function $g(z, t)$, represented by the series (11.4–13), maps

$|z| < 1$ onto $|w| < 1$, except for L_t. In particular $g(z, 0) = f(z)$ and $g(z, t_0) = z$. Next we introduce the function

$$f(z, t) = \check{g}(f(z), t). \tag{11.4–19}$$

Since $\check{g}(w, t)$ maps $|w| < 1$ cut along L_t onto $|z| < 1$, the function $f(z, t)$ maps $|z| < 1$ onto a subset of $|z| < 1$ and is a univalent function for $0 \leqq t \leqq t_0$. From (11.4–19) it follows

$$f(z) = \beta z + \ldots = g(f(z, t), t) = \beta e^t f(z, t) + \ldots.$$

Hence

$$f(z, t) = e^{-t} z + \ldots$$

and so $e^t f(z, t)$ is in $\mathscr{S}$. In particular

$$\beta^{-1} f(z, t_0) = e^{t_0} f(z, t_0) = \beta^{-1} f(z)$$

is in $\mathscr{S}$ and from the above considerations it follows that the functions $e^{t_0} f(z, t_0)$ form a dense subclass of $\mathscr{S}$.

Now we turn our attention to the functions $f(z, t)$. It is clear that $f(z, 0) = \check{g}(g(z, 0), 0) = z$.

The function

$$F(z) = \log \frac{h(z, t', t'')}{z} \tag{11.4–20}$$

is such that $F(0)$ is real and $\leqq 0$. It is holomorphic throughout $|z| < 1$ and continuous in $|z| \leqq 1$. It takes the value zero on $|z| = 1$ except on the arc $B_{t't''}$, where it is negative. Let us denote the endpoints of $B_{t't''}$ by $e^{i\alpha}$ and $e^{i\beta}$ respectively.

Applying Schwarz's formula (2.15–13) for the case that $R < 1$, where z is a point inside $|z| = R$, we obtain

$$F(z) = \frac{1}{2\pi} \int_0^{2\pi} \operatorname{Re} F(Re^{i\varphi}) \frac{Re^{i\varphi} + z}{Re^{i\varphi} - z} \, d\varphi.$$

Since the integrand is continuous on $|z| \leqq 1$ (and, consequently, uniformly continuous) we may let $R \to 1$. It follows that

$$\log \frac{h(z, t', t'')}{z} = \frac{1}{2\pi} \int_\alpha^\beta \operatorname{Re} F(e^{i\varphi}) \frac{e^{i\varphi} + z}{e^{i\varphi} - z} \, d\varphi. \tag{11.4–21}$$

Now

$$\begin{aligned} h(f(z, t'), t', t'') &= \check{g}(g(f(z, t'), t'), t'') \\ &= \check{g}(g(\check{g}(f(z), t'), t'), t'') = \check{g}(f(z), t'') = f(z, t''). \end{aligned}$$

If in (11.4–21) we replace z by $f(z, t')$, we get

$$\log \frac{f(z, t'')}{f(z, t')} = \frac{1}{2\pi} \int_{\alpha}^{\beta} \operatorname{Re} F(e^{i\varphi}) \frac{e^{i\varphi} + f(z, t')}{e^{i\varphi} - f(z, t')} d\varphi. \tag{11.4–22}$$

Referring to the expansion (11.4–14) we see from this by taking $z = 0$ that

$$t'' - t' = -\frac{1}{2\pi} \int_{\alpha}^{\beta} \operatorname{Re} F(e^{i\varphi}) d\varphi. \tag{11.4–23}$$

Now we apply a mean value theorem of integral calculus (stating: if $f(t)$ and $g(t)$ are continuous in the interval $\alpha \leqq t \leqq \beta$ and $g(t)$ does not change sign, then there is a number θ in the interval such that $\int_{\alpha}^{\beta} f(t) g(t) dt = f(\theta) \int_{\alpha}^{\beta} g(t) dt$) and we obtain

$$\log \frac{f(z, t'')}{f(z, t')} = \frac{1}{2\pi} \left(\operatorname{Re} \frac{e^{i\theta_1} + f(z, t')}{e^{i\theta_1} - f(z, t')} + i \operatorname{Im} \frac{e^{i\theta_2} + f(z, t')}{e^{i\theta_2} - f(z, t')} \right) \int_{\alpha}^{\beta} \operatorname{Re} F(e^{i\varphi}) d\varphi, \tag{11.4–24}$$

where θ_1 and θ_2 are in $\alpha \leqq t \leqq \beta$. Dividing corresponding members of (11.4–24) and (11.4–23) we find

$$\frac{\log f(z, t'') - \log f(z, t')}{t'' - t'} = -\operatorname{Re} \frac{e^{i\theta_1} + f(z, t')}{e^{i\theta_1} - f(z, t')} - i \operatorname{Im} \frac{e^{i\theta_2} + f(z, t')}{e^{i\theta_2} - f(z, t')}. \tag{11.4–25}$$

Making $t'' \to t = t'$, then, on account of the fact that $e^{i\alpha} \to \lambda(t)$, $e^{i\beta} \to \lambda(t)$ simultaneously, the expression on the right of (11.4–25) tends to

$$-\frac{\lambda(t) + f(z, t)}{\lambda(t) - f(z, t)} = -\frac{1 + \kappa(t) f(z, t)}{1 - \kappa(t) f(z, t)}, \tag{11.4–26}$$

where $\kappa(t) = 1/\lambda(t)$, $|\kappa(t)| = 1$ and is continuous for $0 \leqq t \leqq t_0$. Since we may interchange the rôles of t' and t'', the expression on the left of (11.4–25) tends to

$$\frac{\partial}{\partial t} \log f(z, t) = \frac{\partial f(z, t)/\partial t}{f(z, t)}.$$

Summing up, we may state the following fundamental theorem

The function $f(z, t)$, defined by (11.4–19), *satisfies the differential equation*

$$\boxed{\frac{\partial w}{\partial t} = -w \frac{1 + \kappa(t) w}{1 - \kappa(t) w},} \tag{11.4–27}$$

with the initial condition $f(z, 0) = z$. The functions $e^{t_0} f(z, t_0)$ for varying positive t_0 and functions $\kappa(t)$ form a dense subclass of $\mathscr{S}$. The functions $\kappa(t)$ are continuous for $0 \leqq t \leqq t_0$ and $|\kappa(t)| = 1$.

The differential equation (11.4–27) will be referred to as *Löwner's differential equation.*

11.4.8 – The existence theorem

We shall complete the foregoing considerations by proving the theorem

Let $\kappa(t)$ be a continuous function of the real variable t throughout an interval $0 \leqq t \leqq t_0$, satisfying the condition $|\kappa(t)| = 1$. There exists a unique solution $w = f(z, t)$ of the differential equation

$$\frac{\partial w}{\partial t} = -w\,\frac{1+\kappa(t)w}{1-\kappa(t)w} \tag{11.4–28}$$

such that $f(z, 0) = z$, $|z| < 1$. For any fixed t the function $e^t f(z, t)$ belongs to the class $\mathscr{S}$.

The differential equation (11.4–28) with initial condition $w = z$ at $t = 0$ is equivalent to the integral equation

$$w = z \exp\left(-\int_0^t \frac{1+\kappa(\tau)w}{1-\kappa(\tau)w}\,d\tau\right). \tag{11.4–29}$$

We may solve this equation by applying Picard's method of successive approximations. To this end we construct a sequence of functions by the recurrent relations

$$w_0 = z,\ w_n = w_n(z, t) = z \exp\left(-\int_0^t \frac{1+\kappa(\tau)w_{n-1}}{1-\kappa(\tau)w_{n-1}}\,d\tau\right), \qquad n > 0. \tag{11.4–30}$$

An easy calculation shows that $\operatorname{Re}(1+\kappa w)/(1-\kappa w) > 0$ if $|w| < 1$. Then it follows by induction that $|w_n| \leqq |z|$, as $|z| < 1$ and that $w_n(z, t)$, $n > 0$ considered as a function of z is holomorphic throughout $|z| < 1$ and considered as a function of t is continuous in the interval $0 \leqq t \leqq t_0$. In particular

$$w_n(0, t) = 0, \qquad w_n'(0, t) = e^{-t}, \quad n > 0.$$

If $t = 0$ all functions reduce to z. Differentiating (11.4–30) we get

$$\frac{\partial w_n}{\partial t} = -w_n\,\frac{1+\kappa w_{n-1}}{1-\kappa w_{n-1}} \tag{11.4–31}$$

and thus

$$\frac{\partial(w_n - w_{n-1})}{\partial t} = -(w_n - w_{n-1})\,\frac{1+\kappa w_{n-1}}{1-\kappa w_{n-1}} + \\ -(w_{n-1} - w_{n-2})\,\frac{2\kappa w_{n-1}}{(1-\kappa w_{n-1})(1-\kappa w_{n-2})}$$

All functions w_n coincide at $t = 0$. Hence, by integrating,

$$w_n - w_{n-1} = -\int_0^t (w_n - w_{n-1}) \frac{1+\kappa w_{n-1}}{1-\kappa w_{n-1}} d\tau + $$
$$- \int_0^t (w_{n-1} - w_{n-2}) \frac{2\kappa w_{n-1}}{(1-\kappa w_{n-1})(1-\kappa w_{n-2})} d\tau.$$

Now we use the additional assumption that $|z| \leqq r < 1$. Since $|\kappa w_n| \leqq |z| \leqq r$, $n = 0, 1, \ldots$, we have evidently

$$|w_n - w_{n-1}| \leqq A \int_0^t |w_n - w_{n-1}| d\tau + B \int_0^t |w_{n-1} - w_{n-2}| d\tau,$$

or

$$|w_n - w_{n-1}| - A \int_0^t |w_n - w_{n-1}| d\tau \leqq B \int_0^t |w_{n-1} - w_{n-2}| d\tau,$$

with

$$A = \frac{1+r}{1-r}, \qquad B = \frac{2r}{(1-r)^2}. \tag{11.4-32}$$

This remains true if we multiply both members on the left by e^{-At}. We obtain

$$\int_0^t \frac{\partial}{\partial \tau} e^{-A\tau} |w_n - w_{n-1}| d\tau \leqq B \int_0^t |w_{n-1} - w_{n-2}| d\tau,$$

or

$$e^{-At} |w_n - w_{n-1}| \leqq B \int_0^t |w_{n-1} - w_{n-2}| d\tau,$$

so that we finally have

$$|w_n - w_{n-1}| \leqq e^{At_0} B \int_0^t |w_{n-1} - w_{n-2}| d\tau.$$

By direct computation we see that $|w_1 - w_0| < K$, where K is again a constant depending only on r. By induction we find

$$|w_n - w_{n-1}| \leqq K \frac{(e^{At_0} B t)^{n-1}}{(n-1)!} \leqq K \frac{(e^{At_0} B t_0)^{n-1}}{(n-1)!}.$$

It follows with reference to Weierstrass's test (section 1.5.4) that the series

$$w_0 + (w_1 - w_0) + (w_2 - w_1) + \ldots$$

is uniformly convergent as regards z in $|z| \leqq r$ and as regards t in the interval $0 \leqq t \leqq t_0$. Hence the partial sums $w_n(z, t)$ tend to a function $w = f(z, t)$ as $n \to \infty$ which is holomorphic throughout $|z| < 1$ (section

2.20.1), and continuous throughout the interval $0 \leqq t \leqq t_0$, since $w_n(z, t)$ is continuous with respect to t (section 1.5.3). In particular

$$f(0, t) = 0, \qquad f'(0, t) = e^{-t}$$

and since $w_n(z, 0) = z$, $|w_n(z, t)| \leqq |z|$, we have also

$$f(z, 0) = z, \qquad |f(z, t)| \leqq |z|.$$

Since $w_n(z, t)$ converges uniformly it is alllowed to let $n \to \infty$ in (11.4–30) and we obtain

$$f(z, t) = z \exp\left(-\int_0^t \frac{1+\kappa(\tau)f(z, \tau)}{1-\kappa(\tau)f(z, \tau)}\, d\tau\right). \tag{11.4–33}$$

Thus it appears that $f(z, t)$ is differentiable with respect to t and it follows that it is a solution of (11.4–28), having the assigned initial value.

Suppose that $f_1(z, t)$ is a solution of (11.4–28) such that $f(z_1, t_1) = f_1(z_2, t_1)$ for some pairs z_1, t_1 and z_2, t_1, with $0 \leqq t_1 \leqq t_0$.

Writing for short $f = f(z_1, t), f_1 = f_1(z_2, t)$, we have in view of (11.4–28)

$$\frac{\partial(f-f_1)}{\partial t} = -(f-f_1)\frac{1+\kappa f}{1-\kappa f} - (f-f_1)\frac{2\kappa f_1}{(1-\kappa f)(1-\kappa f_1)}. \tag{11.4–34}$$

Let r denote a number such that $|z_1| < r < 1$. Then $|f(z_1, t)| \leqq |z_1| < r$. Since $f_1(z_2, t)$ depends continuously on t we also have, if $t_1 < t_0$,

$$|f_1(z_2, t)| < r,$$

provided that $t_1 \leqq t \leqq t_1 + \alpha \leqq t_0$, where α is a suitably chosen positive number. It follows from (11.4–34) that for these values of t

$$|f - f_1| \leqq (A+B)\int_{t_1}^{t} |f-f_1|\, d\tau = C\int_{t_1}^{t} |f-f_1|\, d\tau,$$

where A and B have the values (11.4–32). Let μ denote the maximum of $|f-f_1|$ on the interval $t_1 \leqq t \leqq t_1+\alpha$. Then $|f-f_1| \leqq C\mu\alpha$ and if we take $\alpha < \frac{1}{2}C$ we even have $|f-f_1| \leqq \frac{1}{2}\mu$, in contradiction with the meaning of μ, unless $\mu = 0$. A similar assertion is true in the case that $t_1 > 0$ in an interval $0 \leqq t_1 - \alpha \leqq t \leqq t_1$. Thus we see that the subset of all values t for which $f = f_1$ is an open subset of the interval $0 < t < t_0$. On the other hand, because of the continuity of $f - f_1$ the subset of all t, for which $f \neq f_1$, is likewise an open set. Since the interval $0 < t < t_0$ is a connected set this latter subset must be empty. We conclude, again with reference to continuity, that even for $0 \leqq t \leqq t_0$ the solutions f and f_1 coincide.

A direct consequence is the fact that $f(z, t)$ is univalent. For if $f(z_1, t)$

and $f(z_1, t)$ take equal values for some $t = t_1$ then $f(z_1, t) = f(z_2, t)$ throughout the whole interval $0 \leqq t \leqq t_0$ and in particular $f(z_1, 0) = f(z_2, 0)$, i.e., $z_1 = z_2$. Since $f'(0, t) = e^{-t}$ the function $e^t f(z, t)$ belongs to the class $\mathscr{S}$.

It should be noticed that the mapping of the unit circle as given by $e^t f(z, t)$ is not necessarily always a slit mapping.

We conclude this section with a final remark.

Suppose that the conditions of the above theorem are fulfilled for all. From (11.4–33) we deduce

$$e^t f(z, t) = z \exp \int_0^t \left(1 - \frac{1+\kappa f}{1-\kappa f}\right) d\tau = z \exp\left(-\int_0^t \frac{2\kappa f}{1-\kappa f}\, d\tau\right).$$

Since $|f(z, t)| \leqq |z|$ the integral in the last member is uniformly convergent as $t \to \infty$ and z remains in a closed subset of $|z| < 1$. We may apply Weierstrass's theorem of section 2.20.5 and it follows that the limiting function

$$f(z) = \lim_{t \to \infty} e^t f(z, t) \tag{14.4–35}$$

is holomorphic throughout $|z| < 1$.

11.5 – Applications of Löwner's theory

11.5.1 – The sharp form of Bieberbach's rotation theorem

At the end of section 11.2.4 we obtained a limitation of $\arg f'(z)$, if $f'(z)$ belongs to the class $\mathscr{S}$. But the result (11.2–23) is not sharp. The exact form of the rotation theorem has been obtained by Golusin by using Löwner's theory. We know that the class of the solutions of Löwner's differential equation, by taking all possible functions $\kappa(t)$ includes the slit mappings and with them we may approximate a given function of the class $\mathscr{S}$ as closely as desired. Hence we may confine ourselves to the functions $e^t f(z, t)$, where $f(z, t)$ satisfies a differential equation of Löwner.

In the case that we wish to estimate the argument we can omit the factor e^t and it suffices to consider a function $f(z, t)$ which satifies the equation

$$\frac{\partial f(z, t)}{\partial t} = -f(z, t)\frac{1+\kappa(t) f(z, t)}{1-\kappa(t) f(z, t)} = f + \frac{2}{\kappa} - \frac{2}{\kappa(1-\kappa f)},$$

while $f(z, 0) = z$, writing $f = f(z, t)$, $\kappa = \kappa(t)$ for short. Differentiation with respect to z yields

$$\frac{\partial}{\partial t} f'(z, t) = f'(z, t)\left(1 - \frac{2}{(1-\kappa f)^2}\right),$$

or

$$\frac{\partial}{\partial t}\log f'(z,t) = 1 - \frac{2}{(1-\kappa f)^2}.$$

Equating imaginary parts we obtain

$$\frac{\partial}{\partial t}\arg f'(z,t) = \frac{2\,\mathrm{Im}\,(1-\kappa f)^2}{|1-\kappa f|^4}. \tag{11.5–1}$$

On the other hand we also have

$$\frac{\partial}{\partial t}\log f = -\frac{1+\kappa f}{1-\kappa f} = -\frac{(1+\kappa f)(1-\overline{\kappa f})}{|1-\kappa f|^2}.$$

Equating real and imaginary parts we get

$$\frac{\partial}{\partial t}\log|f| = -\frac{1-|f|^2}{|1-\kappa f|^2} \tag{11.5–2}$$

and

$$\frac{\partial}{\partial t}\arg f = -\frac{2\,\mathrm{Im}\,\kappa f}{|1-\kappa f|^2}. \tag{11.5–3}$$

Combining (11.5–1) and (11.5–2) we deduce

$$\frac{\partial}{\partial t}\arg f' = \frac{2\,\mathrm{Im}\,(1-\kappa f)^2}{|1-\kappa f|^2}\,\frac{-1}{|f|(1-|f|)^2}\,\frac{\partial|f|}{\partial t}.$$

From De Moivre's theorem we deduce, since $|\kappa(t)| = 1$,

$$\frac{\mathrm{Im}\,(1-\kappa f)^2}{|1-\kappa f|^2} = \sin\big(2\arg(1-\kappa f)\big).$$

Hence, since $|\kappa(t)| = 1$, (fig. 11.5.1),

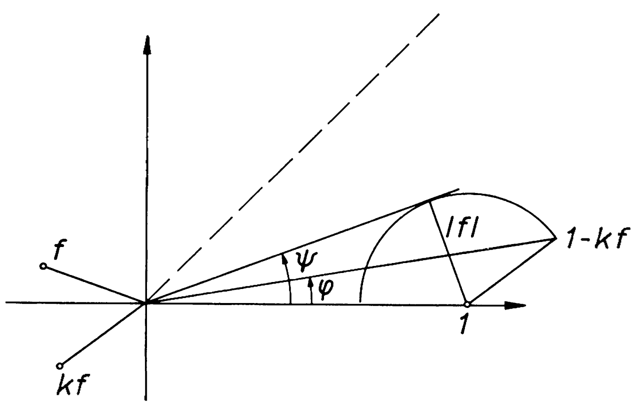


Fig. 11.5–1. The inequalities (11.5–4)

$$\frac{|\mathrm{Im}\,(1-\kappa f)^2|}{|1-\kappa f|^2} \leqq \begin{cases} \sin(2\arcsin|f|) = 2|f|\sqrt{1-|f|^2}, & \text{if } |f| \leqq \dfrac{1}{\sqrt{2}}, \\ 1, & \text{if } |f| \geqq \dfrac{1}{\sqrt{2}}. \end{cases} \tag{11.5–4}$$

Because of $-\partial|f|/\partial t > 0$, we have

$$\frac{\partial}{\partial t}\arg f' \leqq \begin{cases} \dfrac{-4}{\sqrt{1-|f|^2}}\cdot\dfrac{\partial|f|}{\partial t}, & \text{if } |f| \leqq \dfrac{1}{\sqrt{2}}, \\ \dfrac{-2}{|f|(1-|f^2|)}\cdot\dfrac{\partial|f|}{\partial t}, & \text{if } |f| \geqq \dfrac{1}{\sqrt{2}}. \end{cases}$$

Integrating from $t = 0$ to $t = t_0 > 0$, we obtain for $|z| \leqq 1/\sqrt{2}$

$$|\arg f'| \leqq \int_0^{t_0} \frac{-4}{\sqrt{1-|f|^2}}\frac{\partial|f|}{\partial t}\,dt = \int_{|f(z,t_0)|}^{|z|}\frac{4dx}{\sqrt{1-x^2}} \leqq 4\arcsin|z|,$$

and for $|z| \geqq 1/\sqrt{2}$

$$|\arg f'| \leqq \int_{|f(z,t_0)|}^{1/\sqrt{2}}\frac{4dx}{\sqrt{1-x^2}} + \int_{1/\sqrt{2}}^{|z|}\frac{2dx}{x(1-x^2)} \leqq \pi + \log\frac{|z|^2}{1-|z|^2}.$$

Thus we proved:

If $f(z)$ belongs to the class $\mathscr{S}$ then we have the inequalities

$$|\arg f'(z)| \leqq \begin{cases} 4\arcsin|z|, & |z| \leqq 1/\sqrt{2}, \\ \pi + \log\dfrac{|z|^2}{1-|z|^2}, & 1 > |z| \geqq 1/\sqrt{2}. \end{cases} \tag{11.5–5}$$

These inequalities are sharp. To prove this we must find $\kappa(t)$ such that if $f(z_0, t)$ is the solution of (11.4–28) with $f(z_0, t) = z_0$, z_0 being a value of z within the unit circle, then equality holds in (11.5–4). The resulting equation enables us to calculate $\kappa f(z_0, t)$ in terms of $|f(z_0, t)|$, and $|f(z_0, t)|$ in terms of t by means of (11.5–2) and (11.5–3). We can then choose $\kappa(t)$, so that equality holds in (11.5–4). When this is done for $0 \leqq t \leqq t_0$ then (11.5–2) and (11.5–3) and equality in (11.5–4) will hold simultaneously and we shall have

$$\arg f'(z_0, t_0) = \begin{cases} \displaystyle\int_{|f(z_0,t_0)|}^{|z_0|}\frac{4dx}{\sqrt{1-x^2}}, & \text{if } |z_0| \leqq 1/\sqrt{2}, \\ \displaystyle\int_{|f(z_0,t_0)|}^{1/\sqrt{2}}\frac{4dx}{\sqrt{1-x^2}} + \log\frac{|z_0|^2}{1-|z_0|^2}, & \text{if } |z_0| \geqq 1/\sqrt{2}, \end{cases}$$

and if $|f(z_0, t_0)| \leqq 1/\sqrt{2}$, and so for all large t_0. For, $f(z_0, t_0) \to 0$ as

$t_0 \to \infty$, since $e^t f(z, t)$ remains bounded. Thus the upper bounds of the theorem stated above may be approached as closely as desired and so are sharp. By considering $\overline{f(z_0, t)}$ instead of $f(z_0, t)$ we can show in the same way that also the lower bounds are sharp.

By the same method we can also find sharp bounds for $\arg(f(z)/z)$ which we need in the next section. From (11.5–2) and (11.5–3) we deduce

$$\left|\frac{\partial}{\partial t}\arg f\right| = -\frac{|2\operatorname{Im}\kappa f|}{1-|f|^2}\frac{\partial}{\partial t}\log|f| \leq \frac{-2|f|}{1-|f|^2}\frac{\partial}{\partial t}\log|f|.$$

Integrating from 0 to t_0 we find, by noting that $f(z, 0) = z$

$$\left|\arg\frac{f(z,t)}{z}\right| \leq \int_{|f(z,t_0)|}^{|z|}\frac{2dx}{1-x^2} \leq \int_0^{|z|}\frac{2dx}{1-x^2} = \log\frac{1+|z|}{1-|z|}.$$

Hence

If $f(z)$ belongs to the class $\mathscr{S}$ then the inequality

$$\boxed{\left|\arg\frac{f(z)}{z}\right| \leq \log\frac{1+|z|}{1-|z|},} \qquad |z| < 1 \qquad (11.5\text{–}6)$$

holds.

This result is sharp. Let z_0 denote a fixed value of z with $|z_0| < 1$. We have to find $\kappa(t)$ in an assigned range $0 \leq t \leq t_0$ such that the solution $f(z_0, t)$ of (11.4–28) with the initial condition $f(z_0, 0) = z_0$ satisfies $\kappa f = -i|f|$. Since then because of (11.5–2)

$$\frac{\partial|f|}{\partial t}\Big/|f| = \frac{\partial}{\partial t}\log|f| = -\frac{1-|f|^2}{1+|f|^2},$$

we see that $f = f(z_0, t)$ satisfies

$$\frac{|f|}{1-|f|^2} = e^{-t}\frac{|z_0|}{1-|z_0|^2}.$$

This result may be used as a definition of $|f(z_0, t)|$. We calculate $\arg f(z_0, t)$ from

$$\frac{\partial}{\partial t}\arg f = \frac{2|f|}{1+|f|^2} = -\frac{2}{1-|f|^2}\frac{\partial|f|}{\partial t},$$

so that

$$\arg\frac{f(z_0,t)}{z_0} = \log\frac{1+|z_0|}{1-|z_0|} - \log\frac{1+|f|}{1-|f|}.$$

Finally $\kappa(t)$ shall be the function $-i|f(z_0, t)|/f(z_0, t)$. With these definitions (11.5–2) and (11.5–3) are satisfied and $\arg(f(z_0, t)/z_0)$ can be chosen as close as desired to $\log(1+|z_0|)/(1-|z_0|)$, since $f(z_0, t) \to 0$ as $t \to \infty$. This completes the proof of our assertion.

11.5.2 – THE RADII OF CONVEXITY AND STARSHAPEDNESS

In section 11.3.5 we obtained the result that a function $f(z)$ of class $\mathcal{S}$ maps the circumference $|z| = r < 1$ onto a convex curve if and only if

$$1+\operatorname{Re}\frac{zf''(z)}{f'(z)} > 0, \qquad |z| = r < 1. \tag{11.5–7}$$

The inequality (11.2–11) gives for $f(z)$

$$\operatorname{Re}\frac{zf''(z)}{f'(z)} \geqq \frac{r(2r-4)}{1-r^2}.$$

Hence our condition (11.5–7) is satisfied for those values of r for which $2r^2-4r > r^2-1$ holds, or $r^2-4r+1 > 0$, i.e., $r < 2-\sqrt{3}$.
If

$$f(z) = \frac{z}{(1-z)^2}$$

then

$$\frac{zf''(z)}{f'(z)} = \frac{2z^2+4z}{1-z^2}$$

and this is real and less than -1 for $-1 < z < \sqrt{3}-2$. Thus this function does not map $|z| = r$ onto a convex curve for $r \geqq 2-\sqrt{3}$. Summing up we have

By all functions $f(z)$ beloning to the class $\mathcal{S}$ the circumference $|z| = r < 1$ is mapped onto a convex curve if

$$0 < r < r_c,$$

where

$$r_c = 2-\sqrt{3}$$

is called the radius of convexity.

A similar but much more difficult problem arises if we ask for the radius of the largest circle $|z| = r$ such that the image of it by $f(z)$ (this being again any function of class $\mathcal{S}$) bounds a starshaped region with respect to the origin. The condition for this was seen in section 11.3.4 to be (comp. (11.3–7) and (11.3–6))

$$\operatorname{Re}\frac{zf'(z)}{f(z)} > 0, \qquad |z| = r < 1,$$

i.e.,

$$\left|\arg\frac{zf'(z)}{f(z)}\right| < \tfrac{1}{2}\pi. \tag{11.5–8}$$

In order to obtain sharp bounds for the expression on the left of (11.5–8) we apply (11.5–6) to the function

$$g(z) = \frac{f\left(\dfrac{z+z_0}{1+z\bar{z}_0}\right) - f(z_0)}{(1-|z_0|^2)f'(z_0)},$$

this function belonging to the class $\mathscr{S}$ if and only if $f(z)$ belongs to $\mathscr{S}$. At $z = -z_0$ we have

$$\frac{g(-z_0)}{-z_0} = \frac{f(z_0)}{(1-|z_0|^2)z_0 f'(z_0)}$$

and thus (see (11.5–6))

$$\left|\arg \frac{z_0 f'(z_0)}{f(z_0)}\right| = \left|\arg \frac{g(-z_0)}{-z_0}\right| \leqq \log \frac{1+|z_0|}{1-|z_0|},$$

This result is sharp, for replacing in the first member f by g at $z = -z_0$ we obtain (11.5–6) agains, whence:
By all functions $f(z)$ belonging to the class $\mathscr{S}$ the circumference $|z| = r < 1$ is mapped onto a curve which bounds a starshaped region with respect to the origin if

$$0 < r < r_s$$

where the radius of starshapedness r_s satisfies the equation

$$\log \frac{1+r_s}{1-r_s} = \tfrac{1}{2}\pi$$

i.e., r_s has the sharp value

$$r_s = \tanh \tfrac{1}{4}\pi.$$

11.5.3 – The third coefficient

Löwner's theory enables us to obtain an exact upper bound for $|a_3|$ in the Taylor expansion of $f(z)$, this function belonging to the class $\mathscr{S}$. We may confine ourselves to functions of the type $e^t f(z, t)$, where $f(z, t)$ is a solution of Löwner's differential equation with $f(z, 0) = z$.
Let

$$f(z, t_0) = e^{t_0}(z + \sum_{\nu=2}^{\infty} a_\nu z^\nu), \qquad \beta = e^{-t_0}. \tag{11.5–9}$$

It is convenient to introduce the function

$$g_{t_0}(z, t) = f(\check{f}(z, t), t_0) = \beta e^t(z + \sum_{\nu=2}^{\infty} c_\nu(t) z^\nu) \tag{11.5–10}$$

where $\check{f}$ denotes the inverse of f. Thus

$$g_{t_0}(f(z, t), t) = f(z, t_0)$$

and differentiation with respect to t yields

$$\frac{\partial g_{t_0}}{\partial z}\frac{\partial f}{\partial t}+\frac{\partial g_{t_0}}{\partial t} = 0.$$

It follows that in a neighbourhood of $z = 0$ we must have (comp. (11.4–24))

$$\frac{\partial g_{t_0}}{\partial t} = \frac{\partial g_{t_0}}{\partial z} f \frac{1+\kappa(t)f}{1-\kappa(t)f}.$$

In view of (2.16–7) we have

$$e^{t-t_0}c_n(t) = \frac{1}{2\pi i}\int_C \frac{g_{t_0}(\zeta, t)}{\zeta^{n+1}}\,d\zeta, \qquad n = 2, 3, \ldots,$$

C denoting a small circle around the origin. Differentiation with respect to t (section 2.9.1) yields for $n \geqq 2$

$$e^{t-t_0}(c_n(t)+c_n'(t)) = \frac{1}{2\pi i}\int_C \frac{\partial g_{t_0}(\zeta, t)}{\partial t}\frac{d\zeta}{\zeta^{n+1}}$$

$$= \frac{1}{2\pi i}\int_C \frac{\partial g_{t_0}}{\partial \zeta}\zeta\frac{1+\kappa\zeta}{1-\kappa\zeta}\frac{d\zeta}{\zeta^{n+1}} = \frac{e^{t-t_0}}{2\pi i}\int_C \Big(1+\sum_{\nu=2}^{\infty}\nu c_\nu(t)\zeta^{\nu-1}\Big)\Big(1+2\sum_{\nu=1}^{\infty}\kappa^\nu\zeta^\nu\Big)\frac{d\zeta}{\zeta^n}$$

$$= e^{t-t_0}\Big(nc_n(t)+2\kappa^{n-1}(t)+2\sum_{\mu=1}^{n-1}\mu c_\mu(t)\kappa^{n-\mu}(t)\Big), \qquad 0 \leqq t \leqq t_0.$$

Hence

$$c_n'(t) = (n-1)c_n(t)+2\kappa^{n-1}(t)+2\sum_{\mu=2}^{n-1}\mu c_\mu(t)\kappa^{n-\mu}(t). \qquad (11.5\text{–}11)$$

If $t = 0$, then $f(z, 0) = z$, $g_{t_0}(z, 0) = f(z, t_0)$, $c_n(0) = a_n$ and if $t = t_0$, then $g_{t_0}(z, t_0) = z$ and so $c_n(t_0) = 0$, $n \geqq 2$. Thus we have the boundary conditions

$$c_n(0) = a_n, \quad c_n(t_0) = 0, \quad n \geqq 2.$$

From (11.5–11) we can determine successively the coefficients. In fact

$$c_n(t) = -2e^{(n-1)t}\int_t^{t_0} e^{-(n-1)\tau}\Big(\kappa^{n-1}(\tau)+\sum_{\mu=1}^{n-1}\mu c_\mu(\tau)\kappa^{n-\mu}(\tau)\Big)d\tau, \qquad (11.5\text{–}12)$$

as may be verified by differentiating both members. In particular

$$c_2(t) = -2e^t\int_t^{t_0} e^{-\tau}\kappa(\tau)d\tau, \quad c_2(0) = -2\int_0^{t_0} e^{-\tau}\kappa(\tau)dt. \qquad (11.5\text{–}13)$$

Hence for all $t_0 > 0$ we have $|a_2| = |c_2(0)| \leqq \int_0^{t_0} e^{-\tau} d\tau$, whence $|a_2| \leqq 2 \int_0^{\infty} e^{-\tau} d\tau = 2$, which in Bieberbach's result (11.2–1).

Observing that

$$\frac{d}{dt}\left(\int_t^{t_0} e^{-\tau}\kappa(\tau)d\tau\right)^2 = -2\kappa(t)\, e^{-t}\int_t^{t_0} e^{-\tau}\kappa(\tau)d\tau = \kappa(t)e^{-2t}c_2(t)$$

we also have

$$c_3(t) = -2e^{2t}\left(\int_t^{t_0} e^{-2\tau}\kappa^2(\tau)d\tau + 2\int_t^{t_0}\kappa(\tau)e^{-2\tau}c_2(\tau)d\tau\right)$$

$$= -2e^{2t}\left(\int_t^{t_0} e^{-2\tau}\kappa^2(\tau)d\tau - 2\left(\int_t^{t_0} e^{-\tau}\kappa(\tau)d\tau\right)^2\right). \qquad (11.5\text{–}14)$$

Since the function $\eta^{-1}f(\eta z)$ with $|\eta| = 1$ belongs to the class $\mathscr{S}$, if $f(z)$ does so, the maximum of $|a_3|$ with respect to the class $\mathscr{S}$ coincides with the maximum of $\operatorname{Re} a_3$. We take

$$\kappa(t) = e^{i\theta(t)}$$

and consider

$$\operatorname{Re} a_3 = -2\int_0^{t_0} e^{-2t}(2\cos^2\theta(t) - 1)dt + 4\left(\int_0^{t_0} e^{-t}\cos\theta(t)dt\right)^2 + -4\left(\int_0^{t_0} e^{-t}\sin\theta(t)dt\right)^2.$$

We omit the third term on the right. The first term is

$$-4\int_0^{t_0} e^{-2t}\cos^2\theta(t)dt + 1 - e^{-2t_0}$$

and by Schwarz's inequality (10.5–4) the second term does not exceed

$$4\int_0^{t_0} e^{-t}dt \int_0^{t_0} e^{-t}\cos^2\theta(t)dt < 4\int_0^{t_0} e^{-t}\cos^2\theta(t)dt.$$

Thus we obtain

$$\operatorname{Re} a_3 \leqq 1 + 4\int_0^{t_0}(e^{-t} - e^{-2t})\cos^2\theta(t)dt \leqq 1 + 4\int_0^{\infty}(e^{-t} - e^{-2t})dt = 3,$$

i.e.,

$$|a_3| \leqq 3.$$

As we know this result is sharp.

11.5.4 – Coefficients of the inverse functions

A very remarkable success of Löwner's theory is the fact that it enables us to get sharp bounds for all coefficients of the expansions of the inverses of the functions belonging to the family $\mathscr{S}$. The theory is applicable to this type of functions, for the coefficients of their expansions are obtained from those of the expansions of the functions of $\mathscr{S}$ by means of rational operations. Hence if a given function of $\mathscr{S}$ is approximated by a certain class of functions, then the inverse is approximated by the inverses of the functions of the class.

Let, again, $f(z, t)$, $0 \leqq t \leqq t_0$ be a solution of Löwner's differential equation, subject to the initial condition $f(z, 0) = z$. Near $w = 0$ this function is invertible; the inverse will be denoted by $\varphi_t(w)$, so that

$$\varphi_t(f(z, t)) = z. \tag{11.5–15}$$

We may write the expansion of $\varphi_t(w)$ near $w = 0$ as

$$\varphi_t(w) = \mathrm{e}^t\left(w + \sum_{\nu=2}^{\infty} b_\nu(t) w^\nu\right), \tag{11.5–16}$$

for, as we pointed out in section 11.4.7, the function $\mathrm{e}^t f(z, t)$ is in $\mathscr{S}$, that is to say

$$f(z, t) = \mathrm{e}^{-t}z + \ldots .$$

The function

$$\varphi(w) = \varphi_{t_0}(\mathrm{e}^{-t_0}w) = w + \sum_{\nu=2}^{\infty} b_\nu w^\nu, \tag{11.5–17}$$

with

$$b_n = b_n(t_0)\mathrm{e}^{-(n-1)t_0}, \qquad n \geqq 2, \tag{11.5–18}$$

is inverse to $\mathrm{e}^{t_0} f(z, t_0)$ as follows from

$$\varphi(\mathrm{e}^{t_0} f(z, t_0)) = \varphi_{t_0}(f(z, t_0)) = z.$$

Since the functions $\mathrm{e}^{t_0} f(z, t_0)$ for varying t_0 and functions $\kappa(t)$ form a dense subclass of $\mathscr{S}$, it suffices to prove inequalities for the coefficients b_n.

The following discussion is very similar to that of the previous section. Differentiating (11.5–15) yields

$$\frac{\partial \varphi_t}{\partial w} \frac{\partial f(z, t)}{\partial t} + \frac{\partial \varphi_t}{\partial t} = 0$$

and because $f(z, t)$ is a solution of (11.4–27) we have

$$\frac{\partial \varphi_t(w)}{\partial t} = \frac{\partial \varphi_t(w)}{\partial w} w \frac{1 + \kappa(t)w}{1 - \kappa(t)w}, \tag{11.5–19}$$

where w stands for $f(z, t)$.

Let C denote a small circumference about $w = 0$. From

$$e^t b_n(t) = \frac{1}{2\pi i}\int_C \frac{\varphi_t(\zeta)}{\zeta^{n+1}}\, d\zeta$$

we obtain by differentiating with respect to t

$$e^t(b_n'(t)+b_n(t)) = \frac{1}{2\pi i}\int_C \frac{\partial \varphi_t(\zeta)}{\partial t}\frac{d\zeta}{\zeta^{n+1}} = \frac{1}{2\pi i}\int_C \frac{\partial \varphi_t(\zeta)}{\partial \zeta}\frac{1+\kappa(t)\zeta}{1-\kappa(t)\zeta}\frac{d\zeta}{\zeta n}$$

$$= \frac{e^t}{2\pi i}\int_C (1+\sum_{\nu=2}^{\infty}\nu b_\nu(t)\zeta^{\nu-1})(1+2\sum_{\nu=1}^{\infty}\kappa^\nu(t)\zeta^\nu)\frac{d\zeta}{\zeta^n}$$

$$= e^t\big(nb_n(t)+2\kappa^{n-1}(t)+2\sum_{\nu=2}^{n-1}\nu b_\nu(t)\kappa^{n-\nu}(t)\big),$$

whence

$$b_n'(t) = (n-1)b_n(t)+2\kappa^{n-1}(t)+2\sum_{\nu=1}^{n-1}\nu b_\nu(t)\kappa^{n-\nu}(t), \qquad (11.5\text{–}20)$$

throughout the interval $0 \leqq t \leqq t_0$.

Taking into account (11.5–18) we see that the initial conditions for the coefficients are

$$b_n(0) = 0, \qquad b_n(t_0) = e^{(n-1)t_0}b_n, \qquad n \geqq 2.$$

Thus the equations (11.5–20) are equivalent to

$$b_n(t) = 2e^{(n-1)t}\int_0^t e^{-(n-1)\tau}(\kappa^{n-1}(\tau)+\sum_{\nu=2}^{\infty}\nu b_\nu(\tau)\kappa^{n-\nu}(\tau))d\tau. \qquad (11.5\text{–}21)$$

By induction on n we may prove that the $b_n(t)$ attain their maximum possible value for any fixed $t_0 > 0$ if $\kappa(t) = 1$ identically. Then all these coefficients are real and positive. Moreover, $b_n = e^{-(n-1)t_0}\, b_n(t_0)$ increases with increasing t_0, as follows again from (11.5–21). As a consequence of this the upper bounds for the coefficients b_n are obtained in the limit if $t_0 \to \infty$.

The extremal function may be obtained by solving Löwner's differential equation with $\kappa(t) = 1$ identically. In this case it takes the form

$$\frac{\partial w}{\partial t} = w\,\frac{w+1}{w-1},$$

or

$$\frac{\partial w}{\partial t}\left(\frac{1}{w} - \frac{2}{w+1}\right) = -1.$$

A solution $w(z, t)$ with $w(z, 0) = z$ satisfies the relation

$$\frac{w}{(1+w)^2} = e^{-t}c(z),$$

where $c(z)$ does not depend on t. By taking $t = 0$ we find

$$c(z) = \frac{z}{(1+z)^2}$$

and it appears that $f(z, t)$ satisfies the relation

$$\frac{e^t f(z,t)}{(1+f(z,t))^2} = \frac{z}{(1+z)^2}.$$

Since $e^t f(z, t)$ remains bounded as $t \to \infty$, we conclude that $\lim_{t\to\infty} f(z, t) = 0$, whence

$$\lim_{t\to\infty} e^t f(z,t) = \frac{z}{(1+z)^2}.$$

Thus we see that the extremal function for the problem under consideration is, therefore, the inverse $\check{k}(w)$ of this function.

It remains to evaluate the coefficients of the expansion of $\check{k}(w)$ near $w = 0$. Let us write

$$\check{k}(w) = \sum_{\nu=1}^{\infty} k_\nu w^\nu.$$

According to (3.12–8) we have

$$k_n = \frac{1}{2\pi i n}\int_C \frac{(1+\zeta)^{2n}}{\zeta^n}\, d\zeta = \frac{1}{2\pi i n}\int_C \sum_{\nu=0}^{2n} \binom{2n}{\nu} \zeta^{n-\nu} d\zeta = \frac{1}{n}\binom{2n}{n+1}.$$

In particular, $k_1 = 1$. Thus we have proved:

If $w = f(z)$ is a function of the family $\mathcal{S}$ and if

$$z = \check{f}(w) = w + \sum_{\nu=2}^{\infty} b_\nu w^\nu$$

is the expansion of its inverse near $w = 0$, then

$$\boxed{|b_n| \leqq \frac{1}{n}\binom{2n}{n+1}, \quad n \geqq 2.} \tag{11.5–22}$$

Equality holds for the coefficients of the expansion of $\check{k}(w)$, the inverse of the function $k(z) = z/(1+z)^2$. It should be noticed that this function is essentially Koebe's function (11.2–5).

CHAPTER 12

ANALYTIC FUNCTIONS – RIEMANN SURFACES

12.1 – Analytic continuation

12.1.1 – INTRODUCTION

Strictly speaking the notion of a function is a pair of two things, viz. a set $\mathfrak{S}$ of complex numbers (possibly including ∞) and a rule f which assigns to every element of $\mathfrak{S}$ a complex number or ∞, called the *value* $f(z)$ of f. Sometimes it is convenient to exhibit the set $\mathfrak{S}$ and we denote the function defined on it by $(f, \mathfrak{S})$. It is clear that the rule f applies to every subset $\mathfrak{S}'$ of $\mathfrak{S}$. The pair $(f, \mathfrak{S}')$ is then called a *restriction* of the function. It may happen that another rule yields the same values of the restriction. Thus, for instance, the function defined by the rule

$$\frac{1}{1-z} \tag{12.1–1}$$

applies to the whole z-plane, from which $z = 1$ is excluded (supposed we do not admit the value ∞). The restriction to the region $|z| < 1$ can be obtained by means of the rule

$$\sum_{\nu=0}^{\infty} z^{\nu}. \tag{12.1–2}$$

It is natural to consider the inverse problem. Given a function $(f, \mathfrak{S}')$ and a set $\mathfrak{S}$ including $\mathfrak{S}'$, we may ask: is there a function $(g, \mathfrak{S})$ such that its restriction with respect to $\mathfrak{S}'$ is the same as $(f, \mathfrak{S}')$? If so, this new function is called an *extension* of the given function. It is true that this problem can be solved in many ways, e.g., by assigning arbitrary values to the elements of $\mathfrak{S}$ which are not in $\mathfrak{S}'$. It is clear that the interest of the problem lies in the fact that the extension also belongs to a certain class of functions. In our case we consider only holomorphic functions and the process of extension is then called *analytic continuation.*

Let $\mathfrak{R}_1$ and $\mathfrak{R}_2$ be overlapping regions, i.e., the intersection of these regions is not empty. Let $f_1(z)$ be holomorphic in $\mathfrak{R}_1$ and $f_2(z)$ in $\mathfrak{R}_2$. If now f_1 and f_2 define the same restriction in the intersection of $\mathfrak{R}_1$ and $\mathfrak{R}_2$, that is, if $f_1(z) = f_2(z)$ for all z in this intersection, then f_2 is said to be a *direct analytic continuation* of f_1 into $\mathfrak{R}_2$. It is clear that according to this definition f_1 is a direct analytic continuation of f_2 into $\mathfrak{R}_1$.

Now we can define a function f throughout the union $\mathfrak{R}$ of $\mathfrak{R}_1$ and $\mathfrak{R}_2$ such that its restriction with respect to $\mathfrak{R}_1$ is f_1 and that with respect to $\mathfrak{R}_2$ is f_2. On account of the identity principle f, and also f_2 are uniquely defined by f_1.

More generally we consider a sequence of n regions $\mathfrak{R}_1, \ldots, \mathfrak{R}_n$, such that two consecutive members $\mathfrak{R}_k$, $\mathfrak{R}_{k+1}$, $k = 1, \ldots, n-1$, are overlapping. Assume we are given n functions $f_1, \ldots, f_n$, where f_k is holomorphic in $\mathfrak{R}_k$ and f_{k+1} a direct analytic continuation of f_k into $\mathfrak{R}_{k+1}$. Then f_n is said to be an *analytic continuation* of f_1 into $\mathfrak{R}_n$ along the sequence of regions. Is is clear that every one of the functions f_k can be considered as an analytic continuation of every other one.

It may happen that $\mathfrak{R}_1$ and $\mathfrak{R}_n$, $n > 2$, overlap. In that case the functions f_1 and f_n need not be direct analytic continuations, i.e., $f_n(z)$ and $f_1(z)$ may be different in the intersection of $\mathfrak{R}_1$ and $\mathfrak{R}_n$. This phenomenon is the key of a rigorous definition of multiply valued functions, as we shall see later on, (paragraph 12.2).

In the next section we shall describe a method of analytic continuation, employing circular disks instead of more general regions. This enables us to describe the process of continuation in a more concrete manner.

12.1.2 – Rearrangement of power series

Consider a function given by the power series

$$f_0(z) = \sum_{\nu=0}^{\infty} a_\nu (z-z_0)^\nu. \tag{12.1–3}$$

If it is necessary to exhibit also the region $\mathfrak{K}_0$ of convergence, we shall write $(f_0, \mathfrak{K}_0)$. The case $z_0 = \infty$ may be included, for then we have

$$f_0(z) = \sum_{\nu=0}^{\infty} a_\nu z^{-\nu}. \tag{12.1–4}$$

Such a power series, together with its disc of convergence, will often be referred to as a *function element* with centre z_0. Sometimes we understand by this the pair $(f, \mathfrak{R})$ of a function f holomorphic throughout $\mathfrak{R}$.

Let $z_1 \neq z_0$ denote a point inside the disc of convergence $\mathfrak{K}_0$ of the function element (12.1–3).

Since $f_0(z)$ is regular at $z = z_1$ we have the Taylor expansion

$$f_1(z) = \sum_{\nu=0}^{\infty} b_\nu (z-z_1)^\nu, \tag{12.1–5}$$

with

$$b_n = \frac{1}{n!} f_0^{(n)}(z_1) = \sum_{\nu=n}^{\infty} \binom{\nu}{n} a_\nu (z_1-z_0)^{\nu-n}. \tag{12.1–6}$$

It is easy to see that this new series can be obtained by *rearrangement*, that is, by means of the expansion of

$$(z-z_0)^n = (z-z_1+z_1-z_0)^n$$

in powers of $z-z_1$. The series (12.1–5) is convergent at all points of an open disc about z_1 which is inside the circle of convergence of (12.1–3), i.e., for all z satisfying

$$|z-z_1| < r_0-|z_1-z_0|,$$

where r_0 is the radius of convergence of (12.1–3). However, and this is the most interesting case, it may converge in a larger disc, (fig. 12.1–1). It is evident that $f_1(z)$ as given by (12.1–6) is a direct analytic continuation of $f_0(z)$ into the disc of convergence of the series (12.1–5).

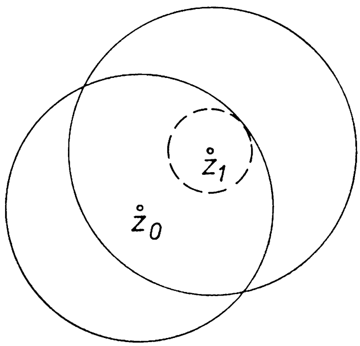


Fig. 12.1–1. Direct analytic continuation

If it is not possible to extend f_0 beyond its disc of convergence by means of this process, we call the circumference of $\mathfrak{K}_0$ a *natural boundary*. Examples are provided by the series (8.2.–4), or (8.2–7). In that case it is in no way possible to extend f_0 beyond $\mathfrak{K}_0$.

12.1.3 – Hyperconvergence

Let $\mathfrak{R}$ denote a region including the origin and f a function holomorphic throughout $\mathfrak{R}$. We assume that there is no larger region in which f is holomorphic. In a neighbourhood of $z = 0$ the function can be represented by the expansion

$$f(z) = \sum_{\nu=0}^{\infty} c_\nu z^\nu, \qquad (12.1\text{–}7)$$

which is convergent within a circumference about $z=0$ with radius equal to the distance of the boundary to the origin. The function f is regular at those points of the circumference which are not boundary points. Without loss of generality we may assume that the radius of convergence of (12.1–7) is unity.

The sequence of partial sums

$$s_n(z) = \sum_{\nu=0}^{n} c_\nu z^\nu, \qquad n = 0, 1, \ldots, \tag{12.1–8}$$

is divergent at every point outside the circumference of convergence. In certain cases, however, we can find a subsequence of partial sums which is also convergent at points of a region which extends the region of convergence of the initial series (12.1–7). A power series with this property is called *hyperconvergent*. It is obvious that a power series can only be hyperconvergent in a neighbourhood of a point of the circumference of convergence where the function $f(z)$ is regular. We call a point of this kind a *regular point* of the series (12.1–7).

As we shall see the phenomenon of hyperconvergence can occur in the case of power series which possess *gaps* i.e., certain coefficients are zero. Thus we consider series of the type

$$\sum_{\nu=0}^{\infty} c_\nu z^{\lambda_\nu}, \tag{12.1–9}$$

where $0 \leqq \lambda_0 < \lambda_1 < \ldots$ is an increasing sequence of integers. Now we shall prove the following theorem due to A. Ostrowski:

Suppose that the power series (12.1–9) *has the radius of convergence* 1 *and that there exists an increasing sequence* $\mu_0, \mu_1, \ldots,$ *of suffixes such that*

$$\lambda_{\mu_n+1} - \lambda_{\mu_n} \geqq \vartheta \lambda_{\mu_n}, \qquad n = 0, 1, 2, \ldots, \tag{12.1–10}$$

where $\vartheta > 0$ *is a fixed number. Then the sequence of partial sums*

$$s_{\mu_n}(z) = \sum_{\nu=0}^{\mu_n} c_\nu z^{\lambda_\nu} \tag{12.1–11}$$

is convergent in a region of which every regular point of (12.1–9) *on the circumference of convergence is an interior point.*

Let $f(z)$ denote a function coinciding with the sum of (12.1–9) for $|z| < 1$ and being holomorphic in a larger region. Without affecting the generality we may assume that it is regular at $z = 1$. If $\delta > 0$ is small enough, it is regular at the points in and on a circle with radius $\frac{1}{2}+\delta$ about the point $z = \frac{1}{2}$. We apply Hadamard's three circles theorem (section 2.13.4) to the function

$$g_n(z) = f(z) - s_{\mu_n}(z) \tag{12.1–12}$$

and the circles with centre $z = \frac{1}{2}$ and radii $r_1 = \frac{1}{2} - \delta$, $r = \frac{1}{2} + \varepsilon$, $r_2 = \frac{1}{2} + \delta$,

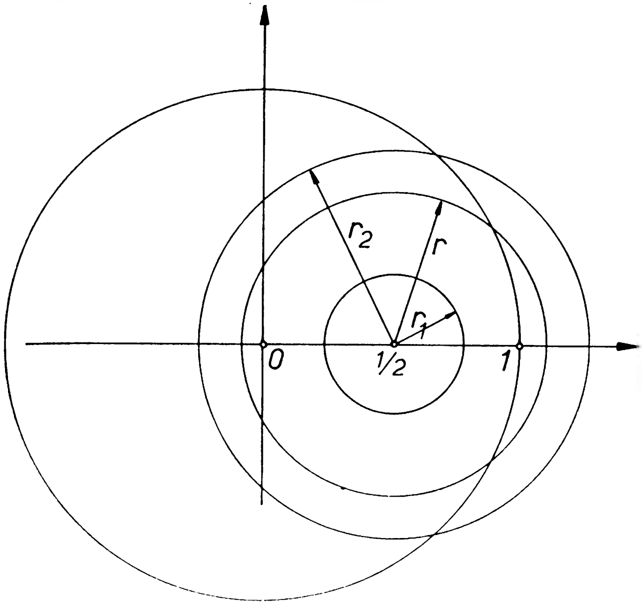


Fig. 12.1–2. Application of Hadamard's three circles theorem

where $0 < \varepsilon < \delta < \frac{1}{2}$, (fig. 12.1–2). If M_1, M and M_2 are the maximum moduli of $g_n(z)$ on these circumferences respectively, then in view of (2.13–14)

$$M^{\log r_2/r_1} \leqq M_1^{\log \frac{1+2\delta}{1+2\varepsilon}} M_2^{\log \frac{1+2\varepsilon}{1-2\delta}}. \tag{12.1–13}$$

We proceed to show that we can take ε so small that $M \to 0$ as $\mu_n \to \infty$. Then the sequence (12.1–11) represents an extension of (12.1–9) beyond its circle of convergence, for inside and on the circle of radius r this sequence tends to $f(z)$.

To this end we take a number η between 0 and δ and denote by K the maximum modulus of $f(z)$ on the circumference $|z| = 1-\eta$. By Cauchy's inequality (2.18–2) we have

$$|c_n| \leqq \frac{K}{(1-\eta)^{\lambda_n}}.$$

If z is on the circumference with radius r_1 we have

$$M_1 \leqq \sum_{\nu=\mu_n+1}^{\infty} |c_\nu||z|^{\lambda_\nu} \leqq \sum_{\nu=\mu_n+1}^{\infty} |c_\nu|(1-\delta)^{\lambda_\nu} \leqq K \sum_{\nu=\lambda_{\mu_n}+1} \left(\frac{1-\delta}{1-\eta}\right)^{\lambda_\nu}$$

$$\leqq K \sum_{\nu=\lambda_{\mu_n}+1} \left(\frac{1-\delta}{1-\eta}\right)^{\nu} = \frac{K}{1-\dfrac{1-\delta}{1-\eta}} \left(\frac{1-\delta}{1-\eta}\right)^{\lambda_{\mu_n}+1} \leqq K_1 \left(\frac{1-\delta}{1-\eta}\right)^{(1+\vartheta)\lambda_{\mu_n}},$$

where we have employed (12.1–10). The constant K_1 depends only on δ and η, but not on r. If M^* is the maximum modulus of $f(z)$ on the circumference with radius r_2, we have for z on this circumference

$$M_2 \leqq M^* + \sum_{\nu=0}^{\mu_n} |c_\nu||z|^{\lambda_\nu} \leqq M^* + K \sum_{\nu=0}^{\mu_n} \left(\frac{1+\delta}{1-\eta}\right)^{\lambda_\nu}$$

$$\leqq \left(M^* + \frac{K}{1 - \dfrac{1-\eta}{1+\delta}}\right)\left(\frac{1+\delta}{1-\eta}\right)^{\lambda_{\mu_n}} = K_2\left(\frac{1+\delta}{1-\eta}\right)^{\lambda_{\mu_n}},$$

where K_2 is again independent of n. Inserting the estimates for M_1 and M_2 thus obtained into Hadamard's inequality (12.1–13) we find

$$M^{\log r_2/r_1} \leqq K_1 K_2 \left\{\left(\frac{1-\delta}{1-\eta}\right)^{(1+\vartheta)\log\frac{1+2\delta}{1+2\varepsilon}}\left(\frac{1+\delta}{1-\eta}\right)^{\log\frac{1+2\varepsilon}{1-2\delta}}\right\}^{\lambda_{\mu_n}}.$$

The expression between the braces tends to

$$(1-\delta)^{(1+\vartheta)\log(1+2\delta)}(1+\delta)^{-\log(1-2\delta)}$$

as $\varepsilon \to 0$, $\eta \to 0$, and the logarithm of this last expression may be written as

$$\left((1+\vartheta)\frac{\log(1+2\delta)}{2\delta}\frac{\log(1-\delta)}{-\delta} - \frac{\log(1-2\delta)}{-2\delta}\frac{\log(1+\delta)}{\delta}\right)(-2\delta^2).$$

For sufficiently small δ this expression is negative. As a consequence we may take ε and η so small that the original expression between braces is less than 1. Since $\lambda_{\mu_n} \to \infty$ as $n \to \infty$ it is possible to make M as small as desired, provided that n is large enough. This concludes the proof of the theorem.

The following example may serve as an illustration. Consider the series of polynomials

$$f(z) = \sum_{\nu=1}^{\infty} \frac{(z+z^2)^{4^\nu}}{A_\nu}, \tag{12.1–14}$$

where A_n is the maximal coefficient in the polynomial $(z+z^2)^{4^n}$. Evidently, in each of the polynomials occurring in (12.1–14) the coefficients do not exceed unity and one of them is actually equal to 1. The highest term in the polynomial corresponding to n is of degree 2×4^n and the lowest of the next polynomial of degree 4^{n+1}. Hence, if we expand $f(z)$ in powers of z, each term is a single term of one of the polynomials occurring in (12.1–14). On applying the Cauchy-Hadamard test (1.6–11) we deduce that the radius of convergence of the power series is unity. As a consequence, the series of polynomials is convergent for $|z| < 1$. But since

$z+z^2 = z(1+z)$ is symmetric with respect to z and $1+z$, it is also convergent for $|z+1| < 1$. The special sequence of partial sums of the power series for $f(z)$ by taking each polynomial as a whole is, therefore, convergent in a region which lies partially outside the unit circle.

In this example we have

$$\mu_0 = 5, \mu_{n+1} = \mu_n + 4^{n+2} + 1, \qquad n \geqq 0$$
$$\lambda_{\mu_n} = 2 \times 4^{n+1}, \qquad \lambda_{\mu_n+1} = 4^{n+2}$$

and so

$$\lambda_{\mu_n+1} - \lambda_{\mu_n} = 2 \times 4^{n+1} = \lambda_{\mu_n},$$

whence $\vartheta = 1$.

12.1.4 – Hadamard's gap theorem

An interesting and famous theorem due to Hadamard is obtained from Ostrowski's theorem if we take $\mu_n = n$. It states

If the sequence of exponents in the power series (12.1–9) *satisfies the condition of Hadamard*

$$\lambda_{n+1} - \lambda_n \geqq \vartheta \lambda_n, \tag{12.1–15}$$

where $\vartheta > 0$ is fixed, then the circumference of convergence is a natural boundary.

Assume that the series were regular at any point of the circumference. Then the series would be hyperconvergent at that point and the sequence

$$s_{\mu_n}(z) = s_n(z) = \sum_{\nu=0}^{n} a_\nu z^{\lambda_\nu} \tag{12.1–16}$$

would be convergent outside the circle. But the sequence (12.1–16) is exactly the whole sequence of partial sums of the series (12.1–9). This leads to a contradiction and our conclusion is that each point of the circumference of convergence is a singularity of the series.

A simple example is provided by the series (8.2–4). In this case $\lambda_n = 2^n$, whence $\lambda_{n+1} - \lambda_n = 2^n = \lambda_n$, i.e., $\vartheta = 1$.

Another example is provided by the series

$$\sum_{\nu=0}^{\infty} z^{\nu!} \tag{12.1–17}$$

Here $\lambda_n = n!$, $\lambda_{n+1} - \lambda_n = n \cdot n! \geqq n!$, whence $\vartheta = 1$.

12.1.5 – The Fatou–Pólya theorem

The following remarkable theorem was conjectured by Fatou and proved by Pólya. It can be considered as a consequence of Hadamard's gap theorem.

If

$$\sum_{\nu=0}^{\infty} c_\nu z^\nu \tag{12.1–18}$$

denotes an arbitrary power series with a finite radius of convergence, then we can find numbers η_n, *being either* $+1$ or -1, *such that the series*

$$\sum_{\nu=0}^{\infty} \eta_\nu c_\nu z^\nu \tag{12.1–19}$$

is not continuable beyond the circumference of convergence.

Without loss of generality we may assume that the radius of convergence is unity. From (12.1–18) we take a subseries satisfying Hadamard's condition (12.1–15). This represents a series $h(z)$ whose circumference of convergence is a natural boundary. We put

$$\sum_{\nu=0}^{\infty} c_\nu z^\nu = f_0(z) + h(z) \tag{12.1–20}$$

and decompose $h(z)$ into infinitely many power series

$$h(z) = f_1(z) + f_2(z) + \ldots, \tag{12.1–21}$$

such that no two of these series contain terms with the same exponent, whereas every series contains infinitely many terms. Consider the collection of all series

$$f_0(z) + \varepsilon_1 f_1(z) + \varepsilon_2 f_2(z) + \ldots, \tag{12.1–22}$$

where every ε_n is either $+1$ or -1. Suppose that all these series are continuable beyond the circumference of convergence. We observe that the roots of unity form an enumerable set which is everywhere dense on the boundary of the unit circle. Hence there is at least one root of unity, say p_1, at which a certain series (12.1–22) is regular. Let this series be

$$f_0(z) + \varepsilon_{11} f_1(z) + \varepsilon_{12} f_2(z) + \ldots. \tag{12.1–23}$$

Next we can find a root of unity, say p_2, at which a second series (12.1–22) is regular. Let this series be

$$f_0(z) + \varepsilon_{21} f_1(z) + \varepsilon_{22} f_2(z) + \ldots. \tag{12.1–24}$$

After n steps we have a series

$$f_0(z) + \varepsilon_{n1} f_1(z) + \varepsilon_{n2} f_2(z) + \ldots, \tag{12.1–25}$$

which is regular at the root p_n. Thus we find an enumerable collection of series and it is easy to see that the series

$$_0(z) - \varepsilon_{11} f_1(z) - \varepsilon_{22} f_2(z) - \ldots - \varepsilon_{nn} f_n(z) - \ldots \tag{12.1–26}$$

does not belong to this collection. In fact, it cannot coincide with (12.1–23), nor with (12.1–24), etc. There is, however, according to assumption, a root p_k at which this series is regular. Hence, among the series (12.1–22) there are at least two which are regular at a point of the circumference of convergence (for the set of roots of unity is enumerable and the collection of series (12.1–22) is not) and the same is true for the difference of these series. In this difference, however, $f_0(z)$ is cancelled, and it is therefore a power series satisfying Hadamard's condition and being still regular at a point of the circumference of convergence. This means a contradiction.

The proof reveals that the collection of the series (12.1–22) is not enumerable. Hence, in a certain sense, the non-continuability of a series beyond the circle of convergence is the most common situation.

12.1.6 – Analytic continuation along a path

Let L denote a path leading from $z = a$ to $z = b$. We may assume that L is represented by the function $z(t)$, $0 \leqq t \leqq 1$, such that $z(0) = a$, $z(1) = b$ and that $z(t)$ is continuously differentiable for $0 < t < 1$ and still continuous at the terminal points. Suppose further that to each t we may associate a power series $f_t(z)$ whose disc of convergence $\mathfrak{K}_t$ has the centre $z(t)$. For any t_0, if t has the property that $z(t)$ lies within $\mathfrak{K}_{t_0}$ for all $t_0 \leqq t \leqq t_1$, we shall require that $f_t(z)$ be a direct analytic continuation of f_{t_0}, (fig. 12.1–3). We then say that f_1 *has been obtained by analytically continuing* f_0 *along the path L.* Now we state

Analytic continuation of a given power series f_0 *along a path L leads to a uniquely determined power series* f_1.

Let f_t and g_t denote two continuations of $f_0 = g_0$ along L. For any t_0 the power series f_{t_0} and g_{t_0} converge in a disc $|z-z(t_0)| < r(t_0)$, where $r(t_0)$ is the smallest of their radii of convergence. There is a $\delta = \delta(t_0) > 0$ such that $|z(t)-z(t_0)| < r(t_0)$ for all $|t-t_0| < \delta$. Hence for these values of t the series f_t and g_t are direct continuations of f_{t_0} and g_{t_0} respectively. Let now t_0 denote the greatest lower bound of the t for which f_t and g_t disagree. Clearly $t_0 > 0$. As a consequence there is a t_1 with $|t_1-t_0| < \delta(t_0)$ such that $f_{t_1} = g_{t_1}$. But the discs of convergence of f_{t_1} and those of f_{t_0} and g_{t_0} overlap. In the overlapping area the power series under consideration are identical. It follows that $f_{t_0} = g_{t_0}$, i.e., the series agree at t_0 and, consequently, for all t satisfying $|t-t_0| < \delta$. It follows that $t_0 < 1$ leads to a contradiction.

The radius of convergence of a power series is either identically infinite, or a continuous function of the centre of the disc of convergence, where it is understood that the elements (f, a) *and* (g, b), *a and b being the centres*

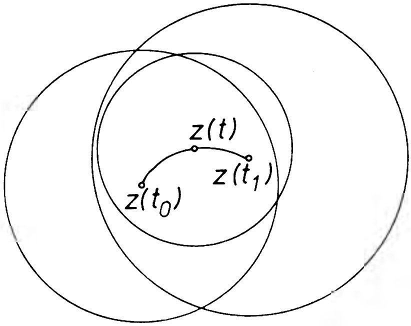


Fig. 12.1–3. Analytic continuation along a path.

of the respective elements, are direct continuations if b is sufficiently near a.

Indeed, if $r(a) < \infty$, $r(a)$ denoting the radius of the disc of convergence with centre a, and $|b-a| < \frac{1}{2}r(a)$, then $r(b) \geqq r(a) - |b-a| \geqq \frac{1}{2}r(a)$. Hence a lies in the circle of convergence of the series about b and also $r(a) \geqq r(b) - |b-a|$. It follows that

$$|r(b) - r(a)| \leqq |b-a|. \tag{12.1–27}$$

An easily proved consequence of this theorem is

If the continuation of the power series f_0 along a path L is possible, it can always be accomplished by analytic continuation along a finite chain of discs.

The radius of convergence $r_t = r(z(t))$ of the series f_t is either identically infinite or is a continuous function of t and possesses a positive minimum ρ. Hence in all cases we may take a positive number ρ and a partition of the interval $0 \leqq t \leqq 1 : 0 = t_0 < t_1 < \ldots < t_n = 1$, such that $|z(t_{k+1}) - z(t_k)| < \rho$ for $k = 0, \ldots, n-1$. Then the discs $\mathfrak{K}_k$: $|z - z(t_k)| < r_{t_k}$ are overlapping and the corresponding functions $f_{t_0}, \ldots, f_{t_n}$ form an analytic continuation.

Finally we prove

If f_0 is analytically continued along L from a to b, then continuation along a path L connecting a and b and sufficiently near L leads to the same element at b.* Let again ρ denote a positive number not exceeding the radii of convergence $r_t = r(z(t))$ of the power series f_t, $0 \leqq t \leqq 1$. Let L^*, defined by $z = z^*(t), 0 \leqq t \leqq 1$, be any other path with $z^*(0) = a$, $z^*(1) = b$ and such that $|z^*(t) - z(t)| < \frac{1}{4}\rho$. If we denote by g_t the power series obtained by continuing $f_0 = g_0$ along the path L^*, then the theorem states that $f_t = g_t$.

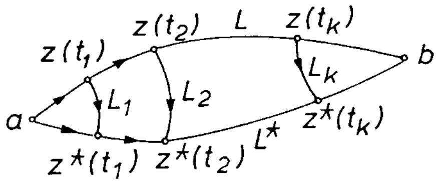


Fig. 12.1–4. Analytic continuation along different paths.

In order to prove this we consider the partition $0 = t_0 < t_1 < \ldots < t_n = 1$, such that for all t in the interval $t_k \leqq t \leqq t_{k+1}$, $k=0, \ldots, n-1$ we have $|z(t)-z(t_k)| < \frac{1}{4}\rho$. The sequence of discs $\mathfrak{K}_k$: $|z-z(t_k)| < r_{t_k}$, gives us the continuation of f_0 to f_1 by a finite succession of direct continuations. Let L_k represent the line segment joining $z(t_k)$ to $z^*(t_k)$. Continuation along any path, lying entirely within the disc of convergence of a power series, leads to the same series, for each direct continuation is merely a rearrangement of the original series. Thus continuation of f_0 from $z = a$ to $z(t_1)$ along L and along the composite path composed of L^* from $z = a$ to $z^*(t_1)$ and then to $z(t_1)$ along $-L_1$ (that is L_1 percorsed in the opposite direction) both lead to the same series f_{t_1}. Now we continue f_{t_1} from $z(t_1)$ to $z(t_2)$ along the path L_1 followed by L^* from $z^*(t_1)$ to $z^*(t_2)$ and then $-L_2$, obtaining the same series f_{t_2}, (fig. 12.1–4). In this process we traced L_1 successively in both directions, the net effect being that we get the same result as we left out L_1. Thus we have continued f_0 along two paths to get f_{t_2}, namely L from $z = a$ to $z(t_2)$ or L^* from $z = a$ to $z^*(t_2)$, followed by $-L_2$. We repeat this stepwise process to b in a finite number of steps and end with $f_1 = g_1$.

This theorem implies that for purposes of analytic continuation we can always replace a path L by a polygon approximating L and which has its vertices on L. We do not need to consider all elements on the broken line, for if the vertices are sufficiently dense it will be sufficient to use only the elements whose centres are vertices. Otherwise we may have to interpolate intermediary points.

12.1.7 – Permanence of functional relations

Consider a collection of function elements

$$(g_1, \mathfrak{K}), \ldots, (g_n, \mathfrak{K})$$

defined throughout a disc $\mathfrak{K}$. Let

$$f(w_1, \ldots, w_n) \tag{12.1–28}$$

denote a function of n variables such that its first order partial derivatives exist and are continuous if $w_1, \ldots, w_n$ are in regions $\mathfrak{R}_1, \ldots, \mathfrak{R}_n$ respectively. Assume that $g_1(\mathfrak{K})$ is in $\mathfrak{R}_1, \ldots, g_n(\mathfrak{K})$ is in $\mathfrak{R}_n$. Then

$$f(g_1(z), \ldots, g_n(z)) \tag{12.1–29}$$

is evidently a holomorphic function of z in $\mathfrak{K}$, for this function is well-defined there and has a unique derivative

$$\frac{df}{dz} = \frac{\partial f}{\partial w_1}\frac{dg_1}{dz} + \ldots + \frac{\partial f}{\partial w_n}\frac{dg_n}{dz}.$$

We assume further that the function (12.1–29) vanishes identically in $\mathfrak{K}$. Let $(\tilde{g}_1, \tilde{\mathfrak{K}}), \ldots, (\tilde{g}_n, \tilde{\mathfrak{K}})$ be a set of direct analytic continuations of the given collection in a region $\mathfrak{R}$. Then we contend that in $\tilde{\mathfrak{K}}$ also $f(\tilde{g}_1(z), \ldots, \tilde{g}_n(z))$ vanishes identically. This is true, for this latter expression vanishes identically in the intersection of $\mathfrak{K}$ and $\tilde{\mathfrak{K}}$. Hence (12.1–29) vanishes for all sets of corresponding function elements. This is the *principle of permanence of functional relations.*

12.2 – Analytic functions

12.2.1 – Weierstrass's definition of an analytic function

Two function elements $(f_1, \mathfrak{K}_1)$ and $(f_2, \mathfrak{K}_2)$ are said to be *equivalent* if the second is an analytic continuation of the first. Hence there exists a finite sequence of power series joining f_1 to f_2. It is easily verified that this equivalence relation has the usual properties in that it is reflexive, symmetric and transitive. As a consequence, all power series can be divided into disjoint classes of equivalent elements. Each class consists of all power series equivalent to any one of its members. Such a member may be taken as a representative of its class and it determines the latter uniquely.

After these preliminary remarks we may introduce a fundamental notion. By an *analytic function* we understand an equivalence class of power series. This notion is due to Weierstrass.

An analytic function is not a function in the usual sense. If $z = a$ belongs to the disc of convergence of a certain power series $f(z)$, this latter assigns a certain value $f(a)$ to $z = a$. But there may be other power series about $z = a$ equivalent to the given one which assign other values to $z = a$. All these values are called *values* of the analytic function at $z = a$ and a function of this type is considered as a multiply valued function.

A remarkable theorem due to Poincaré and Volterra states

If F is an analytic function obtained by continuing analytically a function *element* $(f_0, \mathfrak{K}_0)$ *throughout a region* $\mathfrak{R}$, *then every point of* $\mathfrak{R}$ *is the centre of at most an enumerable set of function elements of* F.

Thus the set of values assigned to a given value of z is at most enumerable.

We start with $(f_0, \mathfrak{K}_0)$, where $\mathfrak{K}_0$ is included in $\mathfrak{R}$. The various elements of F with the same centre $z = a$ as the initial element are obtained by analytic continuation along suitably chosen paths in $\mathfrak{R}$. With each such a path we associate a finite sequence

$$(f_0, \mathfrak{K}_0), (f_1, \mathfrak{K}_1), \ldots, (f_n, \mathfrak{K}_n) \tag{12.2–1}$$

where $\mathfrak{K}_n$ has the same centre as $\mathfrak{K}_0$. Without loss of generality we may assume that the centres of the intermediary discs are rational points The minimum number of elements in (12.1–1) performing the same continuation be m. This number shall be called the length of the sequence. The number of sequences of a given length is enumerable, for we have for the choice of the centres certain subsets of rational points at our disposal. Since the lengths are natural numbers we finally obtain an at most enumerable set of sequences joining two function elements with a given centre.

Let $\mathfrak{R}^*$ denote a subregion of a region $\mathfrak{R}$ on which an analytic function F is defined. Starting with a function element whose centre is in $\mathfrak{R}^*$ and performing only analytic continuation throughout $\mathfrak{R}^*$ it may happen that the analytic function F^* thus obtained assigns only one value to an arbitrary point of $\mathfrak{R}^*$. Then F^* may be interpreted as a single-valued function which is referred to as a *single-valued branch* of F defined on $\mathfrak{R}^*$.

12.2.2 – The derivative and the inverse of an analytic function

If the function elements $(f_1, \mathfrak{K}_1)$ and $(f_2, \mathfrak{K}_2)$ are direct analytic continuations of each other, so are $(f_1', \mathfrak{K}_1)$ and $(f_2', \mathfrak{K}_2)$, where the prime denotes differentiation. Hence

The derivatives of the elements of an analytic function $F(z)$ *are equivalent and constitute again an analytic function.*

This latter function is called the *derivative* $F'(z)$.

The proof of the following theorem requires more attention. We recall that a function element $(f, \mathfrak{K})$ is invertible throughout $\mathfrak{K}$ if it possesses an inverse there. In section 3.12.3 we proved that f is invertible at $z = z_0$ if $f'(z_0) \neq 0$. Otherwise stated, if $f(z)$ assumes its value $f(z_0)$ only once. Now we assert

The inverses of the invertible elements of an analytic function F *are elements of an analytic function* $\check{F}$, *called the inverse of* F.

Excluding the trivial case that F is a constant we observe that every analytic function has invertible elements, for the derivatives of the elements are not identically equal to zero. Let

$$f(z-z_0) = \sum_{\nu=0}^{\infty} c_\nu (z-z_0)^\nu$$

be a power series. Referring to section 3.12.5 we may conclude that $f(z-z_0)$ assumes the value $f(z_1-z_0)$ only once if z_1 is sufficiently near z_0. The point z_0 may be an exception. Hence the rearrangement $g(z-z_1)$ is invertible at z_1 as is shown in section 3.12.3.

Next we consider two invertible elements $(f_0, \mathfrak{K}_0)$ and $(f_1, \mathfrak{K}_1)$ of the same analytic function F, having the centres z_0 and z_1 respectively. We contend that they can be joined by means of a sequence consisting exclusively of invertible elements. There is certainly a collection $(f_t, \mathfrak{K}_t)$, $0 \leqq t \leqq 1$, where $\mathfrak{K}_t$ has its centre at the point $z(t)$ on a path joining $z(0)$ and $z(1)$. By $\mathfrak{T}$ we denote the set of those values of t of the interval $0 \leqq t \leqq 1$ for which the element $(f_t, \mathfrak{K}_t)$ can be joined to $(f_0, \mathfrak{K}_0)$ by a sequence of invertible elements. Let t_0 be the least upper bound of $\mathfrak{T}$ and let $\mathfrak{K}^*$ denote a sufficiently small disc around $z(t_0)$ so that the function f_{t_0} assumes everywhere in $\mathfrak{K}^*$, with possible exception of the point $z(t_0)$, each of its values once, i.e. with multiplicity equal to one. It is our aim to show that $t_0 = 1$.

There is certainly a value t_1 in $\mathfrak{T}$ for which $z(t_1)$ belongs to $\mathfrak{K}^*$ and for which f_{t_1} is a direct analytic continuation of f_{t_0}. Let C_1 denote a path along which a collection of invertible elements joins f_{t_1} to f_{t_0}. Supposing $t_0 < 1$ we should be able to find a value t_2 such that $t_0 < t_2 < 1$, $z(t_2)$ in $\mathfrak{K}^*$ and f_{t_2} a direct continuation of f_{t_0}. Let C_2 denote an arbitrary path in $\mathfrak{K}^*$ joining $z(t_1)$ and $z(t_2)$. If $z(t_1) \neq z(t_0)$ we could assume that C_2 does not pass through $z(t_0)$. Then, continuing along C_1 and C_2 successively, we should obtain a sequence of invertible elements joining f_0 and f_{t_2}. We should therefore have that also t_2 belongs to $\mathfrak{T}$, which is impossible since t_0 is the least upper bound of $\mathfrak{T}$. Hence $t_0 = 1$. The function $f_1(z)$ assumes its values at all points of the segment joining $z(t_1)$ and $z(1)$ in $\mathfrak{K}^*$ only once and the direct continuations of the element f_1 with centres on the segment are invertible. Hence continuing f_0 along the path consisting of C_1 and this segment we obtain a sequence of invertible elements joining f_0 and f_1. Thus we see that there is always a path connecting $z(0)$ with $z(1)$, such that continuation along this path is effected by invertible elements.

Finally we consider the function $w(t) = f_t(z(t))$ and denote by $\check{f}_t(w)$ the inverse of $f_t(z)$ in a neighbourhood of $z(t)$. The elements $\check{f}_t$ form a collection along the path $w(t)$, $0 \leqq t \leqq 1$, joining the elements $\check{f}_0$ and $\check{f}_1$. These elements therefore belong to an analytic function.

12.2.3 – The monodromy theorem

An analytic function F is said to be *arbitrarily continuable* in a region $\mathfrak{R}$ if each element of the function has a continuation along every path emanating from the centre of the element and lying in $\mathfrak{R}$. It is clear that every point of this region is the centre of the same number (finite or infinite) of elements of the function.

Assume further that $\mathfrak{R}$ is not the entire z-plane. Then the disc of convergence $\mathfrak{K}$ of every element of the analytic function F has, inside or on its boundary points of the complement of the region $\mathfrak{R}$. For, let us assume that $\mathfrak{K}$ and also its closure is inside $\mathfrak{R}$. Then the element, having $\mathfrak{K}$ as its disc of convergence, can be continued to every point of the boundary of $\mathfrak{K}$ and by an argument similar to that employed in section 8.2.1. we see that the power series under consideration converges in a disc larger then the disc $\mathfrak{K}$. This is an absurdity. If the closure of $\mathfrak{K}$ is not entirely in $\mathfrak{R}$ the assertion is trivial. If $\mathfrak{R}$ coincides with the whole z-plane then this latter is the disc of convergence of every element of the function F.

An analytic function F which is arbitrarily continuable in a disc $\mathfrak{K}$ is a single-valued function. In order to prove this, we consider an element f_0 of F with the centre of $\mathfrak{K}$ as centre. From the above consideration follows that $\mathfrak{K}$ is contained in the circle of convergence of f_0 and f_0 is holomorphic in $\mathfrak{K}$. Therefore, the function F has for each point in $\mathfrak{K}$ exactly one element with centre at this point.

These preliminary considerations enable us to prove the *general monodromy theorem*:

An analytic function in a simply connected region and arbitrarily continuable in this region, is single-valued in this region.

Let $\mathfrak{K}$ denote an open disc in the w-plane and assume that the region $\mathfrak{R}$ in the z-plane has at least two boundary points. By Riemann's mapping theorem we can find a univalent function $w(z)$, holomorphic throughout $\mathfrak{R}$, which maps $\mathfrak{R}$ onto $\mathfrak{K}$. If $f_t(z)$, $0 \leqq t \leqq 1$, is a collection of elements along a path $z = z(t)$ lying in $\mathfrak{R}$ then the elements $g_t(w) = f_t(z(w))$ form a collection along the path $w = w(z(t))$. Therefore, if $f_2(z)$ is a continuation of $f_1(z)$ effected by the above collection, then $g_2(w) = f_2(z(w))$ is a continuation of $g_1(w) = f_1(z(w))$ in $\mathfrak{K}$. The same is true if we interchange the roles of $\mathfrak{R}$ and $\mathfrak{K}$. Hence, if F is an analytic function in $\mathfrak{R}$ the set of all $g(w) = f(z(w))$, where $f(z)$ is an element of F, is an analytic function G in $\mathfrak{K}$. If $f(z)$ is arbitrarily continuable in $\mathfrak{R}$ then $g(w) = f(z(w))$ is arbitrarily continuable in $\mathfrak{K}$. The consideration in the first part of this section leads to the result that G is single-valued. As a consequence, also F is single-valued. The case that $\mathfrak{R}$ has only one boundary point or is the extended plane needs no comment. Then $\mathfrak{K}$ may be taken as the z-plane or the extended plane. This concludes the proof of the monodromy theorem.

12.2.4 – Continuation by integration

In many cases analytic continuation can be carried out by integrating a single-valued function. Let $f(z)$ denote a function holomorphic throughout a region $\mathfrak{R}$, with the exception of isolated singularities. Consider further a path $z = z(t)$, $0 \leqq t \leqq 1$, which avoids the singular points of the function. Along this path we define the function

$$g(t) = \Phi_0(z_0)+\int_0^t f(z(\tau))z'(\tau)d\tau, \tag{12.2–2}$$

where $\Phi_0(z_0)$ is a certain constant. For a given t there exists in a sufficient small disc $\mathfrak{K}_t$ with centre $z(t)$ a single-valued function $\Phi_t(z)$, which takes the value $g(t)$ at $z = z(t)$, namely

$$\Phi_t(z) = g(t)+\int_{z(t)}^z f(\zeta)d\zeta, \tag{12.2–3}$$

where the integration is performed along a rectilinear path connecting $z(t)$ and z within $\mathfrak{K}_t$. It is clear that $\Phi_0(z_0)$ is the value of $\Phi_0(z)$ at $z = z_0 = z(0)$. It is also evident that the collection of functions (12.2–3) provide an analytic continuation of the element

$$\Phi_0(z) = \int_{z_0}^z f(\zeta)d\zeta+\Phi_0(z_0) \tag{12.2–4}$$

defined in $\mathfrak{K}_0$ around z_0 along the path $z(t)$. By taking all possible paths we have defined an analytic continuation of $\Phi_0(z)$ throughout $\mathfrak{R}$.

Let now $\mathfrak{R}$ be any multiply connected region from which the singularities of $f(z)$ are omitted and assume that $C_1, C_2, \ldots$ constitute a basis for homology. For the sake of convenience we suppose that the connectivity is finite (section 2.2.2). Then every closed circuit is homologous to a linear combination of the cycles $C_1, \ldots, C_r$, where $r+1$ is the connectivity. If

$$\omega_k = \int_{C_k} f(\zeta)d\zeta, \qquad k = 1, \ldots, r \tag{12.2–5}$$

and the path C defined by $z(t)$ satisfies (2.2.–3), then evidently

$$g(1) = g(0)+m_1\omega_1+\ldots+m_r\omega_r, \tag{12.2–6}$$

where $m_1, \ldots, m_r$ are integers. Hence the various values of the analytic function $G(z)$ with elements (12.2–3) differ by linear combinations of the constants ω_k. They are called *moduli of periodicity* of the analytic function $G(z)$. A similar result can be obtained if a basis of homology consists of an enumerable number of cycles.

An interesting illustrative example is provided by the function

$$f(z) = \frac{\psi'(z)}{\psi(z)}, \tag{12.2–7}$$

where $\psi(z)$ has isolated singularities. Additional singularities are at the zeros of $\psi(z)$. In the expression similar to (12.2–2)

$$g(t) = \Phi_0(z_0) + \int_0^t \frac{\psi'(z(\tau))}{\psi(z(\tau))} z'(\tau) d\tau \tag{12.2–8}$$

we take $\Phi_0(z_0)$ in such a way that

$$\exp \Phi_0(z_0) = \psi(z_0). \tag{12.2–9}$$

It is easy to verify that the derivative of

$$\psi(z(t)) \exp(-g(t)) \tag{12.2–10}$$

with respect to t is zero. Hence this function is constant along the path $z(t)$, viz., equal to 1. For a given t there exists in $\mathfrak{K}_t$ a single-valued function $\Phi_t(z)$ such that

$$\exp \Phi_t(z) = \psi(z), \tag{12.2–11}$$

this function being

$$\Phi_t(z) = g(t) + \int_{z(t)}^{z} \frac{\psi'(\zeta)}{\psi(\zeta)} d\zeta. \tag{12.2–12}$$

In view of the principle of the permanence of functional relations the relation (12.2–11) is valid for all analytic continuations of $\Phi_0(z)$. The corresponding analytic function is called the *general logarithm* of $\psi(z)$ and is denoted by

$$G(z) = \text{Log}\, \psi(z). \tag{12.2–13}$$

The particular case $\psi(z) = z$ deserves mention. It refers to the general logarithm of z, already mentioned in section 1.11.2. This function has one modulus of periodicity which is equal to $2\pi i$. This follows from the fact that every closed path in the z-plane punctured at $z = 0$ is homologous to a multiple of the circumference

$$z = e^{2\pi i t}, \qquad 0 \leq t \leq 1,$$

and

$$\int_0^1 \frac{dz(t)}{z(t)} = 2\pi i \int_0^1 dt = 2\pi i.$$

Another elementary example is provided by the *general inverse* *tangent* which may be defined in accordance with (1.12–10). Exceptional points are $z = i$ and $z = -i$, but the corresponding moduli of periodicity are numerically equal to π.

If $(f_1, \mathfrak{K}_1)$ and $(f_2, \mathfrak{K}_2)$ are elements of the same analytic function $F(z)$, then so are $(\exp f_1, \mathfrak{K}_1)$ and $(\exp f_2, \mathfrak{K}_2)$ defining the analytic function $\exp F(z)$. This remark enables us to define the general power $f^\lambda(z)$ of $f(z)$, viz., $\exp(\lambda \operatorname{Log} f(z))$, λ being any complex number. A frequently occurring case is that with $\lambda = \frac{1}{2}$ and $f(z)$ a rational function. Then at each point which is neither a pole nor a zero of $f(z)$ the general root takes two values whose sum is zero.

12.2.5 – Integration of an analytic function

The theory of analytic continuation provides a means to define the integral of an analytic function along a path. Let $z = z(t)$ denote a path in a region $\mathfrak{R}$ and assume that a function element $f_0(z)$ of $F(z)$ defined at $z = z_0 = z(0)$ is continuable along the path. Hence we have a collection $f_t(z)$ of functions which are for a fixed t power series of $z - z(t)$ convergent in a disc $\mathfrak{K}_t$ with centre $z(t)$. If $z(t+h)$ is also in $\mathfrak{K}_t$ we have

$$f_{t+h}(z(t+h)) - f_t(z(t)) = f_t(z(t+h)) - f_t(z(t))$$

and it follows that $f_t(z(t))$ is a continuous function of t along the path. As in the preceding section we introduce the function

$$g(t) = \Phi_0(z_0) + \int_0^t f_t(z(\tau))z'(\tau)d\tau \tag{12.2–14}$$

where $\Phi_0(z_0)$ is a constant selected before. This function may serve to continue the element

$$\Phi_0(z) = \Phi_0(z_0) + \int_{z_0}^{z} f_0(\zeta)d\zeta, \tag{12.2–15}$$

and proceeding as in the previous section we obtain an analytic function $G(z)$. This function may be denoted by

$$G(z) = G(z_0) + \int_C F(\zeta)d\zeta, \tag{12.2–16}$$

but this notation must be handled with care in concrete situations.

A simple example may serve as an illustration. We wish to evaluate the integral

$$\int_C \sqrt{\zeta}\, d\zeta \tag{12.2–17}$$

along the circumference of the unit circle around the origin. We shall start with a function element of $\sqrt{z}$ which takes the value 1 at the point $z = 1$. The path may be represented by

$$z = e^{2\pi it}, \qquad 0 \leqq t \leqq 1.$$

In accordance with (12.2–15) we have

$$g(t) = \int_0^t e^{\pi i\tau} \cdot 2\pi i e^{2\pi i\tau} d\tau = \tfrac{2}{3}(e^{3\pi it}-1) \tag{12.2–18}$$

Hence the value of (12.2–17) is $-\frac{4}{3}$. Percorsing the circumference twice, we must take into account that after one encirclement the integrand has changed sign. Hence the contribution to the integral by performing the integration a second time, starting with the final value of the integrand after the first integration, is $\frac{4}{3}$. The sum is zero, that is to say, (12.2–17) takes the value zero if C is the unit circle percorsed twice.

If in (12.2–16) C is a closed curve and imbedded in a simply connected region such that $F(z)$ is arbitrarily continuable in this region, then the integral vanishes. In fact, starting with a given function element we obtain a single-valued function, as follows from the monodromy theorem, and now Cauchy's theorem is applicable. This result gives us the possibility to modify the path of integration without affecting the value of the integral.

In our illustrative example we may replace the circumference of the unit circle by a path consisting of a rectilinear segment from 1 to ε, with $0 < \varepsilon < 1$, a circumference of radius ε about $z = 0$ and finally a segment from ε to 1. The integral along the small circle tends to zero as $\varepsilon \to 0$, since the integrand remains bounded (Darboux's inequality (2.4–17)). On the second segment it has changed sign and so we have to evaluate

$$\int_1^0 \sqrt{x}\,dx - \int_0^1 \sqrt{x}\,dx = -\tfrac{4}{3},$$

in accordance with the result obtained above.

12.2.6 – The monodromy group of an integral

An extension of the integral considered in the preceding section is an integral of the type

$$\int_{z_0}^{z} (u-a_1)^{\lambda_1} \ldots (u-a_n)^{\lambda_n} du, \tag{12.2–19}$$

the integration being performed along a path connecting z_0 and z, but avoiding the points $a_1, \ldots, a_n$. From the z-plane we omit n half rays emanating from $a_1, \ldots, a_n$, no two of them having a point in common,

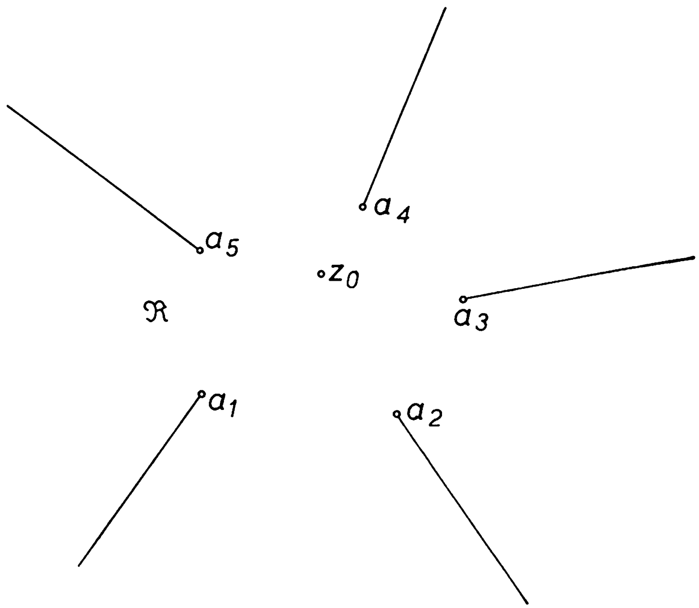


Fig. 12.2–1. Cuts in the z-plane issuing from the singular points

(fig. 12.2–1). There remains a simply connected region $\mathfrak{R}$. We suppose, of course, that none of the rays pass through z_0.

Consider now a simple closed curve starting at z_0 encircling one of the points $a_1, \ldots, a_n$, say a_k, once in the counter-clockwise sense and having only one point in common with the half ray L_k, the half ray beginning at a_k. It arrives at a point z_- on L_k and starts again from the same point, now denoted by z_+, on the "opposite" border of $\mathfrak{R}$ along L_k. The distinction between z_- and z_+ will be clarified presently.

Selecting one of the possible arguments of the integrand at z_0 we obtain a certain function element

$$f_0(z) = (z-a_1)^{\lambda_1} \ldots (z-a_n)^{\lambda_n}$$

which can be continued throughout $\mathfrak{R}$. Continuing along the part connecting z_0 and z_- we arrive with an element $f_-(z)$; continuing along the part connecting z_0 and z_+ we arrive with an element $f_+(z)$. On the other hand, if we start at z_0 with $f_0(z)$ and percorse the closed curve in its full extend in the counter-clockwise sense, we arrive at z_0 with the element

$$e^{2\pi i\lambda_k} f_0(z).$$

It follows that

$$f_-(z) = e^{2\pi i\lambda_k} f_+(z). \tag{12.2–20}$$

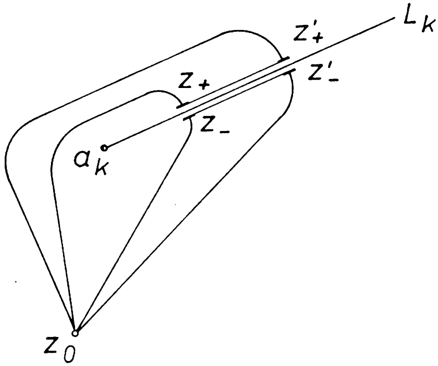


Fig. 12.2–2. Analytic continuation across a cut

Next we consider a similar path, traversing L_k at the point $z'_- = z'_+$, (fig. 12.2–2). Applying Cauchy's theorem we find

$$\int_{z_0}^{z_-} + \int_{z_-}^{z'_-} + \int_{z'_-}^{z_0} = 0 \tag{12.2–21}$$

and

$$\int_{z_0}^{z_+} + \int_{z_+}^{z'_+} + \int_{z'_+}^{z_0} = 0, \tag{12.2–22}$$

where $\int_{z_-}^{z'_-}$ and $\int_{z_+}^{z'_+}$ are evaluated along a linear segment. Since in view of (12.2–20)

$$\int_{z_-}^{z'_-} = e^{2\pi i\lambda_k} \int_{z_+}^{z'_+}$$

we easily find

$$\int_{z_0}^{z_-} + e^{2\pi i\lambda_k} \int_{z_+}^{z_0} = \int_{z_0}^{z'_-} + e^{2\pi i\lambda_k} \int_{z_0}^{z'_+}, \tag{12.2–23}$$

that is to say that the expression

$$\int_{z_0}^{z_-} + e^{2\pi i\lambda_k} \int_{z_+}^{z_0} \tag{12.2–24}$$

is constant along L_k. In particular we may take a closed contour consisting of a segment connecting z_0 with a point near to a_k followed by a small circumference around a_k and completed by a segment going to z_0, which may coincide with the first segment. Since Re $\lambda_k > -1$ the contribution to the value of the integral on the circular arc tends to zero, as the circle shrinks into a_k. It follows that the expression (12.2–24) is equal to

$$(1 - e^{2\pi i\lambda_k})\alpha_k \tag{12.2–25}$$

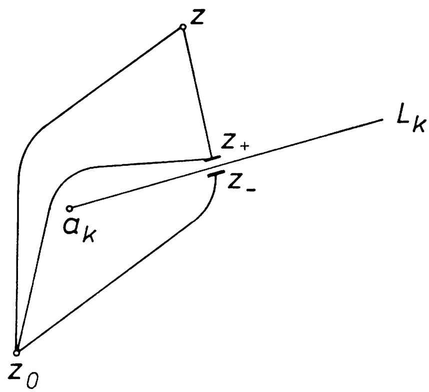


Fig. 12.2–3. The effect on integration along a path of crossing a cut

with

$$\alpha_k = \int_{z_0}^{a_k} (u-a_1)^{\lambda_1} \dots (u-a_n)^{\lambda_n} du. \qquad (12.2\text{–}26)$$

In more general cases α_k may be defined by equating (12.2–24) and (12.2–25).

Consider now a point $z \neq z_0$ and connect z_0 and z by a path which crosses once the half ray L_k in the positive sense, (fig. 12.2–3), that is from the right to the left if we percorse the half ray starting at a_k. Then

$$\int_{z_0}^{z_-} - e^{2\pi i \lambda_k} \int_{z_+}^{z} - e^{2\pi i \lambda_k} \int_{z}^{z_0} = \alpha_k(1-e^{2\pi i \lambda_k}). \qquad (12.2\text{–}27)$$

If we continue $f_0(z)$ along the path from z_0 to z_- and then from z_+ to z we have on the last part function elements which are obtained from those by continuing right-way from z_0 to z and from z to z_+ on multiplying them by $e^{2\pi i \lambda_k}$. Hence we may interpret

$$\int_{z_0}^{z_-} - e^{2\pi i \lambda_k} \int_{z_+}^{z} \qquad (12.2\text{–}28)$$

as the value of the integral taken along the path connecting z_0 and z which crosses the half ray. The integrand is the continuation of $f_0(z)$ remaining in $\mathfrak{R}$. We shall denote the integral from z_0 to z along a path in $\mathfrak{R}$ by $w(z)$ and the expression (12.2–28) by $S_k w(z)$. Then (12.2–27) becomes

$$S_k w(z) = \alpha_k(1-e^{2\pi i \lambda_k}) + e^{2\pi i \lambda_k} w(z). \qquad (12.2\text{–}29)$$

By adding a suitable closed path lying in $\mathfrak{R}$ we may deform the original path into another one, namely a closed circuit beginning and ending at z_0, followed by a path in $\mathfrak{R}$ connecting z_0 and z, (fig. 12.2–4).

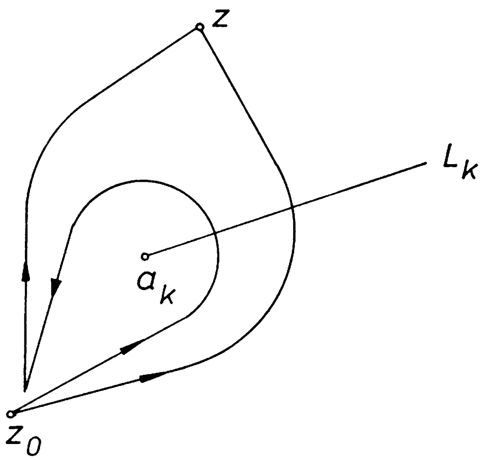


Fig. 12.2–4. Crossing a cut means encircling a singular point

A more general situation presents itself if we consider a path emanating from z_0 which crosses L_k once from the right to the left and then a second half ray L_l, (fig. 12.2–5), beginning at a_l. This is equivalent to two loops beginning and ending at z_0, followed by a path in $\mathfrak{R}$ from z_0 to z. Since we arrive at z_0 with $e^{2\pi i \lambda_k} f_0(z)$, if we continue $f_0(z)$ along the first loop,

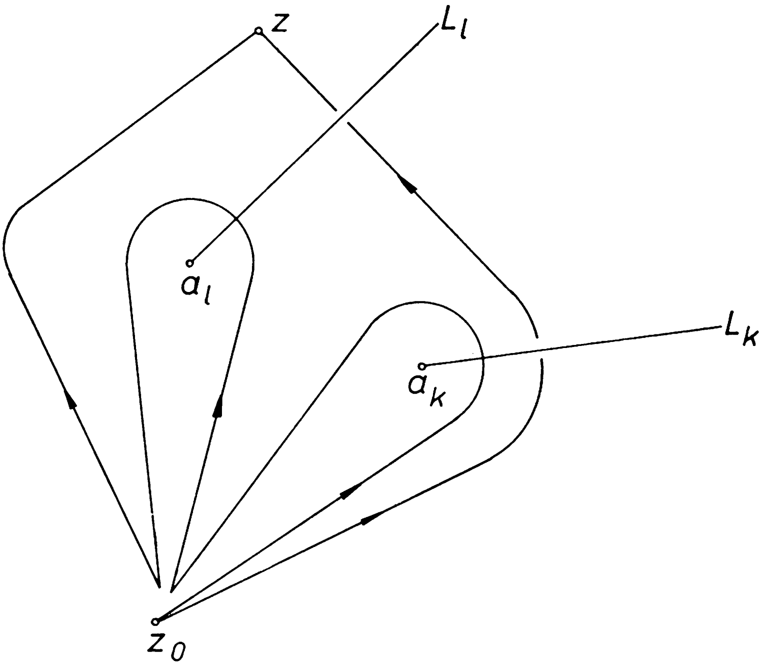


Fig. 12.2–5. Crossing several cuts means encircling singular points in succession.

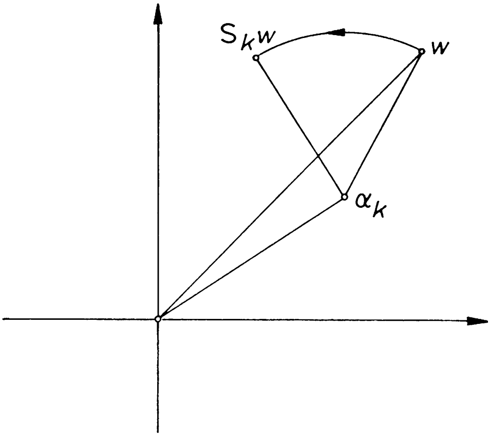


Fig. 12.2–6. Geometric interpretation of the equation (12.2–31)

it is easily seen that, denoting the integral taken along the whole path by $S_{kl}w(z)$,

$$S_{kl}w(z) = \alpha_k(1-e^{2\pi i\lambda_k})+\alpha_l e^{2\pi i\lambda_k}(1-e^{2\pi i\lambda_l})+e^{2\pi i(\lambda_k+\lambda_l)}w(z). \quad (12.2–30)$$

The equations (12.2–29) and (12.2–30) can be interpreted geometrically. Writing the first in the form

$$S_k w-\alpha_k = e^{2\pi i\lambda_k}(w-\alpha_k), \quad (12.2–31)$$

we see that $S_k w$ is obtained from w by rotating the w-plane about the point α_k through the angle $2\pi\lambda_k$, (fig. 12.2–6). The equation (12.2–30) is equivalent to

$$S_{kl}w = \alpha_k(1-e^{2\pi i\lambda_k})+e^{2\pi i\lambda_k}S_l w = S_k S_l w. \quad (12.2–32)$$

Hence $S_{kl}w$ is the result of two successive rotations about the points α_l and α_k respectively through angles $2\pi\lambda_l$ and $2\pi\lambda_k$, (fig. 12.2–7). An alternative form of (12.2–30) is

$$S_{kl}w-S_k\alpha_l = e^{2\pi i\lambda_l}(S_k w-S_k\alpha_l), \quad (12.2–33)$$

hence a rotation of $S_k w$ about $S_k\alpha_l$ through the angle $2\pi\lambda_l$. It is clear that in general S_{kl} and S_{lk} are different operators. Further it is easily seen how still more general situations must be handled. The transformations $S_1, \ldots, S_n$, generate a group, called the *monodromy group* of the integral (12.2–19).

We wish to consider some particular examples. First we turn our

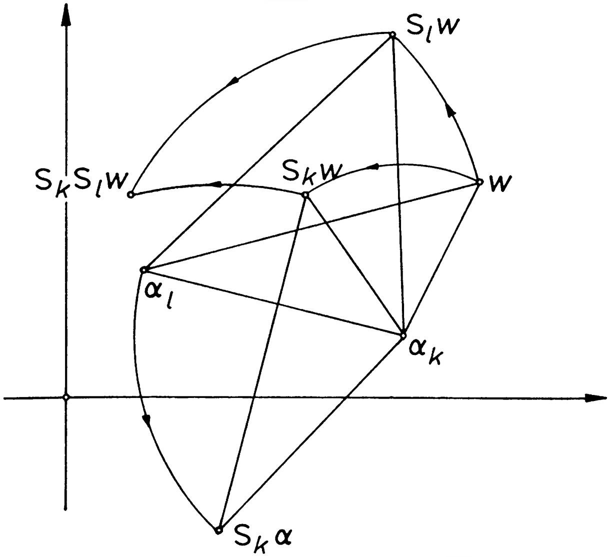


Fig. 12.2–7. Geometric interpretation of the equations (12.2–32) and (12.2–33)

attention to

$$w(z) = \int_0^z \frac{du}{\sqrt{1-u^2}}, \tag{12.2–34}$$

where the square root is taken such that it takes the value 1 at $u = 0$. Further we may take $a_1 = 1$, $a_2 = -1$, and it is clear that now $\lambda_1 = \lambda_2 = -\frac{1}{2}$. Hence, encircling the point $z = 1$ once in the counter-clockwise sense we find

$$S_1 w = \pi - w, \tag{12.2–35}$$

since

$$\alpha_1 = \int_0^1 \frac{du}{\sqrt{1-u^2}} = \tfrac{1}{2}\pi. \tag{12.2–36}$$

Similarly

$$S_2 w = -\pi - w. \tag{12.2–37}$$

Hence

$$S_{12} w = \pi - S_2 w = 2\pi + w. \tag{12.2–38}$$

It should be noticed that in this latter case the integrand continued analytically along the path of integration arrives at its starting point with its initial function element. We shall say that the path is *closed* with respect to the analytic continuation of the integrand.

The analytic function defined by (12.2–34) is the *general arc sine*, Arcsin z. Hence at a given point z its values are $w(z)$ or $\pi - w(z)$ (mod 2π). The *general arc cosine* may be defined by

$$\text{Arccos } z = \tfrac{1}{2}\pi - \text{Arcsin } z. \tag{12.2–39}$$

Putting $\tilde{w} = \frac{1}{2}\pi - w$, we have, evidently,

$$\mathsf{S}_1 \tilde{w} = -\tilde{w}, \qquad \mathsf{S}_{12} \tilde{w} = 2\pi + \tilde{w}. \tag{12.2–40}$$

These results are in accordance with those of section 1.12.1.

A slightly more complicated example is provided by Weierstrass's integral

$$w(z) = \int_{z_0}^{z} \frac{du}{\sqrt{4(u-e_1)(u-e_2)(u-e_3)}}, \tag{12.2–41}$$

where e_1, e_2, e_3 are complex numbers, usually taken such that $e_1 + e_2 + e_3 = 0$. It is easy see that

$$\begin{aligned} \mathsf{S}_{12} w &= 2(\alpha_1 - \alpha_2) + w, \\ \mathsf{S}_{23} w &= 2(\alpha_2 - \alpha_3) + w, \\ \mathsf{S}_{31} w &= 2(\alpha_3 - \alpha_1) + w, \end{aligned}$$

where w is the function (12.2–41) defined throughout the simply connected region $\mathfrak{R}$, starting with a certain function element of the integrand about z_0. It is common practice to put

$$\omega = \int_{e_3}^{e_2} \frac{du}{\sqrt{4(u-e_1)(u-e_2)(u-e_3)}}, \tag{12.2–42}$$

and

$$\omega' = \int_{e_2}^{e_1} \frac{du}{\sqrt{4(u-e_1)(u-e_2)(u-e_3)}} \tag{12.2–43}$$

in accordance with (10.2–55) and (10–2–56). If follows that $\alpha_2 - \alpha_3 = \omega$, $\alpha_1 - \alpha_2 = \omega'$, $\alpha_1 - \alpha_3 = \omega + \omega'$, for it is easily seen that 2ω is equal to the integral taken along a closed path encircling e_2 and e_3 once, (fig. 12.2–8), etc.

Finally we wish to investigate an example which is a counterpart of Hankel's representation of the gamma function (4.7–35) by means of an integral taken along suitably chosen path.

If $\text{Re}\, p > 0$, $\text{Re}\, q > 0$ we have by (4.7–38)

$$B(p, q) = \int_0^1 u^{p-1}(1-u)^{q-1}\, du, \tag{12.2–44}$$

the integrand being such that between 0 and 1 the arguments of u and $1-u$ are both zero. We consider a particular path which is closed with

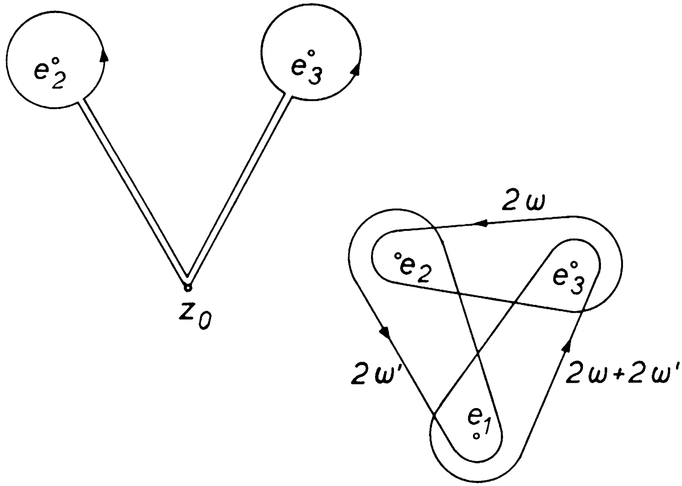


Fig. 12.2–8. The periods of (12.2–41) evaluated along closed paths
$2\omega = 2(\alpha_2-\alpha_3)$, $2\omega' = 2(\alpha_1-\alpha_2)$, $2\omega+2\omega' = 2(\alpha_1-\alpha_3)$

respect to analytic continuation, as depicted in fig. 12.2–9. It starts from a point z_0 between 0 and 1, encircles the point $+1$ once in the positive sense, then 0 in the positive sense, then again $+1$, but now in the negative sense and finally again 0 in the negative sense. It may be replaced, without affecting the value of the integral by four loops beginning and ending at z_0 and percorsed in succession. Since the initial value of the integral

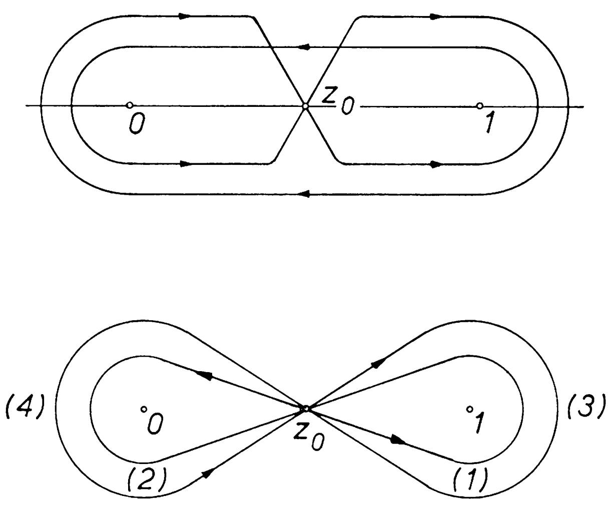


Fig. 12.2–9. The Jordan-Pochhammers contour

at z_1 will be zero we return at this point after percorsing the first loop with the value

$$(1-e^{2\pi i(q-1)})\int_{z_0}^{1} = (1-e^{2\pi iq})\alpha_1 .$$

The second loop yields

$$e^{2\pi iq}(1-e^{2\pi ip})\int_{z_0}^{0} = (e^{2\pi iq}-e^{2\pi i(p+q)})\alpha_2 .$$

The third loop gives

$$e^{2\pi i(p+q)}(1-e^{-2\pi ip})\int_{z_0}^{1} = (e^{2\pi i(p+q)}-e^{2\pi ip})\alpha_1$$

and the contribution of the final loop is

$$e^{2\pi ip}(1-e^{-2\pi ip})\int_{z_0}^{0} = (e^{2\pi ip}-1)\alpha_2 .$$

Hence along the entire path we get

$$(1-e^{2\pi ip}-e^{2\pi iq}+e^{2\pi i(p+q)})(\alpha_1-\alpha_2) = (1-e^{2\pi ip})(1-e^{2\pi iq})\int_0^1 u^{p-1}(1-u)^{q-1}\,du.$$

Denoting the path of integration by $(1_+, 0_+, 1_-, 0_-)$, we have found the *Jordan-Pochhammer's representation* of the Beta function

$$\boxed{\int^{(1+,0+,1-,0-)} u^{p-1}(1-u)^{q-1}\,du = (1-e^{2\pi ip})(1-e^{2\pi iq})B(p, q).} \tag{12.2–45}$$

Since both sides of this equation are holomorphic as regards p and q separately, the result is valid for all values of p and q.

12.2.7 – Singular points

The theory of singularities of analytic function is much less developed than in the single-valued case because of their complexities. The definition requires much more care than in the case of a single-valued function.

Consider a sequence of function elements

$$(f_0, \mathfrak{K}_0), (f_1, \mathfrak{K}_1), \ldots, \tag{12.2–46}$$

with centres $a_0, a_1, \ldots,$ and radii of convergence $r_0, r_1, \ldots.$ This sequence is said to be *singular* if $(f_n, \mathfrak{K}_n)$ is a direct continuation of $(f_{n-1}, \mathfrak{K}_{n-1})$, $n = 1, 2, \ldots$, $\lim_{n\to\infty} a_n = a$ exists and finally $\lim_{n\to\infty} r_n = 0$. The point a is called the *endpoint* of the sequence.

A finite point $z = a$ is a *singular point* of an analytic function $F(z)$ if and only if it is the endpoint of a singular sequence made up of elements of $F(z)$. It is then said to be a singular point for approach to $z = a$ along the sequence in question. It is also said that the sequence determines a singular point *above* $z = a$. Two singular sequences having the same endpoint determine the same singular point if and only if the sequences are equivalent, i.e., if an element of one sequence is equivalent to an element of the other by continuation within a disc $\mathfrak{K}_\varepsilon$ around $z = a$, whose radius ε is sufficiently small.

The point at infinity may be included by making the usual transformation and considering $F(1/z)$ at $z = 0$.

If $F(z)$ is a single-valued function, its singular points are among the boundary points of its region of holomorphism. For this to be the case each must be accessible (section 10.5.3). Hence *the singular points of a single-valued analytic function are the accessible boundary points of its region of holomorphism.*

If $F(z)$ is not single-valued, then it may occur that the same point $z = a$ can be the carrier of several, even of infinitely many singular points, whereas it may also be the centre of regular elements. Consider, e.g., the function $\log \log z$. At $z = 1$ we have infinitely many regular elements which are obtained if we start at $z = 2$ with complex values of $\log 2$. If, however, we take $\log 2$ equal to its real value we find a singular sequence. Above $z = 0$ there are infinitely many singular points. The simplest types of singular points are the isolated singular points. A singular point above $z = a$ is called *isolated* if an element of a singular sequence with endpoint $z = a$ with centre in a sufficiently small disc $\mathfrak{K}_\varepsilon$ around $z = a$ can be continued arbitrarily throughout the region obtained from $\mathfrak{K}_\varepsilon$ by deleting the centre $z = a$.

Some frequently occurring singularities will be considered in somewhat more detail. Let a denote such a singular point. We describe three circles C_1, C_2, C_3 through a which do not contain another singular point in their interior, such that the regions $\mathfrak{K}_1$, $\mathfrak{K}_2$, $\mathfrak{K}_3$ bounded by these circles overlap, (fig. 12.2–10). If we start with a function element f_1 in $\mathfrak{K}_1$ in the first disc we can find an element in the second disc which coincides with f_1 in the common region. Proceeding in this way through $\mathfrak{K}_3$ until we return to $\mathfrak{K}_1$ then it may occur that we get the same function element f_1, but it is also possible that the process of analytic continuation leads to another element, let us say f_2. Continuing we may find by this process in $\mathfrak{K}_1$ an element f_2, and so on. Now it may happen that for a certain natural number h the function f_h coincides with f_1. Then we shall say that the function elements $f_1, \ldots, f_h$ form a *cycle*

$$[f_1, \ldots, f_h]. \tag{12.2–47}$$

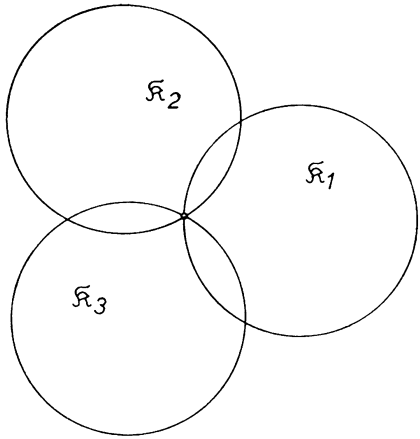


Fig. 12.2–10. Analytic continuation around a singular point

These function elements are connected by analytic continuation and form a part of the given analytic function F. This part may be analyzed by the introduction of a local uniformizing variable t such that

$$z = a + t^h. \qquad (12.2\text{–}48)$$

In the case that $a = \infty$, we put $z = t^{-h}$. Every element in the disc $\mathfrak{K}_i$ and related by analytic continuation to one of the functions of (12.2–47) is holomorphic in z (except possibly at $z = a$) and, therefore, also in t (except possibly at $t = 0$). If t describes a circle about $t = 0$ in the t-plane then z describes a circle about $z = a$ in the z-plane h times. Since $f_1, \ldots, f_h$ are cyclically permutated after one encirclement, we find that they are reproduced after h encirclements and, consequently, they are the elements of a single-valued function of t in a neighbourhood of $t = 0$. Hence these function elements may be expanded in a series of Laurent

$$\begin{aligned} f_1 &= \sum_{\nu=-\infty}^{\infty} c_{\nu 1} t^\nu, \\ &\cdots\cdots\cdots\cdots \\ f_h &= \sum_{\nu=-\infty}^{\infty} c_{\nu h} t^\nu. \end{aligned} \qquad (12.2\text{–}49)$$

If t percorses only an arc $t = re^{i\theta}$, $0 \leqq \theta \leqq 2\pi$, then z percorses a whole circle once and f_1 changes into f_2, f_2 into f_3 etc. Finally f_h into f_1. Hence every series (12.2–49) is obtained from the preceding one by replacing t by $t\eta$, where $\eta = e^{2\pi i/h}$. If we put

$$t = (z-a)^{1/h}$$

the power series takes the form

$$w = \sum_{\nu=-\infty}^{\infty} c_\nu (z-a)^{\nu/h}. \tag{12.2–50}$$

Series of this kind are called *series of Puiseux*. The series (12.2–50) yields a collection of h related developments obtained by choosing different initial values of $(z-a)^{1/h}$.

A common situation is realized if in (12.4–50) only a finite number of terms with negative exponents are present. Then the singularity is called an *algebraic singularity*, being a *pole* if there are terms with negative exponents, and *ordinary* if they are absent. An algebraic singularity is also referred to as an *algebraic branch point* (or *ramification point*) *of order* $h-1$. The order of a regular point is, therefore zero.

These singularities are called algebraic because they are the only singularities exhibited by algebraic functions, to be investigated further on.

Let us now assume that no number h with the above properties exists. Otherwise stated, that f_1 gives rise to an infinite sequence of function elements $f_1, f_2, \ldots$. Then it may happen that $f_k = f_{k-1}+c$, $k = 1, 2, \ldots,$ where c is a constant. It is clear that this set of elements constitute a part of $F(z)$ which behaves at $z = a$ like the function

$$\frac{c}{2\pi i} \log(z-a)$$

and local uniformisation is provided by the substitution

$$z = a + e^t. \tag{12.2–51}$$

A singularity of this kind is called a *logarithmic branch point*.

Finally we mention the case that $f_k = cf_{k-1}$. Then the function behaves like $(z-a)^\lambda$ and it is also uniformized by the substitution (12.2–51).

As already remarked in the beginning of this section a general theory of singularities is rather hopeless. Fortunately many important analytic functions behave rather well, that is to say, their singularities are of the type as described above.

12.3 – Algebraic functions

12.3.1 – The discriminant of a polynomial

Let $w_1, \ldots, w_n$ denote the zeros of the polynomial

$$f(w) = a_0 + a_1 w + \ldots + a_n w^n, \qquad a_n \neq 0, n > 0. \tag{12.3–1}$$

We intend to derive a method to determine which conditions the coefficients must satisfy in order that the polynomial may have repeated zeros. It is clear that $f(w)$ has repeated zeros if and only if *Vandermonde's determinant*

$$P = \begin{vmatrix} 1 & w_1 \dots w_1^{n-1} \\ 1 & w_2 \dots w_2^{n-1} \\ \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \\ 1 & w_n \dots w_n^{n-1} \end{vmatrix} = \prod_{\lambda > \mu} (w_\lambda - w_\mu) \tag{12.3–2}$$

vanishes. If we put

$$\sigma_k = w_1^k + \dots + w_n^k, \qquad k = 0, 1, 2, \dots, \tag{12.3–3}$$

we easily find that the square of this determinant is

$$P^2 = \begin{vmatrix} \sigma_0 & \sigma_1 \dots \sigma_{n-1} \\ \sigma_1 & \sigma_2 \dots \sigma_n \\ \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \sigma_{n-1} & \sigma_n \dots \sigma_{2n-2} \end{vmatrix}. \tag{12.3–4}$$

The elements of this determinant are symmetric expressions of the zeros of (12.3–1) and we shall describe a method for expressing them rationally in terms of the coefficients.

Since

$$f(w) = a_n(w - w_1) \dots (w - w_n)$$

we find by logarithmic differentiation

$$\frac{f'(w)}{f(w)} = \sum_{\nu=1}^{n} \frac{1}{w - w_\nu},$$

or

$$f'(w) = \sum_{\nu=1}^{n} \frac{f(w)}{w - w_\nu}.$$

Now

$$\frac{f(w)}{w - w_k} = \frac{f(w) - f(w_k)}{w - w_k} = \frac{a_n(w^n - w_k^n) + \dots + a_1(w - w_k)}{w - w_k}$$

$$= a_n w^{n-1} + (a_{n-1} + a_n w_k) w^{n-2} + \dots + (a_1 + a_2 w_k + \dots + a_n w_k^{n-1}),$$

where $k = 1, \dots, n$. Hence

$$f'(w) = n a_n w^{n-1} + (n a_{n-1} + a_n \sigma_1) w^{n-2} + \dots + (n a_1 + a_2 \sigma_1 + \dots + a_n \sigma_{n-1}).$$

Since also

$$f'(w) = n a_n w^{n-1} + (n-1) a_{n-1} w^{n-2} + \dots + a_1$$

we find by equating coefficients

$$\begin{aligned} a_{n-1}+a_n\sigma_1 \qquad\qquad &= 0, \\ 2a_{n-2}+a_{n-1}\sigma_1+a_n\sigma_2 \qquad &= 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots &\dots \\ (n-1)a_1+a_2\sigma_1+a_3\sigma_2+\dots+a_n\sigma_{n-1} &= 0. \end{aligned}$$

From these equations we may solve successively $\sigma_1, \dots, \sigma_{n-1}$ and it is easy to verify that

$$\sigma_k = \frac{\psi_k(a_{n-1}, \dots, a_{n-k})}{a_n^k}, \qquad k = 1, \dots, n-1,$$

where ψ_k denotes a polynomial with integral coefficients.

Next we observe that

$$\sum_{\nu=1}^{n} w_\nu^m f(w_\nu) = 0, \qquad m = 0, 1, \dots,$$

and this may be written as

$$\begin{aligned} a_0\sigma_0+a_1\sigma_1+\dots \quad +a_n\sigma_n \quad &= 0, \\ a_0\sigma_1+a_1\sigma_2+\dots \quad +a_n\sigma_{n+1} &= 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots &\dots \\ a_0\sigma_m+a_1\sigma_{m+1}+\dots+a_n\sigma_{m+n} &= 0. \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots &\dots \end{aligned}$$

From these equations we can evaluate σ_k for $k \geqq n$. Summing up we may state

The expression

$$D(a_0, \dots, a_n) = a_n^{2n-2}P^2, \tag{12.3–5}$$

where P^2 *is the determinant* (12.3–3) *is a polynomial in the coefficients of the polynomial* (12.3–1). *This latter polynomial has repeated roots if and only if D vanishes.*

The polynomial (12.3–5) is called the *discriminant* of (12.3–1).

Let us now consider a polynomial of the type

$$f(z, w) = a_0(z)+a_1(z)w+\dots+a_n(z)w^n, \qquad n > 0, \tag{12.3–6}$$

where $a_0(z), a_1(z), \dots, a_n(z)$ *are polynomials in* z and, of course, $a_n(z)$ not identically zero. The discriminant of this polynomial in w is found from $D(a_0, \dots, a_n)$ by replacing $a_0, \dots, a_n$ by $a_0(z), \dots, a_n(z)$ respectively. Thus we find a function $D(z)$, again called the *discriminant.*

If z_0 is a zero of $D(z)$ and $a_n(z_0) \neq 0$ then the polynomial $f(z_0, w)$ has repeated zeros. If, however $a_n(z_0) = 0$ then the equation $w^n f(z_0, 1/w) = 0$ has a root equal to zero. We then say that $f(z_0, w)$ has a root $w = \infty$. Thus we see

If z_0 is a not a zero of $D(z)$ then the equation $f(z_0, w) = 0$ has neither repeated finite roots, nor it has a root at infinity.

The zeros of $D(z)$ are usually called the *critical points* of the polynomial (12.3–6).

12.3.2 – The division transformation

Many properties of polynomials are based on the fact that a process which reduces the degree cannot be repeated more than a finite number of times. By the degree of (12.3–6) we understand the degree of $f(z, w)$ considered as a polynomial in w. If $a_0(z), \ldots, a_n(z)$ are identically zero, then $f(z, w)$ is not regarded as having a degree. We shall denote the degree by $\deg f$.

In the following theorem we establish the existence of a process referred to above, known as the *division transformation.*

If $f(z, w)$ and $g(z, w)$ are polynomials of the type (12.3–6), *with $g(z, w)$ not identically zero, there exists a non-zero polynomial $a(z)$ and two polynomials $q(z, w)$ and $r(z, w)$, with either $r = 0$ identically, or* $\deg r < \deg g$, *such that*

$$a(z)f(z, w) = q(z, w)g(z, w)+r(z, w). \tag{12.3–7}$$

The assertion is trivial if $f = 0$ identically or $\deg f < \deg g$. For then we may take $a(z) = 1$, $q(z, w) = 0$, $r(z, w) = f(z, w)$. If g does not involve w we take $a = g$, $q = f$, $r = 0$. Hence we may assume that $\deg g > 0$ and, using induction, that the theorem is true if f is replaced by a polynomial of smaller degree or by 0.

Let f be the polynomial (12.3–6) and let

$$g(z, w) = b_0(z)+b_1(z)w+ \ldots +b_m(z)w^m, \tag{12.3–8}$$

with $m \leqq n$, $a_n(z) \neq 0$, $b_m(z) \neq 0$. Then either

$$b_m(z)f(z, w)-a_n(z)w^{n-m}g(z, w) = 0$$

identically, or the degree of the expression on the left is less than $\deg f$. Hence there exist polynomials $a^*(z)$, $q^*(z, w)$, $r(z, w)$ such that

$$a^*(z)(b_m(z)f(z, w)-a_n(z)w^{n-m}g(z, w)) = q^*(z, w)g(z, w)+r(z, w),$$

with either $r = 0$ or $\deg r < \deg g$. It follows that (12.3–7) holds with

$$a(z) = a^*(z)b_m(z), \qquad q(z, w) = a^*(z)a_n(z)w^{n-m}+q^*(z, w).$$

This concludes the proof of the theorem.

In describing this process we say that *g has been divided into f* to give the *quotient* q and the *remainder* r.

If $f(z, w)$ and $g(z, w)$ are polynomials of the above considered type we say that g is a *factor* of f, or that f is *divisible* by g if there exists a polynomial $q(z, w)$ such that

$$f(z, w) = q(z, w)g(z, w). \tag{12.3–9}$$

A polynomial is called *irreducible* if it is not divisible by a polynomial of lower degree, except for a constant $\neq 0$.

If in (12.3–9) $q(z, w)$ *does not involve* w *then* q *is a factor of each coefficient of* f.

In fact, if g is the polynomial (12.3–8) then

$$f(z, w) = b_0(z)q(z)+ \dots +b_n(z)q(z)w^n.$$

This proves the assertion.

For many applications the following theorem is useful

If g *is irreducible and if* r *in* (12.3–7) *is identically zero, then* f *is divisible by* g.

We now have a relation

$$a(z)f(z, w) = q(z, w)g(z, w).$$

The polynomial $a(z)$ can be decomposed in factors of the first degree. Let $z-z_0$ be such a factor. Since $g(z, w)$ is irreducible it is not a factor of g. Assume that it is neither a factor of q. If

$$q(z, w) = q_0(z)+q_1(z)w+ \dots +q_p(z)w^p$$

then there is at least one $q_i(z)$ not divisible by $z-z_0$. Let k be the smallest value of the subscript i with this property. As a consequence $q_k(z)$ is not divisible by $z-z_0$. The coefficient of w^{k+l} in qg is

$$b_0(z)q_{k+l}(z)+b_1(z)q_{k+l-1}(z)+ \dots +b_{k+l}(z)q_0(z),$$

with $q_i(z) = b_j(z) = 0$ if $i > p, j > m$. Now $q_k b_l$ is not divisible by $z-z_0$, whereas every other term in this coefficient is. Hence the total sum is not divisible by $z-z_0$. This is a contradiction. We conclude that every linear factor of $a(z)$ is a factor of $q(z, w)$ and we therefore have

$$f(z, w) = q_1(z, w)g(z, w).$$

12.3.3 – Algebraic functions

An analytic function $W(z)$ is called an *algebraic function* if all its function elements $(w(z), \mathfrak{K})$ satisfy a relation

$$f(z, w(z)) = 0 \tag{12.3–10}$$

in $\mathfrak{K}$, where the left member is obtained from a polynomial (12.3–6) by replacing w by $w(z)$. Because of the permanence of functional relations it is sufficient to assume that (12.3–10) be satisfied by one function element of $W(z)$.

Let $\varphi(z, w)$ denote a polynomial of the lowest degree for which the equation $\varphi(z, w(z)) = 0$ is valid throughout $\mathfrak{R}$. Then φ is not divisible by a polynomial involving w of lower degree. In fact, if $\varphi = \varphi_1\varphi_2$ then $\varphi(z, w(z)) = \varphi_1(z, w(z))\varphi_2(z, w(z))$ vanishes for all values of z in $\mathfrak{R}$. Hence at least one of the factors φ_1 or φ_2 vanishes for infinitely many values of z in $\mathfrak{R}$, i.e., either $\varphi_1(z, w(z)) = 0$ or $\varphi_2(z, w(z)) = 0$. If the first case occurs then the degree of $\varphi_1(z, w)$ is equal to the degree of φ and φ_2 does not involve w. As a consequence we can always find an irreducible polynomial $\varphi(z, w)$ such that $\varphi(z, w(z)) = 0$ for all z in $\mathfrak{R}$.

It is easy to see that *the irreducible polynomial φ determined by an algebraic function is unique up to a constant factor.*

For let ψ denote another such polynomial. By the division transformation we can find polynomials $a(z)$, $q(z, w)$, $r(z, w)$ such that

$$a(z)\psi(z, w) = q(z, w)\varphi(z, w) + r(z, w).$$

Since φ and ψ vanish if we replace w by $w(z)$ we find that $r(z, w(z)) = 0$ in $\mathfrak{R}$. Hence $r(z, w) = 0$ identically, for in the contrary case $\deg r < \deg \varphi$ and this is not possible. From the last theorem of the previous section follows that ψ is divisible by φ. By the same arguments we find that φ is divisible by ψ and this proves the assertion.

12.3.4 – Existence of algebraic functions

Suppose that the polynomial (12.3–6) is irreducible. Since then the function $D(z)$ is not identically zero it has only a finite number of zeros, the critical points. Let $z = a$ denote a point which is not critical. Then the equation

$$f(a, w) = 0 \tag{12.3–11}$$

has n different roots $b_1, \ldots, b_n$. The main result of this section will be the following theorem

There exist positive constants δ and ε such that to every z in the disc $\mathfrak{R}$: $|z-a| < \delta$ there correspond n values $w_1, \ldots, w_n$ in discs $\mathfrak{R}_i$: $|w-b_i| < \varepsilon$, $i = 1, \ldots, n$, satisfying $f(z, w_i) = 0$. Every w_i is a holomorphic function of z in $\mathfrak{R}$, taking the value b_i at $z = a$. If $w(z)$ is a holomorphic function of z in $\mathfrak{R}$ such that $f(z, w) = 0$, then $w = w_i$ for some i.

We determine ε such that the discs $|w-b_i| \leq \varepsilon$ do not overlap and we denote the circle $|w-b_i| = \varepsilon$ by C_i. Clearly $f(a, w) \neq 0$ on C_i and by the theorem of the logarithmic derivative (section 3.8.2)

$$1 = \frac{1}{2\pi i} \int_{C_i} \frac{f_w(a, w)}{f(a, w)} dw, \tag{12.3–12}$$

where $f_w(a, w)$ stands for $\partial f(a, w)/\partial w$. If z varies throughout $\mathfrak{K}$, δ being sufficiently small, then

$$\int_{C_i} \frac{f_w(z, w)}{f(z, w)} dw$$

is a holomorphic and *a fortiori* a continuous function of z (section 2.9.1) and, therefore, the right-hand member of (12.3–12) remains equal to 1 if a is replaced by z. Hence $f(z, w) = 0$ has exactly one root in the disc $|w-b_i| < \varepsilon$, this root being denoted by $w_i(z)$. It follows from (3.8–12) that

$$w_i(z) = \frac{1}{2\pi i} \int_{C_i} w \frac{f_w(z, w)}{f(z, w)} dw \tag{12.3–13}$$

and referring again to section 2.9.1, we infer that $w_i(z)$ is holomorphic throughout $\mathfrak{K}$. In particular

$$w_i(a) = \frac{1}{2\pi i} \int_{C_i} w \frac{f_w(a, w)}{f(a, w)} dw = b_i .$$

The last part of the theorem follows from the fact that we have found n different roots for each value of z in $\mathfrak{K}$ and that there cannot be more roots.

Let $w(z)$ be a function holomorphic in a region $\mathfrak{R}$ and let $f(z, w(z)) = 0$ throughout $\mathfrak{R}$. There is certainly a point $z = a$ in $\mathfrak{R}$ such that $D(a) \neq 0$. In a sufficiently small disc $\mathfrak{K}$ about a belonging to $\mathfrak{R}$ the function $w(z)$ coincides with $w_i(z)$ for some i. Hence $w(z)$ belongs to the analytic function W_i generated by w_i.

A function element $w(z)$ which satisfies $f(z, w(z)) = 0$ can be continued along any arc which does not pass through a critical point. For assume there were an arc $z = z(t)$, $0 \leqq t \leqq 1$, such that a given initial element can be continued along all subarcs $0 \leqq t \leqq \alpha < 1$, but not along the whole arc. Let $a = z(1)$ and determine $\mathfrak{K}$ according to the above theorem and take α such that $z(t)$ is in $\mathfrak{K}$, whenever $\alpha \leqq t \leqq 1$. Now the element w obtained by continuation in a neighbourhood of $z(\alpha)$ must coincide with one of the function elements $(w_i, \mathfrak{K})$. But then it can be continued all the way to $z(1)$ and thus we arrive at a contradiction.

It remains to prove that the analytic functions $W_1, \ldots, W_n$ generated by $w_1, \ldots, w_n$ respectively coincide, i.e., that the irreducible polynomial (12.3–6) defines only one algebraic function. This will be done in section 12.3.6. First we have to investigate the behaviour at the critical points.

12.3.5 – Behavior at the critical points

In a neighbourhood of a non-critical point $z = a$ (or $z = \infty$) the elements $w_i(z)$ are regular and may be expanded in power series

$$w_i = c_{0i} + c_{1i}t + \ldots, \qquad i = 1, \ldots, n$$

with $t = z - a$ (or $t = 1/z$). They converge in each circle which does not contain a critical point.

In the case of a critical point the situation is more complicated. Let now $z = a$ denote a critical point. We consider the same circles as described in section 12.2.7, (fig. 12.2–10). Since there are in each disc $\mathfrak{K}_i$, $i = 1, 2, 3$ only n function elements we must get back an initial function element in one of the discs after $h \leqq n$ encirclements about the point $z = a$. Thus we see that the set of n function elements is divided into a certain number of cycles. Suppose that w is defined in $\mathfrak{K}$ and belongs to a cycle of h elements. Introducing the uniformizing variable t by $z = a + t^h$, we see that w is holomorphic in a disc about $t = 0$ in the t-plane, except possibly at $t = 0$. If $a_n(a) \neq 0$ then w is also regular at $t = 0$. This may be seen as follows. Let w denote a root of the equation

$$a_0 + a_1 w + \ldots + a_n w^n = 0$$

and let

$$M = \max\left(\frac{|a_0|}{|a_n|}, \ldots, \frac{|a_{n-1}|}{|a_n|}\right).$$

Then $|w| < 1 + M$. For we have

$$-1 = \frac{a_0}{a_n w^n} + \frac{a_1}{a_n w^{n-1}} + \ldots + \frac{a_{n-1}}{a_n w}$$

and if $|w| \geqq 1 + M$ we should have

$$1 \leqq \left|\frac{a_0}{a_n}\right| \frac{1}{|w|^n} + \ldots + \frac{|a_{n-1}|}{|a_n|} \frac{1}{|w|} \leqq \frac{M}{M+1} < 1.$$

Hence the functions w_i remain bounded as $t \to 0$ and by the Riemann's theorem (section 2.8.3) they are regular at $t = 0$. In this case *the point $z = a$ is an ordinary branch point.*

There is only the need of a slight modification of the reasoning if $a_n(a) = 0$. Then we consider the function $a_n(z)w$ instead of w. The values of this function remain bounded as $z \to 0$ and it follows that $z = a$ *is an algebraic pole.*

Summing up we may state

The singularities of an algebraic function are algebraic and they occur at the critical points.

In the next section we shall prove that a converse of this theorem also holds.

12.3.6 – Uniqueness of an algebraic function

We are now sufficiently prepared to settle the point which was left open in section 12.3.4, namely the fact that an irreducible polynomial defines exactly one analytic function. To this end we consider a somewhat more general problem.

Let F be an analytic function. For each point a we assume the existence of a disc $\mathfrak{K}$ with centre a such that all elements of F which are defined at a point z_0 of $\mathfrak{K}$ can be continued along arcs in $\mathfrak{K}$ avoiding the centre and show algebraic character at this point. This means that an initial element is reproduced after performing a finite (perhaps one) number of encirclements of the point a and tends to a definite limit (which may be ∞) as $z \to a$ along an arbitrary arc. Additionally we assume that the number of elements at z_0 is finite. The assumptions shall be satisfied also at $z = \infty$.

The extended plane can be covered by a finite number of discs $\mathfrak{K}$. It follows that only a finite number of points a_k can be effective singularities. Next we prove that the number of elements having a regular point as centre is constant throughout the plane, from which the singular points a_k are deleted. Indeed, every such point has a neighbourhood in which all elements of F are single-valued and can be continued throughout the neighbourhood. Therefore the set of all points z with exactly n elements is open (n may be infinite). Since the entire plane minus the points a_k is connected, one of these sets is not empty. Hence n is constant and by assumption it cannot be infinite. It cannot be zero, neither, since in that case F would be an empty collection of function elements.

The elements at any point $z \neq a_k$ may be denoted as $w_1(z), \ldots, w_n(z)$. The elementary symmetric functions of the $w_i(z)$ are the coefficients of the polynomial

$$(w-w_1(z)) \ldots (w-w_n(z)).$$

These coefficients represent single-valued holomorphic functions, since analytic continuation merely permutes the n elements involved and leaves their symmetric functions unchanged. The coefficients have only a finite number of singular points and these must be located at the points $z = a_k$. Consequently they are either poles or essential singular points. The latter possibility is easily excluded. For at a point a_k the function elements become infinite as fractional powers of $(z-a_k)^{-1}$, hence the symmetric functions as powers of $(z-a_k)^{-1}$. Since there are only a finite number of terms with negative exponents the point $z = a_k$ is either a pole or possibly a regular point of the coefficients. Thus the coefficients are holomorphic in the extended plane except for poles, that is, each coefficient is a rational function of z. If their common denominator is denoted by $a_n(z)$ we find that all function elements $w_i(z)$ must satisfy a polynomial equation $f(z, w) = 0$, where $f(z, w)$ is an expression of the type (12.3–6). Thus it is proved that F is algebraic.

Suppose now that the function element $(w, \mathfrak{K})$ satisfies the equation $f(z, w) = 0$, where f is an irreducible polynomial of degree n in w. The corresponding analytic function has only algebraic singularities and at each point a finite number of elements. We have just shown that the elements of F will satisfy a polynomial equation whose degree is equal to this number of function elements. It will hence satisfy an irreducible equation whose degree is not higher. But the only irreducible equation it can satisfy is $f(z, w) = 0$ and its degree is n. Therefore the number of function elements centred around a non-critical point is exactly n and it follows that all solutions of $f(z, w)$ belong to the same analytic function.|

Summing up we have established

An analytic function is an algebraic function if at each point it has a finite number of elements centred at this point and no other than algebraic singularities. Every algebraic function satisfies an irreducible polynomial equation $f(z, w) = 0$, unique up to a constant factor and every such equation determines a corresponding algebraic function uniquely.

12.3.7 – Newton's diagram

The problem of determining the expansion of an algebraic function at critical points can be handled with the aid of a device known variously as *Newton's diagram* or as the *method of Puiseux*. Newton introduced the device as an aid in curve tracing. It was adapted to the discussion of algebraic functions by Puiseux.

Without loss of generality we may assume that the point under consideration is at $z = 0$. The equation $f(z, w) = 0$ is written as

$$f(z, w) = a_0(z) + \ldots + a_n(z)w^n = 0, \qquad (12.3\text{–}14)$$

where the $a_0(z), \ldots, a_n(z)$ are polynomials of z. Let the initial term of $a_k(z)$, arranged in ascending order of the power of z, be $c_k z^{\alpha_k}$. Hence we may write our equation as

$$f(z, w) = (c_0 z^{\alpha_0} + \ldots) + \ldots + (c_n z^{\alpha_n} + \ldots)w^n = 0. \qquad (12.3\text{–}15)$$

We insert into this equation a series with indeterminate coefficients and exponents

$$w = b_1 z^{\beta_1} + b_2 z^{\beta_2} + \ldots, \qquad (12.3\text{–}16)$$

assuming that $\beta_1 < \beta_2 < \ldots$ and the coefficients $b_1, b_2, \ldots$ differ from zero. We obtain

$$(c_0 z^{\alpha_0} + \ldots) + \ldots + (c_n z^{\alpha_n} + \ldots)(b_1^n z^{n\beta_1} + \ldots) = 0, \qquad (12.3\text{–}17)$$

where within each pair of brackets only the term of lowest degree has been written. The series (12.3–16) is a solution if the expression on the left of (12.3–17) is identidally zero. Performing the multiplication and arranging the terms in ascending order, the powers with equal exponents must cancel. In particular those terms must cancel with the lowest exponents. This is only possible if at least two such terms occur. We shall see that this condition yields a finite number of exponents β_i. The condition that the terms of lowest order cancel provides us with the coefficient b_1.

The terms of lowest order in (12.3–17) are certainly present among those which we obtain if we take only the initial terms into consideration. Hence among the terms

$$c_0 z^{\alpha_0}, c_1 b_1 z^{\alpha_1+\beta_1}, \ldots, c_n b_1^n z^{\alpha_n+n\beta_1} \qquad (12.3\text{–}18)$$

we have to determine β_1, such that among the exponents

$$\alpha_0 + 0\beta_1, \alpha_1 + 1\beta_1, \ldots, \alpha_n + n\beta_1 \qquad (12.3\text{–}19)$$

the smallest occurs at least twice.

Now we introduce the geometrice device of Newton-Puiseux. In a rectangular coordinate system we plot the $n+1$ points with coordinates, (fig. 12.3–1),

$$(0, \alpha_0), (1, \alpha_1), \ldots, (n, \alpha_n). \qquad (12.3\text{–}20)$$

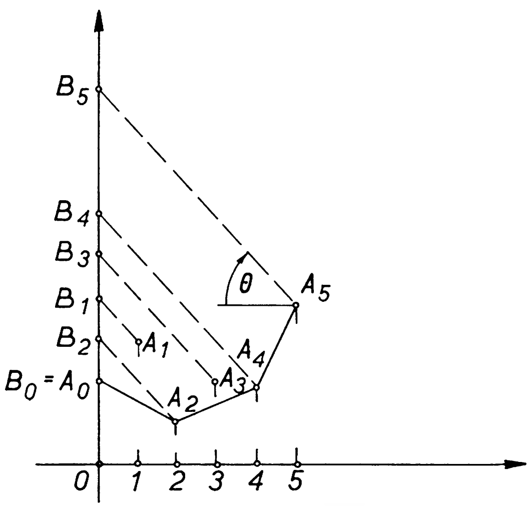


Fig. 12.3–1. The Newton–Puiseux diagram

Let us denote the point (k, α_k) by A_k, $k = 0, \ldots, n$. If in $f(z, w)$ one or more powers of w are absent the corresponding points A_k are omitted. Next we determine an angle θ by the equation

$$\tan \theta = \beta_1 . \tag{12.3–21}$$

The exponents (12.3–19) can be represented geometrically if we draw straight lines through the points A_k which enclose the angle θ with the negative horizontal axis. The line through A_k cuts the vertical axis at B_k. Then $OB_k = \alpha_k + k \tan \theta = \alpha_k + k\beta_1$. Hence the inclination θ of the lines must be chosen in such a way that at least two of the segments OB_k are equal and that none of the other segments OB_k are smaller. If, for instance, B_1 and B_2 coincide, then A_1 and A_2 are on the same line and θ is the angle between $A_1 A_2$ and the negative horizontal axis. Moreover, θ taken in this way, none of the other points A_k is allowed to be situated below the line $A_1 A_2$. All lines of this kind may be found as follows. We take a straight line through A_0 and rotate it about A_0 in the counter-clockwise sense until a second point A_k is on the line, say A_p. If there are more points A_k on this line we denote by A_p the point, which is the farthest to the right. The line $A_0 A_p$ is then one of the desired lines. We rotate again this line in the same sense, but now about A_p, until a point A_q is encountered. If there are other points A_k on this line $A_p A_q$ we denote by A_q the point which is the farthest to the right. The line $A_p A_q$ is again one of the desired lines. It is now clear how we can continue this process. Finally we obtain a line which contains A_n.

For every β_1 which we found in the above manner we find the corresponding value of b_1 if we select among the terms (12.3–18) those which have for this β_1 the same power of z as factor. For this value of β_1 all those exponents $\alpha_k + k\beta_k$ are equal whose corresponding points A_k are on a line $A_p A_q$, in particular $\alpha_p + p\beta_1$ and $\alpha_q + q\beta_1$. This leads to an equation

$$c_p b_1^p + \ldots + c_q b_1^q = 0,$$

or, since $b_1 \neq 0, p < \ldots < q$,

$$c_p + \ldots + c_q b_1^{q-p} = 0,$$

which is an equation of degree $q-p$. Thus we find a finite number of possible exponents β_1 and to every β_1 corresponds a finite number of coefficients b_1.

The determination of more terms of the series expansions may be effected in a similar way. Is $b_1 z^{\beta_1}$ one of the initial terms, we insert

$$w = b_1 z^{\beta_1} + w_1$$

into the relation between z and w. We obtain an equation

$$(c_0' z^{\alpha'_0} + \ldots) + \ldots + (c_n' z^{\alpha'_n} + \ldots) w_1^n = 0, \qquad (12.3\text{–}22)$$

in which possibly fractional exponents of z are involved, but which is similar to (12.3–19). We insert

$$w_1 = b_2 z^{\beta_2} + \ldots$$

into this equation and determine β_2 and b_2 as before. It should be noticed that only those values of β_2 are accepted which are greater than β_1. If β_2 and b_2 are corresponding values we insert

$$w_1 = b_2 z^{\beta_2} + w_2$$

into (12.3–22) and we obtain an equation in terms of w_2 which can be handled in the same way as the equations (12.3–19) and (12.3–22).

In this way we obtain a finite number of series which satisfy (12.3–14), at least formally. Our method, however, yields all expansions of this kind and among them occur the n convergent series whose existence has been proved previously. It might occur, however, that our method produces other developments which satisfy (12.3–14) only formally. It is not difficult to prove that this case does not occur. The reasoning rests on the following remark. If the formal product of two power series is zero, i.e., if all its coefficients obtained formally vanish, then at least one of the series is identically zero. Let $w_1, \ldots, w_n$ denote Puiseux-series corresponding to the roots of the equation $f(z, w) = 0$; we have identically

$$f(z, w) = c_n(w - w_1) \ldots (w - w_n).$$

If now w_0 is a formal Puiseux series satisfying $f(z, w) = 0$ we have identically in z

$$(w_0 - w_1) \ldots (w_0 - w_n) = 0.$$

Hence, for at least one k follows that $w_0 = w_k$.

12.4 – Riemann surfaces

12.4.1 – Introductory examples

The definition of an analytic function as an equivalence class of power series is satisfactory from a logical view-point, but does not visualize the function in concrete cases. Riemann introduced a certain geometric intuitive model of the behaviour of an analytic function which can also be defined in a more abstract way. It is our aim to construct such models in a few simple cases in order to prepare the way for more general considerations.

(i) The function

$$w = \log z$$

maps the upper half of the z-plane onto an infinite strip $0 < \text{Im}\, w < \pi$ of the w-plane and the lower half on the strip $-\pi < \text{Im}\, w < 0$. Since the positive real axis in the z-plane corresponds to the real axis in the w-plane, the z-plane cut along the negative real axis (including the origin) is mapped onto the strip $-\pi < \text{Im}\, w < \pi$. Applying the Schwarz symmetry principle we see that the same region in the z-plane can be mapped onto the strip $\pi < \text{Im}\, w < 3\pi$ and so on. Thus to every point z_0 in the z-plane correspond the points $w_0+2n\pi i$, where n is an arbitrary integer. This is in accordance with the fact that $\log z$ is a single-valued branch of the general logarithm $\text{Log}\, z$, which assigns the values $\log z_0 + 2n\pi i$ to each value $z_0 \neq 0$.

Let us now take a sequence $\{\mathfrak{R}_k\}$, $k = 0, \pm 1, \pm 2, \ldots$, infinite in both directions, of identical replicas of the cut z-plane. The first region $\mathfrak{R}_0$ is related as described above to the strip $\mathfrak{S}_0$: $-\pi < \text{Im}\, w < \pi$, the region $\mathfrak{R}_1$ is related to the strip $\mathfrak{S}_1$: $\pi < \text{Im}\, w < 3\pi$, the region $\mathfrak{R}_{-1}$ to the strip $\mathfrak{S}_{-1}$: $-3\pi < \text{Im}\, w < -\pi$. In general the region $\mathfrak{R}_k$ is related to $\mathfrak{S}_k$: $-\pi+2k\pi < \text{Im}\, w < \pi+2k\pi$. Next we give $\mathfrak{R}_k$ two boundaries. If we approach the negative real axis in the z-plane from below its corresponding point in the w-plane tends to a point whose imaginary part is $(-\pi+2k\pi)i$. The points of the negative real axis in correspondence with the line $w = (-\pi+2k\pi)i$ constitute the *lower boundary* of $\mathfrak{R}_k$. Similarly the points of the negative real axis in correspondence with the line $(\pi+2k\pi)i$ – and those points are obtained if we approach the negative real axis from above – constitute the *upper boundary* of $\mathfrak{R}_k$.

Now we may identify the lower boundary of $\mathfrak{R}_{k+1}$ with the upper boundary of $\mathfrak{R}_k$ for all integral values of k. Then we obtain a connected surface consisting of an infinity of sheets covering the z-plane. If we describe a circle about $z = 0$ we arrive at a point above the intitial point and in the w-plane we move from a point in a certain strip to a point in an adjacent strip along a vertical line, (fig. 12.4–1). Thus we see that to a point of the surface as constructed above corresponds precisely one point in the w-plane. Otherwise stated: *the general logarithmic function is single-valued on the surface*. This configuration is called the *Riemann surface* of the logarithmic function.

Every point of the Riemann surface is above a certain point of the z-plane and above each point of the z-plane (with exception of the points $z = 0$ and $z = \infty$) there are infinitely many points of the surface. Any point of the Riemann surface will be called a *place*, denoted by $\mathfrak{z}$ and the multiply-valued function $\log z$ turns out to be a single-valued

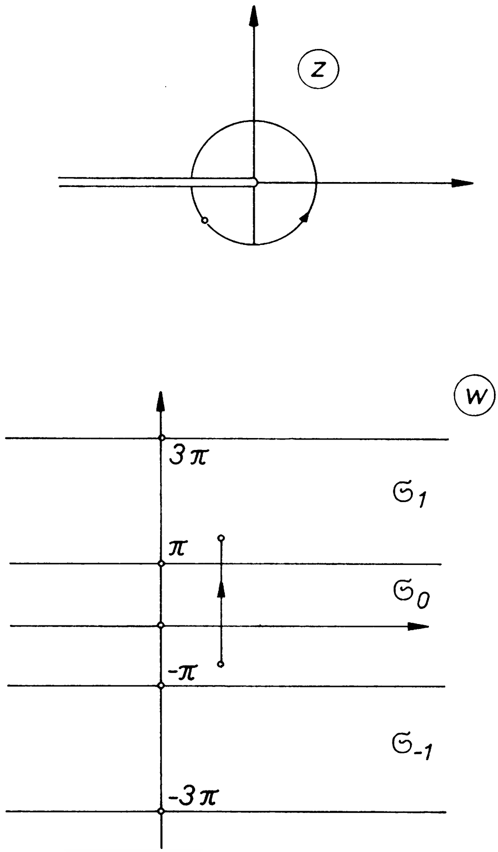


Fig. 12.4–1. The Riemann surface of the logarithm

function of $\mathfrak{z}$. If $\mathfrak{z}$ is above z, we shall say that z is the *trace* of $\mathfrak{z}$.

Let w_0 correspond to the place $\mathfrak{z}_0$. We consider a small disc with centre w_0. The set of all places corresponding to the points of the disc may be called a *neighbourhood* of $\mathfrak{z}_0$ on the surface. It is now rather evident that the Riemann surface is in one-to-one and bicontinuous correspondence with the w-plane (without $w = \infty$) or with the complex sphere from which one point is omitted. We are also able to define a continuous curve on the Riemann surface, being a continuous image of a straight line segment. Moving along such a curve is the pictural representation of analytic continuation in the z-plane.

A region on a Riemann surface is again an open and connected set. A region such that the set of its traces is a region $\mathfrak{R}$ in the z-plane and such that different points have different traces is called a *sheet* if it cannot be enlarged to a region with the same property. It visualizes a *branch* of the logarithm above $\mathfrak{R}$.

(ii) The function

$$w = \sqrt[n]{z},$$

where n is an integer >1, maps the upper half of the z-plane onto the angular region $0 < \arg w < \pi/n$ in the w-plane and the lower half onto the region $-\pi/n < \arg w < 0$. The positive real axis in the z-plane corresponds to the positive real axis in the w-plane such that $z = 0$ corresponds to $w = 0$. By the symmetry principle the z-plane cut along the negative real axis also corresponds to the region $\pi/n < \arg w < 3\pi/n$, etc., (fig. 12.4–2).

Again we consider a sequence $\{\mathfrak{R}_k\}$, $k = 0, \ldots, n-1$, of n identical replicas of the cut z-plane, such that $\mathfrak{R}_0$ is mapped onto $\mathfrak{S}_0$: $-\pi/n < \arg w < \pi/n$, $\mathfrak{R}_1$ onto $\mathfrak{S}_1$: $\pi/n < \arg w < 3\pi/n$, $\mathfrak{R}_{n-1}$ onto $\mathfrak{S}_{n-1}$: $(2n-3)\pi/n < \arg w < (2n-1)\pi/n$. If we approach the negative real axis in the z-plane from below its corresponding point in the w-plane tends to a point on the line $\arg w = (2k-1)\pi/n$ (or $-\pi/n$ if $k = n$). The points of the negative real axis in correspondence with this half ray constitute the lower boundary in the z-plane. Similarly the points of the negative real axis in correspondence with the line $\arg w = (2k+1)\pi/n$ constitute the upper boundary.

Next we identify the lower boundary of $\mathfrak{R}_{k+1}$ with the upper boundary of $\mathfrak{R}_k$ for all values of k, $0 \leqq k \leqq n-1$, where $\mathfrak{R}_n$ means $\mathfrak{R}_0$. If we try to make a material model of the surface thus obtained we see that it must penetrate itself. From a logical point of view this is not of importance. There appears a surface which covers the z-plane with n sheets and above every point of the z-plane (except above $z = 0$ and $z = \infty$) there are n places of the Riemann surface. Above $z = 0$ and $z = \infty$ there is only one place. These places also belong to the surface and are called *ramification points* or *branch points* of order $n-1$. In the case of the logarithm the points $z = 0$ and $z = \infty$ are not traces of corresponding places, i.e., the Riemann surface of the logarithmic function has two boundary points. In the example under consideration the Riemann surface is closed.

If $\mathfrak{z}_0$ is a place on the surface, where the trace is not $z = 0$ or $z = \infty$, then it is clear that a sufficiently small part containing $\mathfrak{z}_0$ can be mapped one-to-one onto a disc in the w-plane around the corresponding point w_0. It is also in one-to-one correspondence with the set of traces of all places belonging to this part.

Consider now the place above the origin. The part of the Riemann

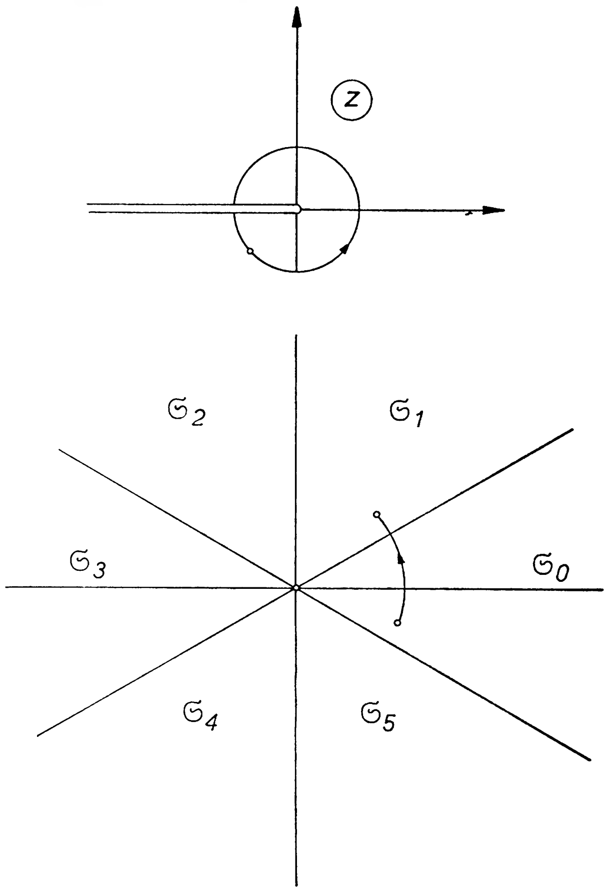


Fig. 12.4–2. The Riemann surface of the function z^6

surface above a small disc around the origin corresponds to a disc in the w-plane. This can be shown explicitly by introducing a parameter t such that $z = t^n$. To the values $t, \eta t, \ldots, \eta^{n-1}t$ with $\eta = \exp(2\pi i/n)$ correspond places above the same point z. This part of the Riemann surface is called a neighbourhood of $z = 0$. Similar considerations are valid for $z = \infty$.

(iii) The function Arcsin z can be defined by means of the integral

$$\int_0^z \frac{dt}{\sqrt{1-t^2}}.$$

The mapping properties were studied in section 10.2.7. We observe that a shaded half strip and an unshaded half-strip, (fig. 10.2–13), together

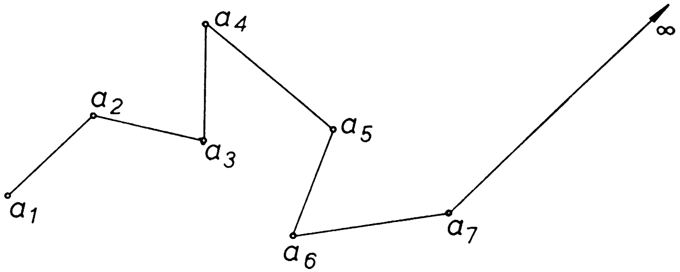


Fig. 12.4–3. Cutting the z-plane to prepare the construction of the Riemann surface for an algebraic function

correspond to the z-plane cut along from 1 to $+\infty$ and from -1 to $-\infty$. Again we take a sequence of identical replicas and identify the various boundaries in accordance with the situation of the images in the w-plane. Thus we see that above $z = 1$ as well as above $z = -1$ there are infinitely many branch points of order one. At $z = \infty$ the function behaves like the logarithm. Hence there is no place above $z = \infty$. We express this by saying that at $z = \infty$ the function presents a *logarithmic branch point.*

In a similar way Riemann surfaces can be constructed for the functions of Schwarz to be studied in Chapter 14. Their mapping properties give a clear insight into the structure of these surfaces and the character of their branch points.

(iv) Finally we wish to describe the Riemann surface of an algebraic function defined by an irreducible polynomial

$$f(z, w)$$

whose degree in w is $n > 1$.

We now imagine the finite critical points $a_1, \ldots, a_s$ to be joined in any order, and then joined to the point ∞, by a simple line L composed of rectilinear segments and a half-line, (fig. 12.4–3). If z_0 is not on L we can find n function elements satisfying $f(z, w) = 0$ which can be continued throughout the cut plane so that, according to the monodromy theorem, each gives rise to a single-valued holomorphic function. We shall denote the resulting functions by $w_1(z), \ldots, w_n(z)$. Corresponding to these functions we take n replicas of the cut plane, whose points bear the values of the functions $w_1(z), \ldots, w_n(z)$, respectively. If we continue these functions one at a time across one of the segments of the cut L, connecting two successive critical points, each of these goes over again into a definite one of these. We join the n replicas to one another in the manner hereby fully uniquely required, whereupon the cut-segment disappears. If we imagine the corresponding process to be carried out for all segments of the cut, including that which extends to ∞ all boundaries disappear and the Riemann surface for the algebraic function is complete.

The critical points are to be made traces of points of the surface by the following method: By continuing around a critical point $z = a$ (which may also be ∞) the n functions $w_1(z), \ldots, w_n(z)$ undergo a definite permutation which can be decomposed into a certain number of cyclical permutations. The sheets corresponding to the functions of one cycle are connected in one and the same place above $z = a$. They are represented by a Puiseux series and we assign the initial coefficient c_0 to this place if the series has no terms with negative exponents and the value ∞ otherwise.

We wish to discuss in some detail a numerical example. Consider the function defined by

$$w = \sqrt[3]{z}+\sqrt{z-1}. \qquad (12.4\text{–}1)$$

Since $\sqrt[3]{z}$ is 3-valued and $\sqrt{z-1}$ is 2-valued we find by combining the values that w is 6-valued. Hence w is defined by an equation of the sixth degree. After some computation we find

$$w^6-3(z-1)w^4-2zw^3+3(z-1)^2w^2-6z(z-1)w-(z^3-4z^2+3z-1) = 0. \qquad (12.4\text{–}2)$$

Critical values are $z = 0$, $z = 1$, $z = \infty$, for at all other points (12.4–1) is regular and can, therefore, be expanded in an ordinary power series at any of these points.

Now we make the following agreement. The functions $w_1(z), \ldots, w_6(z)$ are uniquely defined by their values at a given point not on L, where L is now the segment connecting 0 and $+1$ and the half-ray from 0 to ∞ along the negative real axis, (fig. 12.4–4). We may take $z = 2$. Then we agree

$$w_1(2) = \sqrt[3]{2}+1, \quad w_2(2) = \lambda\sqrt[3]{2}+1, \quad w_3(2) = \lambda^2\sqrt[3]{2}+1,$$
$$w_4(2) = \sqrt[3]{2}-1, \quad w_5(2) = \lambda\sqrt[3]{2}-1, \quad w_6(2) = \lambda^2\sqrt[3]{2}-1$$

where

$$\lambda = \exp(2\pi i/3).$$

In a neighbourhood of $z = 0$ we have

$$w_1 = i+z^{\frac{1}{3}}-\tfrac{1}{2}iz-\tfrac{1}{8}iz^2+\ldots$$

which can be obtained by expanding $\sqrt{z-1}$ by means of the binomial theorem. We obtain w_2 and w_3 by replacing z by λz and $\lambda^2 z$ successively.

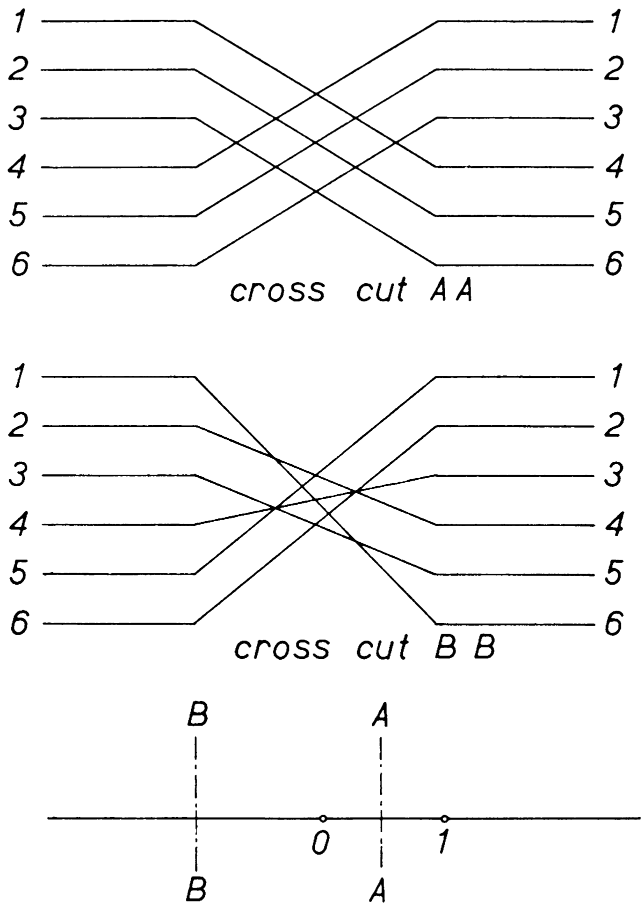


Fig. 12.4–4. Schematic representation of the connection of the sheets of the Riemann surface of (12.4–2) between 0 and 1 (cross cut AA) and between $-\infty$ and 0 (cross cut BB)

Additionally we have the expansion

$$w_4 = -i+z^{\frac{1}{2}}+\tfrac{1}{2}iz+\tfrac{1}{8}iz^2+\ldots$$

which is obtained from the series for w_1, if we take the other value of $\sqrt{z-1}$. In the same way as above we get the expansions for w_5 and w_6. Thus we see that above $z = 0$ there are two branch points of order two.

In a neighbourhood of $z = 1$ we have, if we expand $\sqrt[3]{z}$ in powers of $z-1$,

$$\begin{aligned} w_1 &= 1+(z-1)^{\frac{1}{2}}+\tfrac{1}{3}(z-1)-\tfrac{1}{9}(z-1)^2+\ldots, \\ w_2 &= \lambda+(z-1)^{\frac{1}{2}}+\tfrac{1}{3}\lambda(z-1)-\tfrac{1}{9}\lambda(z-1)^2+\ldots, \\ w_3 &= \lambda^2+(z-1)^{\frac{1}{2}}+\tfrac{1}{3}\lambda^2(z-1)-\tfrac{1}{9}\lambda^2(z-1)^2+\ldots, \end{aligned}$$

and w_4, w_5, w_6 are obtained by replacing $(z-1)^{\frac{1}{2}}$ by $-(z-1)^{\frac{1}{2}}$. Thus we see that above $z = 1$ there are three branch points each of order one.

In a neighbourhood of $z = \infty$ we have

$$w = z^{\frac{1}{3}} + z^{\frac{1}{2}}\sqrt{1-\frac{1}{z}} = z^{\frac{1}{3}} + z^{\frac{1}{2}} - \tfrac{1}{2}z^{-\frac{1}{2}} - \tfrac{1}{8}z^{-\frac{3}{2}} + \ldots$$
$$= z^{\frac{3}{6}} + z^{\frac{2}{6}} - \tfrac{1}{2}z^{-\frac{3}{6}} - \tfrac{1}{8}z^{-\frac{9}{6}} + \ldots,$$

an expansion in terms of $z^{-\frac{1}{6}}$. The others are obtained by multiplying by $\mu, \mu^2, \ldots, \mu^5$, where now

$$\mu = \exp(2\pi i/6).$$

Thus we see that above $z = \infty$ there is a single branch point of order 5. The values of the function at $z = 0$ are i and $-i$ respectively, at $z = 1$ the values are $1, \lambda, \lambda^2$ and at $z = \infty$ the function takes the value ∞. Hence the branch point at $z = \infty$ is an algebraic pole.

If we encircle $z = 1$ once in a positive sense, leaving $z = 0$ on the left, the functions $w_1, \ldots, w_6$ are permutated according to

$$(14)(25)(36).$$

Encircling $z = 0$, leaving $z = 1$ outside, yields the permutation

$$(123)(456).$$

Hence the encirclement of both critical points gives rise to the product of these permutations, viz.,

$$\begin{pmatrix}123456\\564231\end{pmatrix}.$$

Now we take 6 sheets $\mathfrak{R}_1, \ldots, \mathfrak{R}_6$ corresponding to $w_1, \ldots, w_6$. Along the segment between $z = 0$ and $z = 1$ we join $\mathfrak{R}_1$ and $\mathfrak{R}_1$ crosswise, $\mathfrak{R}_2$ and $\mathfrak{R}_5$ crosswise and $\mathfrak{R}_3$ and $\mathfrak{R}_6$ crosswise. Along the segment between $z = 0$ and $z = \infty$ we identify the upper boundary of $\mathfrak{R}_1$ with the lower boundary of $\mathfrak{R}_5$, the upper boundary of $\mathfrak{R}_5$ with the lower boundary of $\mathfrak{R}_3$, etc. A schematic picture is shown in fig. 12.4–4.

12.4.2 – Definition of a Riemann surface

For the construction of a Riemann surface of an analytic function $F(z)$ the following general method suggests itself: We start with a function element which may be assumed to be a power series. We imagine its circle of convergence to be cut out of paper and to its points are assigned the values of the element. If we continue the initial element directly by means of a second power series, we also think of its circle of convergence as being cut out and pasted in the proper position on the first disc. The parts pasted together are counted as a single sheet covered once with values. If we succeed in carrying out another continuation, we paste the new disc on it an entirely similar manner, and so on. Each new disc is pasted on the preceding one from which it was obtained by means of direct continuation in the manner described.

Suppose that, after repeated continuation, we arrive with one of the new circles over a circular disc not immediately preceding. Then the new disc shall be pasted together with the old one if and only if in the overlapping part the function elements take the same values. If, however, this is not the case, then they remain disconnected. Obeying this rule we imagine our procedure to be continued as long as possible. Then there results a surface-like configuration which covers the z-plane with several sheets which can have the most varied forms, and can be joined together in the most varied manners. In the course of pasting sheets together, it is sometimes necessary to join two sheets which are separated by others lying between them. We must imagine this to take place without cutting the intermediary sheets. This is impossible for concrete execution but causes no difficulty for the purely mental construction. As we shall see further on the more rigorous definition of a Riemann surface avoids this difficulty.

It should be noticed that it is immaterial whether we continue by means of circular discs or by means of any other regions, provided only that we adhere to the agreements we have made. The examples dealt with in the previous section may serve as an illustration.

In general the way in which the sheets are joined together may become very complicated. For our purposes the Riemann surface lays no claim to being an end in itself, but is in most cases intended as an aid to the imagination. As far as the general case is concerned, it is sufficient to know that for a given function a Riemann surface can be constructed at all events, on which its values form a single-valued function of position. Every point is covered by as many sheets as there are different elements for a neighbourhood of this point, and these sheets hang together in a perfectly definite manner. This means that if we begin at a certain point $\mathfrak{z}_0$ of a particular sheet and describe any definite path (that is a path whose projection on the z-plane is given), its course on the surface is fully unique, and consequently leads us to a perfectly definite point.

It is possible to give a definition of a Riemann surface which meets all requirements of rigour. But then we must employ a lot of topology and the study of Riemann surface from this point of view is a separate branch of mathematics.

We proceed to give a definition of a Riemann surface in a more abstract way which is quite satisfactory for our purposes. We recall that an analytic function is an equivalence class of function elements, being power series in a variable $z-a$, where a ranges over a region $\mathfrak{R}_0$ of the finite plane. To this set of elements we adjoin the following classes of elements in the case that they exist:

1) polar elements of the form

$$(z-a)^{-k}\sum_{\nu=0}^{\infty} c_\nu (z-a)^\nu, \qquad k \geq 1;$$

2) algebraic elements of the form

$$(z-a)^{k/h}\sum_{\nu=0}^{\infty} c_\nu (z-a)^{\nu/h}, \qquad k \gtreqless 0,\ h > 0;$$

3) infinitary elements of the form

$$z^{k/h}\sum_{\nu=0}^{\infty} c_\nu z^{-\nu/h}, \qquad k \gtreqless 0,\ h > 0,$$

corresponding to the point at infinity. In section 12.2.7 we have seen how these elements can appear. We add these elements to the regular function elements of the analytic function under consideration. The centres of these elements adjoined to $\mathfrak{R}_0$ yield a set $\mathfrak{R}$, *the domain of definition of the analytic function* $F(z)$. The collection of regular elements completed by polar, algebraic and infinitary elements is called an *analytic configuration.* It is clear that the values of $F(z)$ are precisely the values which take its elements at its respective centres.

Now we consider the set of pairs $(a; f(z))$, where $f(z)$ is an element with centre a of the analytic configuration associated with the analytic function $F(z)$. The case $a = \infty$ is included. Such a pair is called a *place* and the set of all places constitute the *Riemann surface* of the analytic function. We shall often denote a place by a small gothic type, if convenient.

We shall say that the place $\mathfrak{p} = (a; f(z))$ is *above* $z = a$ and that $z = a$ is the *trace* of $\mathfrak{p}$ in $\mathfrak{R}$ (or the projection of $\mathfrak{p}$ onto the z-plane). Starting with $f(z)$ we may consider all analytic continuations such that the set of elements obtained in this way represents a single-valued function. This set is maximal in the sense that further continuation yields a function which is no more single-valued. The corresponding places constitute a *sheet* of the Riemann surface. Thus the surface can be decomposed into sheets, but it should be noticed that this decomposition is in no way unique.

It is clear that the abstract definition avoids the difficulty of interpenetration of sheets. However, it might seem to be more natural to define a place as a pair of numbers (a, b), where b is one of the values of

$F(z)$ at $z = a$. But then we should obtain a mathematical entity which does not present adequately the structure of an analytic configuration. For it may happen that different elements take the same value as their common centre and thus essential properties would be obscured.

Our next task is in showing that the abstractly introduced notion of Riemann surface has essential properties in common with a surface in the usual sense. First we shall define a notion of nearness, showing that a Riemann surface has a topological structure.

A *neighbourhood* of the place $\mathfrak{p}_0 = (a_0; f_0(z))$ is the set of all places $\mathfrak{p} = (a; f(z))$, such that a is in a neighbourhood $\mathfrak{U}(a_0)$ of a_0 (and of course in $\mathfrak{R}$) and $f(z)$ is a rearrangement of $f_0(z)$, (section 12.1.2).

This definition does not require comment in the case that $\mathfrak{p}_0$ is a regular place (i.e., a place defined by a regular element) and $\mathfrak{U}(a_0)$ does not contain singular points (i.e., $f_0(z)$ is arbitrarily continuable throughout $\mathfrak{U}(a_0)$). The rearrangement procedure also applies for algebraic elements. Consider an expansion

$$\sum_{\nu=0}^{\infty} c_\nu (z-a_0)^{(k+\nu)/h},$$

where $c_n \neq 0$ for at least one value of n such that h is relatively prime to $k+n$. The series represents a collection of h power series of $(z-a_0)^{1/h}$, obtained by choosing different initial values of $(z-a_0)^{1/h}$. Let now $z = a$ denote a point inside the circle of convergence of the series. We have evidently

$$(z-a_0)^{(k+n)/h} = (a-a_0)^{(k+n)/h} \left(1 + \frac{z-a}{a-a_0}\right)^{(k+n)/h}$$

and a value of $(a-a_0)^{1/h}$ fixes a function element above $z = a$, if the binomial on the right is understood as a principal value and expanded by the binomial theorem as an ordinary power series in $z-a$. In view of Weierstras's double series theorem (section 2.20.4) we obtain an expansion $f(z)$ of $f_0(z)$ in a neighbourhood of $z = a$, being the rearrangement of $f_0(z)$ in the extended sense. Thus we can find h different series about $z = a$. The case of a polar element is included for then we have $k < 0$ and $h = 1$. With obvious modifications we can also handle the case of infinitary elements.

Consider a place $\mathfrak{p}_0$ which is a branch point of order $h-1$. If $h = 1$ the place is ordinary (or regular). Let a_0 be the trace of $\mathfrak{p}_0$. If ρ is positive and sufficiently small then above any point $z = a$ of the disc $|z-a_0| < \rho$ we have h distinct elements

$$(a; f_1(z)), \ldots, (a; f_h(z)),$$

defining h places $\mathfrak{p}_1, \ldots, \mathfrak{p}_h$. Let now t be a point of the unit disc $\mathfrak{K}$

about $t = 0$ in the t-plane. If

$$(k-1)\frac{2\pi}{h} \leqq \arg t \leqq k\frac{2\pi}{h}, \qquad k = 1, \ldots, h,$$

we assign to t the place $\mathfrak{p}_k = (a; f_k(z))$, with

$$a = a_0 + (\rho t)^h$$

and

$$f_k(z) = \sum_{\nu=0}^{\infty} c_\nu (\rho t)^{k+\nu} \sum_{\mu=0}^{\infty} \binom{\frac{k+\nu}{h}}{\mu} \left(\frac{z-a}{(\rho t)^{h\mu}}\right)^{\mu}$$

$$= \sum_{\mu=0}^{\infty} (\rho t)^{-h\mu} (z-a)^\mu \sum_{\nu=0}^{\infty} c_\nu \binom{\frac{k+\nu}{h}}{\mu} (\rho t)^{k+\nu}. \qquad (12.4\text{–}3)$$

This correspondence is one-to-one and continuous in both directions, hence a homeomorphism. This means that the mapping $t \to \mathfrak{p}$ is continuous, for all places $\mathfrak{p}$ which are images of points of a sufficiently small neighbourhood $\mathfrak{U}(t)$ belong to a neighbourhood $\mathfrak{U}(\mathfrak{p}_k)$ of $\mathfrak{p}_k$. In the opposite sense we have that for any given $\mathfrak{U}(t)$ there is a $\mathfrak{U}(\mathfrak{p}_k)$ such that the t corresponding to all $\mathfrak{p}$ in $\mathfrak{U}(\mathfrak{p}_k)$ belong to $\mathfrak{U}(t)$. These statements follow from the definition of neighbourhoods on the Riemann surface. Thus

Every place $\mathfrak{p}$ of the Riemann surface has at least one neighbourhood $\mathfrak{U}(\mathfrak{p})$ which is the homeomorphic image of an open disc of the complex t-plane about the origin.

Formula (12.4–3) gives a parametric representation of the places $\mathfrak{p}$ in a neighbourhood of $\mathfrak{p}_0$ and t is known as a *locally uniformizing parameter*. In many cases it is desirable to consider more general representations where the first coordinate is also given by an infinite series in a parameter t. The choice of the parameter is highly arbitrary as long as it gives the homeomorphic representation of $\mathfrak{K}$: $|t| < 1$ onto a neighbourhood of the place $\mathfrak{p}_0$. Thus if t is a locally uniformizing parameter, so is $t^* = \chi(t), \chi'(t) \neq 0$, where $\chi(t)$ is a power series convergent and univalent in $|t| < 1$.

Continuing our considerations we turn our attention to two places $\mathfrak{p}_1$ and $\mathfrak{p}_2$ with corresponding neighbourhoods $\mathfrak{U}(\mathfrak{p}_1)$ and $\mathfrak{U}(\mathfrak{p}_2)$, each of them being homeomorphic representations of an open disc $\mathfrak{K}$ in the t-plane. We denote the mapping function by φ_1 and φ_2 respectively, i.e.,

$$\mathfrak{U}(\mathfrak{p}_1) = \varphi_1(\mathfrak{K}), \qquad \mathfrak{U}(\mathfrak{p}_2) = \varphi_2(\mathfrak{K}).$$

The inverse mappings are

$$\mathfrak{K} = \check{\varphi}_1(\mathfrak{U}(\mathfrak{p}_1)), \qquad \mathfrak{K} = \check{\varphi}_2(\mathfrak{U}(\mathfrak{p}_2)).$$

It follows that

$$\mathfrak{U}(\mathfrak{p}_2) = \varphi_2(\check{\varphi}_1(\mathfrak{U}(\mathfrak{p}_1))),$$

that is, we can map $\mathfrak{U}(\mathfrak{p}_1)$ onto $\mathfrak{U}(\mathfrak{p}_2)$ in a one-to-one and bicontinuous manner. Suppose that $\mathfrak{U}(\mathfrak{p}_1)$ and $\mathfrak{U}(\mathfrak{p}_2)$ intersect and that $\mathfrak{p}_0$ is a place of the intersection. Then there are values t_1 and t_2 of the parameter t such that

$$\mathfrak{p}_0 = \varphi_1(t_1) = \varphi_2(t_2)$$

and, evidently,

$$t_2 = \varphi_2(\varphi_1(t_1)), \qquad t_1 = \check{\varphi}_1(\varphi_2(t_2)).$$

Assume now that the traces of the places, belonging to $\mathfrak{U}(\mathfrak{p}_1)$ and $\mathfrak{U}(\mathfrak{p}_2)$ respectively, are circular discs $\mathfrak{U}(a_1)$ and $\mathfrak{U}(a_2)$ in the z-plane. The intersection of $\mathfrak{U}(\mathfrak{p}_1)$ and $\mathfrak{U}(\mathfrak{p}_2)$ is a region $\mathfrak{R}$. We put

$$\mathfrak{R}_1 = \check{\varphi}_1(\mathfrak{R}), \qquad \mathfrak{R}_2 = \check{\varphi}_2(\mathfrak{R}).$$

Then $\mathfrak{R}_1$ and $\mathfrak{R}_2$ are subregions of $\mathfrak{K}$ and

$$\mathfrak{R}_2 = \check{\varphi}_2(\varphi_1(\mathfrak{R}_1)), \qquad \mathfrak{R}_1 = \check{\varphi}_1(\varphi_2(\mathfrak{R}_2)).$$

If now $\mathfrak{U}(\mathfrak{p}_1)$ and $\mathfrak{U}(\mathfrak{p}_2)$ have parametric representations of the type (12.4–3) then

$$a_0 = a_1 + (\rho t_1)^{h_1} = a_2 + (\rho t_2)^{h_2},$$

where h_1 and h_2 are positive integers. It follows that t_2 is a holomorphic function of t_1 and conversely, hence the mapping $\check{\varphi}_2\varphi_1$ is directly conformal as is $\check{\varphi}_1\varphi_2$. Thus

If the homeomorphisms φ_1 and φ_2 map the unit disc $\mathfrak{K}$ in the t-plane onto neighbourhoods $\mathfrak{U}(\mathfrak{p}_1)$ and $\mathfrak{U}(\mathfrak{p}_2)$ which have a nonempty connected intersection then the composite transformation $\check{\varphi}_2\varphi_1$ maps a subregion $\mathfrak{R}_1$ of $\mathfrak{K}$ onto another subregion $\mathfrak{R}_2$ in a one-to-one manner.

The above definitions and properties describe adequately a mathematical entity which can be individualized by a surface as introduced before. Since the notion of continuity is available we can define the notions of continuous path, region, etc. It is again clear that the function $F(z)$ to which the Riemann surface is associated is a single-valued function on this surface, i.e., to a given place there corresponds precisely one value of the function.

12.5 – Classification of algebraic Riemann surfaces

12.5.1 – Triangulation

We shall be interested in those properties of Riemann surfaces which can be used to distinguish different surfaces. The most basic of these properties are the ones which determine whether two surfaces are topologically equivalent, i.e., whether there exists a homeomorphism of one of the surfaces onto the other. In the case of surfaces associated to algebraic functions the theory is comparatively simple and brings to light a fundamental invariant, the *genus* of the surface.

A basic property is the *triangulability* of a Riemann surface which we shall explain in more detail. By a *triangle* s^2 on a Riemann surface $\mathfrak{F}$ is understood a closed euclidean triangle e^2 and a one-to-one bicontinuous mapping φ of e^2 into $\mathfrak{F}$. We shall write $s^2 = (e^2, \varphi)$ and call s^2 the image of e^2 under φ. The images of the sides and vertices of e^2 are called *sides* and *vertices* of s^2. Each side or vertex appears when we restrict φ to a side or vertex of e^2. A place $\mathfrak{p}$ of $\mathfrak{F}$ belongs to s^2 if $\mathfrak{p}$ is in the image of e^2 under φ.

A surface $\mathfrak{F}$ is called *triangulated* if on $\mathfrak{F}$ are chosen a finite or enumerably infinite system of triangles such that

(i) every place of $\mathfrak{F}$ is in at least one triangle of the system;

(ii) two triangles have either no point in common, or only one vertex, or only one side;

(iii) a side belongs to precisely two triangles;

(iv) the triangles meeting at the same vertex constitute a finite cycle, in which every triangle has a side in common with the adjacent triangle, (fig. 12.5–1).

It can be proved that every Riemann surface is triangulable, but we shall only prove this theorem in the case of a Riemann surface with a finite number of branch points.

Let $a_1, \ldots, a_r$ denote the traces of the branch points; some of them may be logarithmic branch points. These are then boundary points of

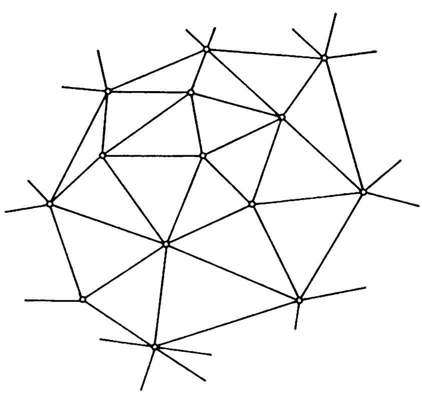


Fig. 12.5–1. Triangulation (schematic)

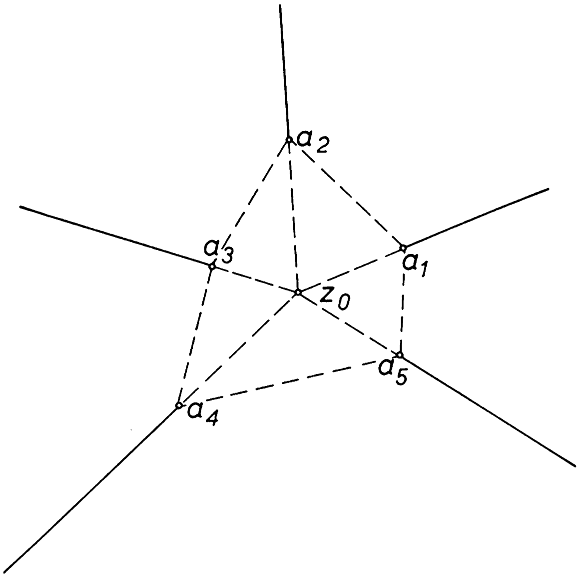


Fig. 12.5–2. Triangulability of the Riemann surface of an algebraic function

the surface. Next we take in the z-plane a point z_0 which is not on a segment connecting two points a_k. We connect z_0 with all these critical points and extend these connecting segments beyond these points to the point ∞ (which may or may not be under a branch point), (fig. 12.5–2). The half-rays are used as cuts in the z-plane. Starting with a function element above z_0 we may continue it analytically throughout the cut plane (which is simply connected) and the function elements thus obtained define the places of a sheet of the Riemann surface. By suitable identification of the various sheets which can be obtained in this way along the half-rays emanating from the final critical points we get the whole Riemann surface. On every sheet we may define segments whose points are places above the segments z_0a_k, $k = 1, \ldots, r$ in the z-plane. In addition we connect $a_1, \ldots, a_r$ in cyclic order and draw corresponding segments in either sheet. In the case that there are only algebraic branch points (and no boundary points at all) every sheet and consequenty also the whole Riemann surface is divided into triangles, i.e. the surface is triangulated. The total number of triangles is finite and the surface is said to be *closed.*

In the case that there are also logarithmic branch points we may assume, after eventual subdividing triangles, that no more than one vertex is a

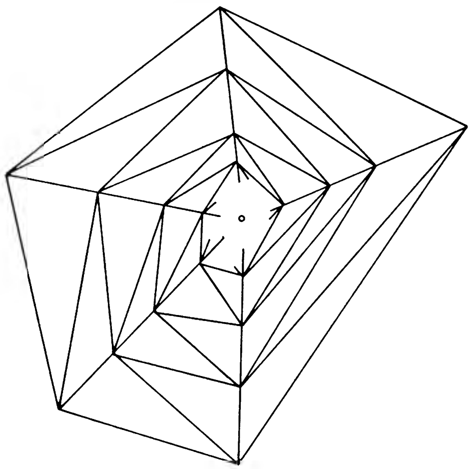


Fig. 12.5–3. Triangulation near a logarithmic singularity

logarithmic branch point. A triangle having one vertex as logarithmic branch point may be subdivided further as depicted in fig. 12.5–3. The adjacent triangles with the same critical vertex are triangulated in such a way that the triangles of the subdivisions are collected into cycles with a finite number of members. The number of subtriangles is now infinite. The whole surface is subdivided into an enumerably infinite number of triangles and a surface of this type is called *open.*

12.5.2 – Normal forms of algebraic Riemann surfaces

In dealing with topological questions about surfaces, it is convenient to be able to visualize a model which is homeomorphic to the surface in question, rather than to proceed abstractly. Such models help our intuitive understanding of the problems we consider and give a fuller meaning to the results. For Riemann surfaces of algebraic functions such models are easily found. By deforming a triangle we understand a one-to-one and bicontinuous mapping of the one triangle onto another. It is clear that by deformation we can transform a sheet into a convex polygon which is divided into triangles. The sides of such a polygon correspond to cuts on the original surface. If we connect two sheets along such a cut in accordance with the identification on the surface, we can connect two polygons in a corresponding manner along a side such that the polygons have no other points in common. It is allowed to suppose that the join of these polygons is again convex, (fig. 12.5–4).

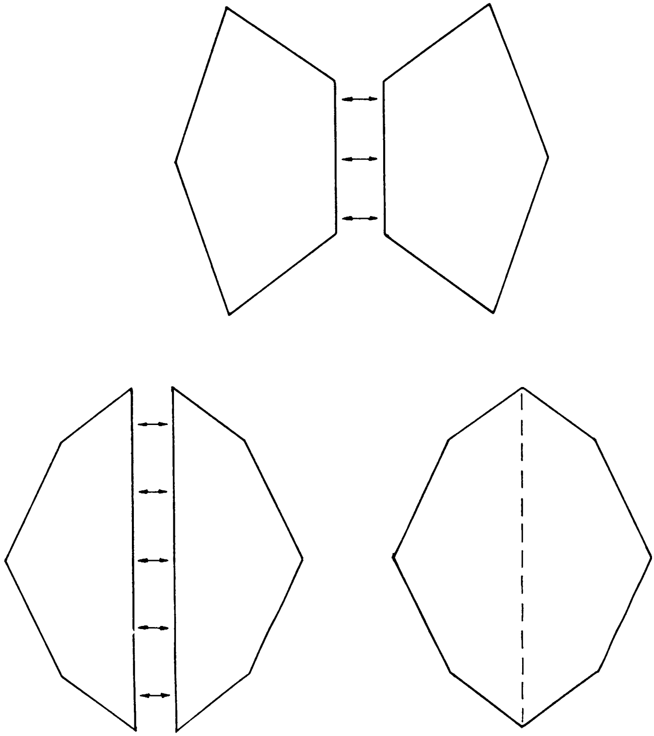


Fig. 12.5–4. The joining of two sheets of a Riemann surface

Continuing in this way we may combine all sheets deformed into convex polygons into a single polygon by identifying appropriate sides. Then we have transformed the original surface into a convex polygon. In general not all identifications along the cuts on the Riemann surface are reproduced by joining the polygons. Every transfer from one sheet to another which is not realized by joining the corresponding polygons corresponds to two sides of the polygon which must be considered as identical if we will get a topological model of the surface. Hence the number of sides of the polygon is even. We shall denote the sides of a pair which must be identified by the same type, from which one is primed.

The classification of Riemann surfaces of algebraic functions will be simplified considerably after the construction of so-called *normal forms* of the polygons representing the surfaces. They are obtained by applying some elementary transformations.

Going around the boundary in the positive sense of the polygon as

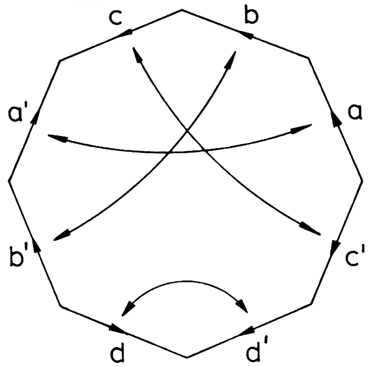


Fig. 12.5–5. Representation of a surface by means of a polygon with sides identified in pairs

constructed previously and associating a letter to each side, such that sides which are to be identified are labelled by the same letter, we obtain a symbol of the polygon by writing these letters in the order in which they are encountered. The symbol of fig. 12.5–5 is

$$abca'b'dd'c'.$$

It is allowed to perform a cyclic permutation of the letters. If the polygon is now cut into two polygons along a line joining two of its vertices, and if the parts are then again attached along a pair of identified sides, and if the two sides of the cut are identified, we obtain a new polygon with pairs of sides identified. The two polygons both represent the same surface, for the identification of points was not changed in this process. Our main task will be the simplification of a given polygon.

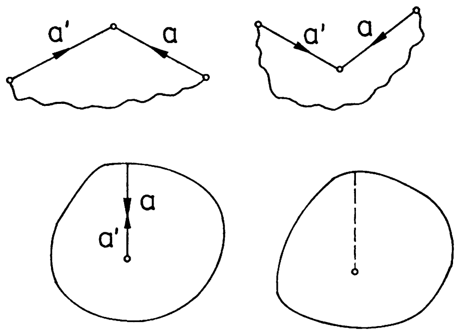


Fig. 12.5–6. The cancelling of two adjacent identified sides

If the letters aa′ appear in the symbol as adjacent sides of the polygon and if the symbol has at least one other letter (hence at least four letters altogether) then the letters can be removed from the symbol to obtain a new symbol for the polygon. Fig. 12.5–6 illustrates this process and suffices as a proof. The polygon aa′ which we could not handle in this way is one of our normal forms. From now on we may assume that the polygon has at least four sides and that pairs as aa′ are suppressed.

The next step will be the transformation of the polygon into a polygon in which all the vertices correspond to the same point on the Riemann surface. We designate a certain vertex of the polygon by P and also label P all other vertices which correspond to the same point as P. If there is a side a of the polygon with one vertex unlabelled, we label it Q together with all other vertices which correspond to the same point on the Riemann surface. Now we show how to reduce the number of vertices labelled Q by one and thereby increase the number of vertices labelled P by one. Let a = PQ and let b denote the side which has the vertex Q in common

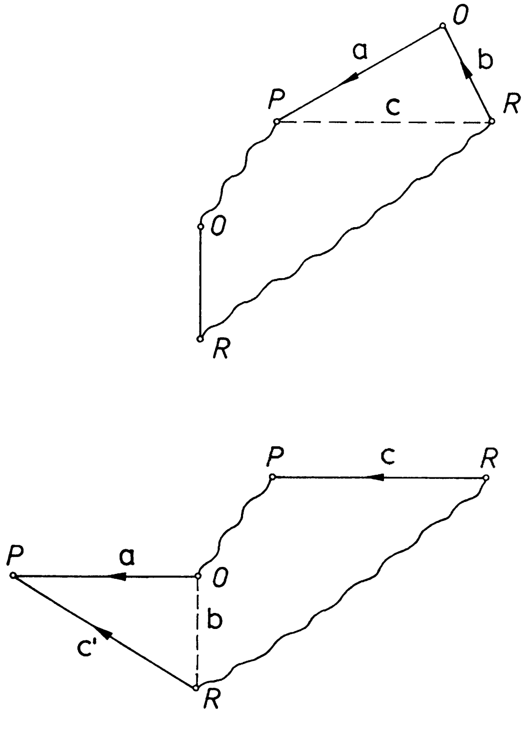


Fig. 12.5–7. Reduction of the number of unidentified vertices

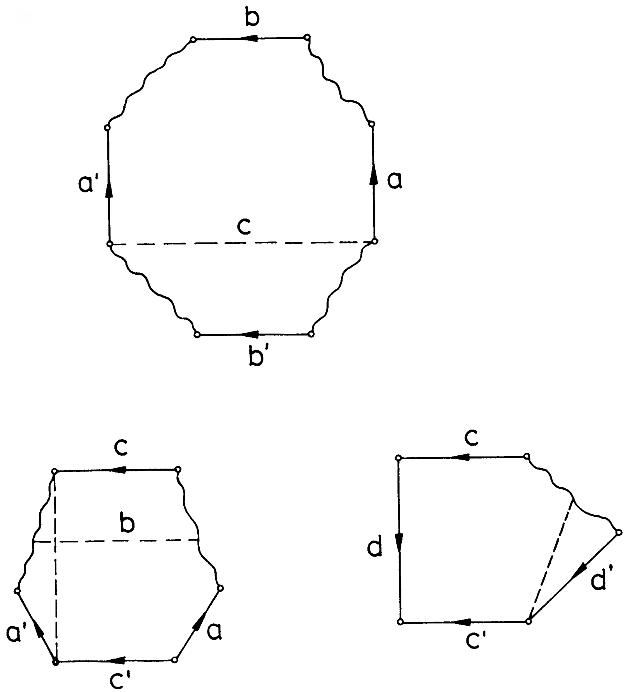


Fig. 12.5–8. Reduction to the normal form

with a. We know that b is not a′, for if it were, we would have had aa′ which was already suppressed. We join the vertex R of b (R may be P or Q) to the vertex P of a by a diagonal to form a triangle with sides a, b and c. We cut out this triangle of the polygon along c and then attach it to the rest of the polygon along the side b of the triangle and the remaining side of the polygon to form a new polygon having the same number of sides as the former one. Whereas the triangle was previously attached to PR exposing the vertex Q it is now attached along RQ, exposing the vertex P, (fig. 12.5–7). Thus the new polygon has one more vertex P and one less vertex Q as did the former polygon. Continuing in this way we obtain a polygon in which all vertices are equivalent (i.e., correspond to the same point on the Riemann surface) and labelled P.

Finally we effect a rearrangement of the sides. A pair of sides are called *linked* if they appear in the symbol in the following order

$$\ldots a \ldots b \ldots a' \ldots b'. \qquad (12.5\text{–}1)$$

We now show that each side is linked with some other side. If this is not

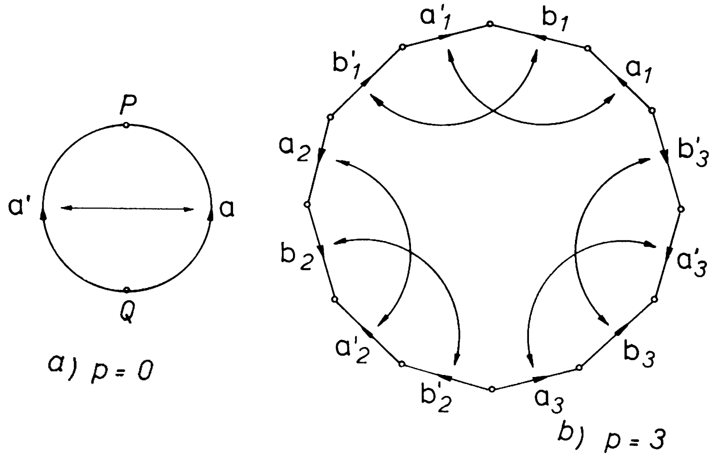


Fig. 12.5–9. Normal form of an algebraic Riemann surface

true, there must be a side c such that all letters between c and c′ are identified among themselves, so that none of the corresponding sides lie outside the letters c . . . c. Then we may select a point on the side c, not the vertex P and join it by a line segment d in the polygon to an equivalent point on the other side c′. This line divides the polygon into two parts which have P and the points on d identified. But one vertex of c lies in the first part and the other vertex P in the second part, which is impossible since P would not have a euclidean neighbourhood in the surface. Thus each side of the polygon is linked with another. By the process shown in fig. 12.5–8 the polygon can be transformed so that the linked pair (12.5–1) is brought together in the sequence cdc′d′. The further combination of linked pairs does not destroy those already combined, so that we finally have:

The normal form of an algebraic Riemann surface is a polygon with symbol

$$\boxed{\begin{array}{ll}\text{(i)} & aa', \\ \text{(ii)} & a_1 b_1 a_1' b_1' a_2 b_2 a_2' b_2' \dots a_p b_p a_p' b_p'.\end{array}} \qquad (12.5\text{–}2)$$

In case (i) we say that the normal form has the *genus zero*, while in case (ii) the normal form has the *genus p*, (fig. 12.5–9).

12.5.3 – Visualization of the normal forms

We now wish to discover what the normal forms look like when we actually paste together the identified sides.

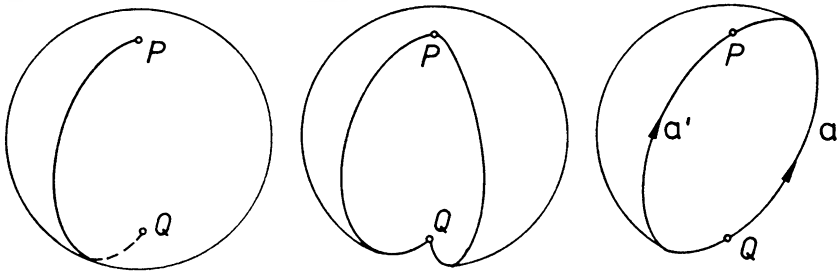


Fig. 12.5–10. Topological model of a Riemann surface of genus zero

Pasting together the sides a and a′ of a surface of genus zero, we get a surface which is topologically a sphere. It is convenient to imagine the sphere made of an elastic flexible material; then cutting it open along the line a as shown in fig. 12.5–10 and flattening it out gives us the polygon which we have taken to be normal form of genus zero.

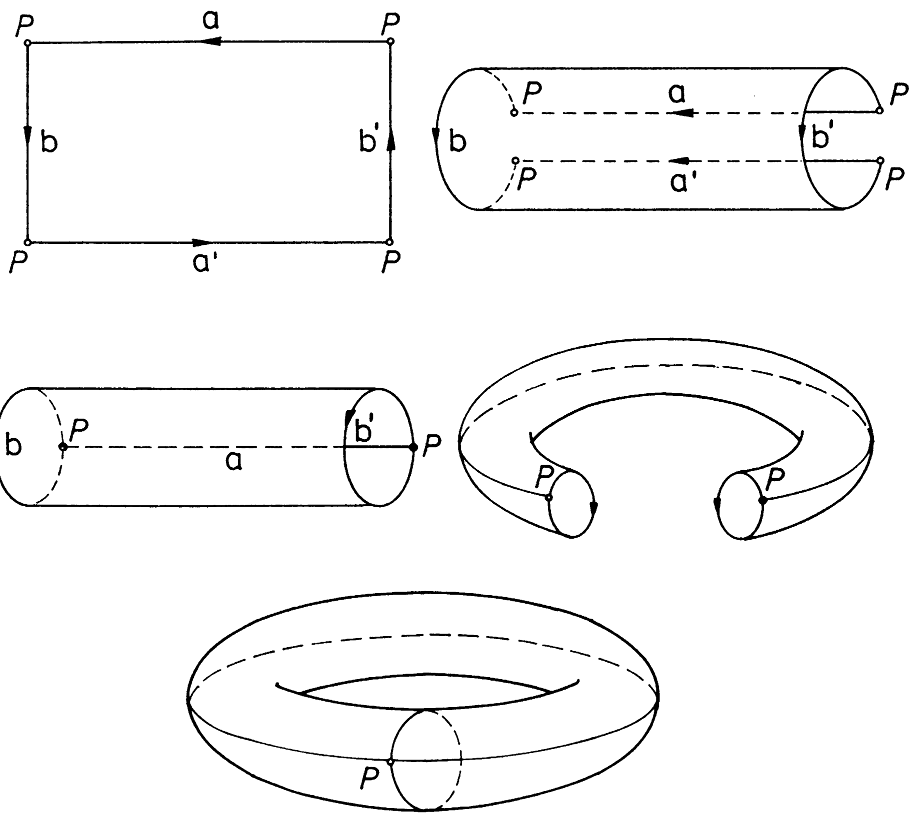


Fig. 12.5–11. Topological model of a Riemann surface of genus one

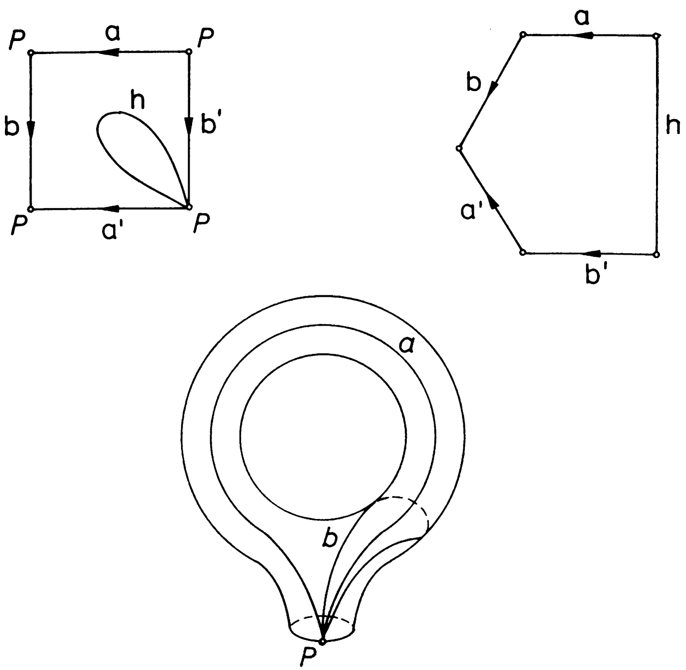


Fig. 12.5–12. The handle

The normal form of genus one is a quadrilateral aba′b′ which is homeomorphic to a quadrangle, (fig. 12.5–11). If we paste together the identified sides a and a′ we obtain a cylinder with its two ends b and b′ identified. Now pasting together these ends, we get a torus as a model for the normal form of genus one.

The torus may be viewed topologically in yet another way. If we cut a disc out of the torus as shown in fig. 12.5–12, we obtain a handle. The hole left after cutting out the disc is bounded by a curve h which may be made to pass through the point corresponding to the vertex P. If we now separate this curve at P, the rectangle with a hole opens into a pentagon with the symbol

$$aba'b'h$$

which is the symbol of a handle. The disc which we cut out of the torus can be deformed into a sphere with a disc removed. Thus the torus may be conceived as a sphere with a handle attached to it, (fig. (12.5–13)).

This leads us to the normal forms of higher genus. Out of a sphere let

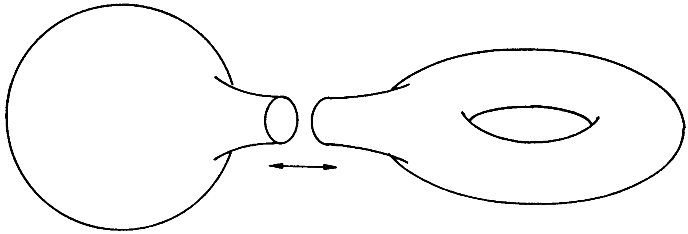


Fig. 12.5–13. Deformation of a torus into a sphere with a handle attached to it

us cut p discs bounded by curves $h_1, \ldots, h_p$, having only the point P in common. By flattening out the resulting surface we obtain a p-sided polygon with the symbol $h_1 \ldots h_p$. If we attach to each h_k the handle $a_k b_k a'_k b'_k h_k$ by pasting together the curves h_k, we obtain the normal form of genus p, (fig. 12.5–14). Thus

The normal form of genus p is topologically equivalent to a sphere with p handles attached.

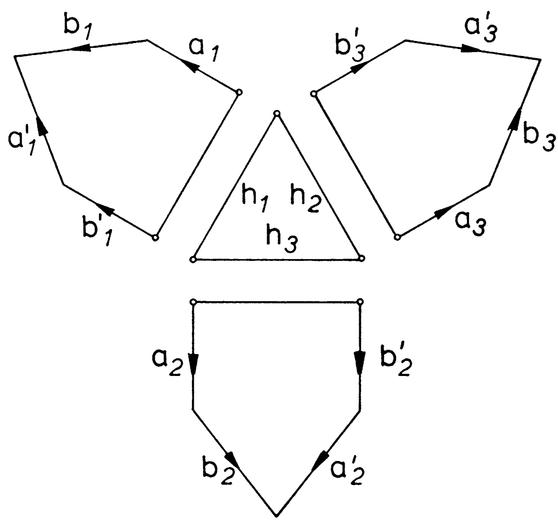


Fig. 12.5–14. Construction of a normal form by means of handles

12.5.4. – The invariance of the genus

We have established that every algebraic Riemann surface is homeomorphic to a normal form of genus p and hence to a sphere with p handles. The genus completely determines the normal form, so that two triangulated surfaces are homeomorphic if their normal forms have the same genus. On the other hand, to find the normal form we used a specific triangulation of the surface. And the question arises, if we would get the normal form of the same genus if we had taken a different triangulation of the same surface. The answer on this question is embodied in the fact that the genus of a normal form depends only upon the surface and not upon the triangulation used, so that homeomorphic normal forms have the same genus. To prove this assertion we shall relate the genus to a topological invariant of the surface, the so-called Euler characteristic.

On a Riemann surface $\mathfrak{F}$ we consider a system of finitely many closed curves, such that each pair of them meet in a finite number of points. These curves may decompose $\mathfrak{F}$ in a finite number of pieces such that each piece is homeomorphic to a region in the z-plane bounded by a simple closed polygon. Such a piece is called a polygon on $\mathfrak{F}$; the images of the sides of the polygon in the z-plane are called the sides of the polygon on $\mathfrak{F}$, their end points are the vertices of the polygon. A polygonal decomposition of the surface is effected if:

1) Each point of $\mathfrak{F}$ is in at least one polygon.

2) Two polygons are either disjoint, or have precisely one vertex, or one side in common.

3) Every side belongs to exactly two polygons.

Every algebraic Riemann surface possesses a polygonal decomposition, e.g., a triangulation.

Consider now two homeomorphic surfaces $\mathfrak{F}$ and $\mathfrak{F}^*$ and denote a polygonal decomposition on the first surface by $\mathfrak{T}$ and a similar decomposition on the other by $\mathfrak{T}^*$. The number of vertices, sides and polygons (which we suppose to be finite) may be denoted by α_0, α_1, α_2 resp.; let α_0^*, α_1^*, α_2^* denote the corresponding numbers on $\mathfrak{F}^*$. Then we assert

$$-\alpha_0+\alpha_1-\alpha_2 = -\alpha_0^*+\alpha_1^*-\alpha_2^*. \tag{12.5–3}$$

The alternating sum

$$\boxed{\chi = -\alpha_0+\alpha_1-\alpha_2} \tag{12.5–4}$$

expresses, therefore, a topological invariant of the surface, the so-called *Euler characteristic.*

In order to prove the statement we map $\mathfrak{F}^*$ topologically onto $\mathfrak{F}$. The decomposition $\mathfrak{T}^*$ is carried into a decomposition of $\mathfrak{F}$ with exactly the same scheme of vertices, sides and polygons as $\mathfrak{F}^*$. We may denote it again by $\mathfrak{T}^*$. It remains to prove that for two decompositions of a surface $\mathfrak{F}$ the alternating sum (12.5–4) is the same.

Consider first a polygon with n vertices in the plane. The alternating sum is here

$$\chi_0 = -n+n-1 = -1.$$

Let $\mathfrak{T}_1$ denote a decomposition of the polygon in subpolygons, χ the corresponding alternating sum. We assert that $\chi = \chi_0$. In fact, if we delete an interior side of the decomposition the number α_0 does not change, while α_1 and α_2 are diminished by 1. Hence χ remains constant after performing this process. If we delete a whole interior system of k sides which is common to the boundaries of two polygons of $\mathfrak{T}_1$ then α_0 is diminished by $k-1$, α_1 by k and α_2 by 1. Again χ remains unchanged. In this way we may delete all interior sides of the polygons. If there remain superfluous vertices on the boundary sides, we omit them. Then every time α_0 and α_1 are diminished by 1, while α_2 remains unchanged. Hence $\chi = \chi_0$.

We now consider two decompositions $\mathfrak{T}$ and $\mathfrak{T}^*$. By shifting eventually a little the sides of $\mathfrak{T}$ we may suppose that $\mathfrak{T}$ and $\mathfrak{T}^*$ present only a finite number of common points, without changing the alternating sum of $\mathfrak{T}$. The decompositions $\mathfrak{T}$ and $\mathfrak{T}^*$ together define a polygonal decomposition $\mathfrak{T}'$ which we get by decomposing every polygon of $\mathfrak{T}$. On account of the homeomorphy between the polygons of $\mathfrak{T}$ and plane polygons we may conclude from the above considerations that the number χ for the decompositions $\mathfrak{T}$ and $\mathfrak{T}'$ are the same. Since $\mathfrak{T}'$ may also be considered as a decomposition of $\mathfrak{T}^*$, we see that χ is also the same for $\mathfrak{T}'$ and $\mathfrak{T}^*$. But this proves that χ as calculated from $\mathfrak{T}$ is the same as calculated from $\mathfrak{T}^*$.

The characteristic of a sphere with p handles is easily found. A normal polygon can be decomposed into polygons. If we delete as in the above proof all interior sides this number does not change. Hence the decomposed polygon has the same characteristic as the original polygon. In the case of genus zero we have $\alpha_0 = 2$, $\alpha_1 = 1$, $\alpha_2 = 1$ and in the case of genus $p > 0$ we have $\alpha_0 = 1$, $\alpha_1 = 2p$, $\alpha_2 = 1$. In all cases we have

$$\boxed{\chi = 2p-2.} \qquad (12.5\text{–}5)$$

It follows that two spheres with different numbers of handles have different Euler characteristics and, therefore, they are not homeomorphic.

12.5.5 – Evaluation of the genus of an algebraic function

We shall say that the genus of an algebraic function defined by the polynomial $f(z, w)$ of degree n in w is the genus of the associated Riemann surface. Assume that there are s ramification points of order $r_1-1, \ldots, r_s-1$, (section 12.2.7), respectively. The number

$$m = \sum_{\lambda=1}^{s} (r_\lambda - 1) \tag{12.5–6}$$

is the *ramification number* of the surface. We triangulate in such a way that the branch points become vertices. First we consider a triangulation for the extended plane (or the sphere). Then

$$\chi_0 = -2.$$

Since we have n sheets above the plane we have for the Riemann surface

$$\beta_2 = n\alpha_2, \qquad \beta_1 = n\alpha_1,$$

β_1 the number of sides and β_2 the number of triangles of the triangulation. The number β_0 of vertices is, however, in general less than $n\alpha_0$, for at every branch point r_i points coincide. Hence

$$\beta_0 = n\alpha_0 - (r_1 - 1 + \ldots + r_s - 1) = n\alpha_0 - m.$$

Thus

$$2p-2 = \chi = -\beta_0 + \beta_1 - \beta_2 = n\chi_0 + m = -2n + m,$$

or

$$\boxed{p = \tfrac{1}{2}m - n + 1.} \tag{12.5–7}$$

In the last example of section 12.4.1 we have

$$n = 6,\ s = 2+3+1 = 6,\ r_1 = r_2 = 3,\ r_3 = r_4 = r_5 = 2,\ r_6 = 6.$$

Hence

$$m = 2 \times (3-1) + 3 \times (2-1) + 6 - 1 = 12.$$

It follows that $p = 1$.

It is easy to show that p can have any integral value ≥ 0. Consider the algebraic function defined by

$$w^2 = (z-a_1) \ldots (z-a_{2k}), \tag{12.5–8}$$

where the a_i are different. If some a_i is ∞ the corresponding factor is omitted. Now we have $n = 2$, $s = 2k$, $m = 2k$, hence $p = k-1$.

12.6 – Uniformization

12.6.1 – The concept of uniformization

On a Riemann surface $\mathfrak{F}$, each point has a neighbourhood which is a topological image of a plane disc. This gives us a system of local coordinates in the neighbourhood of the point. In general this local system cannot be extended to a coordinate system over the whole surface, assigning in a one-to-one fashion a number to each point of the surface.

Simple examples may serve to illustrate the idea of uniformization in the large. Consider e.g. the function

$$w = \sqrt[5]{z^3} + \sqrt[3]{z^5}. \tag{12.6–1}$$

If we introduce the variable s such that $z = s^{15}$ we have evidently

$$z = s^{15}, \qquad w = s^{25} + s^9. \tag{12.6–2}$$

Another example is the following

$$w = z^i. \tag{12.6–3}$$

This is uniformized by putting

$$z = e^s, \qquad w = e^{is}. \tag{12.6–4}$$

Finally we consider the algebraic function defined by

$$w^2 + z^2 = 1. \tag{12.6–5}$$

A uniformization is effected by means of

$$z = \frac{1-s^2}{1+s^2}, \qquad w = \frac{2s}{1+s^2}. \tag{12.6–6}$$

The problem of finding a representation of an arbitrary analytic function $w = F(z)$ at all places of the associated Riemann surface $\mathfrak{F}$ by means of two single-valued meromorphic functions

$$z = \varphi(s), \qquad w = \psi(s), \tag{12.6–7}$$

where s runs through a simply connected region $\mathfrak{R}$ in the s-plane, such that to each s in $\mathfrak{R}$ there corresponds one place $\mathfrak{p}$ of $\mathfrak{F}$ in the following fashion: $\varphi(s)$ is the trace of $\mathfrak{p}$ and $\psi(s)$ coincides with the element defining $\mathfrak{p}$, is called *the problem of uniformization.* The correspondence between $\mathfrak{p}$ and s needs not to be one-to-one. Only to a given s there must correspond precisely one place $\mathfrak{p}$.

The fundamental theorem of the theory of uniformization states that *every analytic function can be uniformized.* A first proof of this theorem has been given by H. Poincaré and P. Koebe.

We shall not prove this theorem in its full extend. The subsequent considerations imply, however, a very important special case, viz. the uniformization of algebraic functions. They are based on some elementary results of combinatiorial topology and classical theorems of the theory of functions.

The idea is the following. Instead of the original Riemann surface we construct a simpler surface which can be mapped onto the Riemann surface. This new surface, the universal covering surface, turns out to be a one-to-one image of a region of the s-plane and provides the uniformizing function.

12.6.2 – Covering surfaces

First we give a combinatiorial definition of a surface. A surface is a system of a finite or enumerably infinite number of euclidean triangles whose sides are pasted together according to a certain rule. This pasting consists in identifying corresponding points on two sides which are topologically (e.g., affinely) related. The end points of one side must be identified with the end points of the other side. After pasting, every side must belong to exactly two triangles and the finitely many triangles meeting at a vertex must form a closed cycle in which each two adjacent triangles have one side in common.

We suppose further that the surface is *connected* in the combinatorial sense i.e., one can reach every triangle from another by traversing a finite sequence of adjacent triangles. The surface is called *closed* if the number of triangles is finite. A Riemann surface of an algebraic function is closed. The surface is called *open* if the number of triangles is enumerably infinite. An example is provided by a Riemann surface associated to the logarithmic function.

A very important notion is that of a covering surface of a given surface. Let $\tilde{\mathfrak{F}}$ denote a surface which can be mapped onto another surface $\mathfrak{F}$ such that every triangle of $\tilde{\mathfrak{F}}$ corresponds affinely to precisly one triangle of $\mathfrak{F}$ and that two triangles of $\tilde{\mathfrak{F}}$ meeting at a side correspond to triangles on $\mathfrak{F}$ with the same property. We shall say that $\tilde{\mathfrak{F}}$ is a *covering surface* of $\mathfrak{F}$. It is easy to see that a Riemann surface of an analytic function is a covering surface of the sphere which is itself a surface in the combinatorial sense.

The covering is called *smooth* (or *unramified*) if the cycle of triangles meeting at a poing $\tilde{\mathfrak{p}}$ of $\tilde{\mathfrak{F}}$ corresponds to a cycle of triangles meeting at the corresponding point $\mathfrak{p}$ on $\mathfrak{F}$ and is percorsed once if we percorse the initial cycle. It is not difficult to verify that the mapping of $\tilde{\mathfrak{F}}$ onto $\mathfrak{F}$ is locally topological. That means: the mapping is a homeomorphism in a suitable neighbourhood of an arbitrary point $\tilde{\mathfrak{p}}$ of $\tilde{\mathfrak{F}}$.

12.6.3 – An illustrative example

The following example may serve as an introduction to subsequent developments. Consider the algebraic function defined by

$$w^2 = (1-z^2)(1-k^2z^2), \quad 0 < k < 1. \qquad (12.6\text{–}8)$$

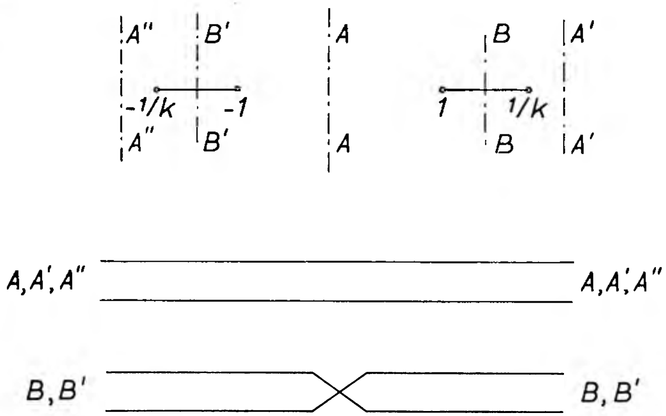


Fig. 12.6–1. The Riemann surface of the function (12.6–8)(schematic)

It has four branch points at $z = \pm 1$, $z = \pm 1/k$ respectively, of order one. Hence, according to the last remark in section 12.5.4 its genus is unity. A Riemann suface may be contructed in the following way: We take two congruent z-planes cut along from $z = +1$ to $+\infty$ and from $z = -1$ to $-\infty$. On the first, denoted by I, the function $w = \sqrt{(1-z^2)(1-k^2z^2)}$ such that $w(0) = 1$, is single-valued; on the second II, the function having the opposite values is also single-valued. We assume that II is placed above I, such that points bearing opposite values have the same projection on the z-plane. Now we past the parts of the borders between $+1$ and $+1/k$ cross-wise together: the same is done for the borders between -1 and $-1/k$. Finally we past the upper and lower borders of the sheet I beyond $1/k$ and $-1/k$ and do the same for sheet II. Hence beyond these points the surface does not penetrate itself, (fig. 12.6–1). Thus we have constructed a two-sheeted Riemann surface of the function (12.6–8).

The topological character of the surface may be visualized by means of the mapping

$$z = \operatorname{sn} s. \tag{12.6–9}$$

The sheet I can be mapped onto a rectangle with the vertices $\pm \mathrm{K} \pm i\mathrm{K}'$ (section 10.2.10). The same function can be employed to map the sheet II onto a rectangle which is obtained from the first one by means of the translation

$$s' = s + 2\mathrm{K}. \tag{12.6–10}$$

The join of these rectangles yields a model of the Riemann surface if

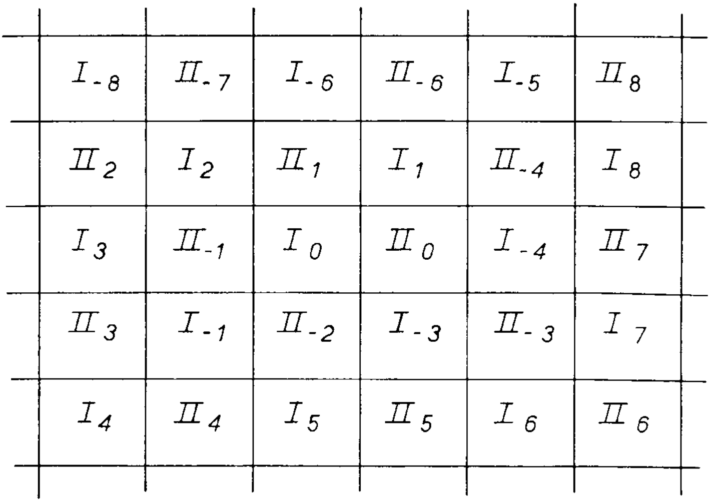


Fig. 12.6–2. A possible enumeration of the images of the sheets of the covering surface of the Riemann surface of (12.6–8)

we identify the opposite sides and thus we see that it is topologically a torus, in accordance with the fact that its genus is unity.

However, we can proceed in another way. We take doubly infinite sequences of sheets

$$I_n, II_n, n = 0, \pm 1, \pm 2, \ldots,$$

each sheet being a cut z-plane as considered above. It is assumed that the sheets are superposed alternately, i.e., I_{n+1} lies above II_n and II_{n+1} above I_{n+1}. By means of (12–6–10) we map these sheets onto rectangles in the s-plane which we enumerate as shown in fig. 12.6–2. The free boundaries of the sheets are pasted together in accordance with the situation in the s-plane. That is to say: if two rectangles have a side in common, the corresponding borders in the z-plane are identified and pasted together. Thus we obtain a Riemann surface of the inverse function of sn s, viz. the integral

$$s = \int_0^z \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}, \qquad (12.6\text{–}11)$$

which covers the surface of (12.6–8). It has an infinity of algebraic branch points of order 3 and a logarithmic branch point at infinity.

To every point of this new surface $\tilde{\mathfrak{F}}$ corresponds a point of $\mathfrak{F}$ with the same projection on the z-plane such that if $\tilde{\mathfrak{p}}$ is in II_n then $\mathfrak{p}$ is in II_0

and if $\tilde{\mathfrak{p}}$ is in I_n then $\mathfrak{p}$ is in I_0. There is no ambiguity as regards the points on the identified borders, for they can be reached from points near any such point along a continuous path. The surface $\tilde{\mathfrak{F}}$ is in one-to-one correspondence with the open s-plane ($s=\infty$ is excluded) and the covering map appears as a map of this plane upon the rectangle consisting of the rectangles I_0, II_0 in such a fashion that points equivalent under the set of translations (n and n' being integers)

$$s' = s+4n\mathrm{K}+2n'i\mathrm{K}' \tag{12.6–12}$$

have the same image. This may also be expressed by saying that the points of the s-plane are reduced modulo (4K, $2i$K′).

The function sn s is single-valued throughout the s-plane and meromorphic. It is easy to see that s is a uniformizing variable, the uniformization of the function (12.6–8) being performed by the functions

$$z = \operatorname{sn} s, \qquad w = \operatorname{sn}' s = \operatorname{cn} s \operatorname{dn} s, \tag{12.6–13}$$

in accordance with (5.14–12).

The surface $\tilde{\mathfrak{F}}$, which is in our case homeomorphic with a punctured sphere, is an example of a universal covering surface. As we shall see the existence of a universal covering surface is the key for the solution of the uniformization problem.

12.6.4 – The universal covering surface

We shall now use the notion of "path" in a rather restricted sense. By a path we understand a sequence of sides

$$a_1, \ldots, a_r$$

of the triangulation, such that a_k and a_{k+1}, $k<r$, have a vertex in common. Two paths a and b connecting the points $\mathfrak{p}$ and $\mathfrak{q}$, are called (combinatorially) *homotopic* if we can transform a into b by applying a finite number of the following steps:

(i) Replacing one side of a triangle by the two others, percorsed in order, and conversely.

(ii) Adding or removing a side which is percorsed in a certain direction and immediately back, (fig. 12.6–3).

All paths homotopic to a path $\mathfrak{pq}$ are the elements of a class $\{\mathfrak{pq}\}$. If to an arbitrary pair of points $\mathfrak{p}$, $\mathfrak{q}$ there corresponds only one class, the surface is called *simply connected* (in the combinatorial sense).

Now we are prepared to construct a *universal covering surface*. Every triangle Δ of the given surface $\mathfrak{F}$ is covered by as many triangles as there are classes of paths which connect a fixed point $\mathfrak{o}$ and a vertex of Δ.

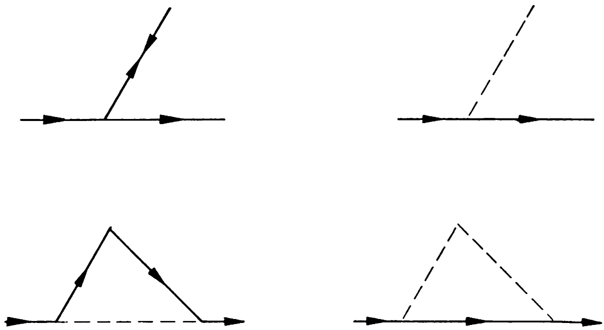


Fig. 12.6–3. Combinatorial deformation of a path

Which of the three vertices is selected is of no importance, for to every path $\mathfrak{op}$ of the class $\{\mathfrak{op}\}$ we can add the side $\mathfrak{pq}$ and so we see that to this class corresponds uniquely the class $\{\mathfrak{oq}\}$. If two triangles of $\mathfrak{F}$ meet along $\mathfrak{pq}$ and if we are given to either of them the same class $\{\mathfrak{op}\}$ or $\{\mathfrak{oq}\}$ then the corresponding triangles are pasted together along this side. It is clear that we obtain a smooth covering surface $\tilde{\mathfrak{F}}$. Every vertex $\tilde{\mathfrak{p}}$ on $\tilde{\mathfrak{F}}$ is defined by a vertex $\mathfrak{p}$ on $\mathfrak{F}$ and a class $\{\mathfrak{op}\}$ on $\mathfrak{F}$. We shall say that $\tilde{\mathfrak{p}}$ is *above* $\mathfrak{p}$ and that $\mathfrak{p}$ is the *trace* of $\tilde{\mathfrak{p}}$.

The surface $\tilde{\mathfrak{F}}$ has the following fundamental property:

The universal covering surface is simply connected.

Given a side $\mathfrak{pq}$ on $\mathfrak{F}$ and to $\mathfrak{p}$ a covering point $\tilde{\mathfrak{p}}$ then by definition precisely one side $\widetilde{\mathfrak{pq}}$ is determined which covers $\mathfrak{pq}$. If we proceed beyond $\mathfrak{q}$ along an arbitrary side then the same is true. Thus, if a path b is given, issuing from $\mathfrak{p}$, then it is covered by a uniquely determind path $\tilde{b}$. Assume that $\tilde{\mathfrak{p}}$ is defined by the path $a = \mathfrak{op}$. Then the end point $\tilde{\mathfrak{r}}$ of $\tilde{b}$ is defined by the path ab which appears if we first percorse a and then b.

Let now $\tilde{c}$ denote a second path from $\tilde{\mathfrak{p}}$ to $\tilde{\mathfrak{r}}$ whose trace on $\mathfrak{F}$ is c. Then, evidently, $\tilde{\mathfrak{r}}$ is also defined by ac. Since ab is homotopic to ac, so is b homotopic to c, for we may multiply on the left by a^{-1}, which is the path a percorsed in the opposite direction. The transformation (i) and (ii) which carry b into c can be effected on $\tilde{\mathfrak{F}}$. Hence also $\tilde{b}$ is homotopic to $\tilde{c}$. Since $\tilde{b}$ and $\tilde{c}$ are arbitrary paths connecting $\tilde{\mathfrak{p}}$ and $\tilde{\mathfrak{q}}$, the proof of the theorem is completed.

12.6.5 – Chains

A path becomes a 1-*chain* if we omit all segments which occurs twice. It is, therefore, a formal sum of segments which may be added modulo 2. Similarly we define a 2-*chain* as a formal sum of a finite number of distinct triangles $\Delta_1, \ldots, \Delta_n$, on $\mathfrak{F}$ (this sum may be empty) which we denote as

$$A = \Delta_1 + \ldots + \Delta_n \tag{12.6–14}$$

Two 2-chains are again added modulo 2, by collecting them to one sum, but omitting all triangles which occur twice.

The *boundary* $\partial\Delta$ of a triangle Δ is the sum of its sides. The boundary ∂A of a 2-chain is the sum modulo 2 of the boundaries of all its triangles. This amounts to: the boundary ∂A consists of those sides which belong to only one of the triangles $\Delta_1, \ldots, \Delta_n$. This is in accordance with the elementary geometric meaning of boundary.

Two 1-chains are called *homologous* (in the combinatorial sense) if one can be transformed into the other by performing the steps (i) and (ii) of the previous section. Another characterization of this relation is the following:

If a is homologous to b then the sum $a+b$ (mod 2) is the boundary of a 2-chain.

In fact, by effecting the transformation (i) the chain a goes over into $a+\partial\Delta$ (mod 2). By (ii) the chain does not undergo any change. After performing a finite number of these steps the path a is replaced by

$$a+\partial\Delta_1+\partial\Delta_2+ \ldots = A+\partial A (\text{mod } 2),$$

where A is the sum $\Delta_1+\Delta_2+ \ldots$ (mod 2). If b is homologous to a then

$$b = a+\partial A \qquad (\text{mod } 2),$$

or

$$a+b = \partial A \qquad (\text{mod } 2).$$

An immediate consequence is

On a simply connected surface every closed 1-chain is homologous to zero.

We shall say that a 1-chain is *homologous to zero* if it bounds a 2-chain. Every closed 1-chain (being a 1-chain such that every vertex is the end point of an even number of sides) can be decomposed into two chains a and b which, considered as paths are homotopic and hence the one can be transformed into the other by the steps (i) and (ii).

Assume now that A and B have the same boundary

$$\partial A = \partial B.$$

Then $\partial(A+B) = 0$, where $A+B$ is the sum modulo 2. Since $\mathfrak{F}$ is connected $A+B$(mod 2) coincides with $\mathfrak{F}$. Since $A+B$ contains only a finite number of different triangles the surface $\mathfrak{F}$ is closed. Conversely, if $\mathfrak{F}$ is closed and A is any 2-chain on it then we may put $\mathfrak{F} = A+B$, where A and B have the same boundary. Thus we see

On an open simply connected surface a closed 1*-chain is the boundary of only one* 2*-chain. On a closed simply connected surface such a* 1*-chain bounds exactly two* 2*-chains which fill the surface entirely.*

12.6.6 – Van der Waerden's lemma

The following lemma due to Van der Waerden simplifies considerably the topological part of the theory of uniformization.

Assume that $\mathfrak{F}$ *is an open connected surface of such a kind that every closed* 1*-chain is homologous to zero. Then we can enumerate its triangles in such a way that every* Δ_{n+1} *has one side or two sides in common with the sum of the preceding triangles*

$$E_n = \Delta_1 + \ldots + \Delta_n,$$

but not a side and the opposite vertex.

The proof is by induction. Assume that $\Delta_1, \ldots, \Delta_n$ are already selected in the prescribed manner. It follows from the fashion in which it is generated that the boundary ∂E_n is a simple closed polygon. Let Δ denote an adjacent triangle. Now three cases are possible:

1) Δ has exactly two sides with E_n in common.
2) Δ has in common with E_n a side, but not the opposite vertex.
3) Δ has in common with E_n a side p and the opposite vertex $\mathfrak{p}$.

Since the surface is open it is not possible that Δ has three sides in common with E_n.

In the first two cases we can take $\Delta_{n+1} = \Delta$. In the third case the vertices $\mathfrak{p}$, $\mathfrak{q}$, $\mathfrak{r}$ of Δ decompose the boundary of E_n into pieces $\mathfrak{q}\mathfrak{r} = p$, $\mathfrak{r}\mathfrak{p} = a$, $\mathfrak{p}\mathfrak{q} = b$. If we denote the sides $\mathfrak{p}\mathfrak{r}$ and $\mathfrak{p}\mathfrak{q}$ of Δ by q and r then $q+a$ and $r+b$ are two closed 1-chains which bound, therefore, two 2-chains A and B, i.e., (fig. 12.6–4).

$$\partial A = q+a, \qquad \partial B = r+b.$$

The sum $A+B$(mod 2) has the same boundary as $E_n+\Delta$. Indeed, $\partial(A+B) = \partial A+\partial B = a+q+b+r = a+b+p+p+q+r = \partial E_n+\partial\Delta = \partial(E_n+\Delta)$. Hence

$$A+B = E_n+\Delta \text{ (mod 2)}. \tag{12.6–15}$$

In the sum on the right occurs the triangle Δ_1, hence either A or B, say B, must contain Δ_1. But then B must also contain Δ_2, for the common side of Δ_1 and Δ_2 does not belong to ∂B. Proceeding in this way we see that B contains all triangles $\Delta_1, \ldots, \Delta_n$ and finally also Δ. On account of (12.6–15) the chain A cannot contain any of these triangles. As a consequence A and $E_n+\Delta$ have no triangles in common and instead of (12.6–15)

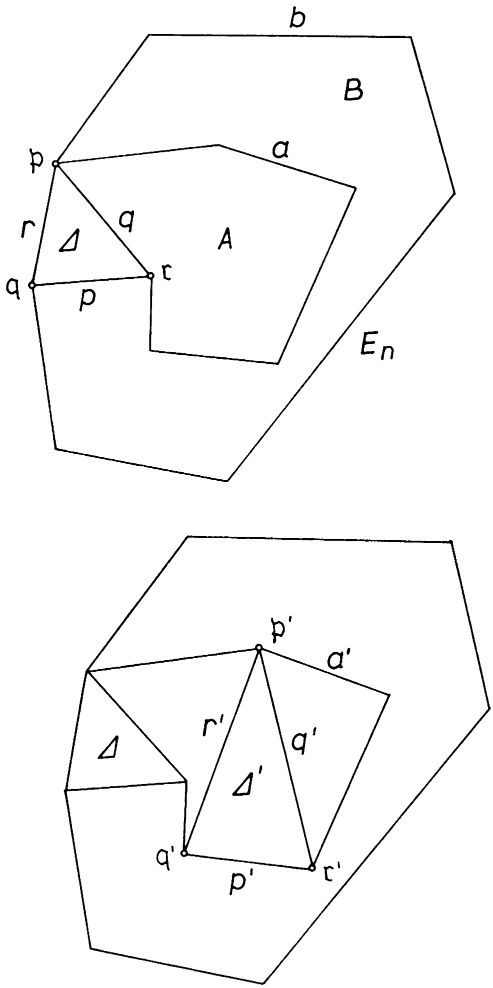


Fig. 12.6–4. Van der Waerden's lemma

we may write

$$B = E_n + \Delta + A. \tag{12.6–16}$$

We shall take Δ_{n+1} from A. Let Δ' denote any triangle adjacent to a. If for Δ' case 1) or case 2) is realized we take $\Delta_{n+1} = \Delta'$. In the case 3) the vertices of Δ' decompose ∂E_n again into three parts p', a' and b'. Since these three vertices belong to a, two of these three parts say p' and a' are entirely contained in a. If we take, similarly, in A a triangle Δ'' adjacent to a', then Δ'' determines in the case 3) a 1-chain a'' which is contained in a', etc. This process must terminate after a finite number of steps, for

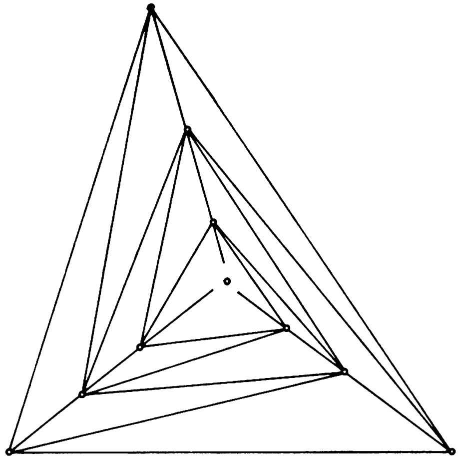


Fig. 12.6–5. Triangulation of a punctured triangle

$a, a', a'', \ldots$ become smaller after each step. Hence for a certain triangle Δ''' case 1) or case 2) occurs. Thus we always find a suitable triangle Δ_{n+1} in A.

If we take Δ_{n+1} from A and add it to E_n then the number of triangles of A has diminished by one. We may repeat the process and take Δ_{n+2} again from A, etc., until A has been exhausted. Then we have

$$E_{n+m} = E_n + \Delta_{n+1} + \ldots + \Delta_{n+m} = E_n + A.$$

Next we may choose $\Delta_{n+m+1} = \Delta$. By means of this method the triangle Δ will get its turn.

Starting from an arbitrary triangle Δ_1, annexing by the described method in turn all adjacent triangles, then all triangles adjacent to these, etc., then every triangle of the surface will get its turn. This concludes the proof.

For a closed surface the same considerations are valid, but now the process terminates, because the last triangle Δ^* is bounded by the last three sides of E_n. Now we puncture the surface by omitting an inner point of Δ^*. Then appears an open surface which may be triangulated by means of infinite many triangles. We take $\Delta_1, \ldots, \Delta_n$ as in the former way and construct in Δ^* a sequence of triangles which shrink into the selected point. The annular parts may be divided into triangles as depicted in fig. 12.6–5.

12.6.7 – General Riemann surfaces

By associating a new structure to a surface as defined in section 12.6.4 we get *a general Riemann surface*. This means that for a neighbourhood of every point we are given a topological map into the plane of a locally uniformizing parameter t which is univalent on $\mathfrak{F}$. It is required that always, when two such neighbourhoods have a region in common, the parameter defined in one neighbourhood depends regularly upon the parameter in the other region throughout the common region.

A function defined throughout a region on the general Riemann surface is said to be holomorphic if it is holomorphic in the locally uniformizing parameters of the points belonging to the region. The mapping as given by such a function is called conformal if it is one-to-one.

It is possible to subdivide the triangles of a general Riemann surface such that every one is included in a neighbourhood in which the locally uniformizing parameter has been defined. These parameters map the triangles as well as the neighbourhoods in which they are contained conformally on regions in a plane. The universal covering surface of a general Riemann surface is again a general Riemann surface. The locally uniformizing parameters of the points of the covering surface $\tilde{\mathfrak{F}}$ are the same as those of their traces on $\mathfrak{F}$.

12.6.8 – Solution of the fundamental uniformization problem

Every single-valued analytic function on the covering surface $\tilde{\mathfrak{F}}$ of a general Riemann surface $\tilde{\mathfrak{F}}$ defines a multiple-valued analytic function on $\mathfrak{F}$ which is unramified. We have solved the fundamental uniformization problem if we succeed in finding such a function which maps $\tilde{\mathfrak{F}}$ conformally onto a region in the s-plane. Since there is a lot of freedom in selecting the region, the variable s is not uniquely defined. We do not destroy the generality of the solution if we take the region as a canonical region, being either the extended s-plane, the open s-plane or the interior of a circle.

We shall suppose that $\mathfrak{F}$ is a Riemann surface of a function having only a finite number of branch points as considered in section 12.5.1. We introduce locally uniformizing parameters t as described in section 12.6.7. These may also be taken as locally uniformizing parameters on $\tilde{\mathfrak{F}}$. If we triangulate $\mathfrak{F}$ as described in section 12.5.1 the triangles on $\mathfrak{F}$ appear in the t-plane as simply connected regions bounded by three analytic arcs.

We consider first the case that $\tilde{\mathfrak{F}}$ is open and we assume that the triangles $\Delta_1, \Delta_2, \ldots,$ of its triangulation are ordered in accordance with Van der Waerden's lemma.

Let Δ_1 be included in a neighbourhood $\mathfrak{U}$ in which the local parameter

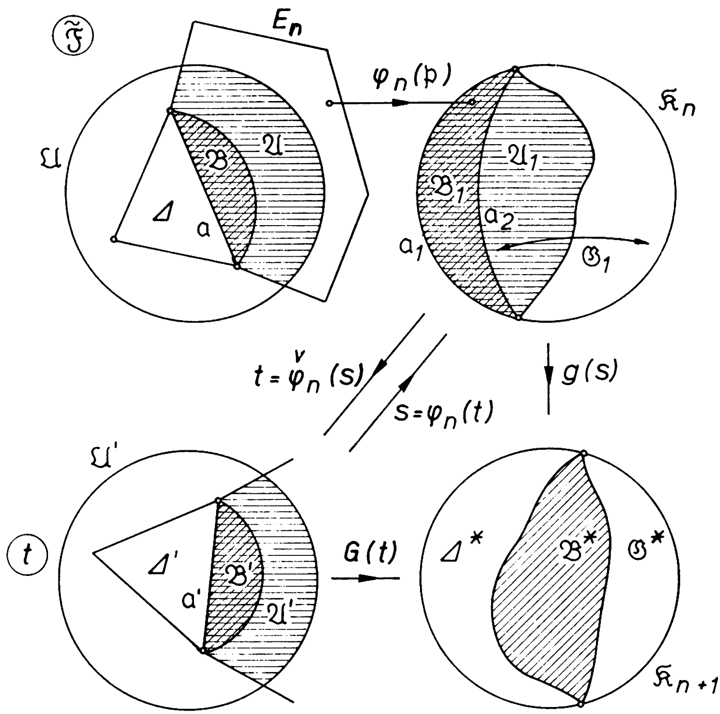


Fig. 12.6–6. The mapping of E_n in a neighbourhood of Δ

t is defined. This triangle corresponds to a simply connected region Δ' in the t-plane, bounded by analytic arcs. Hence Δ' can be mapped onto the interior of a circle in the s-plane such that there is a one-to-one correspondence between the boundaries (section 10.5.6).

We proceed by induction and assume that $E_n = \Delta_1 + \ldots + \Delta_n$ (whose boundary is stripped) is mapped onto an open disc $\mathfrak{K}_n$ in the s-plane, such that there is a one-to-one correspondence between the boundaries. Let $s = \varphi_n(\mathfrak{p})$ denote the mapping function. In order to obtain a map of E_{n+1} we employ a device due to Carathéodory. The region E_{n+1} consists of E_n and a triangle $\Delta = \Delta_{n+1}$ which has one or two sides with E_n in common, (fig. 12.6–6). Since Δ is in the interior of a neighbourhood $\mathfrak{U}$ which corresponds one-to-one to a neighbourhood $\mathfrak{U}'$ in the t-plane, the triangle Δ is mapped onto a region Δ' in $\mathfrak{U}'$. A common side $\mathfrak{a}$ of Δ and E_n is continuously and in a one-to-one manner related to a circular arc $\mathfrak{a}_1$, on the boundary of $\mathfrak{K}_n$ and by means of the local parameter to an arc $\mathfrak{a}'$ of Δ'. A part $\mathfrak{A}$ of E_n with $\mathfrak{a}$ on its boundary is in $\mathfrak{U}$. The image

$\mathfrak{A}_1$ of $\mathfrak{A}$ in the s-plane is a part of the circle $\mathfrak{K}_n$, with $\mathfrak{a}_1$ on its boundary. In the t-plane the image of $\mathfrak{A}$ is a part $\mathfrak{A}'$ of $\mathfrak{U}'$. Hence $\mathfrak{A}'$ and $\mathfrak{A}_1$ are related by a function which we shall denote by $s = \varphi_n(t)$, being the mapping $\varphi_n(\mathfrak{p})$ restricted to $\mathfrak{A}$, where $\mathfrak{p}$ is replaced by its image in the t-plane. Now we erect on $\mathfrak{a}_1$ in $\mathfrak{K}_n$ a circular two-gon $\mathfrak{B}_1$ with second boundary arc $\mathfrak{a}_2$ with such a small angle $\pi/2^r$ that $\mathfrak{B}_1$ is still in $\mathfrak{A}_1$. Then the image $\mathfrak{B}$ of $\mathfrak{B}_1$ on $\tilde{\mathfrak{F}}$ is in $\mathfrak{U}$ and the image $\mathfrak{B}'$ of $\mathfrak{B}_1$ in the t-plane is in $\mathfrak{U}'$. This mapping is given by $\varphi_n(t)$. Now we shall prove that there are holomorphic functions

$$s^* = g(s), \qquad s^* = G(t) \tag{12.6–17}$$

such that the first maps the circle $\mathfrak{K}_n$ onto a region $\mathfrak{B}^*+\mathfrak{G}^*$ and the second $\Delta+\mathfrak{B}'$ onto a region $\Delta^*+\mathfrak{B}^*$ such that

1) $\Delta^*+\mathfrak{B}^*+\mathfrak{G}^*$ constitutes an open disc $\mathfrak{K}_{n+1}$;
2) $s^* = g(s)$ maps $\mathfrak{B}_1$, onto $\mathfrak{B}^*$ and $\mathfrak{G}_1 = \mathfrak{K}_n-\mathfrak{B}_1$ onto $\mathfrak{G}^*$;
3) $s^* = G(t)$ maps $\mathfrak{B}'$ onto $\mathfrak{B}^*$ and Δ' onto Δ^*;
4) in $\mathfrak{B}'$ holds the relation

$$G(t) = g(\varphi_n(t)). \tag{12.6–18}$$

This construction is visualized in the schematic figure (12.6–7).

First we map by means of $t_1 = A_1(t)$ the simply connected region $\Delta'+\mathfrak{B}'$ onto a circular disc $\Delta''+\mathfrak{B}_1''$, such that the centre of the disc is in the interior of Δ''. Then the mapping as given by

$$t_1 = A_1(\check{\varphi}_n(s)) = g_1(s) \tag{12.6–19}$$

relates $\mathfrak{B}_1$ conformally to $\mathfrak{B}_1''$. This mapping may be extended by employing the symmetry-principle of section 10.1.4 to $\mathfrak{B}_2$, which is the reflected image of $\mathfrak{B}_1$ with respect to $\mathfrak{a}_2$. The image of $\mathfrak{B}_2$ in the t_1-plane is $\mathfrak{B}_2''$, the mirror-image of $\mathfrak{B}_1''$ with respect to $\mathfrak{a}''$, the image of $\mathfrak{a}$. The region $\Delta''+\mathfrak{B}_1''+\mathfrak{B}_2''$ is simply connected and can be mapped by means of $t_2 = A_2(t_1)$ onto a circular disc $\Delta'''+\mathfrak{B}_1'''+\mathfrak{B}_2'''$ such that the centre of the disc in Δ'''. Then the mapping

$$t_2 = A_2(g_1(s)) \tag{12.6–20}$$

relates the two-gon $\mathfrak{B}_1+\mathfrak{B}_2$ with angle $\pi/2^{r-1}$ and the sides $\mathfrak{a}_1$ and $\mathfrak{a}_3$ to $\mathfrak{B}_1'''+\mathfrak{B}_2'''$. Applying again the symmetry principle the mapping can be extended to $\mathfrak{B}_3$, the mirror-image of $\mathfrak{B}_1+\mathfrak{B}_2$ with respect to $\mathfrak{a}_3$ and the image is $\mathfrak{B}_3'''$, the mirror-image of $\mathfrak{B}_1'''+\mathfrak{B}_2'''$ with respect to $\mathfrak{a}'''$. The region $\Delta'''+\mathfrak{B}_1'''+\mathfrak{B}_2'''+\mathfrak{B}_3'''$ is simply connected and may be mapped onto a circular disc. We proceed in the described way and after r steps the disc $\mathfrak{K}_n$ is wholly covered by two-gons. As a consequence we have two mappings

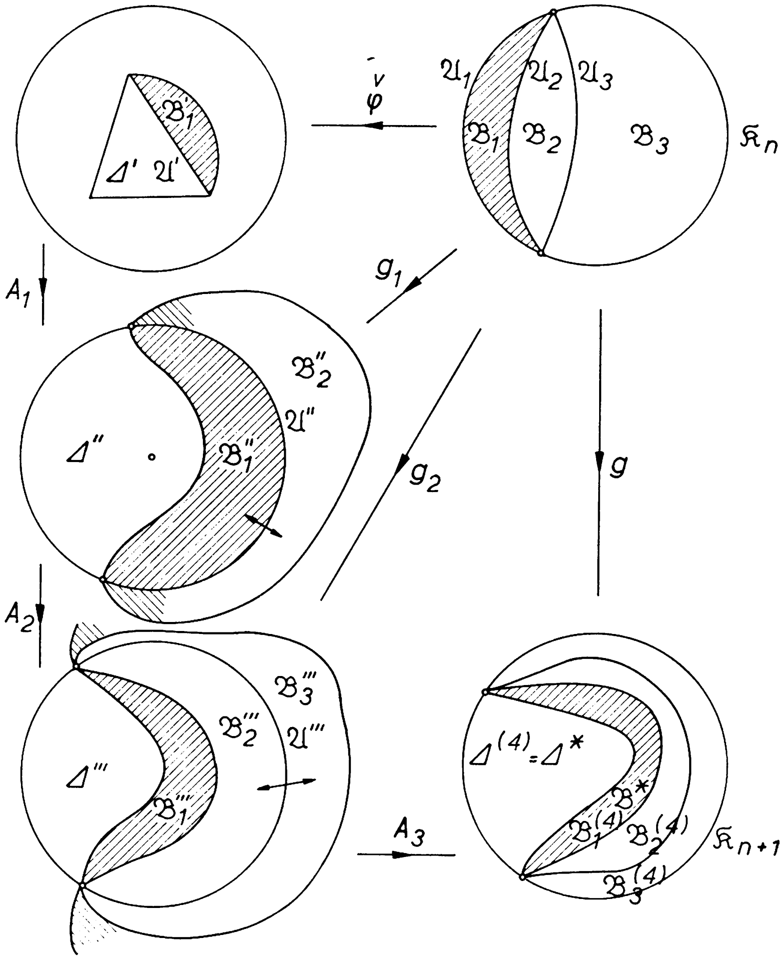


Fig. 12.6–7. The construction of the functions (12.6–17)

$$\begin{aligned} s^* &= A_{r+1} A_r \ldots A_2 A_1(t) = G(t), \\ s^* &= A_{r+1} A_r \ldots A_2 A_1 \check{\varphi}_n(s) = g(s), \end{aligned} \tag{12.6–21}$$

with the properties:

$G(t)$ maps Δ' and $\mathfrak{B}'$ onto Δ^* and $\mathfrak{B}^* = \mathfrak{B}_1^{(r+2)}$; $g(s)$ maps $\mathfrak{B}_1$ and $\mathfrak{G}_1 = \mathfrak{B}_2 + \mathfrak{B}_3 + \ldots + \mathfrak{B}_{r+1}$ onto $\mathfrak{B}^* = \mathfrak{B}_1^{(r+2)}$ and $\mathfrak{G}^* = \mathfrak{B}_2^{(r+2)} + \mathfrak{B}_3^{(r+2)} + \ldots + \mathfrak{B}_{r+1}^{(r+2)}$; $\Delta^* + \mathfrak{B}^* + \mathfrak{G}^*$ is an open disc. Thus the properties 3), 2) and 1) listed above are verified.

From (12.6–21) follows

$$g(\varphi_n(t)) = A_{r+1} A_r \ldots A_2 A_1 \check{\varphi}_n(\varphi_n(t)) = A_{r+1} A_r \ldots A_1(t) = G(t)$$

in $\mathfrak{B}'$. This is the property 4).

The function

$$s^* = g(s) = g(\varphi_n(\mathfrak{p})) = g^*(\mathfrak{p}) \tag{12.6–22}$$

maps E_n onto $\mathfrak{B}^*+\mathfrak{G}^*$ and the function

$$s^* = G(t) = g^{**}(\mathfrak{p}) \tag{12.6–23}$$

maps $\Delta+\mathfrak{B}$ onto $\Delta^*+\mathfrak{B}^*$. If $\mathfrak{p}$ is in $\mathfrak{B}$ then $g^*(\mathfrak{p})$ and $g^{**}(\mathfrak{p})$ are defined and we have

$$g^*(\mathfrak{p}) = g(s) = g(\varphi_n(t)) = G(t) = g^{**}(\mathfrak{p}).$$

Now $g^*(\mathfrak{p})$ is defined throughout E_n, $g^{**}(\mathfrak{p})$ throughout $\Delta+\mathfrak{B}$. Hence the one is an analytic continuation of the other and they combine to a function

$$s^* = \varphi_{n+1}(\mathfrak{p})$$

which maps $E_{n+1} = E_n+\Delta$ onto the disc $\mathfrak{K}_{n+1}$. Thus we proved

The interior of every of the infinitely many of the parts $E_n = \Delta_1+\ldots+\Delta_n$ of an open covering surface $\tilde{\mathfrak{F}}$ can be mapped one-to-one and conformally onto an open disc $\mathfrak{K}_n$. The boundary of E_n corresponds one-to-one and continuously to the circumference of $\mathfrak{K}_n$.

The mapping functions of $E_1, E_2, \ldots$ were denoted by $\varphi_1(\mathfrak{p}), \varphi_2(\mathfrak{p}), \ldots$, respectively. If we take a point $\mathfrak{o}$ in the interior of Δ_1 we can normalize them by

$$\varphi_n(\mathfrak{o}) = 0, \qquad \varphi_n'(\mathfrak{o}) = 1, \tag{12.6–24}$$

where the prime denotes differentiation with respect to the locally uniformizing parameter.

Now we form the functions

$$\varphi_{1,n}(s) = \varphi_n(\check{\varphi}_1(s)), \qquad n = 1, 2, \ldots. \tag{12.6–25}$$

They are holomorphic in $\mathfrak{K}_1$, univalent and normalized at $s = 0$. By virtue of a theorem of section 11.2.6 (consequence of (11.2–28)) they constitute a normal family. Hence we may extract from (12.6–25) a subsequence which converges in the interior of $\mathfrak{K}_1$ to a univalent function. The same may be asserted for the sequence

$$\varphi_1(\mathfrak{p}), \varphi_2(\mathfrak{p}), \ldots, \tag{12.2–26}$$

containing a subsequence

$$\varphi_1^1(\mathfrak{p}), \varphi_2^1(\mathfrak{p}), \ldots \tag{12.6–27}$$

which converges in the interior of E_1 to a univalent holomorphic function $\varphi_0(\mathfrak{p})$. Since we may suppose that $\mathfrak{K}_2$ is again in the s-plane we can construct the functions

$$\varphi_{2,n}(s) = \varphi_n(\check{\varphi}_2(s)), \qquad n = 1, 2, \ldots \tag{12.6–28}$$

As above we may extract a convergent subsequence which yields a sequence

$$\varphi_1^2(\mathfrak{p}), \varphi_2^2(\mathfrak{p}), \ldots, \tag{12.6–29}$$

converging throughout E_2 to a function whose restriction with respect to E_1 is φ_0. Accordingly we denote it also by φ_0.

By repeating the process we obtain by applying the diagonal principle the sequence

$$\varphi_1^1(\mathfrak{p}), \varphi_2^2(\mathfrak{p}), \ldots \tag{12.6–30}$$

where $\varphi_k^k(\mathfrak{p})$ is defined throughout E_n if $k \geqq n$ and converges to φ_0 there. Since the E_n exhaust the surface $\tilde{\mathfrak{F}}$ we see that $\varphi_0(\mathfrak{p})$ is univalent on $\tilde{\mathfrak{F}}$ and maps $\tilde{\mathfrak{F}}$ onto a region $\mathfrak{R}$ in the s-plane. It is easy to see that $\mathfrak{R}$ must be simply connected and is, therefore, conformally equivalent to a normal region. However, the extended plane must be excluded, since $\tilde{\mathfrak{F}}$ is not a closed surface.

It remains to consider the case that $\tilde{\mathfrak{F}}$ is closed. Then $\tilde{\mathfrak{F}}$ consists of E_n and a closing triangle Δ which has three sides in common with E_n. We puncture Δ by an interior point $\mathfrak{q}$. The remaining part of $\tilde{\mathfrak{F}}$ is a simply connected open surface $\tilde{\mathfrak{F}}_0$ which can be mapped onto the finite plane or onto an open disc. We wish to show that the latter case cannot occur.

Suppose that $\tilde{\mathfrak{F}}$ is composed of $E+\Delta$, where E contains the point $\mathfrak{o}$ and Δ the point $\mathfrak{q}$. Let E' be the image of E and Δ' that of Δ minus $\mathfrak{q}$; we assume that $s = 0$ corresponds to $\mathfrak{o}$. By $\mathfrak{K}'$ we denote a disc about $s = 0$ which is included in E'. Then Δ' is outside this disc. The function $w = 1/s$ maps Δ' on a bounded region of the w-plane. It follows from Riemann's theorem (section 2.8.3) that the singularity of a function regular near $\mathfrak{q}$ is removable, for the function is bounded. Let q'' denote the image of $\mathfrak{q}$ in the w-plane. Suppose now that the image of the punctured surface were the interior of a disc $\mathfrak{K}$ in the s-plane, hence the exterior of a disc $\mathfrak{K}''$ in the w-plane. A sequence of points $p_1'', p_2'', \ldots$ outside $\mathfrak{K}''$ and having an accumulation point on the circumference corresponds to a sequence $\mathfrak{p}_1, \mathfrak{p}_2, \ldots$ on $\tilde{\mathfrak{F}}$ and it has also an accumulation point, since $\tilde{\mathfrak{F}}$ is closed. This can only be the point $\mathfrak{q}$, for any other accumulation point on $\tilde{\mathfrak{F}}$ would lead to an accumulation point in the interior of $\mathfrak{K}''$, because the mapping is continuous. Hence $\mathfrak{q} = \lim_{n\to\infty} \mathfrak{p}_n$ and, as a consequence, $q'' = \lim_{n\to\infty} p_n''$. The accumulation point of the sequence $p_1'', p_2'', \ldots$ can be selected arbitrarily on the circumference of $\mathfrak{K}''$. Hence to $\mathfrak{q}$ must correspond infinitely many points q'' and this is impossible. There is no contradiction if the radius of $\mathfrak{K}''$ is zero and q''

the origin. Returning to the s-plane we find that the radius of $\mathfrak{K}$ is infinite and thus $\tilde{\mathfrak{F}}$ corresponds to the extended plane. This concludes the proof of the fundamental uniformization problem.

Consider now a triangulated Riemann surface $\mathfrak{F}$ of an analytic function $w = F(z)$ covering the z-plane. This function is a collection of function elements $w = f(z)$ such that w and z may be interpreted as meromorphic functions

$$w = w(\mathfrak{p}), \quad z = z(\mathfrak{p})$$

on the surface, meromorphic in terms of a locally uniformizing parameter.

Next we consider the universal covering surface $\tilde{\mathfrak{F}}$ and the projection mapping

$$\mathfrak{p} = \pi(\tilde{\mathfrak{p}}).$$

This mapping is conformal. Finally we can find a variable s in a canonical region $\mathfrak{N}$ such that $\tilde{y} = \check{\varphi}_0(s)$ is a conformal mapping of $\mathfrak{N}$ onto $\tilde{\mathfrak{F}}$. Hence we can find two functions

$$\varphi(s) = z(\pi(\check{\varphi}_0(s))), \qquad \psi(s) = w(\pi(\check{\varphi}_0(s)))$$

which are related by $w = F(z)$. Since an algebraic Riemannian surface is triangulable, we see that *every algebraic function can be uniformized.*

12.7 – Deformation of paths

12.7.1 – The parametrized path

The last theorem of section 12.1.6 states that the result of analytic continuation along a curve is not affected by a small deformation of this path. The question rises what happens when an arbitrary deformation is performed. In order to give a satisfactory answer to this question, it is necessary to formulate a precise definition of the conception of deformation, notwithstanding the fact that this conception is intuitively extremely simple. But it should be noticed that a continuous map of a line segment does not always resemble the naive impression of a curve. Indeed, it is possible, for instance, to map a segment continuously onto a square (curve of Peano-Hilbert) and thus caution is needed. It is, however, a pleasant experience that assertions about deformation of curves which are intuitively quite clear, turn out to be true if they are based on logical reasoning. It is our intention to give an idea how this may be done.

First we focus our attention on the notion of path. Consider the set of continuous mappings of closed intervals $a \leqq t \leqq b$ into an open set $\mathfrak{A}$ of the z-plane. We shall say that a continuous mapping γ_1 of the interval $a_1 \leqq t \leqq b_1$ into $\mathfrak{A}$ is equivalent to a continuous mapping γ_2 of the interval $a_2 \leqq t \leqq b_2$ into $\mathfrak{A}$ if

$$\gamma_1(t) = \gamma_2\left(a_2 + \frac{t-a_1}{b_1-a_1}(b_2-a_2)\right), \qquad a_1 \leqq t \leqq b_1. \tag{12.7–1}$$

It is easily verified that a relation of this kind has the usual properties of an equivalence relation. A *parametrized path*, or briefly a *path*, in $\mathfrak{A}$ is defined to be an equivalence class of continuous mappings of closed intervals into $\mathfrak{A}$. That is to say, we shall not distinguish between equivalent mappings.

Given a mapping of $a \leqq t \leqq b$ into $\mathfrak{A}$ we can always find an equivalent mapping of the unit interval $0 \leqq t \leqq 1$ into $\mathfrak{A}$. Unless otherwise stated we shall take as a representative mapping for each path a continuous mapping of the closed unit interval. It is clear that equivalent mappings of this interval are identical. Hence a path is wholly characterized by a function $\gamma(t)$ defined on $0 \leqq t \leqq 1$ with values in $\mathfrak{A}$. Therefore, the path may be denoted by the symbol γ. The set of all points $z = \gamma(t)$ is called the *carrier* of the path. Different paths may have the same carrier. In fact, let $\mu(t)$ denote a non-decreasing continuous function with $\mu(0) = 0$, $\mu(1) = 1$. In general the points $\gamma(t)$ and $\gamma(\mu(t))$ are different, but the whole collection of image points is in both cases the same subset of $\mathfrak{A}$.

The point $\gamma(0)$ is called the *initial point* of the path γ and $\gamma(1)$ the *end point*. A path is said to be *closed*, or a *loop*, if the initial point and the end point coincide, i.e., if $\gamma(0) = \gamma(1)$. A path is said to *join* the points z_0 and z_1 if $\gamma(0) = z_0$ and $\gamma(1) = z_1$.

It is an important fact that a certain algebraic combination can be carried out for paths, viz., the *multiplication*. Consider two paths γ_1 and γ_2, such that the end point of γ_1 coincides with the initial point of γ_2, i.e., $\gamma_1(1) = \gamma_2(0)$. Now we define a path γ by means of the mapping

$$\gamma(t) = \begin{cases} \gamma_1(2t), & 0 \leqq t \leqq \frac{1}{2}, \\ \gamma_2(2t-1), & \frac{1}{2} \leqq t \leqq 1, \end{cases} \tag{12.7–2}$$

where γ_1 and γ_2, originally defined as functions on the interval $0 \leqq t \leqq 1$, are replaced by equivalent mappings of suitably chosen intervals. The mapping (12.7–2) defines a path γ which is called the *product* of the paths γ_1 and γ_2 and written as

$$\gamma = \gamma_2\gamma_1. \tag{12.7–3}$$

Notice that in (12.7–3) the product does not mean the product of two functions in the usual sense. However, there will be no fear for confusion. It is clear that a product is not always defined and, if so, the order of the factors is essential.

If $\gamma(t)$ defines the path γ, then the path defined by $\gamma(1-t)$ is called the *inverse* of γ and denoted by γ^{-1}. It is clear that

$$(\gamma^{-1})^{-1} = \gamma. \tag{12.7–4}$$

In addition we have

$$(\gamma_2\gamma_1)^{-1} = \gamma_1^{-1}\gamma_2^{-1}. \tag{12.7–5}$$

In fact, it follows from (12.7–2) that

$$\gamma(1-t) = \begin{cases} \gamma_2(1-2t), & 0 \leqq t \leqq \frac{1}{2}, \\ \gamma_1(2-2t), & \frac{1}{2} \leqq t \leqq 1, \end{cases}$$

which defines $\gamma_1^{-1}\gamma_2^{-1}$.

A *constant path* is defined by a constant function $\gamma(t)$, i.e., $\gamma(t) = z_0$, $0 \leqq t \leqq 1$; its carrier consists of a single point and, evidently, the path is equal to its inverse.

12.7.2 – Homotopy

Consider two paths γ_1 and γ_2 in an open set $\mathfrak{A}$ joining the points z_0 and z_1. We shall say that γ_1 can be deformed continuously into γ_2 if each point $\gamma_1(t)$ of the first path (t being fixed) can move along a path $\varphi_t(u)$, $0 \leqq u \leqq 1$ to $\gamma_2(t)$ inside $\mathfrak{A}$. The function $\varphi_t(u) = \varphi(t, u)$ will be supposed to be a continuous function of both variables t and u. The function $\varphi(t, u)$ is such that

$$\begin{aligned} &\varphi(t, 0) = \gamma_1(t), \quad &&\varphi(t, 1) = \gamma_2(t), \quad &&0 \leqq t \leqq 1, \\ &\varphi(0, u) = z_0, &&\varphi(1, u) = z_1, &&0 \leqq u \leqq 1. \end{aligned} \tag{12.7–6}$$

The initial point and the end point remain fixed during the process of deformation.

Geometrically the meaning of the deformation of γ_1 into γ_2 is that there is a continuous function $\varphi(t, u)$ which maps the unit square $\mathfrak{E}$: $0 \leqq t \leqq 1$, $0 \leqq u \leqq 1$ into the set $\mathfrak{A}$ in such a way that the lower side of the square is mapped as the path γ_1 and the upper side as the path γ_2; the vertical sides are mapped onto z_0 and z_1 respectively, (fig. 12.7–1). The function $\varphi(t, u)$ will be called the *deformation function*, briefly the

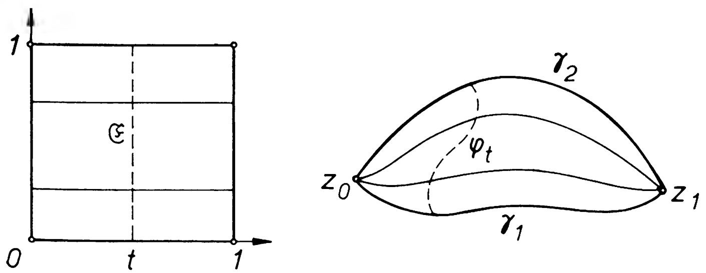


Fig. 12.7–1. Deformation of a path

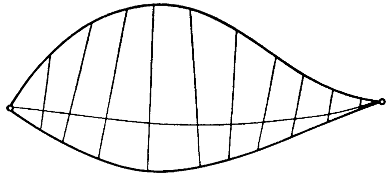


Fig. 12.7–2. Linear deformation of a path

deformation; the path φ_t joining $\gamma_1(t)$ to $\gamma_2(t)$ is a deformation path and the square $\mathfrak{E}$ the *deformation square.* The image of a horizontal segment connecting two points on the vertical sides of $\mathfrak{E}$ is an *intermediate path*; the image of a vertical segment joining two points on the horizontal sides is a deformation path.

A particular, but nevertheless very important instance of a deformation is the *linear deformation* which can be defined if every point $\gamma_1(t)$ can be connected with $\gamma_2(t)$ by a rectilinear segment included in $\mathfrak{A}$ (fig. 12.7–2). Then a deformation is given by

$$\varphi(t, u) = (1-u)\gamma_1(t)+u\gamma_2(t) \tag{12.7–7}$$

and the fact that this function is linear in u explains the name.

If the path γ_1 can be deformed into the path γ_2 we shall say that γ_1 and γ_2 are *homotopic.* We write this as

$$\gamma_1 \approx \gamma_2 . \tag{12.7–8}$$

The homotopy relation between parametrized paths is an equivalence relation.

1) First we have

$$\gamma \approx \gamma. \tag{12.7–9}$$

This is clear from the identity deformation

$$\varphi(t, u) = \gamma(t), \qquad 0 \leqq t \leqq 1, 0 \leqq u \leqq 1,$$

which arises from (12.7–7) if we set $\gamma_1(t) = \gamma_2(t) = \gamma(t)$. In this case the deformation paths are single points.

2) Secondly

$$\gamma_1 \approx \gamma_2 \quad \text{implies} \quad \gamma_2 \approx \gamma_1 . \tag{12.7–10}$$

Indeed, if $\varphi(t, u)$ is the deformation of γ_1 into γ_2, then $\varphi(t, 1-u)$ is the deformation of γ_2 into γ_1. The deformation paths of the second deformation are the inverses of the deformation paths of the first deformation.

3) Finally

$$\gamma_1 \approx \gamma_2, \gamma_2 \approx \gamma_3 \quad \text{implies} \quad \gamma_1 \approx \gamma_3. \tag{12.7–11}$$

Let $\varphi_1(t, u)$ denote the deformation of γ_1 into γ_2 and $\varphi_2(t, u)$ that of γ_2 into γ_3. Then

$$\varphi(t, u) = \begin{cases} \varphi_1(t, 2u), & 0 \leqq u \leqq \frac{1}{2}, \\ \varphi_2(t, 2u-1), & \frac{1}{2} \leqq u \leqq 1 \end{cases}$$

is a deformation of γ_1 into γ_3, for $\varphi(t, 0) = \varphi_1(t, 0) = \gamma_1(t)$, $\varphi(t, 1) = \varphi_2(t_1, 1) = \gamma_3(t)$. The deformation paths of the resulting deformations are the products of the deformation paths of the first and the second deformation.

An intermediate path defined by $\varphi(t, u_0)$, where u_0 is a fixed number, $0 \leqq u_0 \leqq 1$, is homotopic to the original path.

In fact, the path γ defined by $\varphi(t, u_0)$ is homotopic to γ_1 as follows from the deformation $\psi(t, u) = \varphi(t, uu_0)$.

A *reparametrization of a path* γ is another path, obtained if we replace the parameter t by a non-decreasing continuous function $\mu(t)$, $0 \leqq t \leqq 1$, with $\mu(0) = 0$, $\mu(1) = 1$. The reparametrized path is defined by the function $\gamma(\mu(t))$.

A reparametrized path is homotopic to the original path.

In fact, the deformation function may be taken as

$$\varphi(t, u) = \gamma((1-u)t + u\mu(t)). \tag{12.7–12}$$

It is clear the new path is identical with the original path if and only if $\mu(t) = t$.

12.7.3 – Theorems about homotopy

Because of the parametrization of the paths the multiplication as defined in section 12.7.1 is not associative. If γ_1, γ_2 and γ_3 are given paths such that the product $\gamma_3(\gamma_2\gamma_1)$ is defined, then also the product $(\gamma_3\gamma_2)\gamma_1$ has a meaning. But for the constructions of $\gamma_2\gamma_1$ we map the intervals $0 \leqq t \leqq \frac{1}{2}$, $\frac{1}{2} \leqq t \leqq 1$ and for the construction of $\gamma_3(\gamma_2\gamma_1)$ we map the intervals $0 \leqq t \leqq \frac{1}{4}$, $\frac{1}{4} \leqq t \leqq \frac{1}{2}$, $\frac{1}{2} \leqq t \leqq 1$. On the other hand, for the construction of $(\gamma_3\gamma_2)\gamma_1$ we map successively the intervals $0 \leqq t \leqq \frac{1}{2}$, $\frac{1}{2} \leqq t \leqq \frac{3}{4}$ and $\frac{3}{4} \leqq t \leqq 1$. However, since they differ only in parametrization (fig. 12.7–3), they are homotopic, i.e.,

$$\gamma_3(\gamma_2\gamma_1) \approx (\gamma_3\gamma_2)\gamma_1. \tag{12.7–13}$$

This may be proved by writing down explicitly the function which performs the reparametrization. The path on the left is defined by

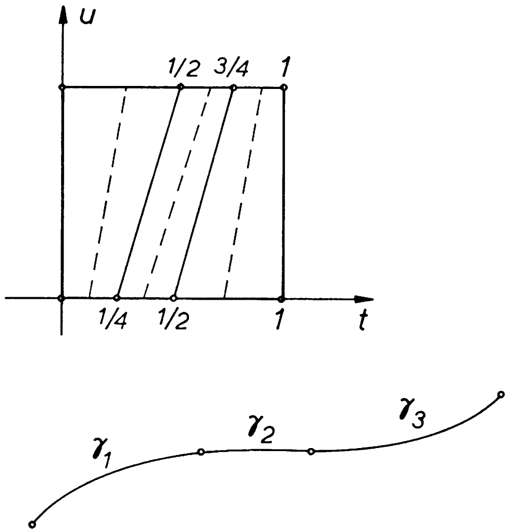


Fig. 12.7–3. The associativity of the product of paths

$$\gamma(t) = \begin{cases} \gamma_1(4t), & 0 \leqq t \leqq \frac{1}{4}, \\ \gamma_2(4t-1), & \frac{1}{4} \leqq t \leqq \frac{1}{2}, \\ \gamma_3(2t-1), & \frac{1}{2} \leqq t \leqq 1. \end{cases}$$

Now we introduce the function, (fig. 12.7–4),

$$\mu(t) = \begin{cases} \frac{1}{2}t, & 0 \leqq t \leqq \frac{1}{2}, \\ t-\frac{1}{4}, & \frac{1}{2} \leqq t \leqq \frac{3}{4}, \\ 2t-1, & \frac{3}{4} \leqq t \leqq 1. \end{cases}$$

Then

$$\gamma(\mu(t)) = \begin{cases} \gamma_1(2t), & 0 \leqq t \leqq \frac{1}{2}, \\ \gamma_2(4t-2), & \frac{1}{2} \leqq t \leqq \frac{3}{4}, \\ \gamma_3(4t-3), & \frac{3}{4} \leqq t \leqq 1, \end{cases}$$

and this function defines the path on the right of (12.7–13).

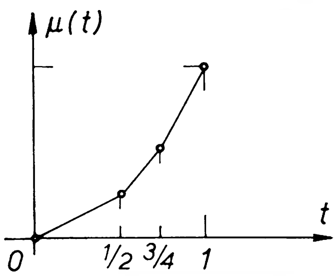


Fig. 12.7–4. The function $\mu(t)$ used for the proof of (12.7–13)

Suppose we are given the paths α_1, α_2 and β_1, β_2 such that the products $\alpha_2\alpha_1$ and $\beta_2\beta_1$ have a meaning. Then we may state

$$\alpha_1 \approx \beta_1, \alpha_2 \approx \beta_2 \quad \text{implies} \quad \alpha_2\alpha_1 \approx \beta_2\beta_1. \tag{12.7–14}$$

Let $\varphi_1(t, u)$ denote the deformation of α_1 into α_2 and $\varphi_2(t, u)$ that of β_1 into β_2. Then

$$\varphi(t, u) = \begin{cases} \varphi_1(2t, u), & 0 \leqq t \leqq \frac{1}{2}, \\ \varphi_2(2t-1, u), & \frac{1}{2} \leqq t \leqq 1 \end{cases}$$

is the deformation of $\alpha_2\alpha_1$, given by $\varphi(t, 0)$, into $\beta_2\beta_1$, given by $\varphi(t, 1)$.

Next we have

$$\gamma_1 \approx \gamma_2 \quad \text{implies} \quad \gamma_1^{-1} \approx \gamma_2^{-1}. \tag{12.7–15}$$

Indeed, if $\varphi(t, u)$ defines the deformation of γ_1 into γ_2 then $\varphi(1-t, u)$ is the deformation of γ_1^{-1} into γ_2^{-1}.

Let 1_{z_0} denote the identity path $\gamma(t) = z_0$, $0 \leqq t \leqq 1$. If γ is a path issuing from z_0 we have

$$\gamma^{-1}\gamma \approx 1_{z_0} \tag{12.7–16}$$

i.e., a path percorsed successively into two directions can be shrunk into its initial point. The deformation is performed by the function

$$\varphi(t, u) = \begin{cases} \gamma(2t(1-u)), & 0 \leqq t \leqq \frac{1}{2}, \\ \gamma(2(1-t)(1-u)), & \frac{1}{2} \leqq t \leqq 1, \end{cases}$$

for $\varphi(t, 0)$ is actually the product $\gamma^{-1}\gamma$ and $\varphi(t, 1) = \gamma(0) = z_0$, for $0 \leqq t \leqq 1$.

In quite the same way we may prove that

$$\gamma\gamma^{-1} \approx 1_{z_1}, \tag{12.7–17}$$

if the carrier of 1_{z_1} is now the end point of γ. Indeed the deformation function is now

$$\varphi(t, u) = \begin{cases} \gamma((1-2t)(1-u)), & 0 \leqq t \leqq \frac{1}{2}, \\ \gamma((2t-1)(1-u)), & \frac{1}{2} \leqq t \leqq 1. \end{cases}$$

Finally we have

$$\gamma 1_{z_0} \approx \gamma \tag{12.7–18}$$

if z_0 is the initial point of γ.

This is a consequence of the fact that the path on the left is a reparametrization of the path on the right, for if we introduce the function, (fig. 12.7–5),

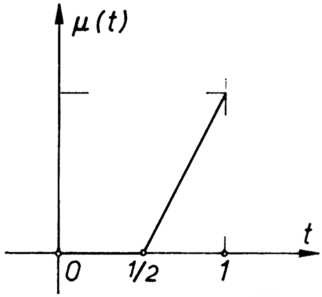


Fig. 12.7–5. The function $\mu(t)$ related to the proof of (12.7–18)

$$\mu(t) = \begin{cases} 0, & 0 \leqq t \leqq \frac{1}{2}, \\ 2t-1, & \frac{1}{2} \leqq t \leqq 1, \end{cases}$$

then

$$\gamma(\mu(t)) = \begin{cases} \gamma(0), & 0 \leqq t \leqq \frac{1}{2}, \\ \gamma(2t-1), & \frac{1}{2} \leqq t \leqq 1, \end{cases}$$

and the function that defines the product of the paths 1_{z_0} and γ is determined by $\gamma_0(t) = \gamma(0)$, $0 \leqq t \leqq \frac{1}{2}$ and $\gamma(2t-1)$, $\frac{1}{2} \leqq t \leqq 1$.

In a similar way we may prove

$$1_{z_1}\gamma \approx \gamma, \tag{12.7–19}$$

if z_1 is the end point of γ. This is performed by the reparametrization

$$\mu(t) = \begin{cases} 2t & 0 \leqq t \leqq \frac{1}{2}, \\ 1, & \frac{1}{2} \leqq t \leqq 1 \end{cases}$$

(fig. 12.7–6).

It should be noticed that the multiplication of paths as defined in section 12.7.1 is applicable not to all pairs of paths in the set $\mathfrak{A}$, but

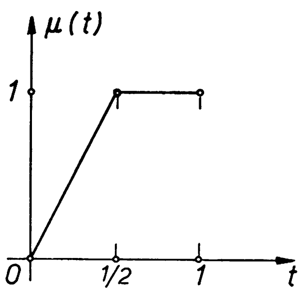


Fig. 12.7–6. The function $\mu(t)$ related to the proof of (12.7–19)

only to those pairs for which the end point of one and the initial point of the other coincide. To make multiplication possible for all pairs of paths under consideration we must constrict ourselves to the set of paths beginning and ending at the same point z_0, the *base point* of the set. In this respect the following theorem is useful.

If γ_1 and γ_2 are any two paths joining the points z_0 and z_1 then γ_1 is homotopic to γ_2 if and only if $\gamma_2^{-1}\gamma_1$ is homotopic to 1_{z_0}.

Applying the previous result we may conclude that if $\gamma_2^{-1}\gamma_1 \approx 1_{z_0}$ then $\gamma_1 \approx \gamma_2 1_{z_0} \approx \gamma_2(\gamma_2^{-1}\gamma_2) \approx (\gamma_2\gamma_2^{-1})\gamma_2 \approx 1_{z_1}\gamma_2 \approx \gamma_2$. Conversely, if $\gamma_1 \approx \gamma_2$ then $\gamma_2^{-1}\gamma_1 \approx \gamma_2^{-1}\gamma_2 \approx 1_{z_0}$.

For this reason it is sufficient to study the homotopy of closed paths.

12.7.4 – Homotopy and analytic continuation

A fundamental theorem in the theory of analytic continuation gives an answer to the question posed in the beginning of section 12.7.1. It states

Let z_0 and z_1 denote two points in an open set $\mathfrak{A}$ and let $(f_0, \mathfrak{K}_0)$ be a function element defined in a disc with centre z_0. Assume that this element is continuable along every path joining z_0 to z_1. If γ_1 and γ_2 are two paths joining these points then analytic continuation along γ_1 leads to the same function element $(f_1, \mathfrak{K}_1)$ at z_1 as analytic continuation along γ_2, provided that γ_1 and γ_2 are homotopic in $\mathfrak{A}$.

Let $\varphi(t, u)$ denote the deformation of the path γ_1 anto γ_2 and denote by γ the path defined by $\varphi(t, u_0)$, where u_0 is fixed. Since $\varphi(t, u)$ is uniformly continuous on the deformation square $\mathfrak{E}$, to a given number $\varepsilon > 0$ corresponds a number $\delta > 0$ such that

$$|\varphi(u, t)-\varphi(u_0, t)| < \varepsilon, \tag{12.7–20}$$

provided that $|u-u_0| < \delta$. By hypothesis continuation is possible along every path $\varphi(t, u)$ (u being a fixed number from the unit interval). With reference to the last theorem of section 12.1.6 we may conclude that analytic continuation along γ and along every path defined by $\varphi(t, u)$, where u satisfies (12.7–20), leads to the same element at z_1, provided ε has been chosen sufficiently small. Hence $\mathfrak{E}$ can be covered by horizontal open strips having the property that all segments within the same strip are mapped onto paths along which continuation leads to the same final element at z_1. By the Heine-Borel theorem already a finite number of such strips cover $\mathfrak{E}$. Thus it appears that continuation along γ_2 can be performed by means of a finite number of intermediate steps, starting with a continuation along γ_1. This concludes the proof of the theorem.

This theorem gives us information about the reason why the product of

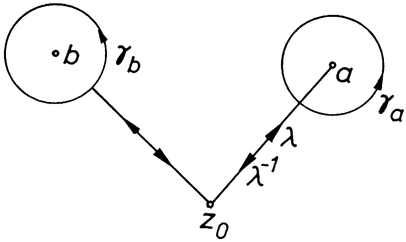


Fig. 12.7–7. Analytic continuation around two singular points

two paths is not necessarily commutative, also if these paths are loops beginning and ending at the same point z_0. We consider an algebraic function defined by $f(z, w) = 0$. Analytic continuation about a path consisting of a small circumference γ_a about a critical point a together with two segments λ, λ^{-1} joining z_0 to the nearest point of γ_a, i.e., the loop $\lambda^{-1}\gamma_a\lambda$, (fig. 12.7–7), effects a permutation π_a of the roots of $f(z, w) = 0$. Similarly, analytic continuation along a path γ_b of the same kind surrounding another critical point b effects a permutation π_b. In general $\pi_a\pi_b \neq \pi_b\pi_a$, i.e., the product of the paths in a certain order needs not to be homotopic to the product with factors interchanged.

We apply the above theorem to a special function. We assume *that the set $\mathfrak{A}$ does not contain the point at infinity*. If a is a finite point not belonging to $\mathfrak{A}$, a branch of $\log(z-a)$ defined in a neighbourhood of a point z_0 of $\mathfrak{A}$ is continuable along every path in $\mathfrak{A}$. If γ begins at z_0 and ends at z_1 then the analytic continuation along γ gives rise to a continuous function $\log(\gamma(t)-a)$ and so is $\arg(\gamma(t)-a)$, being the imaginary part of this function. It is clear that

$$\arg(\gamma(1)-a)-\arg(\gamma(0)-a) \tag{12.7–21}$$

is the increment $2\pi i\Omega_\gamma(a)$ of the argument of $z-a$ along γ (section 2.2.1) and $\Omega_\gamma(a)$ is the winding number of a loop γ with respect to a if γ begins and ends at z_0. In view of the previous theorem we have

The winding number of two homotopic loops with respect to a point a outside $\mathfrak{A}$ are equal.

This assertion fails to be valid if $\mathfrak{A}$ contains the point at infinity, for then $\log(z-a)$ cannot be continued along every path in $\mathfrak{A}$.

Since homotopy is an equivalence relation it is natural to collect all paths homotopic to a given path γ into a class C, the *homotopy class* of γ. The class is determined by any of its elements. If we wish to exhibit the generating path γ we prefer to write $\{\gamma\}$. It is clear that the number denoted by $\Omega_C(a)$ makes sense if we define it as $\Omega_\gamma(a)$ where γ is any generator of C. Thus it is not important whether C denotes a path or a homotopy class. It depends on the particular circumstances which interpretation is to be preferred.

By the same reason analytic continuation along C can be defined as analytic continuation along any path belonging to C. It must be supposed, of course, that analytic continuation is possible along each path of C.

12.7.5 – The homotopy group of a region

In order to make multiplication possible for all pairs of paths under consideration we shall restrict ourselves to the set of paths beginning and ending at a given point z_0 of the open set $\mathfrak{A}$. We have already introduced an algebraic structure in this set of paths but this will become more manageable if we consider homotopy classes rather than separate loops.

The *product* $C_2 C_1$ of two classes C_1 and C_2 generated by the paths γ_1 and γ_2 respectively is defined as the class defined by $\gamma_2\gamma_1$. It follows from (12.7–14) that the product is determined by the classes C_1 and C_2 only.

The associativity is a consequence of (12.7–13). It is not claimed that the product is commutative. The class generated by a loop homotopic to 1_{z_0} will also be denoted by 1. Then (12.7–18) and (12.7–19) assure that this class is a neutral element with respect to multiplication, or a unit. Finally it is clear that every class C has an inverse C^{-1} generated by γ^{-1} if C is generated by γ. This assertion is based on (12.7–16) and (12.7–17). Thus we have

The homotopy classes of the loops beginning and ending at a point z_0 of the set $\mathfrak{A}$ constitute a group F_{z_0} with respect to the multiplication as defined above.

It appears as though the group depends on z_0 selected as base point. This is not true if we assume that $\mathfrak{A}$ is a region $\mathfrak{R}$. Then we can prove that, if another point z_1 is selected as base point, the group F_{z_0} and F_{z_1} are isomorphic. Since $\mathfrak{R}$ is arcwise connected there is a path λ joining z_0 to z_1. If C is a class of F_{z_0} generated by the path γ we associate to it a class $\lambda C\lambda^{-1}$ being the class of F_{z_1} generated by $\lambda\gamma\lambda^{-1}$. This correspondence is a homomorphism, for if C_1 and C_2 are in F_{z_0} then the product $\lambda C_2\lambda^{-1}\lambda C_1\lambda^{-1} = \lambda C_2 C_1\lambda^{-1}$ is in F_{z_1} and corresponds to $C_2 C_1$. Interchanging the roles of z_0 and z_1 we have to replace λ by λ^{-1} and thus it appears that the correspondence between F_{z_0} and F_{z_1} is one-to-one, i.e., an isomorphism.

The isomorphism between F_{z_0} and F_{z_1} depends on the path joining z_0 to z_1. If $z_0 = z_1$ and λ a loop beginning and ending at z_0 then

$$\lambda C\lambda^{-1} = \{\lambda\gamma\lambda^{-1}\} = LCL^{-1}$$

where L is the class generated by λ. The correspondence between C and LCL^{-1} is an inner automorphism.

Since F_{z_0} as an abstract group does not depend on z_0 we may denote it simply by F. It is called *the homotopy group of the region* $\mathfrak{R}$.

If f is a continuous mapping of a region $\mathfrak{R}$ into a region $\mathfrak{R}'$ then f induces a homomorphism of the homotopy group F of $\mathfrak{R}$ into the homotopy group F' of $\mathfrak{R}'$.

To every path γ in $\mathfrak{R}$ corresponds a path $f(\gamma)$ defined by the function $f(\gamma(t))$ in $\mathfrak{R}'$ and if γ is a loop beginning and ending at z_0 then $f(\gamma)$ is a loop beginning and ending at $f(z_0)$. Further, if $\varphi(t, u)$ is a deformation of γ_1 into γ_2 in $\mathfrak{R}$ then $f(\varphi(t, u))$ is a deformation of $f(\gamma_1)$ into $f(\gamma_2)$ in $\mathfrak{R}'$. Finally $f(\gamma_2\gamma_1) = f(\gamma_2)f(\gamma_1)$ as follows immediately from (12.7–2). Hence the correspondence induced by f is a homomorphism of F_{z_0} into $F'_{f(z_0)}$.

The homotopy group of the region $\mathfrak{R}$ is a topological invariant.

This means that if the regions $\mathfrak{R}$ and $\mathfrak{R}'$ are homeomorphic their homotopy groups are isomorphic. This is a direct consequence of the previous theorem, for if f is a homeomorphism then f^{-1} is a continuous mapping of $\mathfrak{R}'$ onto $\mathfrak{R}$ and the correspondence between F_{z_0} and $F'_{f(z_0)}$ is one-to-one.

12.7.6 – Homotopy and homology

In the theory of integration the relation of homology between chains plays a dominating part. Moreover, chains are formal sums of paths. It is our intention to show that there is an intimate connection between the conceptions of homotopy and homology.

Let $\mathfrak{R}$ denote a region in the finite z-plane. By N we denote the set of all homotopy classes C whose winding number $\Omega_C(a)$ with respect to every point a outside $\mathfrak{R}$ is zero. If A and B are two elements of the set then it follows from (12.7–21) that $\Omega_{B^{-1}A} = \Omega_A - \Omega_B = 0$, that is, $B^{-1}A$ is also an element of the set. Hence

The set N of all homotopy classes of a region $\mathfrak{R}$ whose winding numbers with respect to the points outside $\mathfrak{R}$ are zero is a subgroup of the homotopy group.

In addition we have

The group N is an invariant subgroup of the homotopy group F.

In fact, if C is an arbitrary element of the homotopy group F and A an element of N then $\Omega_{CAC^{-1}} = \Omega_C + \Omega_A - \Omega_C = \Omega_A = 0$ i.e., CAC^{-1} belongs to N.

Now we shall say that *the classes A and B are homologous* if $B^{-1}A$ belongs to N, that is to say that $\Omega_A = \Omega_B$. We may write this as $A \sim B$.

The multiplication of classes is commutative with respect to the homology relation.

This is a direct consequence of

$$\Omega_{BA} = \Omega_A + \Omega_B = \Omega_B + \Omega_A = \Omega_{AB}$$

i.e.,

$$AB \sim BA.$$

If α and β are generators of the classes A and B respectively this relation states that the winding number $\Omega_{\beta\alpha}$ of the product $\beta\alpha$ is equal to the winding number of $\alpha\beta$ and we may say that $\beta\alpha \sim \alpha\beta$. By this reason the group operation may be written additively. Since the group N is an invariant subgroup, it is natural to collect all homotopy classes which are homologous to a given class C into a class $[C]$ and we may define

$$[A]+[B] = [AB].$$

These classes constitute a commutative group the factor group F/N of F modulo N. It is called the *homology group of the region* $\mathfrak{R}$.

Thus we have found a more abstract approach to the notion of homology and the question rises whether the theorems about integration may be interpreted in terms of this new notion. This will be the subject matter of the next section.

12.7.7 – Integration

First we recall (section 2.11.4) that if $\mathfrak{R}$ is a simply connected region then every function $f(z)$ holomorphic throughout $\mathfrak{R}$ possesses a primitive, i.e. there exists a (single-valued) function $F(z)$ with $F'(z) = f(z)$. This assertion fails to remain valid if $\mathfrak{R}$ is not simply connected. But we can introduce a similar notion which has a meaning in every region.

Let γ denote a path joining the points z_0 and z_1 of a region $\mathfrak{R}$ and let $f(z)$ be a function holomorphic throughout $\mathfrak{R}$. We shall say that the funtion $g(t)$, $0 \leq t \leq 1$, is *a primitive of* $f(z)$ *along* γ if it satisfies the following condition: for every t_0 in the unit interval there exists in a neighbourhood of the point $\gamma(t_0)$ in $\mathfrak{R}$ a primitive $F(z)$ such that

$$g(t) = F(\gamma(t)), \qquad (12.7\text{–}22)$$

if t is sufficiently near t_0. We shall prove that such a primitive always exists.

Because of the uniform continuity of $\gamma(t)$ we can find a finite number of points in the unit interval

$$t_0 = 0 < t_1 < \ldots < t_n = 1,$$

such that for every t_k, $k = 1, \ldots, n$, the function $\gamma(t)$ maps the segment

$t_{k-1} \leqq t \leqq t_k$ into an open disc $\mathfrak{K}_k$. Since this disc is simply connected there exists a primitive $F_k(z)$ of $f(z)$. The intersection of two consecutive discs $\mathfrak{K}_{k-1}$ and $\mathfrak{K}_k$, $k > 1$, is not empty, for it contains the point $\gamma(t_{k-1})$ and it is again simply connected. It follows that $F_k(z) - F_{k-1}(z)$ is constant throughout this intersection. By adding a suitable constant we even have $F_k(z) = F_{k-1}(z)$ throughout the intersection. Now we define a function $g(t)$ by means of

$$g(t) = F_k(\gamma(t)), \qquad t_{k-1} \leqq t \leqq t_k,\, k = 1, \ldots, n. \tag{12.7–23}$$

It is an easy matter to verify that this function is continuous on the unit interval. This proves the existence.

A primitive of this kind is determined up to an additive constant. Let $g_1(t)$ and $g_2(t)$ denote two primitives of $f(z)$ along the path γ. Since two primitives of $f(z)$ in a disc differ only in a constant we may conclude that $g_2(t) - g_1(t)$ is constant in a neighbourhood of every t_0 of the unit interval. We express this by saying that $g_2(t) - g_1(t)$ is *locally constant*. The set of values of t for which $g_2(t) - g_1(t) = g_2(0) - g_1(0)$ is evidently open and closed. On the other hand, because the unit interval is connected the set under consideration coincides with the unit interval. Summing up

A function $f(z)$ holomorphic throughout a region $\mathfrak{R}$ possesses along every path γ a primitive $g(t)$, being a function such that for every t_0 there is a primitive $F(z)$ of $f(z)$ in a suitably chosen neighbourhood of $\gamma(t_0)$ for which (12.7–22) *holds.*

Let now γ denote a smooth path joining $z_0 = \gamma(0)$ to $z_1 = \gamma(1)$ According to (2.3–25) the integral of $f(z)$ along a path γ is

$$\int_\gamma f(\zeta)d\zeta = \int_0^1 f(\gamma(t))\gamma'(t)dt = \sum_{k=1}^n \int_{t_{k-1}}^{t_k} F_k'(\gamma(t))\gamma'(t)dt$$

$$= \sum_{k=1}^n (F_k(\gamma(t_k)) - F_k(\gamma(t_{k-1}))) = F_n(\gamma(t_n)) - F_1(\gamma(t_0)) = g(1) - g(0).$$

If γ is piecewise smooth we obtain the same result by adding the contributions of the separate parts.

The condition of differentiability does not enter into the definition of primitive along a path. Hence we may define

$$\int_\gamma f(\zeta)d\zeta = g(1) - g(0), \tag{12.7–24}$$

where γ is any path joining $z_0 = \gamma(0)$ to $z_1 = \gamma(1)$, not necessarily smooth

In this connection it is interesting to investigate the behaviour of an integral if the path of integration undergoes a deformation. To this end we introduce the conception of *primitive* $g(t, u)$ of $f(z)$ *with respect to the deformation* $\varphi(t, u)$, satisfying the condition:

For every point (t_0, u_0) in the deformation square $\mathfrak{E}$ there exists a neighbourhood of $\varphi(t_0, u_0)$ in $\mathfrak{R}$ such that, if $F(z)$ is a primitive of $f(z)$ in this neighbourhood, we have

$$g(t, u) = F(\varphi(t, u)) \tag{12.7–25}$$

for all (t, u) sufficiently near (t_0, u_0).

The existence of a primitive of this kind may be proved in quite the same way as the existence of a primitive along a path. We divide the deformation square into small squares $\mathfrak{E}_{k,l}$: $t_{k-1} \leqq t \leqq t_k$, $u_{l-1} \leqq u \leqq u_l$, $k = 1, \ldots, n$, $l = 1, \ldots, n$. Since $\varphi(t, u)$ is uniformly continuous we can take these subsquares so small that $\mathfrak{E}_{k,l}$ can be mapped into an open disc $\mathfrak{K}_{k,l}$ within $\mathfrak{R}$, on which a primitive $F_{k,l}(z)$ of $f(z)$ is defined. Keeping l fixed, we observe that the intersection of $\mathfrak{K}_{k-1,l}$ and $\mathfrak{K}_{k,l}$ is not empty. By adjusting constants we may conclude that $F_{k-1,l}(z) = F_{k,l}(z)$ within this intersection and for $u_{l-1} \leqq u \leqq u_l$ we obtain a function $g_l(t, u)$ such that for every k we have

$$g_l(t, u) = F_{k,l}(t, u), \qquad t_{k-1} \leqq t \leqq t_k, u_{l-1} \leqq u \leqq u_l.$$

It is readily shown that $g_l(t, u)$ is continuous in the rectangle $0 \leqq t \leqq 1$, $u_{l-1} \leqq u \leqq u_l$. Every function g_l is determined up to an additive constant and by adjusting constants we may conclude that $g_{l-1}(t, u) = g_l(t, u)$ if $u = u_l$. Let now $g(t, u)$ denote the function defined by the condition that

$$g(t, u) = g_l(t, u), \qquad 1 \leqq l \leqq n,$$

if $u_{l-1} \leqq u \leqq u_l$. It is the desired function. Now we are sufficiently prepared to prove

If $\gamma_1 \approx \gamma_2$ *then*

$$\int_{\gamma_1} f(\zeta)d\zeta = \int_{\gamma_2} f(\zeta)d\zeta. \tag{12.7–26}$$

In fact, according to (12.7–24),

$$\int_{\gamma_1} f(\zeta)d\zeta = g(1, 0) - g(0, 0),$$

$$\int_{\gamma_2} f(\zeta)d\zeta = g(1, 1) - g(0, 1).$$

The truth of the assertions follows from

$$g(1, 0) = g(1, 1), g(0, 0) = g(0, 1).$$

The following theorem brings the theory of this section in line with the considerations of section (2.5.2).

The value of the integral

$$\int_{\gamma} f(\zeta)d\zeta \tag{12.7–27}$$

of the holomorphic function $f(z)$ in a region $\mathfrak{R}$ in the finite plane along a closed path γ is zero if γ is homologous zero.

As we pointed out in section 2.4.5 the path γ can be approximated by a polygon γ_1, whose vertices are on γ. Since this polygon can be linearly deformed into γ its winding number, being the integral

$$\frac{1}{2\pi i}\int_{\gamma_1} \frac{d\zeta}{\zeta - a}, \tag{12.7–28}$$

is equal to the winding number of $\mathfrak{S}$ with respect to every a outside $\mathfrak{R}$. Hence $\gamma_1 \sim 0$ and from Cauchy's theorem of section 2.5.2 follows the truth of the assertion.

If γ generates the homotopy class C and the homology class $[C]$ the integrals

$$\int_C f(\zeta)d\zeta, \qquad \int_{[C]} f(\zeta)d\zeta$$

can be defined by (12.7–24). Hence it is rather indifferent whether C denotes a path or a class.

As we pointed out in section 12.2.5 integration of an analytic function may be defined by analytic continuation. Since analytic continuation is possible along any continuous path we need not suppose that the path is (sectionally) differentiable. We wish to clarify some details which will find application in subsequent sections.

Let $F(z)$ denote an analytic function defined throughout a region $\mathfrak{R}$ and let z_0, z_1 denote two (not necessarily different) points connected by a path γ in $\mathfrak{R}$. We consider a function element $f_0(z)$ of $F(z)$ in a neighbourhood of z_0. It possesses a primitive $\Phi_0(z)$ which is uniquely determined if we assign a given value to it at z_0. Starting with $\Phi_0(z)$ we are led by analytic continuation along γ to a uniquely defined element $\Phi_1(z)$ in the vicinity of z_1. According to the first theorem of section 12.2.2 its derivative $f_1(z)$ is obtained by continuing analytically $f_0(z)$ along γ. With this in mind we define

$$\int_\gamma F(\zeta)d\zeta = \Phi_1(z_1) - \Phi_0(z_0). \tag{12.7–29}$$

It should be noticed that the expression on the left makes sense, provided we have assigned the function element at z_0.

Now we may ask what happens if γ is replaced by its inverse γ^{-1}, assuming that γ is a loop beginning and ending at z_0. Analytic continuation of $\Phi_0(z)$ along γ^{-1} leads to $\Phi_{-1}(z)$ in the vicinity of z_0, and, therefore,

$$\int_{\gamma^{-1}} F(\zeta)d\zeta = \Phi_{-1}(z_0) - \Phi_0(z_0), \tag{12.7–30}$$

whereas the integral on the left of (12.7–9) is now $\Phi_1(z_0)-\Phi_0(z_0)$.

At first sight there does not seem to exist a simple relation between the two integrals under consideration. The situation is, however, extremely simple if γ is such that by analytic continuation of $f_0(z)$ we are led to $f_0(z)$ again, i.e., if γ is closed with respect to analytic continuation of this element. In this case we may state

If the loop γ is closed with respect to analytic continuation of the function elements of the analytic function $F(z)$ then

$$\int_{\gamma^{-1}} F(\zeta)d\zeta = -\int_{\gamma} F(\zeta)d\zeta. \tag{12.7–31}$$

Indeed, $\Phi_1 = \Phi_0+c$, where c is a constant and $\Phi_{-1} = \Phi_0-c$. The integral on the left has the value $\Phi_{-1}(z_0)-\Phi_0(z_0) = -c = -(\Phi_1(z_0)-\Phi_0(z_0))$. This concludes the proof.

In the case that $F(z)$ is a single-valued holomorphic function $f(z)$ the equality (12.7–31) remains valid if γ is not necessarily a closed path. Indeed, if $g(t)$ is a primitive of $f(z)$ along γ then $h(t) = g(1-t)$ is a primitive along γ^{-1} and

$$\int_{\gamma^{-1}} f(\zeta)d\zeta = h(1)-h(0) = g(0)-g(1) = -\int_{\gamma} f(\zeta)d\zeta.$$

A somewhat more general situation occurs if γ is an elementary loop about $z = z$ and $F(z)$ admits of a *multiplier* there. This means that continuation of any element of $F(z)$ leads to an element obtained from the one we started with by multiplying it by a certain constant ξ. It is clear that then $\Phi_1(z) = \xi\Phi_0(z)$ and $\Phi_{-1}(z) = \xi^{-1}\Phi_0(z)$. In this case we have

If γ denotes an elementary loop about $z = a$ and if $F(z)$ possesses a multiplier there, then

$$\int_{\gamma^{-1}} F(\zeta)d\zeta = -\xi^{-1}\int_{\gamma} F(\zeta)d\zeta. \tag{12.7–32}$$

This follows from

$$\int_{\gamma^{-1}} F(\zeta)d\zeta = \xi^{-1}\Phi_0(z_0)-\Phi_0(z_0) = (\xi^{-1}-1)\Phi_0(z_0)$$
$$= -\xi^{-1}(\xi-1)\Phi_0(z_0) = -\xi^{-1}\int_{\gamma} F(\zeta)d\zeta.$$

Examples of functions with multipliers at isolated points have been considered in section 12.2.5.

We proceed with the investigation of an integral taken along the product $\beta\alpha$ of two loops β and α. If α is subject to a restrictive condition as stated below we have

If the loop α is closed with respect to analytic continuation of the function elements of the analytic function $F(z)$, then

$$\int_{\beta\alpha} F(\zeta)d\zeta = \int_{\beta} F(\zeta)d\zeta + \int_{\alpha} F(\zeta)d\zeta. \tag{12.7–33}$$

Let $\Phi_1(z)$ denote the function element obtained from $\Phi_0(z)$ by analytic continuation along α and $\Phi_2(z)$ the element obtained from $\Phi_0(z)$ by analytic continuation along β. Then $\Phi_2(z)$ is obtained by analytic continuation along $\beta\alpha$ from $\Phi_0(z)$. Hence

$$\int_{\beta\alpha} F(\zeta)d\zeta = \Phi_2(z_0)-\Phi_0(z_0) = \Phi_2(z_0)-\Phi_1(z_0)+\Phi_1(z_0)-\Phi_0(z_0).$$

Now

$$\Phi_1(z_0)-\Phi_0(z_0) = \int_{\alpha} F(\zeta)d\zeta.$$

Let $\Psi(z)$ be the element obtained from $\Phi_0(z)$ by analytic continuation along β. Then

$$\Psi(z_0)-\Phi_0(z_0) = \int_{\beta} F(\zeta)d\zeta.$$

Since α is closed with respect to analytic continuation of the elements of $F(z)$ we have $\Phi_1(z) = \Phi_0(z)+c$ and $\Phi_2(z) = \Psi(z)+c$. Hence

$$\Phi_2(z_0)-\Phi_1(z_0) = \Psi(z_0)-\Phi_0(z_0) = \int_{\beta} F(\zeta)d\zeta.$$

This concludes the proof of the assertion.

In the case that $F(z)$ is a single-valued holomorphic function $f(z)$ the equality (12.7–33) remains true if β and α are not necessarily closed paths. Indeed if $g_\alpha(t)$ is a primitive of $f(z)$ along α and $g_\beta(t)$ a primitive of $f(z)$ along β then

$$g(t) = \begin{cases} g_\alpha(2t), & 0 \leqq t \leqq \frac{1}{2}, \\ g_\beta(2t-1), & \frac{1}{2} \leqq t \leqq 1, \end{cases}$$

is a primitive of $f(z)$ along $\beta\alpha$. According to (12.7–24) we have

$$\begin{aligned} \int_{\beta\alpha} f(\zeta)d\zeta &= g(1)-g(0) = g(1)-g(\tfrac{1}{2})+g(\tfrac{1}{2})-g(0) \\ &= g_\beta(1)-g_\beta(0)+g_\alpha(1)-g_\alpha(0) = \int_{\beta} f(\zeta)d\zeta + \int_{\alpha} f(\zeta)d\zeta. \end{aligned}$$

Next we suppose that α and β are elementary loops about the points $z = a$ and $z = b$ and that $F(z)$ possesses multipliers ξ and η at these points respectively. In this case we have

If α denotes an elementary loop about $z = a$ and β an elementary loop about $z = b$, if further $F(z)$ possesses a multiplier ξ at a and a multiplier η at b then

$$\int_{\beta\alpha} F(\zeta)d\zeta = \int_{\alpha} F(\zeta)d\zeta + \xi \int_{\beta} F(\zeta)d\zeta,$$
$$\int_{\alpha\beta} F(\zeta)d\zeta = \int_{\beta} F(\zeta)d\zeta + \eta \int_{\alpha} F(\zeta)d\zeta. \tag{12.7–34}$$

By assumption $f_1(z) = \xi f_0(z)$, whence $\Phi_1(z) = \xi\Phi_0(z)$ and $\Phi_2(z) = \xi\Psi(z)$. It follows that

$$\Phi_2(z_0) - \Phi_1(z_0) = \xi(\Psi(z_0) - \Phi_0(z_0)) = \xi \int_{\beta} F(\zeta)d\zeta,$$

whereas

$$\Phi_1(z_0) - \Phi_0(z_0) = \int_{\alpha} F(\zeta)d\zeta.$$

The proof of the second equation (12.7–34) is quite the same.

The following assertion is of particular interest.

Under the assumptions of the previous theorem, if (b_+, a_+, b_-, a_-) denotes a Jordan-Pochhammer contour about a and b within a region $\mathfrak{R}$ punctured at the points a and b and in which $F(z)$ is arbitrarily continuable then

$$\int^{(b_+, a_+, b_-, a_-)} F(\zeta)d\zeta = (1-\xi)\int_{\beta} F(\zeta)d\zeta - (1-\eta)\int_{\alpha} F(\zeta)d\zeta. \tag{12.7–35}$$

It is not difficult to see that the Jordan-Pochhammer contour (fig. 12.2–9) is homotopic in $\mathfrak{R}$ to the path $\alpha^{-1}\beta^{-1}\alpha\beta$. By virtue of (12.7–34) and (12.7–32) we have

$$\begin{aligned}\int_{\alpha^{-1}\beta^{-1}\alpha\beta} &= \int_{\beta} + \eta\int_{\alpha^{-1}\beta^{-1}\alpha} = \int_{\beta} + \eta\int_{\alpha} + \xi\eta\int_{\alpha^{-1}\beta^{-1}}\\ &= \int_{\beta} + \eta\int_{\alpha} + \xi\eta\int_{\beta^{-1}} + \xi\int_{\alpha^{-1}}\\ &= \int_{\beta} + \eta\int_{\alpha} - \xi\int_{\beta} - \int_{\alpha} = (1-\xi)\int_{\beta} - (1-\eta)\int_{\alpha}.\end{aligned}$$

A particular case is of some interest. Suppose that

$$\int_{z_0}^{a} F(\zeta)d\zeta, \qquad \int_{z_0}^{b} F(\zeta)d\zeta$$

exist, the first integral being taken along a path that does not encircle b, etc. Then, evidently,

$$\int_{\alpha} F(\zeta)d\zeta = \int_{z_0}^{a} F(\zeta)d\zeta + \xi \int_{a}^{z_0} F(\zeta)d\zeta = (1-\xi)\int_{z_0}^{a} F(\zeta)d\zeta,$$

$$\int_{\beta} F(\zeta)d\zeta \qquad = (1-\eta)\int_{z_0}^{b} F(\zeta)d\zeta.$$

It follows that

$$(1-\xi)\int_{\beta} -(1-\eta)\int_{\alpha} = (1-\xi)(1-\eta)\left(\int_{z_0}^{b} - \int_{z_0}^{a}\right) = (1-\xi)(1-\eta)\int_{a}^{b},$$

whence

$$\int^{(b+,a+,b-,a-)} F(\zeta)d\zeta = (1-\xi)(1-\eta)\int_{a}^{b} F(\zeta)d\zeta. \qquad (12.7\text{–}36)$$

A particular case of this formula is (12.2–45).

12.7.8 – Determination of the homotopy groups of certain regions

For the applications we need the knowledge of the homotopy groups of certain regions which we shall consider in more detail. Let us first focus our attention on a convex region (section 2.2.1).

The homotopy group of a convex region reduces to the unit element.

This is almost trivial, for any loop γ can be deformed linearly into the base point z_0; indeed the base point can be joined to each point by means of a linear segment within the region.

In particular *the homotopy group of an open disc or of the finite plane or of the extended plane reduces to the unit element.*

Now we state the important theorem

A region is simply connected if and only if its homotopy group reduces to the unit element.

By Riemann's mapping theorem of section 10.5.2 any simply connected region with at least two boundary points can be mapped conformly (and hence topologically) onto an open disc. If there is only one boundary point or if there is no boundary point then by a linear fractional transformation the region can be mapped onto the open plane or into the extended plane. In view of the last theorem of section 12.7.5 we conclude that the homotopy group reduces to the unit element.

Suppose now that $\mathfrak{R}$ is not simply connected. Then its complement in the extended plane is not connected (section 9.1.3) and it has at least one finite component. As in the proof of the first theorem of section 9.1.3 we may construct a path whose winding number with respect to a point in this finite component is not zero. Without loss of generality we may assume that $\mathfrak{R}$ does not contain the point at infinity. Hence the path mentioned above is not homotopic to 1, in contradiction to the assumption. This concludes the proof of the theorem.

In particular we have again the *monodromy theorem* of section 12.2.3: *If a function element* $(f, \mathfrak{K})$ *can be continued along every path in a simply connected region then it generates a single-valued function defined throughout the region.* In fact in such a region every two paths connecting the points z_0 and z_1 are homotopic.

Next we consider in the finite plane a convex region $\mathfrak{R}$ from which one point is deleted. On applying, if necessary, a linear transformation, we may suppose that the region is punctured at the origin. We consider a circumference of radius r around the origin belonging to $\mathfrak{R}$. Let z_0 denote a point on the circumference. The circumference may be parametrized as

$$z(t) = re^{2\pi it}, \qquad 0 \leqq t \leqq 1. \tag{12.7–37}$$

Let now γ denote a loop within the region beginning and ending at z_0. The deformation

$$\varphi(t, u) = (1-u)\gamma(t) + ru \frac{\gamma(t)}{|\gamma(t)|}$$

deforms γ into a path β whose carrier belongs to the circumference. By continuity every point t_0 in the unit interval $0 \leqq t \leqq 1$ has a neighbourhood such that for all points within the neighbourhood

$$|\beta(t) - \beta(t_0)| < \tfrac{1}{2}r.$$

Hence $\beta(t)$ cannot take both the values z_0 and $-z_0$ for these values of t.

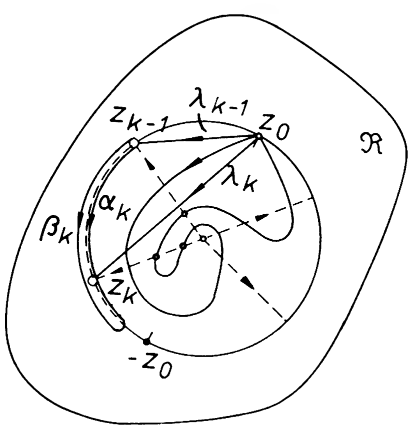


Fig. 12.7–8. Determination of the homotopy group of a convex region from which one point is deleted

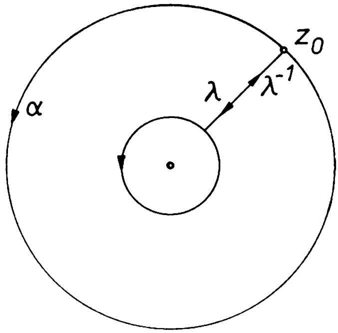


Fig. 12.7–9. Elementary loop about a point

By uniform continuity we may divide β into subpaths $\beta_1, \ldots, \beta_n$ such that

$$\beta = \beta_n \beta_{n-1} \cdots \beta_1,$$

where β_k, $k = 1, \ldots, n$, either does not contain z_0 or $-z_0$. Let the terminal points of β_k be denoted by z_{k-1} and z_k, (fig. 12.7–8). Now β_k is in a simply connected region obtained if we delete from $\mathfrak{R}$ the points of the positive or the negative axis; hence it may be deformed into one of the circular arcs joining z_{k-1} to z_k. Thus γ is homotopic to a product of circular arcs $\alpha_1, \ldots, \alpha_n$. Introducing rectilinear segments $\lambda_1, \ldots, \lambda_{n-1}$, joining z_0 to $z_1, \ldots, z_{n-1}$ respectively we readily see that

$$\gamma \approx \alpha_n \lambda_{n-1} \lambda_{n-1}^{-1} \alpha_{n-1} \cdots \lambda_2^{-1} \alpha_2 \lambda_1 \lambda_1^{-1} \alpha_1,$$

that is to say γ is homotopic to a product of paths $\lambda_k^{-1} \alpha_k \lambda_{k-1}$, where λ_0 and λ_n are constant paths with carrier z_0. Now by linear deformation such a path is either homotopic to 1_{z_0} or to $\alpha^{\pm 1}$, where α denotes the circumference described in the positive direction beginning and ending at z_0. It is clear that α is not homotopic to 1_{z_0}, for $\Omega_\alpha(0) = 1$. Hence γ is homotopic to a power α^m of α. Since $\alpha^m \approx 1_{z_0}$ entails $m_\alpha \Omega_\alpha(0) = 0$, we find that $m = 0$ and we may infer that

The homotopy group of a punctured convex region in the finite plane is the infinite cyclic group.

Sometimes it is convenient to replace α by a so-called *elementary loop*. We connect the points $z_0 = r$ and $z = \varepsilon > 0$ by means of a straight line segment λ, then describe a circle about $z = 0$ with radius ε beginning and ending at ε and go back along λ^{-1} to z_0, (fig. 12.7–9). Since the winding number of this path with respect to the origin is again equal to 1, it is homotopic to α, and may, therefore, be taken as a generating element of the class A which generated the cyclic group.

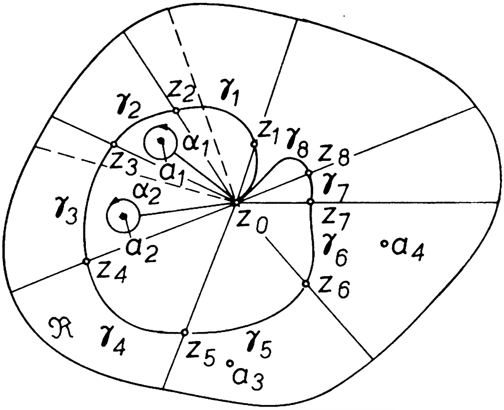


Fig. 12.7–10. Determination of the homotopy group of a convex region punctured at several points

We conclude this section by determining the homotopy group of a convex region from which the points $a_1, \ldots, a_n, n > 1$, are omitted. We take a point such that no two of these are collinear with z_0, (fig. 12.7–10). Next we take in the unit interval $n+1$ points

$$t_0 = 0 < t_1 < \ldots < t_n = 1, \tag{12.7–38}$$

dividing this interval into n parts. The path, being a map of this interval beginning and ending at z_0, is divided into n subpaths $\gamma_1, \ldots, \gamma_n$, where the terminal points of γ_k are $z_{k-1} = \gamma(t_{k-1})$, $z_k = \gamma(t_k)$, $k = 1, \ldots, n$. Next we construct half rays issuing from z_0 and passing through $z_1, \ldots, z_{n-1}$. By uniform continuity we can make the subdivision (12.7–38) so fine that between two consecutive half rays there is no more than one point $a_1, \ldots, a_n$ (if a half ray through r_k passes through one of these points we omit it) and that they bound a convex region which is punctured at at most one point.

Introducing the segments which join z_0 to $z_1, \ldots, z_{n-1}$, calling them $\lambda_1, \ldots, \lambda_{n-1}$, we evidently have

$$\gamma \approx \gamma_n \lambda_{n-1} \lambda_{n-1}^{-1} \gamma_{n-1} \cdots \lambda_2^{-1} \gamma_2 \lambda_1 \lambda_1^{-1} \gamma_1,$$

i.e., γ is a product of paths $\lambda_k^{-1} \gamma_k \lambda_{k-1}$ where λ_0 and λ_n are constant paths at z_0. It is clear that any path of this type is included in a convex angular region, containing z_0 and two consecutive half rays and no more than one of the points $a_1, \ldots, a_n$. Hence, by the previous results either path is homotopic to a power of α_l, where α_l is an elementary path corresponding to a_l. Since no power of such an elementary path is homotopic to z_0

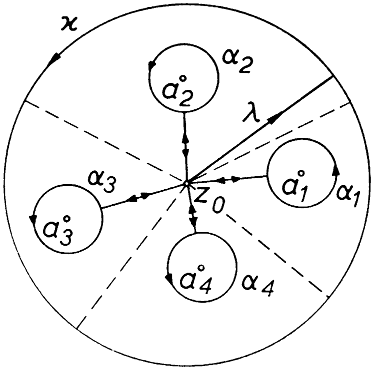


Fig. 12.7–11. Decomposition of analytic continuation along a circumference enclosing singular points into analytic continuations along elementary paths

we see that γ is homotopic to a product of factors $\alpha_i^{m_i}$ and we conclude

The homotopy group of a convex region in the finite plane punctured at n points $a_1, \ldots, a_n$ is the free group with n generators.

By a similar reasoning we find the same result if, instead of points, small open discs are omitted from the region.

A particular case deserves mention. Let κ denote a circumference about z_0 and including the n points deleted from the finite plane. This circumference gives rise to a path $\lambda^{-1}\kappa\lambda$ where λ denotes any segment joining z_0 to a point of the circumference. By dividing it into n parts, (fig. 12.7–11), we readily see that

$$\lambda^{-1}\kappa\lambda \approx \alpha_n \ldots \alpha_1. \tag{12.7–39}$$

Thus analytic continuation along the elementary paths $\alpha_1, \ldots, \alpha_n$ taken in cyclic order yields the same result as analytic continuation along $\lambda^{-1}\kappa\lambda$.

12.7.9 – Symmetric expression for the Eulerian beta function

In section 12.2.6 we obtained a general expression for the Eulerian beta function (4.7–38), viz. the formula (12.2–45), valid for all values of p and q. If $\operatorname{Re} p > 0$, $\operatorname{Re} q > 0$ it reduces to (4.7–38). This formula is, however, not symmetric with respect to the winding points of the integrand, for these are at $z = 0$, $z = 1$ and $z = \infty$ and do not enter in a symmetric way in the formula. In order to find an expression in which these points play the same part, we perform a linear fractional transformation which carries these points into the finite points a, b and c.

The desired transformation is

$$u = \frac{(t-a)(b-c)}{(t-c)(b-a)}. \tag{12.7-40}$$

Then, evidently,

$$1-u = \frac{(t-b)(c-a)}{(t-c)(b-a)}$$

and

$$du = \frac{(a-b)(b-c)(c-a)}{(t-c)^2(b-a)^2}\,dt.$$

The integral on the left of (12.2–44) takes the form

$$-(b-c)^p(c-a)^q(b-a)^{-p-q+1}\times$$
$$\times\int^{(b+,a+,b-,a-)}(t-a)^{p-1}(t-b)^{q-1}(t-c)^{-(p+q)}dt$$
$$=(b-c)^p(c-a)^q(a-b)^{-p-q+1}\,e^{-\pi i(p+q)}\times$$
$$\times\int^{(b+,a+,b-,a-)}(t-a)^{p-1}(t-b)^{q-1}(t-c)^{-(p+q)}\,dt.$$

The expression on the right of (12.2–44) is

$$e^{\pi ip}(e^{\pi ip}-e^{-\pi ip})e^{\pi iq}(e^{\pi iq}-e^{-\pi iq})B(p,q) = -4e^{\pi i(p+q)}\sin\pi p\sin\pi q\,\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

With reference to (4.6–13) we may bring this into the form

$$\frac{-4\pi^2 e^{\pi i(p+q)}}{\Gamma(1-p)\Gamma(1-q)\Gamma(p+q)}.$$

Let now r denote the number satisfying

$$p+q+r = 1. \tag{12.7–41}$$

Then (12.2–44) takes the form

$$\boxed{(b-c)^p(c-a)^q(a-b)^r\int_C(t-a)^{p-1}(t-b)^{q-1}(t-c)^{r-1}\,dt = \frac{-4\pi^2}{\Gamma(1-p)\Gamma(1-q)\Gamma(1-r)},} \tag{12.7–42}$$

where, for the time being, C denotes the image of the Jordan-Pochhammer loop (b_+, a_+, b_-, a_-). The symmetry will be complete if we can deform this loop suitably in the extended plane, punctured at the points a, b and c. In fact, the integrand occurring in (12.7–32) is regular at infinity.

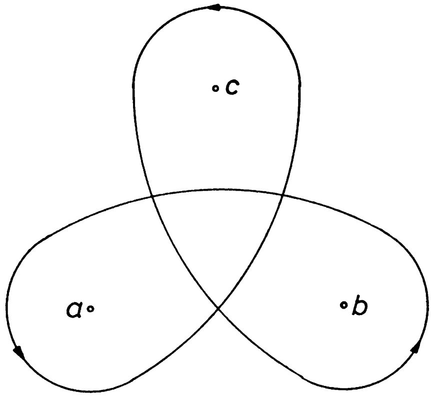


Fig. 12.7–12. The path of integration for the symmetric expression for the bêta function

Let α, β and γ denote elementary loops around a, b and c respectively. It is clear that the path C is homotopic to the product $\alpha^{-1}\beta^{-1}\alpha\beta$. On the other hand the result obtained at the end of the previous section states that $\beta\alpha\gamma$ is homotopic to a closed path encircling the points a, b and c once. Since the exterior of a closed disc in the extended plane is simply connected this loop can be shrunk into a point and it follows that $\gamma \approx \alpha^{-1}\beta^{-1}$. Hence C is homotopic to $\gamma\alpha\beta$ and by a slight modification we may deform it into a path having the form of a clover leaf, encircling the points a, b and c once, (fig. 12.7–12).

CHAPTER 13

AUTOMORPHIC FUNCTIONS

13.1 – Groups of linear transformations

13.1.1 – COVERING TRANSFORMATIONS

The functions which solve the problem of the uniformization of an analytic function have a remarkable property which can be brought to the fore if we consider the so-called covering transformations of the universal covering surface of a Riemann surface.

A *covering transformation* of a triangulated universal covering surface $\tilde{\mathfrak{F}}$ of a Riemann surface $\mathfrak{F}$ is a homeomorphism of $\tilde{\mathfrak{F}}$ into itself which maps each vertex $\tilde{\mathfrak{p}}$ over $\mathfrak{p}$ onto another vertex lying over the same point $\mathfrak{p}$. Thus a covering transformation interchanges the vertices having the same trace on $\mathfrak{F}$. Moreover, it permutes sides lying above the same side of a triangle on $\mathfrak{F}$.

The covering transformation which maps $\mathfrak{p}_1$ *onto* $\mathfrak{p}_2$ *is unique.*

Let $\mathfrak{p}$ denote an arbitrary vertex on $\tilde{\mathfrak{F}}$ and let $\tilde{C}_1$ be a path from $\tilde{\mathfrak{p}}_1$ to $\tilde{\mathfrak{p}}_2$. The path $\tilde{C}_1$ lies above a path C on $\mathfrak{F}$. When $\tilde{\mathfrak{p}}_1$ transforms into $\tilde{\mathfrak{p}}_2$, then $\tilde{C}_1$ goes into a path $\tilde{C}_2$ on $\tilde{\mathfrak{F}}$ which starts at $\tilde{\mathfrak{p}}_2$ and also lies over C. Thus the end point of $\tilde{C}_2$ must be the uniquely defined image of $\tilde{\mathfrak{p}}$.

The set of covering transformations of $\tilde{\mathfrak{F}}$ clearly forms a group under the operation of composition of mappings. It is called the *fundamental group* of $\mathfrak{F}$ and it can be proved that it is a topological invariant of $\mathfrak{F}$, i.e., fundamental groups of homeomorphic surfaces are isomorphic.

A direct consequence of the above assertion is

A covering transformation which is not the identity has no fixed points.

The fundamental group is *properly discontinuous*, i.e., there is a point on $\tilde{\mathfrak{F}}$ which possesses a neighbourhood $\mathfrak{U}(\tilde{\mathfrak{q}})$ such that every non-identical transformation of the fundamental group transforms $\tilde{\mathfrak{q}}$ into a point which is outside $\mathfrak{U}(\tilde{\mathfrak{q}})$. This is clear if we take for $\tilde{\mathfrak{q}}$ a point which is inside a triangle of the triangulation, for the transformations interchanges triangles.

Let us now map the surface $\tilde{\mathfrak{F}}$ onto a canonical region $\mathfrak{C}$ in the s-plane in conformation with the considerations of section 12.6.7. To every covering transformation corresponds a topological transformation of the normal region onto itself. As we proved in section 9.3.3 and in section 9.5.1 such a transformation is a linear fractional transformation and none of these transformations have a fixed point in the canonical region. Since the mapping of $\tilde{\mathfrak{F}}$ onto $\mathfrak{C}$ is one-to-one the group G of these automorphisms of the canonical region is also properly discontinuous.

By means of the mapping of $\tilde{\mathfrak{F}}$ onto $\mathfrak{C}$ a function on $\tilde{\mathfrak{F}}$ can be considered as a function on $\mathfrak{C}$ with the property that it takes the same value at points of $\mathfrak{C}$ which correspond with respect to the transformations of G. Hence a function of this kind is automorphic with respect to this group. Conversely an automorphic function with respect to G gives rise to a function on $\tilde{\mathfrak{F}}$.

This chapter is devoted to the study of properly discontinuous groups and the associate (single-valued) automorphic functions.

13.1.2 – Properly discontinuous groups

A group of linear (fractional) transformations in the s-plane is called *discontinuous* if *it contains no infinitesimal transformations*. That means that it does not contain a sequence of different transformations which converge to the identity. Otherwise stated: in the group is no sequence

$$\frac{a_n s+b_n}{c_n s+d_n}, \qquad n = 1, 2, \ldots, \begin{vmatrix} a_n & b_n \\ c_n & d_n \end{vmatrix} \neq 0, \quad d_n \neq 0, \qquad (13.1\text{–}1)$$

with

$$\lim_{n\to\infty} \frac{a_n}{d_n} = 1, \lim_{n\to\infty} \frac{b_n}{d_n} = \lim_{n\to\infty} \frac{c_n}{d_n} = 0.$$

A group is said to be *properly discontinuous* if there exists a point s_0 and a neighbourhood $\mathfrak{U}(s_0)$ of this point such that all transformations of the group carry s_1 of $\mathfrak{U}(s_0)$ outside $\mathfrak{U}(s_0)$. The theory of automorphic functions is founded on the theory of these groups.

A properly discontinuous group is discontinuous. The converse is not true. A famous counter example is *Picard's group* consisting of the transformations

$$s \to \frac{as+b}{cs+d}, \qquad \begin{vmatrix} a & b \\ c & d \end{vmatrix} = 1, \qquad (13.1\text{–}2)$$

where a, b, c and d are *Gaussian integers*, i.e., numbers of the form $p+qi$, where p and q are real integers.

Suppose the group contains a sequence converging to the identity. We may write

$$s_n = \frac{\alpha_n s+\beta_n}{\gamma_n s+1}, \qquad n = 1, 2, \ldots,$$

with $\alpha_n = a_n/d_n$, $\beta_n = b_n/d_n$, $\gamma_n = c_n/d_n$ and make the additional assumption $a_n d_n - b_n c_n = 1$. Now we have

$$\lim_{n\to\infty} \alpha_n = 1, \qquad \lim_{n\to\infty} \beta_n = \lim_{n\to\infty} \gamma_n = 0.$$

Since

$$\lim_{n\to\infty} (\alpha_n - \beta_n \gamma_n) = \lim_{n\to\infty} \frac{1}{d_n^2} = 1,$$

we have for sufficiently large n that $d_n = \pm 1$, $a_n = d_n$, $b_n = c_n = 0$. Hence, from a certain index upwards all transformations in the sequence are equal to the identical transformation; the group is discontinuous.

The group is, however, not properly discontinuous. To substantiate this statement we observe that it is possible to define an Euclidean algorithm in the integral domain of the Gaussian integers and by well-known arguments it turns out that every Gaussian number can be decomposed into a finite number of prime numbers.

Suppose we are given a rational number $r_1 + ir_2$, which we may represent as

$$r_1 + ir_2 = \frac{b}{d},$$

where b and d are relatively prime Gaussian integers. Hence the greatest common divisor of b and d is unity and we can find integers a and c such that

$$1 = ad - bc.$$

It follows that

$$s \to \frac{as+b}{cs+d}$$

belongs to the group and it transforms $s = 0$ into $b/d = r_1 + ir_2$. Hence all Gaussian rational numbers are congruent to zero and, therefore, there are images of this point within an arbitrary small neighbourhood of $s = 0$.

13.1.3 – The isometric circles

The study of the properly discontinuous groups is much facilitated by the introduction of the conception of the isometric circle of a transformation. We are interested in those points at which the transformation

$$\mathsf{A}s = \frac{as+b}{cs+d}, \qquad \begin{vmatrix} a & b \\ c & d \end{vmatrix} = 1, \tag{13.1–3}$$

leaves the infinitesimal Euclidean lengths unchanged, at which, therefore,

$$|\mathsf{A}'s| = 1,$$

the prime denoting differentiation, i.e., $\mathsf{A}'s = d(\mathsf{A}s)/ds$. The necessary and sufficient condition is expressed by

$$|\mathsf{A}'s| = \frac{1}{|cs+d|^2} = 1, \tag{13.1–4}$$

or

$$|cs+d| = 1. \tag{13.1–5}$$

If $c \neq 0$ the set of points satisfying this equation is a circle I with centre

$$s_{\mathsf{A}} = -\frac{d}{c} \tag{13.1–6}$$

and radius

$$r_{\mathsf{A}} = \frac{1}{|c|}. \tag{13.1–7}$$

It is called the *isometric circle* of the transformation (13.1–3).

If $c = 0$, then $s = \infty$ is a fixed point for the transformation (13.1–3). In this case (13.1–5) is only satisfied for $|d| = 1$. But then the whole open plane is an isometric set. If $c = 0$ and $|d| \neq 1$ there are no isometric sets of points.

It is natural to exclude first the case that $c = 0$. We assert

Lengths within the isometric circle are increased in magnitude and lengths outside the isometric circle are decreased in magnitude.

The proof is easy. If s is inside I, then

$$\left| s+\frac{d}{c} \right| < \frac{1}{|c|},$$

or

$$|cs+d| < 1,$$

i.e., $|\mathsf{A}'s| > 1$. Similarly, if s is outside I then $|\mathsf{A}'s| < 1$. Next we prove

A transformation carries its isometric circle into the isometric circle of the inverse transformation.

The inverse transformation of (13.1–3), being

$$\mathsf{B}s = \frac{-ds+b}{cs-a}. \tag{13.1–8}$$

has the isometric circle I' represented by

$$|cs-a| = 1.$$

Its centre is a/c, its radius $1/|c|$. Now we have for all s, t, related by $t = \mathsf{A}s$

$$|\mathsf{B}'t|\ |\mathsf{A}'s| = 1,$$

where the primes attached to A and B denote differentiation. Hence the interior of I is transformed into the exterior of I' and the exterior of I into the interior of I'. It follows that I is transformed into I'.

The proof can also given by straight-forward computation. In fact, inserting (13.1–8) into (13.1–5) we get

$$\left| \frac{-cds+cb}{cs-a} + d \right| = \left| \frac{cb-ad}{cs-a} \right| = \frac{1}{|cs-a|} = 1.$$

Finally we shall prove

Given two circles I and I' with equal radii and two points s_1 on I and s_1' on I', there is precisely one linear fractional transformation

$$s' = \mathsf{A}s$$

such that I is its isometric circle, I' the image of I and $s_1' = \mathsf{A}s_1$.

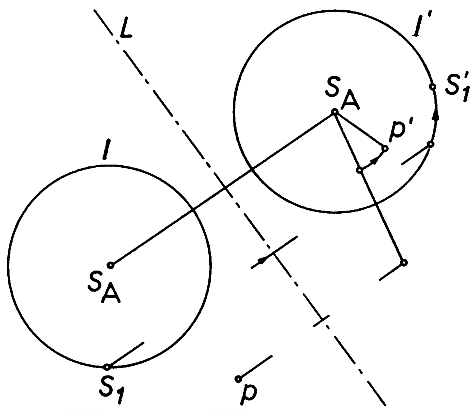


Fig. 13.1–1. Linear fractional transformation with given different isometric circles

Assume first that the centres s_{A} and s_{A}' of I and I' are different, (fig. 13.1–1). We reflect the s-plane in the line L, the perpendicular bisector of the segment connecting s_{A} and s_{A}'. Then we reflect with respect to I' and finally we rotate the plane about s_{A}' such that we obtain the image s_1' of the point s_1. It is clear that we thus obtain the required transformation. If s_{A} and s_{A}' coincide we first reflect in the bisector L of the angle $s_1 s_{\mathsf{A}} s_1'$, (fig. 13.1–2), then in $I = I'$. The uniqueness of the mapping is proved by the following arguments. Since the exterior of I is transformed into the interior of I' the circumferences are percorsed in the same sense. Because of the isometry the corresponding arcs of I and I' are equal.

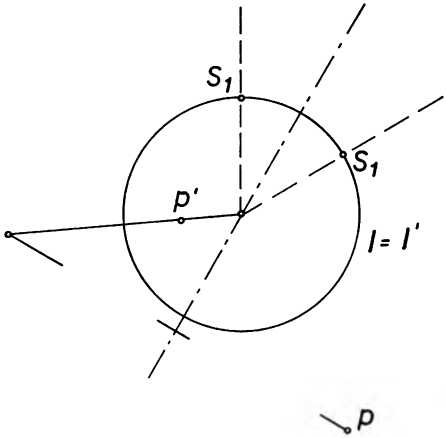


Fig. 13.1–2. Linear fractional transformation with given coincident isometric circles

Hence by assigning the image s_1' on I' of a point s_1 on I to s_1 all images of the points of I are determined. Referring to section 9.3.4 we may conclude to the truth of the assertion.

13.1.4 – Fixed circles

Assume that the set $\mathfrak{A}$ is invariant with respect to a transformation A, i.e., if s is in $\mathfrak{A}$ so is $\mathsf{A}s$ and, conversely, every point of $\mathfrak{A}$ is the image of a point of $\mathfrak{A}$. If P is any transformation we denote the image of $\mathfrak{A}$ as given by P by $\mathfrak{A}^* = \mathsf{P}\mathfrak{A}$. Then we have in a notation which requires no comment

$$\mathsf{PAP}^{-1}\mathfrak{A}^* = \mathsf{PA}\mathfrak{A} = \mathsf{P}\mathfrak{A} = \mathfrak{A}^*.$$

Hence

If $\mathfrak{A}$ *is invariant with respect to* A, *then* $\mathsf{P}\mathfrak{A}$ *is invariant with respect to the conjugate transformation* PAP^{-1} (compare section 9.3.5).

By a *fixed circle* of a transformation A we understand a circle which is (as a whole) invariant with respect to A as well as its interior. Otherwise stated, it is invariant preserving its orientation.

In view of the above remark we may study the configuration of fixed circles for transformations represented by their canonical forms (section 9.3.6), viz.,

$$s' = \kappa s, \qquad s' = s + p, \qquad p > 0.$$

If $\kappa = e^{i\theta}$, $0 < \theta < 2\pi$, the transformation is elliptic and the concentric circles about $s = 0$ are fixed circles. If $\theta = \pi$, the lines through the origin are also invariant, but they are not fixed circles, because the two half planes separated by such a line are interchanged.

If κ is real, positive and $\neq 1$, then the transformation is hyperbolic and the lines through the origin are fixed circles.

In the case of a parabolic transformation $s' = s+p$ the horizontal lines $\operatorname{Im} s =$ constant are fixed circles.

If, finally, $\kappa = re^{i\theta}$, $0 < \theta < 2\pi$, $r \neq 1$, the transformation is loxodromic and there are no fixed circles. In the case $\theta = \pi$ the lines through the origin are invariant, but they are not fixed circles by the reason as above.

Summing up

In the elliptic case the fixed circles are the circles orthogonal to the pencil of the circles through the fixed points.

In the hyperbolic case the fixed circles are the circles through the fixed points.

In the parabolic case the fixed circles constitute a pencil of circles which are tangent at the fixed point to a line through this fixed point.

In order to study the relation between fixed circles and the isometric circle of a transformation we assume that $c \neq 0$. Then the fixed points s_1 and s_2 are finite and in the elliptic or hyperbolic case we may represent the transformation as

$$\frac{s'-s_1}{s'-s_2} = \kappa \frac{s-s_1}{s-s_2}. \qquad (13.1\text{–}9)$$

Making $s \to s_1$ (whence $s' \to s_1$) we find that the value of ds'/ds at s_1 is κ. By a similar argument we find that the value of this derivative at s_2 is $1/\kappa$. Thus we see that if $|\kappa| \neq 1$ there is an increase of length at one fixed point and a decrease at the other. Hence

For the hyberbolic and loxodromic transformations one fixed point is inside the isometric circle, the other outside.

If $|\kappa| = 1$ there is no alteration. Hence

For the elliptic transformations both fixed points are on the isometric circle.

A parabolic transformation with finite fixed point s_0 can be represented by

$$\frac{1}{s'-s_0} = \frac{1}{s-s_0} + p, \qquad p > 0, \qquad (13.1\text{–}10)$$

and reasoning as above we see that ds'/ds is 1 at the fixed point. Hence

For the parabolic transformation the fixed point is on the isometric circle.

In the elliptic and parabolic case the fixed points are also on I'.

In the elliptic transformation with $a+d \neq 0$ the circles I and I' intersect and the line L used in the construction at the end of section 13.1.3 is the common chord. In the parabolic transformation L is the common tangent to I and I' at their point of tangency. If $a+d = 0$ so that I and I' coincide the line L is the line joining the fixed points which are then at the end of a diameter.

A non-loxodromic transformation has a one-parameter family of fixed circles. It consists of the circles with centres on L orthogonal to I (and also to I'). Indeed, being orthogonal to I, such a circle is transformed into itself by an inversion in I; and a reflection in L, a diameter, transforms it again into itself. Thus we see

In a non-loxodromic transformation the isometric circle is orthogonal to the fixed circles.

13.1.5 – The isometric circles of a group

For a properly discontinuous group there exists, by hypothesis, at least one point s_0 such that there are no transforms of s_0 in a sufficiently small neighbourhood of s_0. By applying a suitable transformation and considering a conjugate group (section 9.3.5) we may suppose that $s_0 = \infty$. A neighbourhood of ∞ is the exterior of a circle about the origin. It follows that we can find a number ρ such that all transforms of ∞ with respect to the group are inside an open disc $\mathfrak{K}_\rho$ of radius ρ and centred at the origin.

Since ∞ is not a fixed point for any transformation of the group we have $c \neq 0$. The centre of an isometric circle, being $-d/c$, is congruent to ∞ and lies inside $\mathfrak{K}_\rho$.

Certain relations between the isometric circles of two transformations and the isometric circle of their product will be of use. Employing the notation (9.3–8) we consider two transformations

$$\mathsf{A}_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}, \qquad \mathsf{A}_2 = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}, \tag{13.1–11}$$

with determinants equal to unity. Then

$$\mathsf{A}_2\,\mathsf{A}_1 = \begin{bmatrix} a_2 a_1 + b_2 c_1 & a_2 b_1 + b_2 d_1 \\ c_2 a_1 + d_2 c_1 & c_2 b_1 + d_2 d_1 \end{bmatrix}. \tag{13.1–12}$$

We assume further that $\mathsf{A}_1 \neq \mathsf{A}_2^{-1}$, i.e., $\mathsf{A}_2\,\mathsf{A}_1$ is not the identity. Then the isometric circle of $\mathsf{A}_2\,\mathsf{A}_1$ is given by

$$|(c_2 a_1 + d_2 c_1)s + c_2 b_1 + d_2 d_1| = 1. \tag{13.1–13}$$

The isometric circles of A_1, A_1^{-1}, A_2, A_2^{-1}, $\mathsf{A}_2\mathsf{A}_1$ will be denoted by I_{A_1}, I'_{A_1}, I_{A_2}, I'_{A_2}, $I_{\mathsf{A}_2\mathsf{A}_1}$ respectively; their centres by s_{A_1}, s'_{A_1}, s_{A_2}, s'_{A_2}, $s_{\mathsf{A}_2\mathsf{A}_1}$ respectively. Then, evidently,

$$s_{A_1} = -\frac{d_1}{c_1}, \quad s'_{A_1} = \frac{a_1}{c_1}, \quad s_{A_2} = -\frac{d_2}{c_2},$$
$$s'_{A_2} = \frac{a_2}{c_2}, \quad s_{A_2A_1} = -\frac{c_2 b_1 + d_2 d_1}{c_2 a_1 + d_2 c_1}, \tag{13.1–14}$$

and the corresponding radii

$$r_{A_1} = \frac{1}{|c_1|}, \quad r_{A_2} = \frac{1}{|c_2|}, \quad r_{A_2A_1} = \frac{1}{|c_2 a_1 + d_2 c_1|}. \tag{13.1–15}$$

It follows that

$$r_{A_2A_1} = \frac{1}{|c_1 c_2| \left| \dfrac{a_1}{c_1} + \dfrac{d_1}{c_1} \right|} = \frac{r_{A_2} r_{A_1}}{|s'_{A_1} - s_{A_2}|}, \tag{13.1–16}$$

and

$$s_{A_2A_1} - s_{A_1} = -\frac{c_2 b_1 + d_2 d_1}{c_2 a_1 + d_2 c_1} + \frac{d_1}{c_1} = \frac{c_2}{c_1(c_2 a_1 + d_2 c_1)}, \tag{13.1–17}$$

whence

$$|s_{A_2A_1} - s_{A_1}| = \frac{r_{A_2A_1} r_{A_1}}{r_{A_2}} = \frac{r^2_{A_1}}{|s'_{A_1} - s_{A_2}|}, \tag{13.1–18}$$

or

$$r^2_{A_1} = |s_{A_2A_1} - s_{A_1}||s'_{A_1} - s_{A_2}|. \tag{13.1–19}$$

Each factor in the second member, being the distance between two points of $\mathfrak{R}_\rho$, is less than 2ρ. Hence $r^2_{A_1} < 4\rho^2$,

$$r_{A_1} < 2\rho. \tag{13.1–20}$$

Thus we proved:

The radii of the isometric circles are bounded.

From (13.1–16) we deduce

$$|s'_{A_1} - s_{A_2}| = \frac{r_{A_2} r_{A_1}}{r_{A_2A_1}}.$$

Let I_{A_2} and I'_{A_1} be two different isometric circles with radii $> k > 0$. Then $A_2 A_1$ is not the identity and

$$|s'_{A_1} - s_{A_2}| > \frac{k^2}{2\rho}. \tag{13.1–21}$$

The distance between the centres of two isometric circles with radii exceeding k has thus a positive lower bound. Since the distance of the centres of all such circles from the origin does not exceed ρ their number must be finite. It follows that

The number of isometric circles with radii exceeding a given number is finite. Hence the number of transformations of a group with infinitely many members is enumerable.

Finally we shall prove

Let the properly discontinuous group G contain the infinitely many linear transformations $\mathsf{A}_0 = \mathsf{E}, \mathsf{A}_1, \mathsf{A}_2, \ldots$. *Then the radii* r_n *of the isometric circles of the transformations* A_n *tend to zero as* $n \to \infty$;

$$\lim_{n\to\infty} r_n = 0. \tag{13.1–22}$$

We assume, as always, that there is a neighbourhood about ∞ such that the transforms of ∞ are in $\mathfrak{K}_\rho$. Hence all isometric circles are in a disc $\mathfrak{K}_{3\rho}$, as follows from (13.1–20). Since the exterior of an isometric circle is transformed into the interior of another circle, every point outside $\mathfrak{K}_{3\rho}$ is transformed into a point inside $\mathfrak{K}_{3\rho}$. Hence the images of the exterior of $\mathfrak{K}_{3\rho}$ form a set of discs inside $\mathfrak{K}_{3\rho}$. We contend that these discs are disjunct. For if two discs $\mathfrak{K}_\mathsf{A}$ and $\mathfrak{K}_\mathsf{B}$, being the images of the exterior of $\mathfrak{K}_{3\rho}$ as given by A and B, should have a point s_0 in common then there would be two points s_1 and s_2 outside $\mathfrak{K}_{3\rho}$ such that

$$s_0 = \mathsf{A}s_1, \quad s_0 = \mathsf{B}s_2$$

whence

$$s_2 = \mathsf{B}^{-1}\mathsf{A}s_1,$$

in contradiction to the property of the exterior of $\mathfrak{K}_{3\rho}$.

Consider now a transformation A_n. Its isometric circle be I_n and I_n' be the image of I_n as given by A_n. The radius of these two circles will be denoted by r_n. A disc $\mathfrak{K}_n$ with radius 4ρ about the centre of I_n contains $\mathfrak{K}_{3\rho}$ in its interior. Applying the construction as described at the end of section 13.1.3 the disc $\mathfrak{K}_n$ is transformed into a disc $\mathfrak{K}_n'$ about the centre of I_n', with the radius $\rho_n = r_n^2/4\rho$. The interior of this disc is in the interior of the image of the exterior of $\mathfrak{K}_{3\rho}$. Hence the discs $\mathfrak{K}_n'$ are also in $\mathfrak{K}_{3\rho}$ and are disjoint. Their radii ρ_n must, therefore tend to zero, and the same is true for r_n. This proves the assertion.

Another consequence is the following:

The series

$$\sum_{\nu=1}^{\infty} r_\nu^4 \tag{13.1–23}$$

is convergent.

Since the sum of the areas of the circles $\mathfrak{K}_n'$ does not exceed the area of $\mathfrak{K}_{3\rho}$, we have

$$\sum_{\nu=1}^{\infty} \pi\rho_\nu^2 \leqq 9\pi\rho^2,$$

whence

$$\sum_{\nu=1}^{\infty} r_\nu^4 = 16\rho^2 \sum_{\nu=1}^{\infty} \rho_\nu^2 \leqq 144\rho^4.$$

This proves the assertion.

The result remains true if the point at infinity is a fixed point of an elliptic transformation of order g. The isometric circles are confined to a finite region as before, and the constant ρ exists. There is the difference that a point outside $\mathfrak{K}_n$ may have $g-1$ points outside $\mathfrak{K}_n$ congruent to it. The circles $\mathfrak{K}_n'$ can overlap, but no point can be interior to more than g such circles. Hence

$$\sum_{\nu=1}^{\infty} \pi\rho_\nu^2 \leqq 9g\pi\rho^2$$

and the reasoning is as that before.

13.1.6 – Limit points

We assume again that the group G is such that there are no points congruent to ∞ in a suitable neighbourhood of infinity. By a *limit point* of the group we understand an accumulation point of the centres of the isometric circles of the elements of the group. A point which is not a limit point is called *ordinary*.

The set of limit points is transformed into itself by any transformation of the group.

We observe that the centre of the isometric circle of a transformation of the group is the image of the point at infinity as given by this transformation. Hence the transform of a centre by any transformation of the group is the centre of another isometric circle or is the point at infinity.

Let p be an accumulation point of the centres $s_1, s_2, \ldots$. A transformation A which carries p into p' carries these centres into the points $s_1', s_2', \ldots$, with accumulation point p'. The points of the latter set, with possible exception of the point at infinity, are centres of isometric circles and, therefore, p' is also a limit point. No point which is not a limit point is carried by A into a limit point, since otherwise A^{-1} would carry a limit point into an ordinary point.

The following theorem is of fundamental importance.

In every neighbourhood of a limit point p there are infinitely many points equivalent to an arbitrary point of the plane, except perhaps to p itself and to one other point.

That these exceptions are possible may be shown by the group of transformations

$$s' = \frac{2^n s}{s(2^n-1)+1}, \qquad n = 0, \pm 1, \pm 2, \ldots \tag{13.1–24}$$

The centres of the isometric circles are $-1/(2^n-1)$ and tend to 0 as $n \to \infty$ and to 1 as $n \to -\infty$. Hence 0 and 1 are the limit points. But to $s = 0$ corresponds always $s' = 0$ and to $s = 1$ corresponds $s' = 1$. Thus these points are exceptional. The theorem is true for every properly discontinuous group, for the statement is invariant with respect to the replacing of a group by a conjugate group. Hence it is sufficient to prove it under the restrictive assumption made in the beginning of this section.

In order to prove the theorem we consider a sequence $I_1', I_2', \ldots$ of isometric circles whose centres converge to p. We know that their radii converge to zero. Let $\mathsf{A}_1, \mathsf{A}_2, \ldots$ denote transformations of the group whose isometric circles $I_1, I_2, \ldots$ are transformed into $I_1', I_2', \ldots$ by these transformations respectively. The centres of these circles have at least one accumulation point p' and we may assume that this is the only one, for otherwise we could extract a subsequence having this property. Two cases must be considered:

a) Let $p' \neq p$. Take any point q different from p and p'. For sufficiently large n the circles I_n and I_n' are separated, I_n' in a given neighbourhood of p, p exterior to I_n and q exterior to I_n' and I_n. The transformation A_n carries q into the interior of I_n', that is into the neighbourhood of p. It may happen that $\mathsf{A}_n q = p$. Then p is not a fixed point of A_n, i.e., $\mathsf{A}_n p \neq \mathsf{A}_n q$ is different from p. Since p is outside I_n, $\mathsf{A}_n p$ is inside I_n'. Hence either $\mathsf{A}_n q \neq p$, or $\mathsf{A}_n \mathsf{A}_n q \neq p$ is in I_n', i.e., in a neighbourhood of p.

b) Let $p' = p$ and let q_1, q_2 be different from p, and $q_1 \neq q_2$. If n is sufficiently large, then q_1 and q_2 are outside the isometric circle I_n'. Then $\mathsf{A}_n q_1$ and $\mathsf{A}_n q_2$ are inside these circles. For at most one of the points q_1, q_2 it is possible that $\mathsf{A}_n q_1 = p$ or $\mathsf{A}_n q_2 = p$, if n is sufficiently large. Assuming this for q_1 we have that $\mathsf{A}_n q_2 \neq p$ for all $q_2 \neq q_1$ and thus we see that the theorem holds also in this case.

A first application of this theorem is

If a closed set $\mathfrak{C}$ of points, consisting of more than two points, is transformed into itself by all transformations of the group, then $\mathfrak{C}$ contains all limit points.

Let p denote a limit point and p_1, p_2 two points of $\mathfrak{C}$. At least one has transforms in the vicinity of p, these transforms being also points of $\mathfrak{C}$, by hypothesis. Hence p is an accumulation point of $\mathfrak{C}$ and belongs to $\mathfrak{C}$.

Another application is the theorem

The set of limit points, containing more than two points, is a perfect set.

This means that the set is closed and that every point of the set is an accumulation point. The first property follows at once. Indeed, since each limit point contains an infinity of centres of isometric circles in any of its neighbourhoods, an accumulation point of limit points has also an infinite number of centres of isometric circles in each of its neighbourhoods. It remains to show that any limit point β has an infinity of limit points in its vicinity. If p_1 and p_2 are two other limit points, at least one has an infinite number of transforms in a neighbourhood of p. Since these transforms are again limit points (second theorem of this section) the second property is now also clear.

The fixed points of a parabolic or hyperbolic transformation of the group are limit points.

If A is such a transformation then the powers A^n, $n = \pm 1, \pm 2, \ldots$ are all different. In the hyperbolic case the fixed points are within I_n or I_n', the isometric circles of A^n or A^{-n} respectively and their radii tend to zero. In the parabolic case the fixed point is on I_n and I_n' and the same argument may be employed.

13.2 – The fundamental domain

13.2.1 – Definition and simple properties

An open set is called a *fundamental domain* of a group G, if it does not contain two equivalent points, and such that every neighbourhood of any point on the boundary contains points equivalent to points in the given set. Latei we shall adjoin to the domain a part of its boundary, but for the present it shall consist only of inteiior points. It may be a region or it may comprise two or more disconnected parts. It is clear that a fundamental domain is in no wise unique. Any set congruent to it will serve as a fundamental domain. We can replace any part of it by a congruent part and still have a fundamental domain.

The transforms of a fundamental domain by two distinct transformations of the group do not overlap.

It is sufficient to prove the following assertion: If no two points of an open set $\mathfrak{A}$ are equivalent, the transforms of the set by two distinct transformations do not overlap. Suppose that the transformations P and Q carry $\mathfrak{A}$ into two overlapping sets. Any point s_0 in the common part is the transform by P of a point s_1 of $\mathfrak{A}$ and by Q of a point s_2 of $\mathfrak{A}$. If s_1 and s_2 are different for any s_0, then s_1 and s_2 are equivalent points of $\mathfrak{A}$ by the transformation QP^{-1}, which is impossible. If s_1 and s_2 coincide for every s_0 in the common part, then QP^{-1} has infinitely many fixed points and is, therefore the identity, i.e., $\mathsf{P} = \mathsf{Q}$.

13.2.2 – Construction of a fundamental domain

We assume again that the group G is such that within a suitable neighbourhood of ∞ there is no point equivalent to ∞. Our further considerations are based on the following theorem

Let $\mathfrak{R}$ denote the largest open set outside all isometric circles of the group. Then the join of $\mathfrak{R}$ and the sets congruent to $\mathfrak{R}$ extend into the neighbourhood of any point of the plane.

The set $\mathfrak{R}$ can be characterized as follows: If s is a point of $\mathfrak{R}$ then there is a neighbourhood $\mathfrak{U}(s)$ which is outside all isometric circles and every point having this property belongs to $\mathfrak{R}$.

Suppose that s_0 is a point having a neighbourhood $\mathfrak{U}(s_0)$ which contains no point of $\mathfrak{R}$, nor points equivalent to points of $\mathfrak{R}$. Then all transforms of $\mathfrak{U}$ contain neither points of $\mathfrak{R}$, nor points equivalent to points of $\mathfrak{R}$. Since ∞ is a point of $\mathfrak{R}$ and the centres of the isometric circles are equivalent to ∞ it follows that $\mathfrak{U}$ and its transforms contain centres of no isometric circles. Hence the points of $\mathfrak{U}$ and of its transforms are ordinary points. We assume that s_0 is not a point of $\mathfrak{R}$, i.e., s_0 is inside or on the boundary of an isometric circle. Consider a circle C about s_0 within $\mathfrak{U}$. Let A denote a transformation whose isometric circle I_A has s_0 for an interior or boundary point. The centre of I_A is exterior to C. Such a transformation exists, for the interior points of C and all its transforms are ordinary points. We know (section 13.1.3) that A is equivalent to an inversion in I_A followed by a reflection in a line (and possibly a rotation). The magnitude of C_1, the inverse of C, is determined by inversion. Let r_1 denote the radius of C_1, r that of C and r_A that of I_A. The largest and the smallest distance of the points of the circumference to the centre of I_A are $a+r$ and $a-r$, respectively, where a is the distance between the centres of I_A and C. These points are transformed into others, with distances

$$\frac{r_\mathsf{A}^2}{a+r}, \quad \frac{r_\mathsf{A}^2}{a-r}$$

to the centre of I_A. A simple geometric consideration reveals that the radius of the transformed circle C_1 is

$$r_1 = \tfrac{1}{2} r_\mathsf{A}^2 \left(\frac{1}{a-r} - \frac{1}{a+r} \right) = r_\mathsf{A}^2 \frac{r}{a^2 - r^2}.$$

Since the centre of C is within or on the boundary of I_A, we have

$$r_1 \geqq r_\mathsf{A}^2 \frac{r}{r_\mathsf{A}^2 - r^2} = \frac{r}{1 - \left(\dfrac{r}{r_\mathsf{A}}\right)^2}.$$

Taking into account (13.1–20), we finally have

$$r_1 \geqq kr, \qquad k = \frac{1}{1 - \dfrac{r^2}{4\rho^2}} > 1.$$

If we apply to C_1 a transformation whose isometric circle has the centre of C_1 as an interior or a boundary point, we get a circle C_2 of radius r_2, with

$$r_2 \geqq kr_1 \geqq k^2 r.$$

Continuing in this way we prove the existence of a circle C_n congruent to C and of radius $\geqq k^n r$. By taking n sufficiently large the circle C_n will contain points of $\mathfrak{R}$ exterior to the finite disc $\mathfrak{K}_{3\rho}$ in which the isometric circles lie. These points are equivalent to points of $\mathfrak{U}(s_0)$. This contradiction proves the theorem.

A direct consequence is

The largest open set outside all isometric circles is a fundamental domain.

It is clear that no two points of $\mathfrak{R}$ are equivalent, for otherwise one should lie within an isometric circle. It is not possible to extend $\mathfrak{R}$ to a larger set satisfying the definition of a fundamental domain, for the previous theorem states that in the vicinity of every point which does not belong to $\mathfrak{R}$ there are points equivalent to points of $\mathfrak{R}$.

From the definition of $\mathfrak{R}$ we may infer that

To any open set enclosing a limit point belongs an infinite number of transforms of the domain $\mathfrak{R}$.

This is evident, for there are infinitely many isometric circles within the open set containing a limit point. Each of these circles encloses a domain congruent to $\mathfrak{R}$ and the various transforms are different domains.

Any closed set $\mathfrak{C}$ *not containing limit points of the group is covered by a finite number of transforms of* $\mathfrak{R}$. *These domains fit together without gaps.*

There are a finite number of isometric circles containing points of $\mathfrak{C}$. Indeed, if not, there are circles of arbitrarily small radius and their centres have a point of $\mathfrak{C}$ as accumulation point. A transformation A carries $\mathfrak{R}$ into $\mathfrak{R}_{\mathsf{A}}$, lying inside I'_{A}, the isometric circle of A^{-1}. If I'_{A} contains points of $\mathfrak{C}$, $\mathfrak{R}_{\mathsf{A}}$ may contain points of $\mathfrak{C}$. If $\mathfrak{C}$ is exterior to I'_{A} then $\mathfrak{R}_{\mathsf{A}}$ contains no points of $\mathfrak{C}$. Hence the number of transforms of $\mathfrak{R}$ which have points with $\mathfrak{C}$ in common is not greater than the number of isometric circles which contain points of $\mathfrak{C}$. According to the first theorem of this section there are points of some transform of $\mathfrak{R}$ in a neighbourhood of an arbitrary point of $\mathfrak{C}$. It follows that the transforms of $\mathfrak{R}$ having points in common with $\mathfrak{C}$ fit together without gaps. $\mathfrak{C}$ is completely covered, except for the boundaries separating the various transforms.

13.2.3 – Boundary points of the fundamental domain

We consider the fundamental domain as constructed in the previous section. A point on the boundary of $\mathfrak{R}$ is a point p not belonging to $\mathfrak{R}$, but such that in any neighbourhood of p there are points of $\mathfrak{R}$. It is clear that p cannot lie inside an isometric circle. If p is an ordinary point it lies on one or more isometric circles. There is but a finite number of isometric circles penetrating into a neighbourhood of an ordinary point. Hence p possesses a neighbourhood exterior to all isometric circles other than those which pass through p.

We may divide the boundary points into three classes:

1) p is a limit point of the group;

2) p is an ordinary point and lies on a single isometric circle;

3) p is an ordinary point and lies on two or more isometric circles. In this case p is called a *vertex*.

It is convenient to include under 3) the following special case: if p is the fixed point of an elliptic transformation of period two (i.e., a transformation whose square is the identity), so that, although p lies on a single isometric circle, it separates two congruent arcs on the circle, we shall classify p under 3)

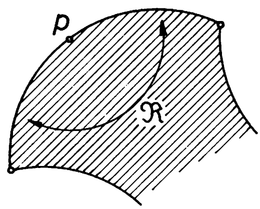


Fig. 13.2–1. Vertex of a fundamental domain on a single isometric circle

Let p denote a point of the second class. It lies on an isometric circle I_{A} and we shall denote by p' the point on I'_{A} into which A carries p. We assert that p' is also a boudary point of the second class. The case $p = p'$ can arise only if I_{A} and I'_{A} coincide and p is a fixed point of the resulting elliptic transformation of period two. This case, however, has been included in 3), (fig. 13.2–1).

Suppose that p' is within an isometric circle I_{B}. Then B magnifies lengths at p'. Since A carries p into p' without alteration of lengths, BA magnifies lengths at p. Hence p is inside I_{BA} which is contrary to the hypothesis that p lies on an isometric circle. It follows that also p' belongs to the class 2)

It is clear that in a sufficiently small neighbourhood of p there are no points of isometric circles other than on I_A, for there is no other isometric circle in the vicinity of p other than I_A. Hence the points on I_A near p are also boundary points of the second class and so are the congruent points on I'_A. Thus we proved

The boundaty points of $\mathfrak{R}$ of the second class form a set of bounding circular arcs which are congruent in pairs. Two such congruent arcs are equal in length.

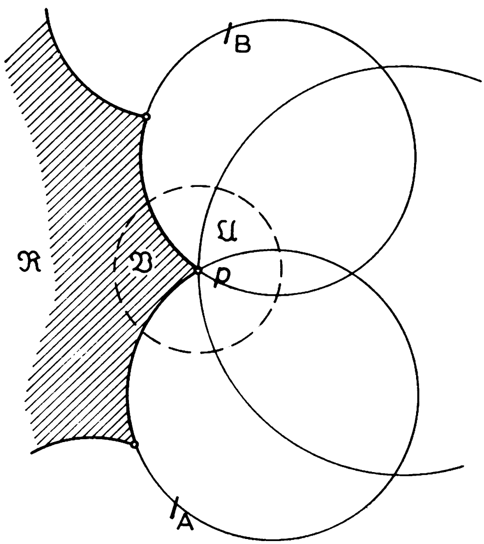


Fig. 13.2–2. At a vertex two sides of the fundamental domain meet

These arcs may consist of entire circles or they may terminate at points of the first or the third class. They are called the *sides* of $\mathfrak{R}$.

Now we turn our attention to the points of the third class, the vertices. Through a point p of this class there pass a finite number of isometric circles, (fig. 13.2–2). Let $\mathfrak{U}$ be a neighbourhood of p, such that all isometric circles other than those through p are outside $\mathfrak{U}$. The isometric circles through p divide $\mathfrak{U}$ into a finite number of parts. One of these parts, say $\mathfrak{V}$, belongs to $\mathfrak{R}$, since p is a boundary point. The two arcs which bound $\mathfrak{V}$ are on isometric circles I_A and I_B and are a part of the boundary. The points on these arcs other than p belong to the second class. Hence

At a vertex of $\mathfrak{R}$ two sides meet.

For the sake of convenience we shall add to $\mathfrak{R}$ certain points of its boundary. Of two congruent sides, one, exclusively of its end points, may be added without including points congruent to points previously in $\mathfrak{R}$. A vertex where bounding arcs meet may be congruent to several other vertices, one of them may be adjoined. The resulting set will be considered as a fundamental domain in a somewhat wider sense.

13.2.4 – Cycles of boundary points

Let us think of the boundary of $\mathfrak{R}$ as being traced in the positive sense. In passing through a vertex, we proceed along one side to the vertex and then proceed along a second side from the vertex. We consider the vertex as the end of the former and the beginning of the latter. When a side is carried into its congruent side, the beginning and the end of the former are carried into the end and the beginning, respectively, of the latter, for the direction around the isometric circle is reversed.

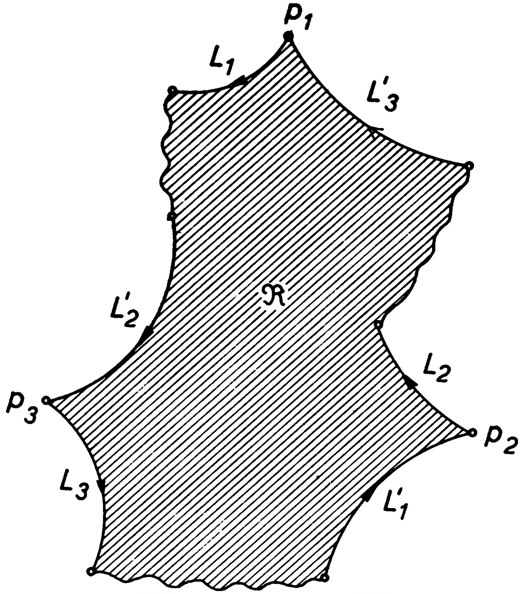


Fig. 13.2–3. A cycle of vertices of a fundamental domain

Let L_1 be the side beginning at a vertex p_1. This side is carried by a transformation A_1 into the congruent side L'_1, p_1 being carried into p_2 at the end of L'_1, (fig. 13.2–3). There is a side L_2 beginning at p_2 which is carried by some transformation A_2 into the congruent side L'_2, p_2 being carried into p_3 at the end of L_2'. Continuing in this way we shall return to p_1 after a finite number of steps. Suppose, on the contrary, that an infinite number of vertices $p_2, p_3, \ldots$, are equivalent to p_1. The transformation B_n which carries p_1 to p_n has an isometric circle passing through p_1; for, if p_1 is outside the isometric circle of B_n, then p_n is inside the isometric circle of B_n^{-1}, which is impossible. Then an infinite number of the isometric circles of $\mathsf{A}_2, \mathsf{A}_3, \ldots$, pass through p_1, which is contrary to the fact that p_1 is an ordinary point. A complete set of equivalent vertices of a fundamental domain is called an *ordinary cycle.*

Let now $\mathsf{A}_1, \ldots, \mathsf{A}_m$ be the transformations which carry p_1 to $p_2, \ldots, p_m$ to p_1, respectively. The transformation

$$\mathsf{B} = \mathsf{A}_m \ldots \mathsf{A}_1$$

carries p_1 into itself. Hence either B is the identity, or B is an elliptic transformation, for fixed points of hyperbolic and loxodromic transformations lie within isometric circles and the fixed point of a parabolic transformation is a limit point.

The transformations

$$\begin{array}{l} \mathsf{A}_m, \\ \mathsf{A}_m \mathsf{A}_{m-1}, \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \mathsf{A}_m \mathsf{A}_{m-1} \ldots \mathsf{A}_1 \end{array} \tag{13.2–1}$$

carry $p_m, p_{m-1}, \ldots, p_2, p_1$, respectively, into p_1. The domains $\mathfrak{R}_m$, $\mathfrak{R}_{m-1}, \ldots, \mathfrak{R}_2, \mathfrak{R}_1$, respectively, in which these transformations carry $\mathfrak{R}$, fit together at p_1, (fig. 13.2–4). Thus A_m carries p_m to p_i, $\mathfrak{R}_m$ being adjacent to $\mathfrak{R}$ along the arc L_m' ending at p_1. In general A_i carries p_i into p_{i+1}

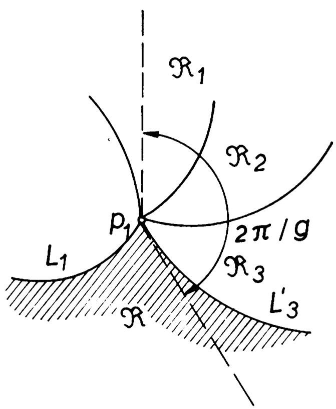


Fig. 13.2–4. Transformation of all vertices of a cycle into one vertex

and carries $\mathfrak{R}$ into a domain abutting on $\mathfrak{R}$ along a side ending in p_i, whence the transform of these two domains by $\mathsf{A}_m \ldots \mathsf{A}_{i+1}$, namely $\mathfrak{R}_i$ and $\mathfrak{R}_{i+1}$ are adjacent along an arc issuing from p_i. In proceeding counterclockwise around p_1, starting from $\mathfrak{R}$, we encounter the adjacent domains $\mathfrak{R}_m, \mathfrak{R}_{m-1}, \ldots, \mathfrak{R}_2, \mathfrak{R}_1$. The curvilinear angle of $\mathfrak{R}$ at p_n is carried into an equal angle of $\mathfrak{R}_n$ at p_1.

Since there can be no overlapping of congruent domains, there are two possibilities. First $\mathfrak{R}_1$ may coincide with $\mathfrak{R}$ and the domains $\mathfrak{R}$, $\mathfrak{R}_m, \ldots, \mathfrak{R}_2$ completely fill up the angle about p_1. Then B is the identity and the sum of the angles at the vertices $p_1, \ldots, p_m$ is equal to 2π. If $\mathfrak{R}_1$ does not coincide with $\mathfrak{R}$, then B is an elliptic transformation. An elliptic transformation amounts to a rotation in the neighbourhood of the fixed point through a certain angle, which is not zero. Carrying $\mathfrak{R}$ into $\mathfrak{R}_1$ requires that this angle be equal to the sum of the angles at the vertices. On applying B the domains $\mathfrak{R}, \mathfrak{R}_m, \ldots, \mathfrak{R}_2$ are carried into adjacent domains, filling out more of the angle about p_1 counterclockwise from $\mathfrak{R}_1$. After a finite number, g, of applications of B the angle about p_1 is completely filled up and $\mathsf{B}^g\mathfrak{R}$ coincides with $\mathfrak{R}$. It follows that there are no vertices of $\mathfrak{R}$ equivalent to p_1 other than $p_2, \ldots, p_m$. Indeed, a transformation carrying any other vertex to p_1 carries $\mathfrak{R}$ into a domain overlapping the domains which fill out the angle about p_1, which is impossible. Hence

The sum of the angles at the vertices of an ordinary cycle is $2\pi/g$, where g is an integer. If $g > 1$ each vertex of the cycle is a fixed point of an elliptic transformation of period g.

If a side of $\mathfrak{R}$ terminates at a limit point various situations may arise.

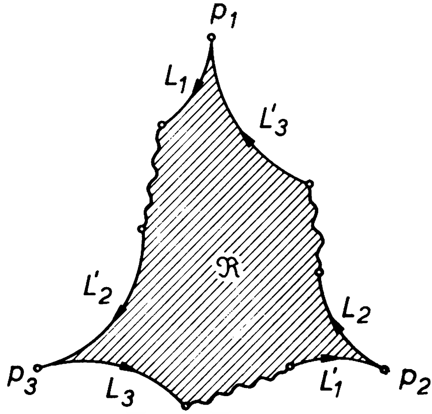


Fig. 13.2–5. A cycle of parabolic vertices

We consider only a particular case. We suppose that two sides meet in p_1. Let L_1 be the side beginning at p_1 and be carried by a transformation A_1 into a side L_1'. Let L_2 the side, if any, begin at p_2 and let A_2 carry this into L_2', etc. We now make the assumption that after arriving at a finite number of equivalent points $p_2, \ldots, p_m$ we may return to p_1 and the side L_1, (fig. 13.2–5). If this assumption holds, we say that $p_1, \ldots, p_m$ constitute a *parabolic cycle* and each point of the cycle is called a *parabolic point.*

The transformation $\mathsf{B} = \mathsf{A}_m \ldots \mathsf{A}_1$ carries p_1 into itself, whence B is either elliptic or parabolic. The above reasoning can be repeated in quite the same way to show that the transformations (13.2–1) carry $p_m, p_{m-1}, \ldots, p_1$ respectively, into p_1 and carry $\mathfrak{R}$ into regions $\mathfrak{R}_m, \ldots, \mathfrak{R}_2, \mathfrak{R}_1$ fitting together along arcs issuing from p_1, (fig. 13.2–6). The

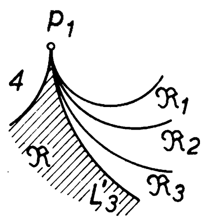


Fig. 13.2–6. Transformation of all vertices of a parabolic cycle into one vertex

angle between the sides of $\mathfrak{R}_i$ which meet at p_1 is equal to the angle between the sides of $\mathfrak{R}$ which meet at p_i. B carries $\mathfrak{R}$ into $\mathfrak{R}_1$, the side L_1 being carried into the side L_1' which separates $\mathfrak{R}_1$ and $\mathfrak{R}_2$. If B is elliptic then L_1 and L_1' meet at an angle different from zero. By repeating B a finite number of times, the angle about p_1 is covered by a finite number of domains, which is contrary to the third theorem of section 13.2.2. Thus we may infer that B is parabolic. It follows that L_1 and L_1' are tangent at the fixed point p_1. Then the arcs bounding the intervening domains are also tangent to L_1. The angle between the sides which meet at each point of a parabolic cycle is zero. Summing up we have

The sides of $\mathfrak{R}$ which meet at parabolic points are tangent. There are an infinite number of domains congruent to $\mathfrak{R}$, each having two sides which meet at the parabolic point and are tangent to the sides of $\mathfrak{R}$.

13.2.5 – Fixed points at infinity

We have supposed hitherto that a given group has been replaced by a conjugate group, such that there is a neighbourhood of infinity which does not contain points equivalent to ∞. This involves no loss of generality. As a consequence *every properly discontinuous group possesses a fundamental domain.*

In practical situations we are often faced with groups having ∞ as a fixed point of some of their transformations and the replacing of the group by a conjugate may introduce complications. We shall describe a method which enables us to find a fundamental domain in this case.

It is clear that all transformations of a group G which leave the point ∞ invariant form a subgroup G_∞ of G.

A transformation U *of the subgroup* G_∞ *carries an isometric circle into an isometric circle.*

Let I_A be the isometric circle of a transformation A and let U denote the transformation

$$s' = \kappa s + c.$$

Since $ds'/ds = \kappa$ this transformation multiplies lengths by $|\kappa|$. Let p denote a point on the image of I_A as given by U. If $p' = \mathsf{U}^{-1}p$, then p' is on I_A, lengths at p being multiplied by $|\kappa|^{-1}$. The transformation A carries p' to p'' without altering lengths. Finally U multiplies lengths at p'' by $|\kappa|$. The result is that UAU^{-1} has not altered lengths at p; hence p is on the isometric circle of UAU^{-1}.

The announced construction may be stated as follows

Let $\mathfrak{R}_\infty$ *be a fundamental domain for the subgroup* G_∞ *of all transformations of* G *having a fixed point at infinity. We assume that the sides of* $\mathfrak{R}_\infty$ *are congruent in pairs and that the transforms of* $\mathfrak{R}_\infty$ *cover the open plane. Let* $\mathfrak{R}$ *consist of all that part of* $\mathfrak{R}_\infty$ *which is exterior to all isometric circles of the group. Then* $\mathfrak{R}$ *is a fundamental domain of the group.*

A transformation of G_∞ carries a point of $\mathfrak{R}_\infty$ into a point of a domain congruent to $\mathfrak{R}_\infty$ and hence outside $\mathfrak{R}$. Any other transformation of the group carries a point of $\mathfrak{R}$ into the interior of an isometric circle and hence outside $\mathfrak{R}$.

That part of a side of $\mathfrak{R}_\infty$ which is outside to all isometric circles, and so is also a side of $\mathfrak{R}$, is, from the previous theorem, congruent to a side which is also exterior to all isometric circles. An ordinary boundary point of $\mathfrak{R}$ lying on an isometric circle I_A is carried by A into a point p' on I'_A. It is easily shown that p' is interior to no isometric circle (section 13.2.3). If p' lies in $\mathfrak{R}_\infty$ it is a boundary point of $\mathfrak{R}$; if not it lies in a congruent domain $\mathsf{U}\mathfrak{R}_\infty$ and the equivalent point $p'' = \mathsf{U}^{-1}p'$ lies in $\mathfrak{R}_\infty$ and is a boundary point of $\mathfrak{R}$. It follows that the sides of $\mathfrak{R}$ which lie on isometric circles are congruent in pairs. A neighbourhood of a limit point contains points congruent to any point of the plane (with the possible exception of two points), hence also to points of $\mathfrak{R}$. This follows from the second theorem of section 13.1.6.

13.3 – **Fuchsian groups**

13.3.1 – Function groups

If $\mathfrak{D}$ is a region in the s-plane which is transformed into itself by all transformations of a properly discontinuous group, then the group is called a *function group*. Groups of this kind play an important part in subsequent considerations. The region $\mathfrak{D}$ is called a *region of discontinuity* of the group.

The intersection $\mathfrak{R}_0$ of a region of discontinuity $\mathfrak{D}$ and the fundamental domain $\mathfrak{R}$ is not empty and is again a fundamental domain.

Not all of $\mathfrak{R}$ can be exterior to $\mathfrak{D}$, for then the transforms of $\mathfrak{R}$ would be exterior to $\mathfrak{D}$, contrary to the first theorem of section 13.2.2. If the boundary point p is a limit point, there are points equivalent to points of $\mathfrak{R}_0$ in any neighbourhood of p, by the second theorem of section 13.1.6. If p is an ordinary point on the boundary of $\mathfrak{R}_0$, it lies in $\mathfrak{D}$ and there are points equivalent to points of $\mathfrak{R}$ in its vicinity, for it is also a boundary point of $\mathfrak{R}$. But these points are equivalent to points of $\mathfrak{R}_0$ since the transforms of all other points of $\mathfrak{R}$ are exterior to $\mathfrak{D}$. Since no points of $\mathfrak{R}_0$ are congruent, $\mathfrak{R}_0$ is a fundamental domain.

The sides of $\mathfrak{R}_0$ are congruent in pairs.

If $\mathfrak{R}_0$ should be bounded entirely by limit points (or consists of the whole open plane if there are no limit points) it would coincide with $\mathfrak{D}$. Then no two points of $\mathfrak{D}$ would be equivalent, which is impossible, provided the group does not reduce to the identity. Hence $\mathfrak{R}_0$ is bounded in part by sides. Each side is congruent to some side of $\mathfrak{R}$ which is also in $\mathfrak{D}$ and so it is a side of $\mathfrak{R}_0$.

Among the function groups those with *a principal circle* are very important. These groups carry the interior of a circle into itself.

Accordingly the function groups are divided into the following classes:

(i) *Elementary groups.* These are the finite groups and the groups with one or two limit points.
(ii) *Fuchsian groups.* These are the groups with a principal circle.
(iii) *Kleinian groups.* A function group belongs to this class if it does not belong to one of the preceding classes.

There is a certain amount of overlapping between (i) and (ii). Certain of the elementary groups possess principal circles, for example the non-loxodromic cyclic groups.

13.3.2 – Fuchsian groups

A Fuchsian group is a properly discontinuous group each of whose transformations carries a certain circle Ω into itself and carries each of the parts into which the circle Ω divides the plane into itself. The circle is then a *principal circle.*

We may interpret the transformations of a Fuchsian group as the automorphisms of the interior of the unit circle. It has been pointed out in section 9.5.1 that there are no loxodromic transformations and in 9.5.3 we stated that the transformations are elliptic with their fixed points inside and outside the circle Ω, parabolic with fixed point on Ω and hyperbolic with both fixed points on Ω.

It follows from the last theorem of section 13.1.4 that the *isometric circles* are orthogonal to Ω. Hence their part within Ω may be considered as hyperbolic straight lines.

From the third theorem of section 13.1.6 we may infer that

The limit points of a Fuchsian group are on the principal circle.

When we make an inversion in the principal circle, each isometric circle, being orthogonal to Ω, is carried into itself and its interior and exterior go, respectively, onto its interior and exterior, (fig. 13.3–1). A point of $\mathfrak{R}$ being exterior to all isometric circles, is carried into another point of $\mathfrak{R}$. Hence

An inversion in the principal circle carries the fundamental domain into itself.

It follows that points in a neighbourhood of the centre of the principal circle (provided this is not a straight line) belong to $\mathfrak{R}$.

Let $\mathfrak{R}_0$ denote the part of $\mathfrak{R}$ inside the principal circle and $\mathfrak{R}_0'$ the part of $\mathfrak{R}$ outside. $\mathfrak{R}_0'$ is the inverse of $\mathfrak{R}$ in the principal circle; its sides and vertices are the inverses of the sides and vertices of $\mathfrak{R}_0$. If a transformation of the group is made, $\mathfrak{R}_0$ is carried into a domain in the interior of the principal circle and $\mathfrak{R}_0'$ into a domain in the exterior. As a consequence of the last theorem of section 9.4.2 we can state that the two transformed domains are inverse with respect to the principal circle.

Thus, in investigating the fundamental domain it will suffice to study the domain $\mathfrak{R}_0$ inside the principal circle. An inversion in the principal circle will then furnish the corresponding results for $\mathfrak{R}_0'$. In the sequel we suppose that the principal circle is not a straight line.

The domain $\mathfrak{R}_0$ is simply connected.

A straight line segment from the centre of the principal circle to any point of $\mathfrak{R}_0$ lies entirely inside $\mathfrak{R}_0$. Hence $\mathfrak{R}_0$ is connected. Any closed curve inside $\mathfrak{R}_0$ can be shrunk continuously to an interior point without crossing the boundary. Hence its winding number with respect to any exterior point is zero.

Any closed set lying entirely within the principal circle is covered by a finite number of transforms of $\mathfrak{R}_0$. These domains fit together without gaps.

This is a direct consequence of the fourth theorem of section 13.2.2.

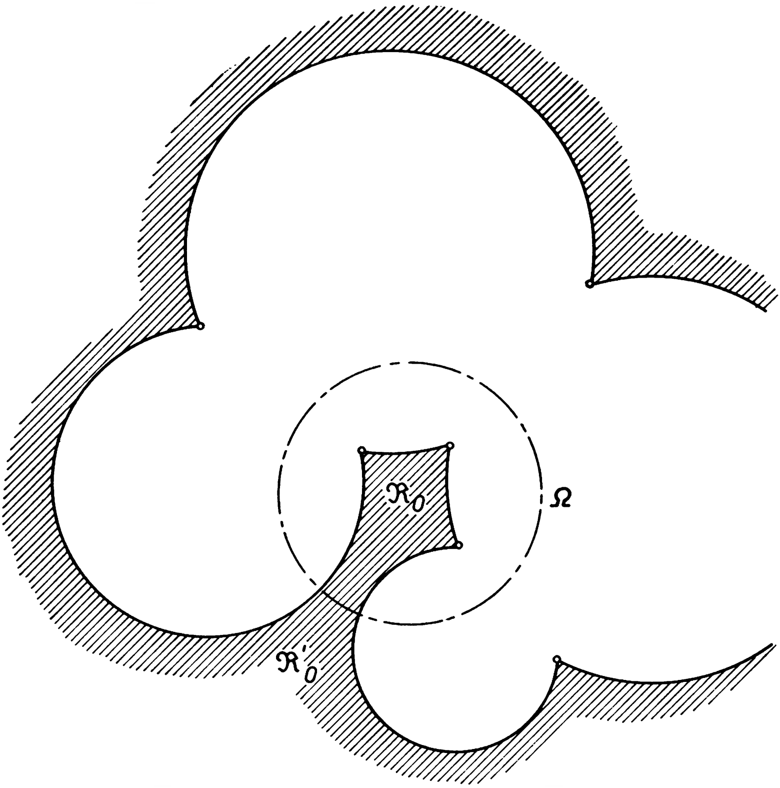


Fig. 13.3–1. The fundamental domain of a Fuchsian group

The transforms of $\mathfrak{R}_0$ fill up, without gaps, the whole interior of the principal circle. They cluster in infinite number about each limit point of the group.

A circle C concentric with the principal circle and of smaller radius is covered by $\mathfrak{R}_0$ and a finite number of domains congruent to $\mathfrak{R}_0$, (third theorem of section 13.2.2). By taking the radius of C arbitrarily near that of Ω any given interior point of the principal circle can be enclosed. The second part of the theorem is a consequence of the third theorem of section 3.2.2.

If a Fuchsian group possesses a fundamental domain $\mathfrak{F}$ whose transforms cover the neighbourhood of each of its boundary points and which, together with its boundary consist of interior points of the principal circle, then the transforms of $\mathfrak{F}$ fill up, without overlapping, the whole interior of the principal circle.

A finite number of transforms of $\mathfrak{R}_0$ cover $\mathfrak{F}$ and its boundary completely (section 13.2.2.). We may carry the portion of $\mathfrak{F}$ in each domain $\mathfrak{R}_i$ into $\mathfrak{R}_0$ by means of a transformation which carries $\mathfrak{R}_i$ into $\mathfrak{R}_0$. These transforms of parts of $\mathfrak{F}$ do not overlap, since no two points of $\mathfrak{F}$ are congruent. Suppose there is a point s_0 in the interior of $\mathfrak{R}_0$ exterior to all the transformed parts of $\mathfrak{F}$. Let s_1 be the nearest point of one of the parts in $\mathfrak{R}_0$. Then in any neighbourhood of s_1 are points which are not covered and which are, therefore, not equivalent to points of $\mathfrak{F}$. On carrying this part back to $\mathfrak{F}$, the point s_1 goes into a boundary point of $\mathfrak{F}$ which has in its neighbourhood points not equivalent to points of $\mathfrak{F}$. This is a contradiction. Hence the transforms of $\mathfrak{F}$ cover $\mathfrak{R}_0$ completely and the transforms of $\mathfrak{F}$, then, cover all transforms of $\mathfrak{R}_0$. These fill up the interior of the principal circle.

In addition we have:

Under the assumptions of the previous theorem the domain $\mathfrak{R}_0$ *lies in the interior of the principal circle.*

Indeed, if $\mathfrak{R}_0$ has a boundary point on Ω, so has one of the parts covering $\mathfrak{R}_0$. But in $\mathfrak{F}$ and on the boundary of $\mathfrak{F}$ there are no points equivalent to points on Ω.

Finally we shall prove

If there are more than two limit points, either the set of limit points consists of all points of the principal circle, or the set of limit points is nowhere dense on the principal circle.

It follows from the fourth theorem of section 13.1.6. that the set of limit points is perfect. Assume now that not all points of Ω are limit points. Then we have to show that not all points on any arc of Ω are limit points. Let s_0 be a point of Ω which is not a limit point. Then the points near s_0 are also ordinary points, i.e., the points of a sufficiently small arc $\mathfrak{a}$ on Ω through s_0 are ordinary. Let p be a limit point. According to the second theorem of section 13.1.6 there are points equivalent to points of $\mathfrak{a}$ in any neighbourhood of p. These points are ordinary and lie on Ω. This proves the theorem.

On the basis of this theorem we may classify the Fuchsian groups as follows:

(i) *Fuchsian groups of the first kind.* For these groups every point on the principal circle is a limit point.

(ii) *Fuchsian groups of the second kind.* The limit points of these groups are nowhere dense on the principal circle.

It is clear that a Fuchsian group is of the second kind, if the fundamental domain $\mathfrak{R}$ contains in its interior a point of the principal circle, for a point inside $\mathfrak{R}$ is ordinary.

13.3.3 – Classification of an algebraic Riemann surface

In section 12.6.7 we proved that the universal covering surface $\tilde{\mathfrak{F}}$ of an algebraic Riemannn surface $\mathfrak{F}$ is homeomorphic, either to the extended plane, or to the whole open s-plane, or to the interior of a circle. In the first case the Riemannian surface is called *elliptic*, in the second case *parabolic* and in the third case *hyperbolic*.

As we pointed out in section 13.1.1 the covering transformations constitute a group which is isomorphic to a properly discontinuous group of automorphisms of the canonical region $\mathfrak{C}$.

1) Let us consider first an elliptic surface. Then $\mathfrak{C}$ is the whole sphere and the automorphisms are fractional linear transformations. Since the covering group has no fixed points, the group G of covering transformations can only be the identity. That means that $\tilde{\mathfrak{F}}$ and $\mathfrak{F}$ are homeomorphic, i.e., $\mathfrak{F}$ is homeomorphic to a sphere and, therefore, of genus 0 (section 12.5.3). Conversely, let $\mathfrak{F}$ be of genus 0. If $\mathfrak{F}$ is an n-sheeted covering surface of the z-plane, and if its ramification number is m, then, according to (12.5–7),

$$m+2 = 2n. \tag{13.3–1}$$

Since $\tilde{\mathfrak{F}}$ is closed, there are only a finite number of points of $\tilde{\mathfrak{F}}$ above a point of $\mathfrak{F}$. If this number is k then also

$$km+2 = 2kn, \tag{13.3–2}$$

for the surface $\tilde{\mathfrak{F}}$ lies unramified over $\mathfrak{F}$.

Subtracting corresponding members of (13.3–1) and (13.3–2) yields

$$(k-1)(m-2n) = 0$$

and it follows that $k = 1$. Hence $\tilde{\mathfrak{F}}$ is homeomorphic to $\mathfrak{F}$, i.e., to a sphere and we infer that $\mathfrak{F}$ is elliptic. Thus

An algebraic Riemann surface is elliptic if and only if it is of genus zero.

2) Next we assume that $\tilde{\mathfrak{F}}$ is homeomorphic to the open plane. The covering transformations are isomorphic to a group of automorphisms of the s-plane having one fixed point. Without loss of generality we may assume that this is the point ∞. Hence G is either the identity, or the cyclic group

$$s' = s+n,\ n = 0,\ \pm 1, \ldots, \tag{13.3–3}$$

or the group

$$s' = s+m_1\omega_1+m_2\omega_2 \tag{13.3–4}$$

where m_1, m_2 are integers and ω_2/ω_1 is complex.

The first case rules out, for $\tilde{\mathfrak{F}}$ is closed if G is the identity. In the second case a fundamental domain is a parallel strip and identifying congruent sides, we obtain a cylinder. Since the fundamental domain is homeomorphic to $\mathfrak{F}$, and $\mathfrak{F}$ is closed, also this case rules out. There remains the third case. Then a fundamental domain is a period parallelogram, and, identifying congruent sides, we obtain a torus. Thus it turns out that *a parabolic surface is of genus* 1. The converse is also true, but the proof is not straightforward. It may be deduced from consideration about hyperbolic surfaces. First we shall prove

The genus of a hyperbolic surface exceeds one.

In this case the canonical region is the interior of a circle Ω and hence the fundamental domain $\mathfrak{R}_0$ of the group G of covering transformations is inside Ω. It follows that $\mathfrak{R}_0$ has a finite number of sides. If not we can find an infinite sequence $s_1, s_2, \ldots,$ of points on sides of $\mathfrak{R}_0$ and no two

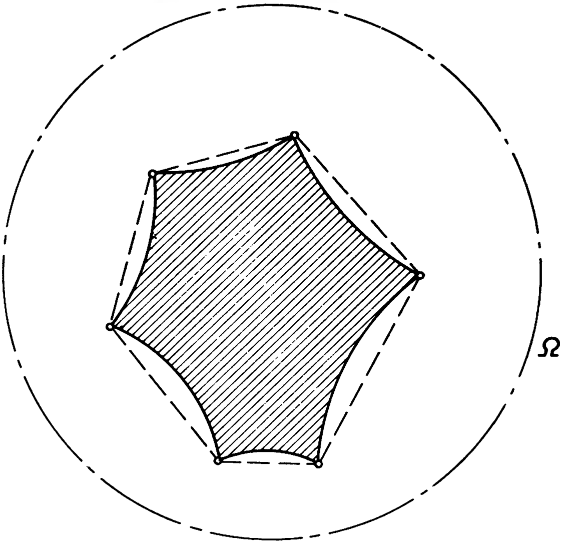


Fig. 13.3–2. Proof that the genus of a hyperbolic Riemann surface exceeds one

lie on the same side. These points have at least one accumulation point which is on the boundary. This is impossible, for the boundary points are ordinary points on the sides, or vertices and, therefore, no limit points.

Let us assume that the number of cycles of vertices is k. They correspond to k points on the Riemann surface. The characteristic of the Riemann surface is

$$\chi = -k+n-1,$$

if $2n$ is the number of sides of $\mathfrak{R}_0$. We may enclose $\mathfrak{R}_0$ in a rectilinear polygon having the same vertices, (fig. 13.3–2). The sum of its angles is

$2\pi(n-1)$. The sum of the angle of $\mathfrak{R}_0$ is $2\pi k$ (section 13.2.4; observe that $g = 1$, for there are no elliptic transformations in the group) and it is geometrically clear that

$$2\pi k < 2\pi(n-1),$$

whence

$$-k+n-1 > 0,$$

or $\chi > 0$. Then we deduce from (12.5–5) that $p > 1$. This proves the theorem.

Now we can complete the former statement. If $p = 1$ then $\tilde{\mathfrak{F}}$ cannot be homeomorphic to the interior of a circle, thus

An algebraic Riemann surface is parabolic if and only if its genus is 1.

Finally

An algebraic Riemann surface is hyperbolic if and only if its genus exceeds 1.

13.4 – Automorphic functions

13.4.1 – Simple automorphic functions

A function $f(s)$ is called *automorphic* with respect to a group G of linear transformations if

(i) the function is meromorphic in the region of discontinuity of the group;

(ii) $f(\mathsf{A}s) = f(s)$ for every element A of the group.

If a group admits non-constant automorphic functions, then the group is properly discontinuous.

Assume that the group is not properly discontinuous. Let s_0 denote a point at which $f(s)$ is regular. Since there are infinitely many points equivalent to s_0 in an arbitrary neighbourhood of s_0, the function $f(s)$ takes the value $f(s_0)$ infinitely often. But then $f(s)$ is a constant (section 2.11.1).

An automorphic function $f(s)$ is called *simple* if the following conditions are fulfilled:

(i) the group G possesses a fundamental domain $\mathfrak{R}_0$ whose boundary consists of a finite number of pairs of congruent sides;

(ii) the region of existence of the function is bounded by limit points of the group;

(iii) at every parabolic point the function tends to a definite limiting value (which may be finite or infinite) along every sequence of points in $\mathfrak{R}_0$ which tend to this parabolic point.

It should be noticed that if the point counts as two parabolic points as in fig. 13.4–1, the approach shall be from one side only. There will be a limit when the approach is from either side, but the two limits may be different.

The region of existence of an automorphic function extends into any neighbourhood of every limit point of the group.

This follows from the fact that in a neighbourhood of a limit point lie points equivalent to points in the region of existence of the functions and these points belong to the region of existence.

If an automorphic function is not a constant, each limit point of the group is an essential singularity of the function.

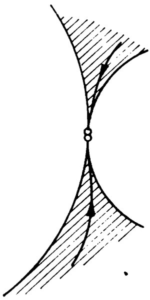


Fig. 13.4–1. Parabolic point counted twice

In a neighbourhood of a regular point or of a pole a function can take on any value only a finite number of times. In a neighbourhood of a limit point there are an infinite number of equivalent points at which the function takes the same value. It follows that a limit point is an essential singularity.

A direct consequence is

If a group possesses a non-constant automorphic function, it is a function group.

Let G be a group having a non-constant automorphic function, existing in a region $\mathfrak{S}$. Denote by $\mathfrak{S}^*$ the connected part of the plane, bounded by limit points, in which $\mathfrak{S}$ lies. The set $\mathfrak{S}^*$ consists of all ordinary points which can be joined to a point of $\mathfrak{S}$ by curves not passing through limit points. Any point s of $\mathfrak{S}^*$ and a curve C joining it to a point of $\mathfrak{S}$ are carried by any transformation A of the group into a point s' and a curve C' joining s' to a point of $\mathfrak{S}$, whilst C' consists of ordinary points. Then s' belongs to $\mathfrak{S}^*$ and it follows that $\mathfrak{S}^*$ is carried into itself by the transformations of the group. This proves that G is a function group.

13.4.2 – Behaviour at the vertices and parabolic points

Of the fixed points of the transformations of a group, only those belonging to elliptic transformations can lie within the region in which the function is meromorphic; all other fixed points are limit points.

We investigate first the behaviour of an automorphic function at the fixed point of an elliptic transformation. Let s_1 be such a fixed point and let g be the period of the elliptic transformation. For the sake of simplicity we shall assume that the fixed points are finite. Then there is a transformation A defined by

$$\frac{\mathrm{A}s-s_1}{\mathrm{A}s-s_2} = \mathrm{e}^{2\pi i/g}\,\frac{s-s_1}{s-s_2}, \qquad (13.4\text{–}1)$$

where s_2 is the other fixed point distinct from s_1. If $f(s)$ is regular at $s = s_1$, we can write

$$f(s) = f(s_1)+(s-s_1)^n g(s) = f(s_1)+\left(\frac{s-s_1}{s-s_2}\right)^n h(s), \qquad (13.4\text{–}2)$$

with $h(s) = (s-s_2)^n g(s)$, $g(s_1) \neq 0$, $g(s)$ is also regular at $s = s_1$. Given a neighbourhood of s_1 we can find another neighbourhood such that if s is in the second neighbourhood, then As is in the first. Hence

$$f(\mathrm{A}s) = f(s_1)+\left(\frac{\mathrm{A}s-s_1}{\mathrm{A}s-s_2}\right)^n h(\mathrm{A}s) = f(s_1)+\left(\frac{s-s_1}{s-s_2}\right)^n \mathrm{e}^{2\pi in/g} h(\mathrm{A}s).$$

Since $f(s)$ is automorphic we have $f(\mathrm{A}s) = f(s)$, whence

$$\mathrm{e}^{2\pi in/g} = \frac{h(s)}{h(\mathrm{A}s)}.$$

The first member of this equation is a constant. Making s approach s_1, the point As approaches s_1 and it follows that

$$\mathrm{e}^{2\pi in/g} = 1.$$

As a consequence n is a multiple of g and $f(s)-f(s_1)$ has a zero of an order which is a multiple of g. If s_1 is a pole of $f(s)$, then $1/f(s)$ has a zero at this point. Hence $f(s)$ has a pole whose order is again a multiple of g. Thus

A non-constant automorphic function takes its value g times or a multiple thereof, at a fixed point of an elliptic transformation of period g within the region of existence of the function.

A direct corollary is the following theorem

A non-constant automorphic function takes its value g times, or some multiple thereof, at a vertex belonging to a cycle, the sum of whose angles *is* $2\pi/g$. This assertion is trivial if $g = 1$. If $g > 1$ the vertex is a fixed point of an elliptic transformation of period g (section 13.2.4).

Next we investigate the behaviour at a parabolic point s_0 and we make the additional assumption that $f(s)$ is a simple automorphic function. We can carry the equivalent points of the cycle to which s_0 belongs to s_0, the transforms of the fundamental domain fitting together at s_0 as in fig. 13.4–2. It is clear that $f(s)$ tends to a definite or infinite value as s

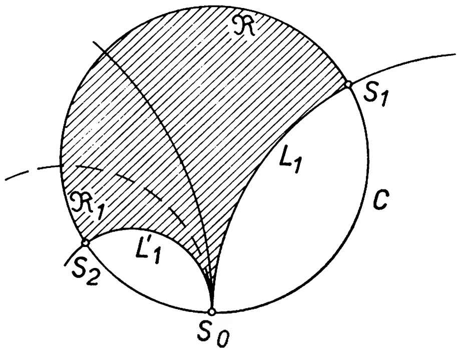


Fig. 13.4–2. Investigation of the behaviour of a simple automorphic function at a parabolic point

tends to s_0 within or on the boundary of each domain. The transformation A which carries $\mathfrak{R}$ to $\mathfrak{R}_1$ and the side L_1 to L'_1 is parabolic; it is determined by

$$\frac{1}{\mathsf{A}s - s_0} = \frac{1}{s - s_0} + c. \tag{13.4–3}$$

Consider the triangular region $s_0 s_1 s_2$ formed by $L_1 L'_1$ and a small circle through s_0 orthogonal to the sides which meet at s_0. This circle is a fixed circle for A. By repeated application of A the transforms of the domains mentioned fill up the circle C and the values of $f(s)$ repeat themselves in the transformed domains.

Now we introduce the variables

$$w = \frac{2\pi i}{c(s - s_0)}, \qquad t = \exp w. \tag{13.4–4}$$

The first substitution is a linear transformation; it carries s_0 to ∞ and the circles through s_0 into straight lines. Let w_1 and w_2 be the transforms of s_1 and s_2. Then

$$w_1 = \frac{2\pi i}{c(s_1 - s_0)}, \qquad w_2 = \frac{2\pi i}{c(s_2 - s_0)} = \frac{2\pi i}{c}\left(\frac{1}{s_1 - s_0} + c\right),$$

as follows from (13.4–3), since $s_2 = As_1$. Hence

$$w_2 = w_1 + 2\pi i.$$

The arc $s_1 s_2$ of C is transformed into a straight line segment parallel to the imaginary axis, (fig. 13.4–3) The sides $s_0 s_1$ and $s_0 s_2$ are carried into straight segments perpendicular to $w_1 w_2$ and, therefore, parallel to the real axis. Congruent points of L_1' and L_1 are carried into points in the w-plane which differ by $2\pi i$. These latter points are carried into coincident points by the second transformation (13.4–4). The side $s_1 s_2$ is carried

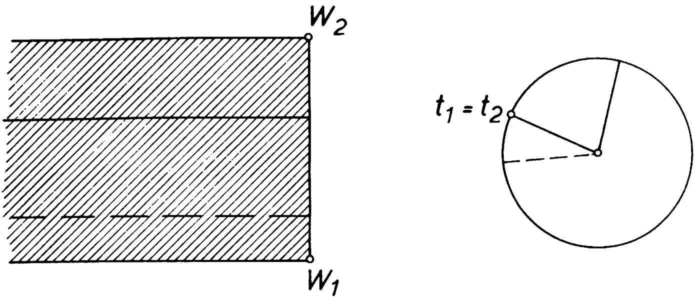


Fig. 13.4–3. The mapping of the triangular domain of fig. 14.4–2 onto a circular disc

into a circular arc with centre at the origin. The original triangle in the s-plane is mapped onto a circular disc slit along a radius.

The function $f(s)$ becomes a function $\varphi(t)$, meromorphic inside and on the boundary of the circular region in the t-plane, except possibly at $t = 0$. The function $\varphi(t)$ takes the same value at a point on the radius along which the region is slit, when it is approached from either side. It follows that this slit can be removed, the function turns out to be single-valued. Since $\varphi(t)$ approaches a definite value as $t \to 0$ (being either finite or infinite) it is regular at $t = 0$, or it has a pole there.

Thus we have proved

At a parabolic point a simple automorphic function $f(s)$ is a function of t regular or having a pole at $t = 0$, where

$$t = \exp \frac{2\pi i}{c(s-s_0)}, \tag{13.4–5}$$

if s_0 is finite and

$$t = \exp \frac{2\pi i}{c} s, \tag{13.4–6}$$

if the parabolic point is at infinity.

13.4.3 – The zeros and the poles of a simple automorphic function

In counting the zeros and the poles of a simple automorphic function which lie in a fundamental domain, certain conventions are necessary in the case of zeros or poles lying on the boundary.

If there is a zero or a pole at a side L there is a zero or a pole at the equivalent point on the congruent side L'. Only one of these will be counted as belonging to the domain.

The order n of a zero or a pole at a vertex will be partitioned equally among the domains which meet there. If there are m vertices in the cycle and the sum of the angles is $2\pi/g$, then gm regions meet at each vertex. Counting n/gm zeros or poles at each vertex, we have n/g zeros or poles at the m vertices of the cycle. This number is an integer as follows from the first theorem of the previous section.

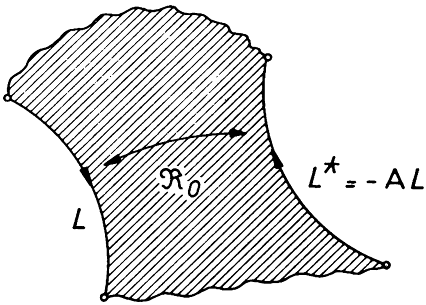


Fig. 13.4–4. Integration along congruent sides of the fundamental domain

If $f(s)$ approaches zero or becomes infinite at a parabolic point we shall determine the number of zeros of poles of the auxiliary functions $\varphi(t)$, introduced in the previous section, from the behaviour at $t = 0$. The number of zeros or poles at $t = 0$ of this function $\varphi(t)$ will be the number of zeros of poles of $f(s)$ at the parabolic point s_0, taken all-together, of the cycle to which s_0 belongs.

After these introductory remarks we can state the following important theorem

A simple automorphic function which is not identically zero has an equal number of zeros and poles in the fundamental domain.

We employ Cauchy's theorem (3.8–8). If the function has neither zeros nor poles on the boundary, then

$$N_0 - N_\infty = \frac{1}{2\pi i} \int \frac{f'(s)}{f(s)}\, ds, \tag{13.4–7}$$

the integral being taken in a positive sense around the boundary. Let L and L^* denote two congruent sides, (fig. 13.4–4). The contribution to the integral arising from these sider is

$$\int_L \frac{f'(s)}{f(s)}\,ds + \int_{L^*} \frac{f'(s)}{f(s)}\,ds, \qquad L^* = -\mathsf{A}L.$$

At the congruent points s and $\mathsf{A}s$ we have $f(s) = f(\mathsf{A}s)$. From

$$f'(\mathsf{A}s)d\mathsf{A}s = \frac{df(\mathsf{A}s)}{d\mathsf{A}s}\frac{d\mathsf{A}s}{ds}\,ds = \frac{d}{ds}f(\mathsf{A}s)ds = f'(s)ds$$

we deduce

$$\int_{L^*} \frac{f'(s)}{f(s)}\,ds = \int_{-L} \frac{f'(\mathsf{A}s)}{f(\mathsf{A}s)}\,d\mathsf{A}s = -\int_L \frac{f'(s)}{f(s)}\,ds.$$

It appears that the integrals along each pair of congruent sides cancel and we have $N_0 = N_\infty$.

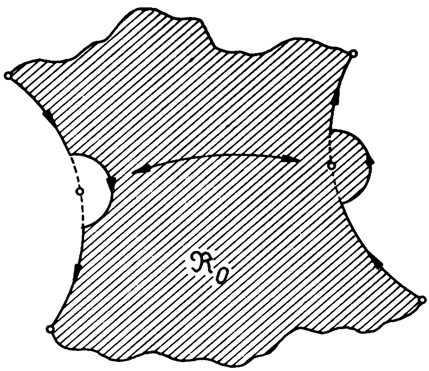


Fig. 13.4–5. Modification of the path of integration if there are zeros or poles on a side of the fundamental domain

If there is a zero or a pole on the side L, we deform this side slightly (fig. 13.4–5), so as to include the zero or pole, and we make the corresponding alteration in the congruent side L^*. The integrals along the new congruent sides cancel as before. Only one of the two zeros or poles now lie within the contour. Since but one of the pair should be counted as belonging to the domain, the theorem holds as before.

In the case that $f(s)$ has a zero or a pole of order n at a vertex we alter the path of integration to exclude each vertex of the cycle as in fig. 13.4–6, the radius of the small circular arcs being ρ. The parts of the sides that remain are congruent in pairs. At s_0 we have

$$f(s) = (s-s_0)^n g(s)$$

$g(s_0) \neq 0$ and $g(s)$ regular at $s = s_0$. Along the small arc we have

$$\int \frac{f'(s)}{f(s)}\, ds = n \int \frac{ds}{s - s_0} + \int \frac{g'(s)}{g(s)}\, ds.$$

Making $\rho \to 0$ the last integral tends to zero, since the integrand remains finite. The first integral on the right-hand side approaches $-i\theta$, where θ is the angle at s_0. Summing for all vertices of a cycle we have

$$N_0 - N_\infty = -\frac{n}{2\pi}\Sigma\theta = -\frac{n}{2\pi}\frac{2\pi}{g} = -\frac{n}{g}.$$

In the case of a zero we have $n > 0$ and

$$N_0 + \frac{n}{g} = N_\infty .$$

As n/g is the number of zeros we are to count at the vertices of a cycle,

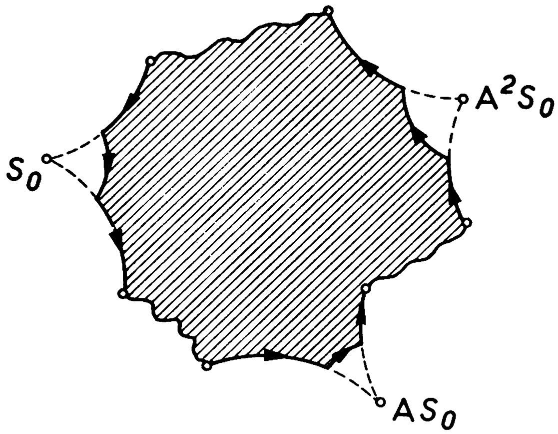


Fig. 13.4–6. Modification of the path of integration if there are zeros or poles at the vertices

the number of zeros is equal to the number of poles. In the case of a pole we have $n < 0$ and we may write

$$N_0 = N_\infty - \frac{n}{g}.$$

Thus the theorem holds also in this case.

It remains to discuss the case, where $f(s)$ has a zero or a pole at a parabolic point. We draw a circle C through s_0 as in fig. 13.4–3, sufficiently small that there are no zeros or poles within the region $s_0 s_1 s_2$ of that figure. The arc cuts off from the regions that lie between $s_0 s_1$ and $s_0 s_2$ certain parts $\mathfrak{A}_1, \ldots, \mathfrak{A}_m$. The congruent parts, $\mathfrak{A}'_1, \ldots, \mathfrak{A}'_m$

lying in the fundamental domain make up the neighbourhoods of the cycle. We shall remove these parts from the fundamental domain and integrate along the boundary of the remainder. The integrals of pairs of congruent sides cancel. The integrals over the circular arcs cutting off the parabolic points may be replaced by the integrals

$$N_0 - N_\infty = \frac{1}{2\pi i}\int_{s_2}^{s_1} \frac{f'(s)}{f(s)}\,ds = \frac{1}{2\pi i}\int \frac{\varphi'(t)}{\varphi(t)}\,dt,$$

where the last integral is taken around the circle of fig. 13.4–3 in the clockwise sense. It has the value $-n$ if $\varphi(t)$ has a zero of order n or a pole of order $-n$ at $t = 0$. Hence we either have $N_0 + n = N_\infty$ or $N_0 = N_\infty - n$.

The function $f(s)$ cannot have an infinite number of poles in the fundamental domain, for the poles would then have an accumulation point. If this is an ordinary point, $f(s)$ has an essential singularity there. If it is a parabolic point, $\varphi(t)$ has an essential singularity at the origin, both of which are contrary to hypothesis. Similarly $f(s)$ cannot have an infinite number of zeros, unless it is identically zero, for an accumulation point of zeros is likewise an essential singularity. This concludes the proof of the theorem.

The number of poles of a simple automorphic function in a fundamental domain is called the *degree* of the function.

Some direct consequences of this theorem deserve mention.

A simple automorphic function which has no poles in the fundamental domain is a constant.

Let $f(s_0)$ be the value of $f(s)$ at $s = s_0$. Since $f(s)$ and $f(s)-f(s_0)$ have the same number of poles, the number of zeros of $f(s)-f(s_0)$ is zero, if $f(s)$ is not a constant. Thus we arrive at a contradiction.

A simple automorphic function which is not a constant takes on every value the same number of times, being equal to the degree of the function.

This is clear, for the number of zeros of $f(s)-f(s_0)$ is equal to the number of poles of $f(s)$.

A more deeper consequence is the following theorem

Between two simple automorphic functions belonging to the same group and having the same region of existence, there exists an algebraic relation.

Let f_1 and f_2 be such functions. Consider a polynomial

$$F(z, w) = \sum_{\mu=0}^{m} \sum_{\nu=0}^{n} a_{\mu\nu} z^\mu w^\nu, \tag{13.4–8}$$

where $a_{00}, \ldots, a_{mn}$ are constants. If the number of poles of f_1 is k_1 and the number of poles of f_2 is k_2 then $F(f_1, f_2)$ is a simple automorphic function whose number of poles does not exceed $mk_1 + nk_2$. The most general polynomial (13.4–8) contains $(m+1)(n+1)$ constants. We can choose these constants in such a way that $F(f_1, f_2)$ will have $(m+1)\times \times(n+1)-1$ zeros at assigned points in the fundamental domain. This gives rise to $(m+1)(n+1)-1$ linear equations for the coefficients and since there is one more constant than equations to be satisfied it is possible to find constants satisfying these equations not all being zero. If m and n are large enough then

$$(m+1)(n+1)-1 > mk_1+nk_2$$

and $F(f_1, f_2)$ has more zeros than poles. This is only possible if $F(f_1, f_2)$ is identically zero.

A particular case is

If there exists a simple automorphic function $f(s)$, having a single pole in the fundamental domain, then any simple automorphic function connected with the same group and having the same region of definition is a rational function of f.

Let $g(s)$ have k poles. Consider the polynomial

$$F(z, w) = Q(z)w-P(z),$$

where P and Q are polynomials of degree at most k. The number of poles of $F(f, g)$ does not exceed $2k$; the number of coefficients is $2k+2$. Hence we can prescribe $2k+1$ zeros and it follows, as above, that there is a relation of the form

$$Q(f)g-P(f) = 0,$$

identically in s. The coefficients of Q are not all zero, for otherwise $P(f)=0$ identically, and this implies that f would be a constant. Hence

$$g = \frac{P(f)}{Q(f)} \tag{13.4–9}$$

and this concludes the proof of the assertion.

13.4.4 – The Schwarzian derivative

Let $f(s)$ and $g(s)$ denote two simple automorphic functions of the first degree, connected with the same group and having the same region of existence. Then, according to the last theorem of the previous section, g is a rational function of f and f is a rational function of g. Hence the two functions are related as

$$g(s) = \frac{Af(s)+B}{Cf(s)+D}, \qquad \begin{vmatrix} A & B \\ C & D \end{vmatrix} \neq 0, \tag{13.4–10}$$

where A, B, C and D are constants. Conversely, if $f(s)$ is a simple automorphic function of the first degree, so is $g(s)$, for the values of g correspond in a one-to-one manner to those of f; hence g takes each value once in the fundamental domain.

Now we ask whether there is an expression having the same value at a given point for all functions related as in (13.4–10). It is clear that we can find such an expression by differentiating (13.4–10) a sufficient number of times and eliminating the constants. First we have, primes denoting differentiating with respect to s,

$$g' = \frac{AD-BC}{(Cf+D)^2}f'.$$

Taking the logarithmic derivative we find

$$\frac{g''}{g'} = \frac{f''}{f'} - 2\frac{Cf'}{Cf+D} \tag{13.4–11}$$

and differentiating again

$$\frac{d}{ds}\frac{g''}{g'} = \frac{d}{ds}\frac{f''}{f'} - \frac{2Cf''}{Cf+D} + \frac{2C^2f'^2}{(Cf+D)^2}. \tag{13.4–12}$$

By squaring both members of (13.4–10) we get

$$\left(\frac{g''}{g'}\right)^2 = \left(\frac{f''}{f'}\right)^2 - \frac{4Cf''}{Cf+D} + \frac{4C^2f'^2}{(Cf+D)^2}.$$

Hence the expression

$$\boxed{[f]_s = \frac{d}{ds}\frac{f''}{f'} - \tfrac{1}{2}\left(\frac{f''}{f'}\right)^2} \tag{13.4–13}$$

is the desired invariant, for

$$[f]_s = [g]_s.$$

The expression (13.4–12) is called the *Schwarzian derivative of f.*

A very useful result is obtained if we investigate the behaviour of this derivative under a change of the independent variable. Let $t(s)$ denote a function of s. Then

$$\frac{df}{ds} = \frac{df}{dt}\frac{dt}{ds},$$

$$\frac{d^2f/ds^2}{df/ds} = \frac{d^2f/dt^2}{df/dt}\frac{dt}{ds} + \frac{d^2t/ds^2}{dt/ds}$$

and, differentiating again

$$\frac{d}{ds}\frac{d^2f/ds^2}{df/ds} = \frac{d}{dt}\frac{d^2f/dt^2}{df/dt}\left(\frac{dt}{ds}\right)^2 + \frac{d^2f/dt^2}{df/dt}\frac{d^2t}{ds^2} + \frac{d}{ds}\frac{d^2t/ds^2}{dt/ds}.$$

It follows that

$$\boxed{[f]_s = [f]_t\left(\frac{dt}{ds}\right)^2 + [t]_s.} \tag{13.4–14}$$

In particular, if t is a fractional linear transformation

$$t = \frac{as+b}{cs+d}, \tag{13.4–15}$$

then $[t]_s = [s]_s = 0$ and

$$[f]_s = [f]_t\frac{(ad-bc)^2}{(cs+d)^4}. \tag{13.4–16}$$

Consider now a function $z(s)$ and let $s(z)$ denote a function element of the inverse function in some neighbourhood of point s where $z'(s) \neq 0$. Since $[s]_s = 0$, we find from (13.4–14)

$$0 = [s]_z\left(\frac{dz}{ds}\right)^2 + [z]_s,$$

whence

$$[s]_z = -[z]_s\Big/\left(\frac{dz}{ds}\right)^2. \tag{13.4–17}$$

It is apparent from (13.4–16) that the Schwarzian derivative of an automorphic function is not automorphic. But we have

If z is a simple automorphic function then also

$$\frac{[z]_s}{\left(\dfrac{dz}{ds}\right)^2} \tag{13.4–18}$$

is a simple automorphic function.

The left hand member of (13.4–17) is invariant with respect to a fractional linear transformation of s. Hence (13.4–18) is automorphic. If $z(s)$ is meromorphic the same is true for its derivatives. The function (13.4–18) is, therefore, meromorphic throughout its region of existence. There remains the question of its behaviour at the parabolic points. As in section 13.4.2 we introduce the variable t by

$$t = e^u, \qquad u = \frac{2\pi i}{c(s-s_0)}$$

or

$$u = \frac{2\pi i s}{c},$$

according as the parabolic point is finite or infinite.

We have

$$\frac{[s]_z}{\left(\dfrac{ds}{dz}\right)^2} = \frac{[z]_u}{\left(\dfrac{dz}{du}\right)^2} = \frac{[z]_t\left(\dfrac{dt}{du}\right)^2+[t]_u}{\left(\dfrac{dz}{dt}\right)^2\left(\dfrac{dt}{du}\right)^2} = \frac{[z]_t}{\left(\dfrac{dz}{dt}\right)^2} + \frac{1}{2t^2\left(\dfrac{dz}{dt}\right)^2}.$$

The function z, expressed in terms of t, has at most a pole at $t = 0$. Hence (13.4–18) is either regular at $t = 0$ or has a pole there. This concludes the proof of the assertion.

According to the fourth theorem of section 13.4.3 there is an algebraic relation between the functions $z(s)$ and the function (13.4–18). Hence in view of (13.4–1) we may assert that $R(z) = [s]_z$, $s(z)$ being the inverse of $z(s)$, depends algebraically on z. Suppose that $z(s)$ is of the first degree. Then z takes its values once and this means that $R(z)$ is rational. Thus we have

If $z(s)$ is a simple automorphic function of the first degree and $s(z)$ its inverse, then the Schwarzian derivative

$$R(z) = [s]_z$$

is a rational function of z.

The function $R(z)$ is called *the invariant* of the simple automorphic function $z(s)$.

13.4.5 – The differential equation of the second order

We consider again a function $s(z)$ in some neighbourhood of a point, where $s'(z) \neq 0$. We introduce two functions $w_0(z)$, $w_1(z)$ such that

$$s(z) = \frac{w_1(z)}{w_0(z)}. \tag{13.4–19}$$

The derivative with respect to z is

$$s' = \frac{w_1' w_0 - w_1 w_0'}{w_0^2}.$$

Since we are free in the choice of w_0 we can take this function in such a way that

$$w_0^2 = \frac{1}{ds/dz} \tag{13.4–20}$$

in some neighbourhood of the above mentioned point. Then from (13.4–20)

$$w_1' w_0 - w_1 w_0' = 1 \tag{13.4–21}$$

and it follows that

$$w_1'' w_0 - w_1 w_0'' = 0,$$

or

$$\frac{w_0''}{w_0} = \frac{w_1''}{w_1}. \tag{13.4–22}$$

Now we express the Schwarzian derivative of $f(z)$ in terms of w_0 and w_1 (and its derivatives). We have from (13.4–20)

$$\frac{s''}{s'} = -\frac{2w_0'}{w_0}$$

and so

$$\frac{d}{dz}\frac{s''}{s'} = -2\frac{w_0'' w_0 - w_0'^2}{w_0^2},$$

whence

$$[s]_z = -2\frac{w_0''}{w_0} = -2\frac{w_1''}{w_1}.$$

This result can be stated as follows. Denoting $[s]_z$ by $R(z)$ we have

The functions w_0 and w_1 satisfying (13.4–19) *and* (13.4–21) *are solutions of the linear differential equation of the second order*

$$\boxed{w'' + \tfrac{1}{2}R(z)w = 0.} \tag{13.4–23}$$

If $[s]_z = 0$ we have $w_0'' = 0$, $w_0 = cz+d$, and $w_1'' = 0$, $w_1 = az+b$, whence $s = (az+b)/(cz+d)$. This enables us to prove

From

$$[f]_z = [g]_z \tag{13.4–24}$$

follows that

$$g = \frac{Af+B}{Cf+D}, \tag{13.4–25}$$

where A B, C and D are constants.

We employ (13.4–14). It is assumed, of course, that $f(z)$ is not a constant. Introducing the new variable $f = f(z)$, we have

$$[g]_z = [g]_f \left(\frac{df}{dz}\right)^2 + [f]_z$$

whence $[g]_f = 0$, since $df/dz \neq 0$. Then the truth of the assertion follows from the above remark.

Let now $z(s)$ denote a simple automorphic function of the first degree and $s(z)$ its inverse: then we have evidently

If $z(s)$ is a non-constant simple automorphic function of the first degree then its inverse can be expressed as the quotient of two solutions of a linear differential equation (13.4–23) *of the second order, where $R(z)$ is a rational function.*

13.4.6 – The invariant of a simple automorphic function of the first degree

In the previous section we obtained the result, that the invariant of a simple automorphic function $z(s)$ of the first degree is a rational function.

First we consider an ordinary point s_0 of $z(s)$. Let a denote the value of this function at s_0. Then we have the expansion

$$z = a + a_1(s - s_0) + \ldots, \tag{13.4–26}$$

with $a_1 \neq 0$, for $z(s)$ is of the first degree (and hence univalent). Introducing the variable

$$t = s - s_0 \tag{13.4–27}$$

we have

$$z = a + a_1 t + \ldots \tag{13.4–28}$$

and this series is invertible giving

$$t = b_1(z - a) + \ldots. \tag{13.4–29}$$

Now

$$R(z) = [s]_z = [s]_t \left(\frac{dt}{dz}\right)^2 + [t]_z = [t]_z,$$

for $[s]_t = [t + s_0]_t = 0$. Since $[t]_z$ is clearly regular at $z = a$, we find

The function $R(z)$ is regular at a point $z = a$ corresponding to an ordinary point of $s(z)$.

Next we consider the case that s_0 is a pole not at a vertex or at a parabolic point. Hence we have an expansion

$$\frac{1}{z} = a_1(s - s_0) + \ldots \tag{13.4–30}$$

and the substitution (13.4–27) leads to

$$t = b_1\left(\frac{1}{z}\right) + \ldots. \tag{13.4–31}$$

As above we find

$$R(z) = [t]_z.$$

But

$$[t]_z = [t]_{1/z}\left(\frac{d(1/z)}{dz}\right)^2 = \frac{1}{z^4}[t]_{1/z},$$

where $[t]_{1/z}$ is an ordinary power series in terms of $1/z$. Thus we find

If the point at infinity of the z-plane corresponds to a pole in the interior of the fundamental domain of the function $z(s)$ then $R(z)$ is regular at infinity, having a zero of order at least four there.

Suppose now that s_0 is a vertex of the fundamental domain and that z takes the value a at this point. Then $z-a$ has a zero of certain order n at $s = s_0$. As we pointed out in section 13.4.3 this point must be counted with a multiplicity n/g, if g is the period of the elliptic transformation associated to this point. Hence we must take $n = g$ and the expansion of z is

$$z = a+a_1(s-s_0)^g+\ldots, \; a_1 \neq 0, \tag{13.4–32}$$

which transforms into (13.4–28) after performing the substitution

$$t = (s-s_0)^g. \tag{13.4–33}$$

The series (13.4–28) can be inverted, yielding the series (13.4–29). From (13.4–33) follows that

$$s = s_0+t^{1/g}, \tag{13.4–34}$$

where t is a power series in $z-a$. In some neighbourhood of $t \neq 0$ the function $t^{1/g}$ represents a single-valued branch.

In

$$R(z) = [s]_t\left(\frac{dt}{dz}\right)^2+[t]_z \tag{13.4–35}$$

the last term is an ordinary power series. An easy calculation shows that

$$[s]_t = \left(1-\frac{1}{g^2}\right)\frac{1}{2t^2}, \tag{13.4–36}$$

whence

$$[s]_t\left(\frac{dt}{dz}\right)^2 = \left(1-\frac{1}{g^2}\right)\frac{1}{2(b_1(z-a)+\ldots)^2}(b_1+\ldots)^2$$

$$= \left(1-\frac{1}{g^2}\right)\frac{1}{2(z-a)^2}+\frac{C'}{z-a}+\ldots,$$

where the omitted terms represent an ordinary power series in $z-a$. Thus

The function $R(z)$ has a pole of the second order at a point $z = a$ corresponding to a vertex of the fundamental domain.

If the vertex is a pole we may write

$$\frac{1}{z} = a_1(s-s_0)^{1/g} + \ldots$$

and this yields (13.4–31) on applying the substitution (13.4–33) and inverting the series. Now we have

$$[s]_t \left(\frac{dt}{dz}\right)^2 = \left(1-\frac{1}{g^2}\right) \frac{1}{2\left(b_1\left(\frac{1}{z}\right)+\ldots\right)^2}\left(-\frac{b_1}{z^2}+\ldots\right)^2$$

$$= \left(1-\frac{1}{g^2}\right)\frac{1}{2z^2}+\frac{C}{z^3}+\ldots$$

and since $[t]_z$ starts with a term involving z^{-4} we obtain

The function $R(z)$ has a zero of the second order at the point corresponding to a pole at a vertex of the fundamental domain.

A parabolic point s_0 is investigated by the substitution (13.4–5). Suppose that z takes the value a at $s = s_0$. The auxiliary function introduced in section 13.4.2 has a zero or a pole of the first order at $t = 0$. In our case it has a zero and from $z-a = \varphi(t)$ we find t as an ordinary power series in terms of $z-a$. Since

$$\frac{1}{s-s_0} = \frac{c}{2\pi i} \log t$$

we find

$$[s]_t = \left[\frac{1}{s-s_0}\right]_t = \frac{1}{2t^2},$$

the same result as (13.4–36), but with $g = \infty$. As a consequence, since the case of a pole can be treated along similar lines,

The function $R(z)$ has a pole of the second order at a point corresponding to a parabolic point and a zero of the second order at infinity, if the parabolic point is a pole.

We tacitly assumed that s_0 is finite. But if s_0 is infinite we replace $s-s_0$ by $1/s$ in our formulas and make use of the fact that $[s]_z = [1/s]_z$.

Summing up we see that $R(z)$ *has only poles of the second order and a zero of at least the second order or of at least the fourth order at infinity.*

13.5 – The Poincaré theta series

13.5.1 – The theta series

A fundamental question in the theory of the automorphic function is whether there are non constant functions which are automorphic with respect to a given properly discontinuous group. In section 13.1.5 we pιoved that a properly discontinuous group is either finite or consists of an enumerable number of elements. Hence we can place them in a certain order

$$\mathsf{A}_0, \mathsf{A}_1, \mathsf{A}_2, \ldots, \tag{13.5–1}$$

where A_0 always denotes the identity E.

If the group is finite and $H(s)$ is any rational function then,

$$\sum_{\nu=0}^{n-1} H(\mathsf{A}_\nu s), \tag{13.5–2}$$

where n is the number of elements of the group is automorphic, for replacing s by $\mathsf{A}_k s$ means only a permutation of the terms in the sum (13.5–2).

If, however, the group is infinite we have instead of (13.5–2) an infinite series and there is no reason to believe that such a series is convergent, for the general term $H(\mathsf{A}_n s)$ will even not approach zero unless $H(s)$ is zero in the limit points to which $\mathsf{A}_n s$ cluster.

The situation is comparable by that of the problem of Mittag–Leffler (section 4.11.1) of constructing a function with prescribed singularities. The problem was solved by the introduction of convergence factors. The same idea led Poincaré to the construction of a certain class of series which have similar properties as the Jacobian theta series. These series can serve to construct automorphic functions.

Let $H(s)$ be a rational function none of whose poles is at a limit point of the group. By a *theta series of Poincaré* we understand a series of the type

$$\boxed{\theta(s) = \sum_{\nu=0}^{\infty} H(\mathsf{A}_\nu s) \left(\frac{d\mathsf{A}_\nu s}{ds}\right)^m} \tag{13.5–3}$$

where m is a positive integer. In the next section we shall prove the unconditional convergence of this series.

Anticipating this result we shall first derive a fundamental property of this series. Let

$$\mathsf{A}_n s = \frac{a_n s + b_n}{c_n s + d_n}, \qquad n = 0, 1, 2, \ldots, \begin{vmatrix} a_n & b_n \\ c_n & d_n \end{vmatrix} = 1. \tag{13.5–4}$$

Then

$$\theta(\mathsf{A}_k s) = \sum_{\nu=0}^{\infty} H(\mathsf{A}_\nu \mathsf{A}_k s) \left(\frac{d\mathsf{A}_\nu \mathsf{A}_k s}{d\mathsf{A}_k s}\right)^m = \sum_{\nu=0}^{\infty} H(\mathsf{A}_\nu \mathsf{A}_k s) \left(\frac{d\mathsf{A}_\nu \mathsf{A}_k s}{ds}\right)^m \left(\frac{ds}{d\mathsf{A}_k s}\right)^m$$

If n runs through all numbers 0, 1, 2, . . ., then $\mathsf{A}_n \mathsf{A}_k s$ runs through all transformations of the group and since – as we shall prove – the series (13.5–3) is unconditionally convergent we may conclude that

$$\theta(\mathsf{A}_k s) = \theta(s) \left(\frac{d\mathsf{A}_k s}{ds}\right)^{-m}$$

or

$$\boxed{\theta(\mathsf{A}_k s) = \theta(s)(c_k s + d_k)^{2m}.} \tag{13.5–5}$$

This is *Poincaré's relation*. The number m will be called the *weight* of the series.

Now we can set up functions which are unaltered when a transformation of the group is applied. Let $\theta_1(s)$ and $\theta_2(s)$ be two theta series of the same weight. If $f(s)$ denotes the quotient

$$\boxed{f(s) = \frac{\theta_1(s)}{\theta_2(s)},} \tag{13.5–6}$$

then

$$f(\mathsf{A}_k s) = \frac{\theta_1(\mathsf{A}_k s)}{\theta_2(\mathsf{A}_k s)} = \frac{\theta_1(s)(c_k s + d_k)^{2m}}{\theta_2(s)(c_k s + d_k)^{2m}} = f(s).$$

It will appear subsequently that, for a function group, $\mathsf{A}_k s$ is in the domain of existence of the theta series. Then $f(s)$ is an automorphic function.

13.5.2 – Convergence of the theta series

We proceed to prove the convergence of the theta series (13.5–3). We state

If $m \geqq 2$ and if the point at infinity is an ordinary point of the group, then the theta series (13.6–3) *defines a function which is meromorphic in any region not containing limit points of the group in its interior.*

The restriction that $s = \infty$ can only be an ordinary point is not very serious. It has the advantage that we can apply the results obtained in paragraph 13.1.

It will suffice to prove the theorem for a region $\mathfrak{R}'$ such that there are no limit points of the group inside or on the boundary of $\mathfrak{R}'$, since such a region can be made large enough to include any given interior point of a region with limit points on its boundary.

Certain terms of (13.5–3) may have poles in $\mathfrak{R}'$. At $s = -d_k/c_k$, the centre of the isometric circle I_k, the factor $(c_k s + d_k)^{-2m}$ becomes infinite. If s is such that $\mathsf{A}_k s = a$, where a is a pole of $H(s)$, then $H(\mathsf{A}_k s)$ has a pole. It is clear, however, that only a finite number of terms of the series have poles in $\mathfrak{R}'$; for $\mathfrak{R}'$ contains in its interior and on its boundary only a finite number of centres of isometric circles and only a finite number of points congruent to each of the poles of $H(s)$.

We now omit the finite number of terms having poles inside $\mathfrak{R}'$ and on the boundary of $\mathfrak{R}'$. The minimum distance from the boundary of $\mathfrak{R}'$ to the centres of those isometric circles whose centres are exterior to $\mathfrak{R}'$ exceeds a positive number d. We have then for all the terms we are considering and for all s in $\mathfrak{R}'$

$$\left| s + \frac{d_k}{c_k} \right| \geqq d.$$

Further we can include the poles of $H(s)$ in small closed discs such that, when s is in $\mathfrak{R}'$, all points $\mathsf{A}_k s$ in the terms considered are outside these discs. Outside these discs the function $H(s)$ is bounded, so that we have

$$|H(\mathsf{A}_k s)| < M.$$

Excluding further the finite number of terms for which $c_k = 0$, we have

$$|H(\mathsf{A}_k s)||c_k s + d_k|^{-2m} = \left| \frac{H(\mathsf{A}_k s)}{c_k^{2m}\left(s + \dfrac{d_k}{c_k}\right)^{2m}} \right| < \frac{M}{d^{2m}} |c_k|^{-2m}.$$

The series

$$\sum_{\nu=1}^{\infty} |c_\nu|^{-4} = \sum_{\nu=1}^{\infty} r_\nu^4,$$

being the series (13.1–23), is convergent under the assumptions made above. Hence the series

$$\sum_{\nu=1}^{\infty} |c_\nu|^{-2m}$$

is certainly convergent if $m \geqq 2$. Thus the absolute and uniform convergence of the series under consideration is established and its sum is a holomorphic function in $\mathfrak{R}'$. It follows that (13.5–3) is meromorphic in $\mathfrak{R}'$ and, consequently, also in $\mathfrak{R}$.

Let us now consider, for example, the Fuchsian group of the first kind. The limit points consist of all points of the principal circle and (13.5–3) defines a meromorphic function inside the principal circle. The poles of $\theta(s)$ arise from the poles of the individual terms of the series. If $H(s)$ has a pole at $s = a$ inside the principal circle, then $H(\mathsf{A}_k s)$ becomes infinite when $A_k s = a$. That is, $\theta(s)$ has poles at the poles of $H(s)$ – except that, in special cases, $H(s)$ may have poles at congruent points of such a character that the singularities arising from two or more terms of the series cancel. Leaving this special case aside we may state that the number of poles of $\theta(s)$ in a fundamental domain $\mathfrak{R}_0$ is exactly equal to the number of poles of $H(s)$ inside the principal circle. If $\theta(s)$ has poles inside the principal circle, these points cluster about each point of the principal circle. *The principal circle is thus a natural boundary of the function.*

The proof of the convergence of the theta series does not require that the group be a function group. Poincaré's relation (13.5–5) connects two distinct functions, unless $\theta(s)$ can be continued analytically from s to $A_k s$. In order that (13.5–5) express a property of a single one of the functions defined by the series, it is necessary that the region of existence of the function can be carried into itself by the transformations of the group. The group is then a function group.

In setting up functions by means of theta series, the poles of $H(s)$ are at our disposal. By placing a pole at a desired point we can be sure that $\theta(s)$ in a region $\mathfrak{R}$ under consideration has a singularity and, hence, it is not identically zero. In forming automorphic functions for a function group by means of theta functions, as in (13.5–6), we can place the poles of the numerator and the denominator at different points of $\mathfrak{R}$ and, thus, be assured that the automorphic function does not reduce to the trivial case of a constant.

13.5.3 – Behaviour of the theta series at vertices and at parabolic points

Let, as in section 13.4.2, s_1 and s_2 be fixed points of an elliptic transformation A of period g. Then A is defined by (13.4–1). Differentiating logarithmically we get

$$\frac{dAs}{ds} = \frac{(As-s_1)(As-s_2)}{(s-s_1)(s-s_2)}. \tag{13.5–7}$$

We assume that s_1 is in the region of existence of a theta function $\theta(s)$, such that A belongs to the group defining this function.

Poincaré's relation (13.5–5) may be written in the form

$$\theta(As)\left(\frac{dAs}{ds}\right)^m = \theta(s). \tag{13.5–8}$$

Let us put

$$F(s) = (s-s_1)^m(s-s_2)^m\theta(s). \tag{13.5–9}$$

Then, by virtue of (13.5–7) and (13.5–8)

$$F(\mathsf{A}s) = (s-s_1)^m(s-s_2)^m \left(\frac{d\mathsf{A}s}{ds}\right)^m \theta(\mathsf{A}s) = F(s)$$

and it appears that $F(s)$ is unaltered by the transformation A; otherwise stated: $F(s)$ is automorphic with respect to the cyclic group generated by A.

Applying the first theorem of section 13.4.2 we may infer that $F(s)$ takes its value g times, or a multiple thereof, at $s = s_1$.

If $\theta(s)$ is regular at $s = s_1$ then $F(s)$ has a zero of order at least m at $s = s_1$. If m is not a multiple of g then $\theta(s)$ must have a zero at s_1 also, the order n_0 of the zero being such that

$$m + n_0 \equiv 0 (\text{mod } g). \tag{13.5–10}$$

Hence

If $\theta(s)$ is regular at $s = s_1$, it is necessarily zero there, unless the weight m is a multiple of g.

If $\theta(s)$ has a pole of order n_∞ at s_1, then $F(s)$ has a pole of order $n_\infty - m$ or a zero of order $m - n_\infty$, unless $n_\infty = m$. An equation

$$m - n_\infty \equiv 0 (\text{mod } g) \tag{13.5–11}$$

holds in this case.

Next we suppose that s_0 is a parabolic point and that there is a parabolic subgroup generated by a transformation A defined by (13.4–3). Now we have

$$\frac{d\mathsf{A}s}{ds} = \left(\frac{\mathsf{A}s - s_0}{s - s_0}\right)^2 \tag{13.5–12}$$

and if we introduce the function

$$G(s) = (s-s_0)^{2m}\theta(s) \tag{13.5–13}$$

we find

$$G(\mathsf{A}s) = (s-s_0)^{2m} \left(\frac{d\mathsf{A}s}{ds}\right)^m \theta(\mathsf{A}s) = G(s),$$

meaning that $G(s)$ is automorphic with respect to the subgroup generated by A.

Introducing the variables (13.4–4) we have

$$g(s) = \frac{G(s)}{(s-s_0)^{2m}} = \left(\frac{c \log t}{2\pi i}\right)^m \psi(t),$$

where $\psi(t)$ is single-valued in a neighbourhood of $t = 0$. The theta function shows a logarithmic singularity at $t = 0$.

We can take the circle C, considered in section 13.4.2, small enough that it contains no point $-d_k/c_k$ and no point congruent to a pole of $H(s)$. In fact, C, when small enough, contains only points congruent to points of the fundamental domain which lie in the neighbourhoods of the parabolic points of the cycle; and these neighbourhoods can sufficiently restricted to exclude ∞ and the finite number of the points of the fundamental domain congruent to poles of $H(s)$. Taking C sufficiently small we have bounds for the expressions

$$\left|\frac{s-s_0}{s+\dfrac{d_k}{c_k}}\right| < K, \quad |H(\mathrm{A}_k s)| < M,$$

where s lies within or on the boundary of the triangle $s_1 s_0 s_2$ considered in the above mentioned section. In the series

$$G(s) = \sum_{\nu=0}^{\infty} \left(\frac{s-s_0}{s+\dfrac{d_\nu}{c_\nu}}\right)^{2m} \frac{1}{c_\nu^{2m}} H(\mathrm{A}_\nu s),$$

the general term is less in absolute value than the corresponding term of the convergent series of positive terms

$$K^{2m} M \sum_{\nu=1}^{\infty} |c_\nu|^{-2m}.$$

$G(s)$ thus remains finite if s tends to s_0 from the interior of the triangle. In the t-plane, then, $\psi(t)$ is regular at $t = 0$, for it is bounded and holomorphic in a neighbourhood of $t = 0$ (section 2.8.3).

An automorphic function formed as a quotient of two theta series may be written as

$$f(s) = \frac{(s-s_0)^{2m}\theta_1(s)}{(s-s_0)^{2m}\theta_2(s)} = \frac{\psi_1(t)}{\psi_2(t)}.$$

This function, as a function of t, is regular or has a pole at $t = 0$. Then, as s tends to s_0 from the interior of the fundamental domain, $f(s)$ has a finite or infinite limit. It thus satisfies the requirements for the behaviour of a simple automorphic function (section 13.4.1).

CHAPTER 14

THE SCHWARZIAN TRIANGLE FUNCTIONS AND THEIR INVERSES

14.1 – The mapping of a curvilinear polygon

14.1.1 – THE DIFFERENTIAL EQUATION OF SCHWARZ FOR THE MAPPING FUNCTION

An interesting class of automorphic functions is intimately related to the problem of the mapping of the upper half of the z-plane onto a curvilinear triangle in the w-plane.

The Schwarzian derivative provides the key for the solution of the problem of mapping a circular disc or a half plane onto a polygon with circular sides. We understand by such a polygon a region whose boundary consists of circular arcs. The case that there are rectilinear segments among the sides need not to be excluded.

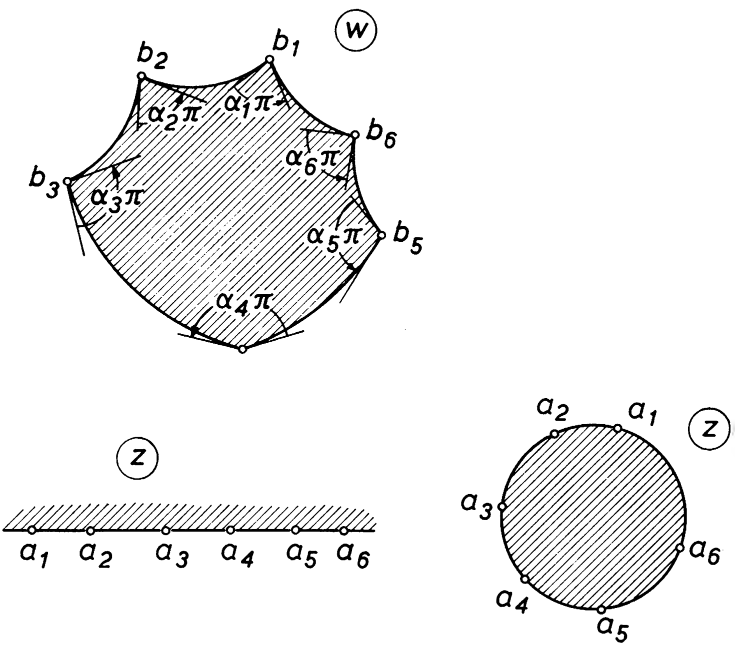


Fig. 14.1–1. The mapping of a curvilinear polygon onto a half plane or onto a circular disc

As we shall see a linear transformation has no influence on the method for characterizing the mapping function. Hence we may assume that the polygon is situated in the w-plane such that $w = \infty$ is an external point. Its vertices will be denoted by $b_1, \ldots, b_n$ and the interior angles by $\alpha_1 \pi, \ldots, \alpha_n \pi$ respectively, (fig. 14.1–1), will $0 \leq \alpha_k \leq 2$, $k = 1, \ldots, n$.

By the Riemann mapping theorem and the supplementary discussion of section 10.5.6 we may conclude that there is a univalent function $f(z)$, holomorphic throughout the upper half of the z-plane or in the interior of the unit disc around $z = 0$ which maps this region onto the interior of the polygon. This function is continuous on the real axis or on the circumference of the disc.

Let $a_1, \ldots, a_n$ denote the points in the z-plane corresponding to the vertices of the polygon. For the time being we suppose that all these points are finite. Since rectilinear segments or circular arcs correspond, we may apply Schwarz's reflection principle and conclude that $f(z)$ is also regular at each point of the boundary, except at the points $a_1, \ldots,$ a_n. The function $f(z)$ can be continued analytically across a line segment (or an arc) between two successive points a_i, a_{i+1} as to be defined in the lower half plane or the exterior of the unit circle. The values of $f(z)$ in this new region make up the interior of a polygon adjacent to the given polygon and symmetric with it with respect to a side. Reflecting again with respect to a part of the boundary between two points a_j, a_{j+1} (different from the above ones) we obtain a function $g(z)$ which is defined in the original region and maps this onto a polygon which is obtained by two reflections, (fig. 14.1–2), i.e., a fractional linear transformation. This new mapping function is related to the first function by means of the equation

$$g(z) = \frac{Af(z)+B}{Cf(z)+D}, \tag{14.1–1}$$

where A, B, C and D are constants such that $AD-BC \neq 0$. It is, therefore, natural not to consider the mapping function f itself, but the Schwarzian derivative $[f]_z$ which is the same for all mapping functions as obtained by means of the reflection principle. It follows that $[f]_z$ is a single-valued function which exists throughout the extended z-plane, except possibly at the singular points $a_1, \ldots, a_n$, for $f'(z) \neq 0$ everywhere in the region outside these points.

By a suitable linear transformation (14.1–1) in the w-plane we can map a given polygon onto a polygon which has a rectilinear side on the real w-plane axis. Hence $g(z)$ and so $[g]_z$ is real between the points corresponding to the endpoints of this segment. As a consequence

The Schwarzian derivative $[f]_z$ of the mapping function is real on

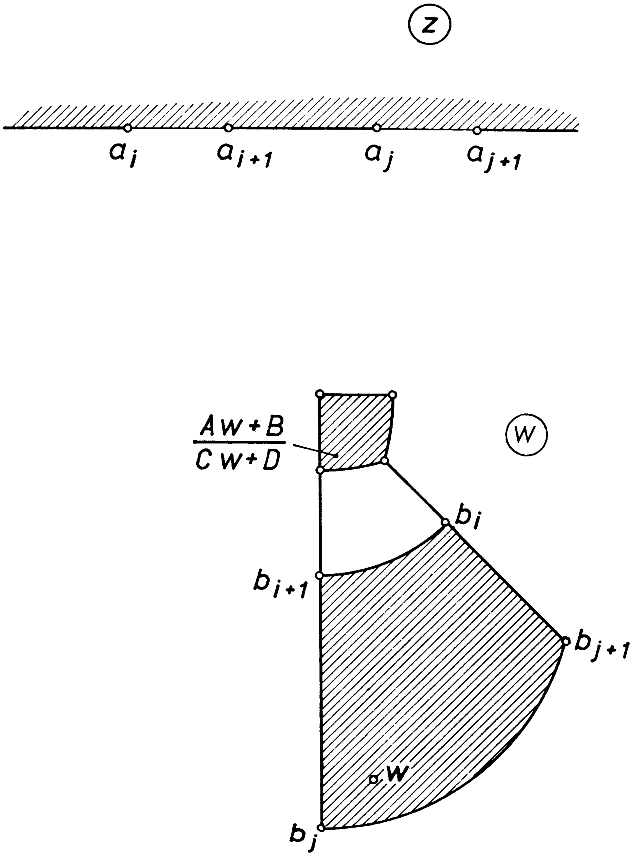


Fig. 14.1–2. Successive reflections of a polygon in its sides

the boundary of the half z-plane or on the circumference of the unit disc.

Now we turn our attention to the discussion of the singularities. Let a denote a singular point and b its image. The angle at b be $\alpha\pi$. Three cases can occur;

a) The circumferences, on which the two sides issuing from b lie, meet in a second point b' which is a finite point. If $b = \infty$ then the circumferences are straight lines, of course.

b) Both sides are rectilinear and issue from a finite point.

c) The sides are tangent at b. In this case the internal angle is 0 or π, (fig. 14.1–3), for we exclude the case $\alpha = 2$.

Case a) can be reduced to case b) on applying a linear transformation which brings the point b' to infinity. We consider this case first.

A suitable movement brings b into the point $w = 0$ and one side along the real w-axis, such that the other side is in the upper half of the w-plane. The resulting mapping function will be denoted by $g(z)$.

It is clear (section 10.2.2) that the transformation

$$t^{\alpha} = w \tag{14.1-2}$$

maps the angular region onto the upper half of the t-plane. Thus we obtain a correspondence between the z-plane and the t-plane such that points in a neighbourhood of $z = a$ and lying on a segment of the boundary of the mapped region in the z-plane correspond to those of a part of the real axis in the t-plane containing $t = 0$. As a consequence the function $t(z)$ is regular at $z = a$ (Schwarz's reflection principle) and can be continued throughout a neighbourhood of $z = a$. Hence t can be expanded as

$$t = c_1(z-a)+c_2(z-a)^2+ \ldots, c_1 \neq 0. \tag{14.1-3}$$

On applying the transformation formula (13.4–14) of the Schwarzian derivative we may evaluate $[g]_z$, for

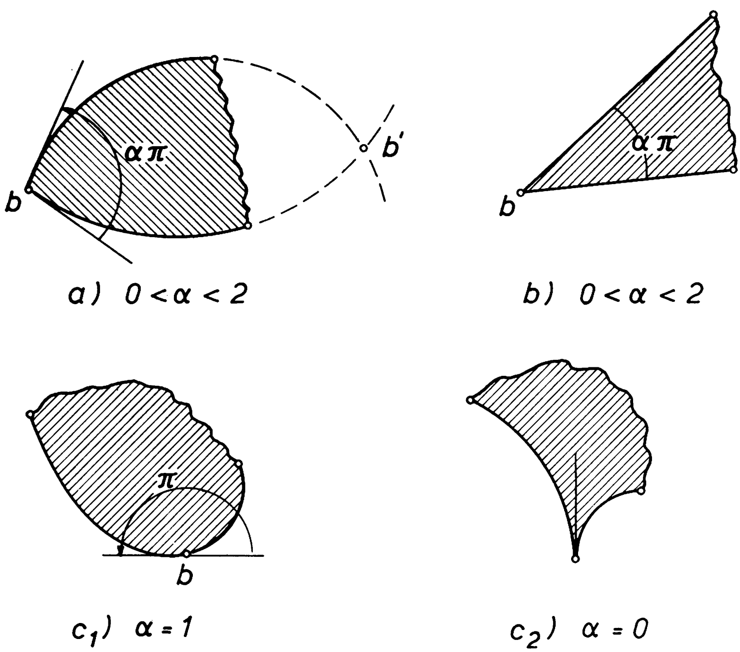


Fig. 14.1–3. The various possible situations at a vertex of a curvilinear polygon if $0 \leq \alpha < 2$

$$[g]_z = [w]_t \left(\frac{dt}{dz}\right)^2 + [t]_z .$$

Now $t = w^{1/\alpha}$ represents a single-valued branch in some neighbourhood of $w = 0$. A simple calculation yields

$$[w]_t = \frac{1-\alpha^2}{2t^2},$$

whence

$$[w]_t \left(\frac{dt}{dz}\right)^2 = \frac{1-\alpha^2}{2(c_1(z-a)+\ldots)^2}(c_1+\ldots)^2 = \frac{1-\alpha^2}{2(z-a)^2} + \frac{C}{z-a} + \ldots,$$

where the omitted terms constitute an ordinary power series. Observing that $[t]_z$ is regular, we finally have, since $[f]_z = [g]_z$:

$$[f]_z = \frac{1-\alpha^2}{2(z-a)^2} + \frac{C}{z-a} + h(z), \qquad (14.1\text{–}4)$$

where $h(z)$ is regular at $z = a$.

There remains the discussion of the case $\alpha = 0$, 1. First we apply a transformation (if necessary) which brings b to infinity in such a way that we obtain figures as considered in section 10.3.6. If $\alpha = 0$ we use the transformation

$$w = \log t$$

and from

$$[w]_t = \frac{1}{2t^2}$$

we easily find

$$[f]_z = \frac{1}{2(z-a)^2} + \frac{C}{z-a} + h(z),$$

being the expression (14.1–4) with $\alpha = 0$.

If $\alpha = 1$ we apply the substitution

$$w = t^{-1} + \log t$$

which does not differ essentially from (10.3–33).

A simple calculation shows

$$[w]_t = \tfrac{1}{2}\frac{t-4}{t(t-1)^2} = -\frac{2}{t} + \ldots .$$

Inserting the series for t we obtain (14.1–4) with $\alpha = 1$.

It is clear that the function

$$\chi(z) = [f]_z - \tfrac{1}{2}\sum_{\nu=1}^{n} \frac{1-\alpha_\nu^2}{(z-a_\nu)^2} - \sum_{\nu=1}^{n} \frac{C_\nu}{z-a_\nu}$$

has no singularities in the finite plane. Let w_0 denote the value of $f(z)$ at $z = \infty$. Then

$$w - w_0 = c_1 \left(\frac{1}{z}\right) + c_2 \left(\frac{1}{z}\right)^2 + \ldots, \; c_1 \neq 0.$$

It is evident that $[w]_{1/z}$ is an ordinary power series in $1/z$ and so

$$[f]_z = [f]_{1/z} \left(\frac{d(1/z)}{dz}\right)^2 = [f]_{1/z} \frac{1}{z^4}.$$

Hence

The function $[f]_z$ is regular at $z = \infty$ and has a zero of multiplicity at least four there, provided that $z = \infty$ does not correspond to a vertex of the polygon.

By the Cauchy-Liouville theorem (section 2.12.1) the function $\chi(z)$ is a constant, in this case the constant zero. Thus

If $a_1, \ldots, a_n$ are finite, then the mapping function satisfies Schwarz's differential equation

$$\boxed{[f]_z = \tfrac{1}{2} \sum_{\nu=1}^{n} \frac{1-\alpha_\nu^2}{(z-a_\nu)^2} + \sum_{\nu=1}^{n} \frac{C_\nu}{z-a_\nu}}. \tag{14.1–5}$$

If is often convenient, in the case that we map the upper half of the x-plane onto a curvilinear polygon, to admit that $z = \infty$ is singular, i.e., that it corresponds to the vertex b_n of the n-gon. Reasoning along the same lines as in the cases of a finite singularity and observing that now

$$t = c_1 \left(\frac{1}{z}\right) + c_2 \left(\frac{1}{z}\right)^2 + \ldots$$

we easily find that

$$[f]_z = \tfrac{1}{2} \frac{1-\alpha_n^2}{z^2} + \ldots, \tag{14.1.6}$$

where the omitted terms constitute a power series in $1/z$, beginning with $1/z^3$. Thus

If $z = \infty$ corresponds to a vertex of the polygon, then $[f]_z$ has a zero of multiplicity at least two at $z = \infty$. The function f satisfies the differential equation (14.1–5), *where n is replaced by $n-1$.*

14.1.2 – Fundamental relations

The various constants occurring in (14.1–5) are not independent. Expanding $[f]_z$ in a neighbourhood of $z = \infty$ we get

$$[f]_z = \frac{1}{z}\sum_{\nu=1}^{n} C_\nu + \frac{1}{z^2}\sum_{\nu=1}^{n}\left(\frac{1-\alpha_\nu^2}{2}+C_\nu a_\nu\right) + $$
$$+\frac{1}{z^3}\sum_{\nu=1}^{n}((1-\alpha_\nu^2)a_\nu + C_\nu a_\nu^2) + \ldots .$$

Assuming that $f(z)$ is regular at $z=\infty$ the function $[f]_z$ has a zero of at least the fourth order there and we conclude that

$$\sum_{\nu=1}^{n} C_\nu = 0, \tag{14.1–7}$$

$$\sum_{\nu=1}^{n} (\tfrac{1}{2}(1-\alpha_\nu^2)+C_\nu a_\nu) = 0, \tag{14.1–8}$$

and

$$\sum_{\nu=1}^{n} ((1-\alpha_\nu^2)a_\nu + C_\nu a_\nu^2) = 0. \tag{14.1–9}$$

By means of these fundamental relations $n-3$ among the constants $C_1,\ldots,C_n$ can be evaluated, provided that $a_1,\ldots,a_n,\alpha_1,\ldots,\alpha_n$ are known. Thus the mapping problem depends on $n-3$ undetermined constants, the so-called *accessory parameters* of the problem.

It is clear that if the vertices and the angles of the polygon are given, the numbers $a_1,\ldots,a_n, C_1,\ldots, C_n$ are determined. The actual evaluation of these constants is, however, an extremely difficult task, except in a few special cases.

It is instructive to investigate what happens if we add a singularity a_{n+1} and let $a_{n+1}\to\infty$. The equation (14.1–5) contains again the terms corresponding to the finite singularities. Writing (14.1–7), (14.1–8) and (14.1–9) as

$$\sum_{\nu=1}^{n} C_\nu + C_{\nu+1} = 0,$$

$$\sum_{\nu=1}^{n} (\tfrac{1}{2}(1-\alpha_\nu^2)+C_\nu a_\nu) + \tfrac{1}{2}(1-\alpha_{n+1}^2) + C_{n+1}a_{n+1} = 0,$$

$$\sum_{\nu=1}^{n} ((1-\alpha_\nu^2)a_\nu + C_\nu a_\nu^2) + (1-\alpha_{n+1}^2)a_{n+1} + C_{n+1}a_{n+1}^2 = 0,$$

we see immediately that from the second of these equations follows

$$\lim_{a_{n+1}\to\infty} C_{n+1}a_{n+1} = -\tfrac{1}{2}(1-\alpha_{n+1}^2) - \sum_{\nu=1}^{n}(\tfrac{1}{2}(1-\alpha_\nu^2)+C_\nu a_\nu)$$

and from the last

$$\lim_{a_{n+1}\to\infty} C_{n+1}a_{n+1} = -(1-\alpha_{n+1}^2).$$

Hence $C_{n+1} \to 0$ and so we get the relation (14.1–7). In addition we now have

$$\sum_{\nu=1}^{n} \left(\tfrac{1}{2}(1-\alpha_\nu^2)+C_\nu a_\nu\right) = \tfrac{1}{2}(1-\alpha_{n+1}^2). \tag{14.1–10}$$

14.1.3 – THE SCHWARZ–CHRISTOFFEL FORMULA

The method for obtaining Schwarz's differential equation can also be applied to the case of a rectilinear polygon. But now things simplify considerably, for we have only to consider transformations of the type

$$w \to Aw+B$$

and for these transformations already

$$\frac{w''}{w'}$$

remains invariant. Evaluating this invariant in the various cases we obtain the result that it has simple poles at $a_1, \ldots, a_n$ with residues $\alpha_1 - 1, \ldots, \alpha_n - 1$ respectively, assuming that these points are finite. There is a double zero at $z = \infty$ and this becomes a simple zero if $a_n = \infty$. In the case of finite singularities we have

$$\frac{w''}{w'} = \sum_{\nu=1}^{\infty} \frac{\alpha_\nu - 1}{z-a_\nu},$$

whence

$$w' = c\prod_{\nu=1}^{n}(z-a_\nu)^{\alpha_\nu - 1} = c\prod_{\nu=1}^{n}(z-a_\nu)^{-\lambda_\nu}.$$

This yields the Schwarz-Christoffel formula (10.3–9).

It is tacitly assumed that $w = \infty$ is not an interior point of the polygon. If $w(z)$ maps the upper half of the z-plane onto the exterior of the polygon, then with the conventions of section 10.3.5 we find that

$$\frac{w''}{w'} = \sum_{\nu=1}^{n} \frac{\lambda_\nu}{z-a_\nu} - \frac{2}{z-a} - \frac{2}{z-\bar{a}},$$

where $z = a$ corresponds to $w = \infty$. By adjusting constants we may take $a = i$. Then (10.3–17) follows. The fundamental relations reduce to

$$\sum_{\nu=1}^{n} \lambda_\nu = 2,$$

which is geometrically evident. There are no accessory parameters.

14.2 – The Schwarzian triangles and their associated groups

14.2.1 – The problem of inversion

An interesting class of functions is related to the problem of the mapping of the upper half of the z-plane onto a curvilinear triangle in the s-plane if we ask under which circumstances the mapping function is extendible throughout the whole s-plane as a single-valued function. By

$$s(\alpha_1, \alpha_2, \alpha_3; z) \tag{14.2–1}$$

we shall denote a function which maps the upper half of the z-plane onto a curvilinear triangle with angles $\alpha_1\pi$, $\alpha_2\pi$, $\alpha_3\pi$, such that the points $z = 0, 1, \infty$ correspond to the vertices of the triangle respectively. The function (14.2–1) is a solution of Schwarz's differential equation

$$[s]_z = R(z), \tag{14.2–2}$$

with

$$R(z) = \frac{1-\alpha_1^2}{2z^2} + \frac{1-\alpha_2^2}{2(z-1)^2} + \frac{C_1}{z} + \frac{C_2}{z-1}.$$

In view of (14.1–7) (with $n = 2$) and (14.1–10) (with $n = 3$) we have

$$C_1 + C_2 = 0,$$
$$\tfrac{1}{2}(1-\alpha_1^2) + \tfrac{1}{2}(1-\alpha_2^2) + C_2 = \tfrac{1}{2}(1-\alpha_3^2),$$

whence

$$\boxed{R(z) = \frac{1-\alpha_1^2}{2z^2} + \frac{1-\alpha_2^2}{2(z-1)^2} + \frac{\alpha_1^2+\alpha_2^2-\alpha_3^2-1}{2z(z-1)}.} \tag{14.2–3}$$

In this case there are no accessory parameters.

Throughout the closed triangle the function $s(z)$ is invertible, its inverse being denoted by $z(s)$. This latter function can be continued analytically beyond any side of the triangle by applying the reflection principle and we obtain another triangle which corresponds to the lower half of the z-plane. On repeating this process in all possible ways we obtain a system of triangles corresponding the upper or the lower half of the z-plane alternately. In a figure we shall shade the triangles corresponding to the upper half plane.

In general the function thus obtained is not a single-valued function in the s-plane, for triangles of the system may overlap. Assume now that none of the numbers α_1, α_2, α_3 is equal to 0 or 1. By a Möbius transformation we can carry the triangle into another triangle with its vertex at the origin and two rectilinear sides issuing from this vertex.

Reflecting in these sides and repeating this process we get a star of triangles around the origin and these triangles fill up a neighbourhood of the origin without gaps or overlappings if and only if the angle at the origin is π/γ, where γ is an integer > 1. Hence the analytic function $z(s)$ takes again its original value if s describes a small circle about the origin. If the same condition holds for the other vertices we have

The inverse $z(s)$ of the function which maps the upper half of the z-plane onto a curvilinear triangle with angles $\alpha_1\pi$, $\alpha_2\pi$, $\alpha_3\pi$ is continuable as a single-valued function beyond the sides of the triangle if and only if the numbers α_1, α_2, α_3 are the reciprocals of integers > 1, assuming that none of the angles are zero.

The case that one or more of the angles is zero need not be excluded. Then we interpret the reciprocal of α as ∞.

The functions (14.2–1) with $\alpha_1 = 1/\gamma_1$, $\alpha_2 = 1/\gamma_2$, $\alpha_3 = 1/\gamma_3$, where the γ's are integers or ∞, are called *Schwarzian triangle functions.*

On applying repeatedly the process of reflection with respect to the sides of a triangle and to those of the triangles thus obtained we get a group of linear fractional transformations which carry any shaded triangle into any other shaded triangle. We shall list the groups further on and it will turn out that they are properly discontinuous. The inverse function $z(s)$ of the corresponding Schwarzian function is a simple automorphic function. We shall see that the region of discontinuity is either the extended plane, the finite plane or the interior of a circle. A shaded and an unshaded triangle together form a fundamental domain for the group. Two vertices opposite to the common side form a cycle. The other cycles consist of only one vertex.

Summing up we have

The Schwarzian functions are the inverses of simple automorphic functions with respect to a group of linear transformations obtained by carrying any shaded triangle into any other shaded triangle by a sequence of reflections in sides, provided the angles of the triangles are π/γ_1, π/γ_2 and π/γ_3, where γ_1, γ_2, γ_3 are certain integers >1 or ∞.

14.2.2 – The three kinds of Schwarzian functions

The Schwarzian functions can be divided into three classes:

A) *The Schwarzian functions of the first kind* are characterized by

$$\frac{1}{\gamma_1} + \frac{1}{\gamma_2} + \frac{1}{\gamma_3} > 1. \tag{14.2–4}$$

Since $\gamma_1 \geqq 3$, $\gamma_2 \geqq 3$, $\gamma_3 \geqq 3$ simultaneously is impossible, at least one of the numbers γ_1, γ_2, γ_3 is equal to 2, say γ_1. Then remains the inequality

$$\frac{1}{\gamma_2} + \frac{1}{\gamma_3} > \tfrac{1}{2}$$

and it is clear that $\gamma_2 \geqq 4$, $\gamma_3 \geqq 4$ is impossible.

Let $\gamma_2 < 4$, i.e., either $\gamma_2 = 2$ or $\gamma_2 = 3$. If $\gamma_2 = 2$ we have

$$\frac{1}{\gamma_3} > 0$$

and γ_3 may be any integer > 1. If $\gamma_2 = 3$ then

$$\frac{1}{\gamma_3} > \tfrac{1}{6},$$

whence $\gamma_2 = 2, 3, 4$ or 5. The case $\gamma_3 = 2$ is already covered by $\gamma_1 = \gamma_2 = 2$. Summing up we have

A1) $\gamma_1 = 2, \quad \gamma_2 = 2, \quad \gamma_3 = n \geqq 2,$

A2) $\gamma_1 = 2, \quad \gamma_2 = 3, \quad \gamma_3 = 3,$

A3) $\gamma_1 = 2, \quad \gamma_2 = 3, \quad \gamma_3 = 4,$

A4) $\gamma_1 = 2, \quad \gamma_2 = 3, \quad \gamma_3 = 5.$

B) *The Schwarzian functions of the second kind* are those characterized by

$$\frac{1}{\gamma_1} + \frac{1}{\gamma_2} + \frac{1}{\gamma_3} = 1. \tag{14.2–5}$$

If we suppose that $\gamma_1 \leqq \gamma_2 \leqq \gamma_3$, we find

$$\frac{3}{\gamma_1} \geqq 1.$$

Hence $\gamma_1 = 2$ or $\gamma_1 = 3$. If $\gamma_1 = 2$ then

$$\frac{1}{\gamma_2} + \frac{1}{\gamma_3} = \tfrac{1}{2},$$

whence

$$\frac{2}{\gamma_2} \geqq \tfrac{1}{2}.$$

Hence $\gamma_2 = 3$ and $\gamma_3 = 6$, or $\gamma_2 = 4$, $\gamma_3 = 4$. If $\gamma_1 = 3$ then

$$\frac{1}{\gamma_2} + \frac{1}{\gamma_3} = \tfrac{2}{3},$$

whence

$$\frac{2}{\gamma_2} \geqq \tfrac{2}{3}.$$

This yields $\gamma_3 = 3$.

If we take $\gamma_1 = \gamma_2 = 2$, then $1/\gamma_3 = 0$. We shall adopt the improper solution $\gamma_3 = \infty$, corresponding to a triangle having a zero angle. Summing up we have

B1) $\gamma_1 = 2,\quad \gamma_2 = 2,\quad \gamma_3 = \infty,$

B2) $\gamma_1 = 2,\quad \gamma_2 = 3,\quad \gamma_3 = 6,$

B3) $\gamma_1 = 2,\quad \gamma_2 = 4,\quad \gamma_3 = 4,$

B4) $\gamma_1 = 3,\quad \gamma_2 = 3,\quad \gamma_3 = 3.$

C) *The Schwarzian functions of the third kind* are those characterized by

$$\frac{1}{\gamma_1} + \frac{1}{\gamma_2} + \frac{1}{\gamma_3} < 1. \tag{14.2–6}$$

There are infinitely many sets of solutions, for if (14.2–6) is satisfied by some values of γ_1, γ_2 and γ_3, then also by larger values. We may adopt improper solutions, corresponding to triangles with one, two or three zero angles.

14.2.3 – The patterns of Schwarzian triangles and their group of automorphisms

Consider a shaded triangle Δ in the s-plane with vertices s_1, s_2, s_3

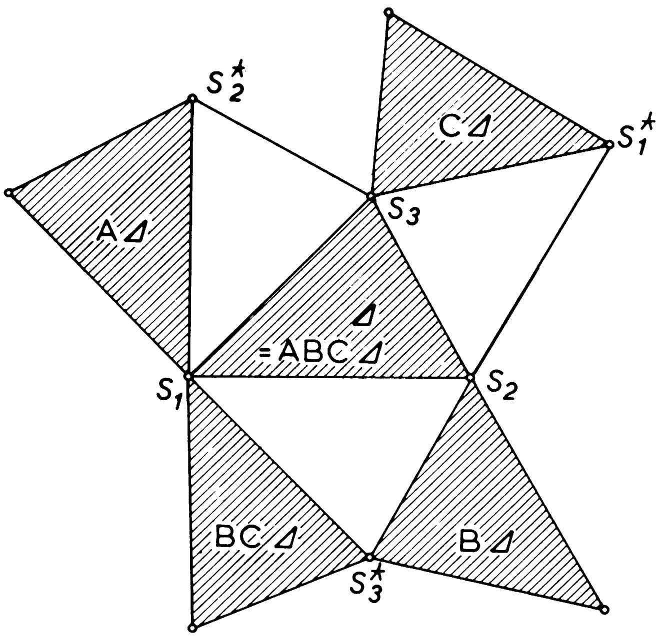


Fig. 14.2–1. Successive reflections of a triangle in its sides (schematic; the situation for curvilinear triangles in similar)

and angles π/γ_1, π/γ_2 and π/γ_3 respectively, wheie γ_1, γ_2 and γ_3 have values listed in the previous section. A triangle of this kind will be called a *Schwarzian triangle*. Reflecting Δ in the side connecting s_3 and s_1 we obtain an unshaded triangle Δ^* with vertices s_1, s_2^* and s_3, (fig. 14.2–1). Reflecting Δ^* in the side connecting s_1 and s_2^* we get a shaded triangle which is obtained from Δ by means of an elliptic or parabolic transformation A, with fixed point s_1. The order of A is γ_1. In a similar way we can define a transformation B with fixed point s_2 and order γ_2 and a transformation C with fixed point s_3 and order γ_3. The triangles obtained from Δ by means of the transformation A, B, C will be denoted by AΔ, BΔ, CΔ respectively.

It is easy to see that the same construction applied to the triangle AΔ yields triangles AAΔ, BAΔ, CAΔ, etc. Hence the shaded triangles obtained by repeated reflections in the sides may be denoted by

$$\mathsf{A}^{\kappa_1}\mathsf{B}^{\lambda_1}\mathsf{C}^{\mu_1}\ldots\mathsf{A}^{\kappa_p}\mathsf{B}^{\lambda_p}\mathsf{C}^{\mu_p}\Delta \tag{14.2–7}$$

and the group of automorphisms of the pattern of all shaded triangles is generated by the transformations A, B and C.

Denoting by A^{-1} the inverse of the transformation A, a simple geometric consideration reveals that $\mathsf{A}^{-1}\Delta$ and BCΔ coincide. Thus we have

The group of linear fractional transformation, carrying any shaded Schwarzian triangle into any other shaded triangle by a sequence of reflections in sides, is generated by three transformations A, B, C, *satisfying the defining relations*

$$\boxed{\mathsf{A}^{\gamma_1} = \mathsf{E}, \quad \mathsf{B}^{\gamma_2} = \mathsf{E}, \quad \mathsf{C}^{\gamma_3} = \mathsf{E}, \quad \mathsf{ABC} = \mathsf{E},} \tag{14.2–8}$$

E *denoting the identity transformation.*

14.2.4 – The Schwarzian triangles of the first kind

Let the sum of the angles of a Schwarzian triangle with vertices s_1, s_2 and s_3 be greater than π. There is no zero angle and the two sides which issue from a vertex, say s_3, meet again at a point s_3'. A linear transformation can be made which carries s_3' to ∞ and s_3 to the origin O in the s-plane. The new triangle has two rectilinaer sides issuing from O and making there an angle π/γ_3. Since the sum of the angles of the triangle is greater than π, the third side is a circular arc concave towards O. Through O we draw a chord of the circle of the third side which is bisected by O, (fig. 14.2–2). Without loss of generality we may assume that the length of this chord is 2. The circle with this chord as a diameter is intersected by each side of the triangle at diametrically situated points.

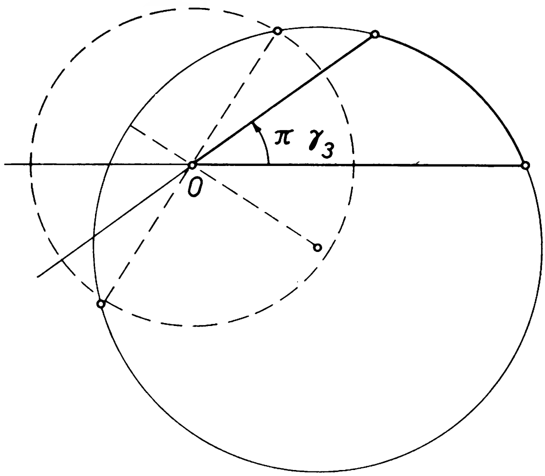


Fig. 14.2–2. Triangle with two rectilinear sides and its projection onto a sphere

Now let the s-plane be projected stereographically onto the sphere, having this circle as equator, (section 1.1.3). The sides of the triangle are carried into three circular arcs on the sphere, lying on circles which pass through opposite ends of a diameter of the equator. Hence the sides of the corresponding triangle on the sphere are arcs of great circles.

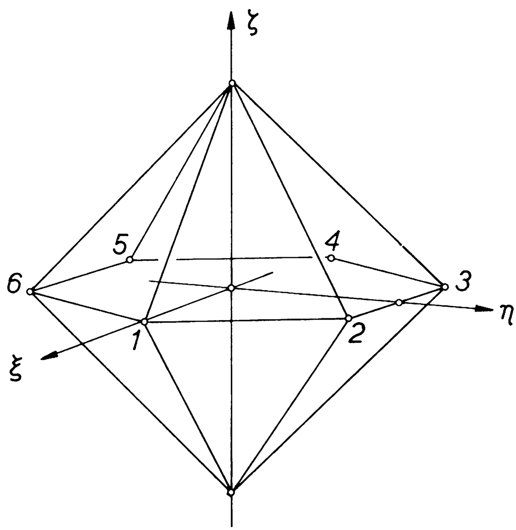


Fig. 14.2–3. Bipyramid

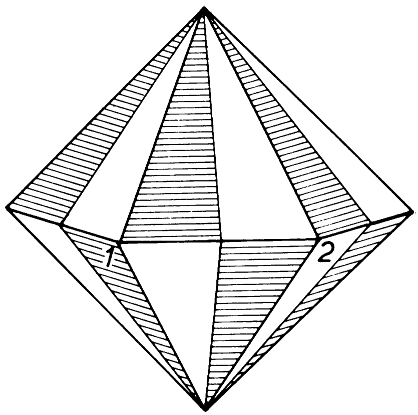


Fig. 14.2–4. The shaded and unshaded triangles on the faces of a bipyramid

A reflection in a side of the triangle in the plane corresponds to a reflection in a diametrical plane of the sphere. Thus it appears that the group associated with the triangle is isomorphic to a group of rotations of the sphere. Since the area of the sphere is a finite number and all triangles obtained by stereographic projection have equal positive area, we find that the group is finite. Evaluating in each case the spherical excess, we find that the number N of the shaded triangles is

$$N = 2n, 12, 24, 60 \qquad (14.2\text{–}9)$$

in the various cases listed in section 14.2.2 under A). Thus the numbers N are the orders of the groups corresponding to these cases. The triangles fit together without overlappings or gaps and fill up the entire sphere.

A1) We describe in the equator of the sphere a regular n-gon; its vertices will be denoted by *1*, *2*, . . ., n. Connecting these vertices with the north pole and the south pole we obtain a bipyramid having $2n$ triangular faces, (fig. 14.2–3). Each of these faces can be divided into shaded and unshaded triangles, having the midpoint of a side of the n-gon in common, (fig. 14.2–4). Projecting this figure from O onto the sphere we get a division of the sphere into $2n$ pairs of triangles by the equator and n complete meridians. The angles of these triangles are $\pi/2$, $\pi/2$ and π/n and stereographic projection onto the s-plane yields the pattern as depicted in fig. 14.2–5. It is understood that the point *1* is on the positive ξ-axis.

We proceed to investigate the group of automorphism of this pattern. The regular polygon inscribed in the equator can be looked upon as a regular polyhedron with two coincident faces, n vertices and n edges.

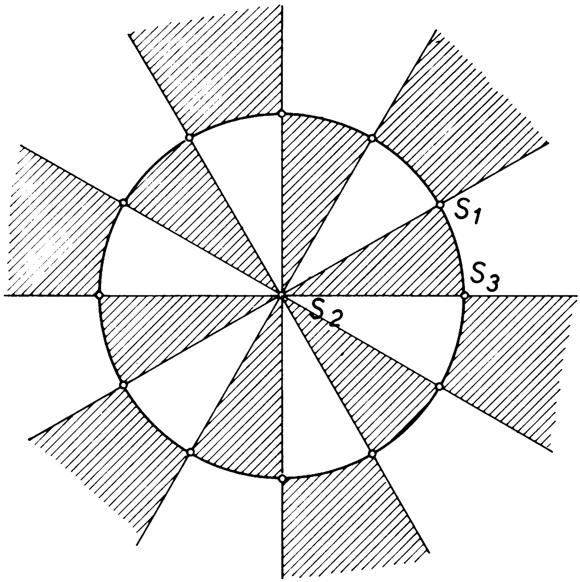


Fig. 14.2–5. The pattern of the dihedral group

The rotations of the sphere carrying this n-gon into itself carries also the bipiramid into itself and yield, therefore, automorphisms of the pattern of triangles. Excluding the identity, there are $n-1$ rotations of the sphere about the vertical axis through multiples of the angle $2\pi/n$ and bringing the n-gon into coincidence with itself. If n is even there are $n/2$ diameters connecting diametral vertices and $n/2$ diameters bisecting sides of the n-gon. If n is odd there are n diameters through vertices which bisect sides. A half turn about such a diameter brings the n-gon into coincidence with itself, but interchanges the upper and the lower side. Together with the identity we find at least $n-1+n+1 = 2n$ rotations of the sphere which carries the n-gon into itself. On the other hand there

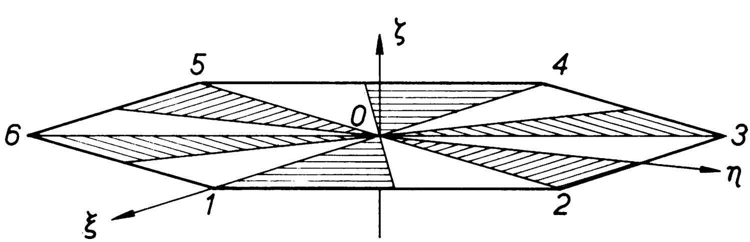


Fig. 14.2–6. Dihedron with shaded and unshaded triangles

are $2n$ such rotations at most. For there are two rotations, the identity included, which leaves a given vertex unchanged. Since by rotations of the sphere every vertex can be moved into any other vertex the number of rotations does not exceed $2n$. Hence the number is exactly $2n$.

The figure discussed above is often referred to as a *dihedron*, a regular polyhedron with two coincident faces, (fig. 14.2–6). The group of rotations of the sphere carrying this figure into itself is the *dihedral group*.

Let A denote the rotation through π about the axis bisecting the edge between the points *1* and *2*. It is clear that A induces a permutation

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 2 & 1 & n & \dots & 3 \end{pmatrix}$$

of the vertices of the n-gon. If B denotes a similar rotation about the axis through the point *2*, then B induces the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & \dots & n \\ 3 & 2 & 1 & n & \dots & 4 \end{pmatrix}.$$

Hence $\mathsf{C} = \mathsf{B}^{-1}\mathsf{A} = \mathsf{BA}$ induces the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & \dots & n \\ 3 & 2 & 1 & n & \dots & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 2 & 1 & n & \dots & 3 \end{pmatrix} = \begin{pmatrix} 2 & 1 & n & \dots & 3 \\ 2 & 3 & 4 & \dots & n \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 2 & 1 & n & \dots & 3 \end{pmatrix}$$
$$= (1\ 2 \dots n),$$

a cyclic permutation of order n. It is induced by a rotation about the vertical axis through the angle $2\pi/n$.

Obviously $\mathsf{C} = \mathsf{B}^{-1}\mathsf{A}$ is equivalent to $\mathsf{BC} = \mathsf{A}$ or $\mathsf{ABC} = \mathsf{A}^2 = \mathsf{E}$. Thus we may conclude

The dihedral group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$\mathsf{A}^2 = \mathsf{E}, \qquad \mathsf{B}^2 = \mathsf{E}, \qquad \mathsf{C}^n = \mathsf{E}, \qquad \mathsf{ABC} = \mathsf{E}. \qquad (14.2\text{–}10)$$

The limiting case of the dihedral group in which $n = 2$ is the *four group*. The corresponding division of the sphere consists of four shaded and four unshaded triangles, being octants of the sphere. In the case $n = 3$ the dihedral group is isomorphic to the symmetric group of 3 symbols.

A2) In the sphere we inscribe a cube whose faces are parallel to the coordinate planes, (fig. 14.2–7). Its vertices constitute the vertices of two desmic tetrahedra *1234* and *1'2'3'4'*. We take *12* above the ξ, η-plane and *1* in the first octant. It is understood that *1* and *1'* are diametrical points as are *2* and *2'*, etc. The points *1'*, *2'*, *3'* and *4'* can also be considered as the projection from O of the centres of the faces of the tetrahedron *1234*.

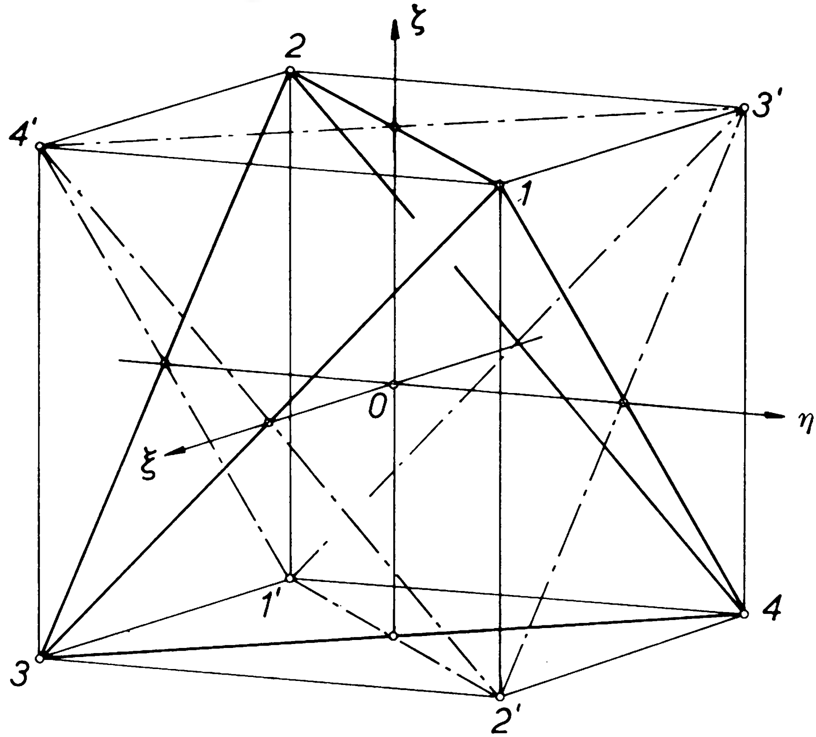


Fig. 14.2–7. The desmic tetrahedrons inscribed in a cube

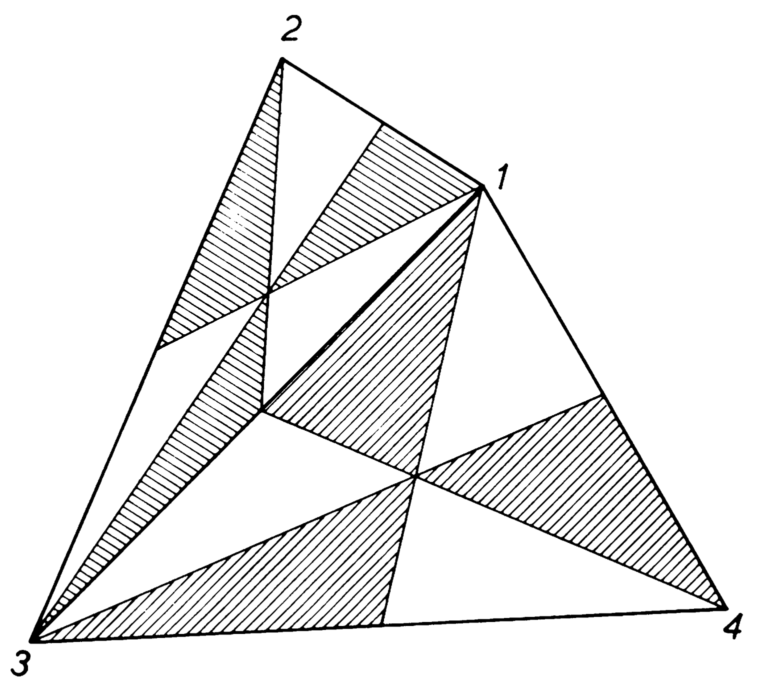


Fig. 14.2–8. The shaded and unshaded triangles on the faces of a tetrahedron

Each face can be divided into 3 shaded and unshaded triangles, these triangles being congruent, (fig. 14.2–8). Projecting this figure from O onto the sphere we obtain a pattern of *12* shaded and *12* unshaded triangles, the angles of the spherical triangles being $\pi/2$, $\pi/3$, $\pi/4$. We have the case A1) of section 14.2.2. The stereographic projection onto the s-plane is depicted in fig. 14.2–9. The same pattern is obtained by division of the twin tetrahedron *1'2'3'4'*.

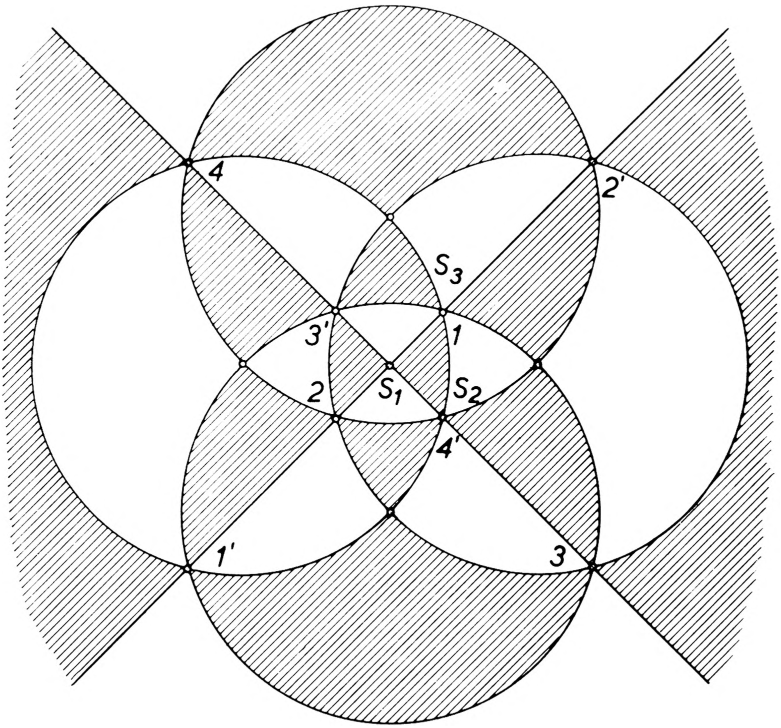


Fig. 14.2–9. Stereographic projection of the tetrahedral pattern

We proceed to investigate the group of covering transformations of the tetrahedron. This polyhedron is transformed into itself by rotation about an axis through a vertex through angles $2\pi/3$ and $4\pi/3$. Also by rotating about an axis bisecting two opposite edges through the angle π. Together with the identity we obtain at least $4\times 2+3+1 = 12$ rotations, bringing the tetrahedron *1234* (and also the tetrahedron *1'2'3'4'*) into coincidence with itself. On the other hand there are 3 rotations (including the identity) leaving a vertex invariant. Since each vertex can be carried into any other vertex by a rotation of the sphere, the number of rotations does not exceed $4 \times 3 = 12$.

The group of rotations bringing the tetrahedron into coincidence with itself is called the *tetrahedral group.* Its order is 12.

Let A denote the rotation through the angle π about the axis bisecting the edges *12* and *34*, and B the rotation through the angle $2\pi/3$ about the axis through the point *2.* It is clear that A induces the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = (1\ 2)(3\ 4)$$

of the vertices of the tetrahedron and that B induces the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix} = (1\ 3\ 4)$$

Hence $\mathsf{B}^{-1}\mathsf{A}$ induces the permutation

$$(1\ 4\ 3)(1\ 2)(3\ 4) = (1\ 2\ 4)$$

being induced by a rotation about the axis through the point *3.* Thus we have

The tetrahedral group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$\mathsf{A}^2 = \mathsf{E}, \qquad \mathsf{B}^3 = \mathsf{E}, \qquad \mathsf{C}^3 = \mathsf{E}, \qquad \mathsf{ABC} = \mathsf{E}. \qquad (14.2\text{–}11)$$

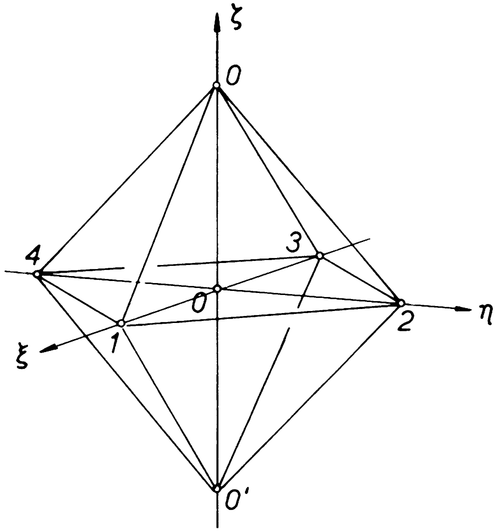


Fig. 14.2–10. The octahedron

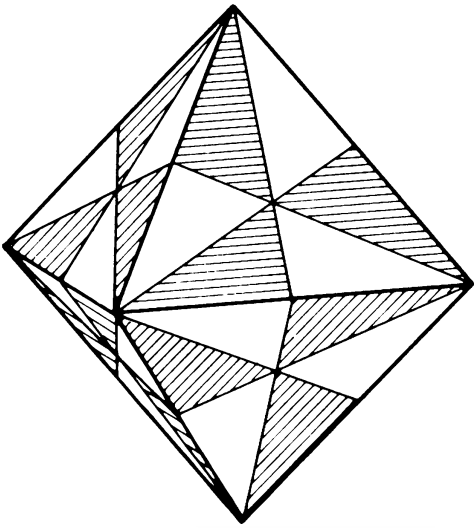


Fig. 14.2–11. The shaded and unshaded triangels on the faces of an octahedron

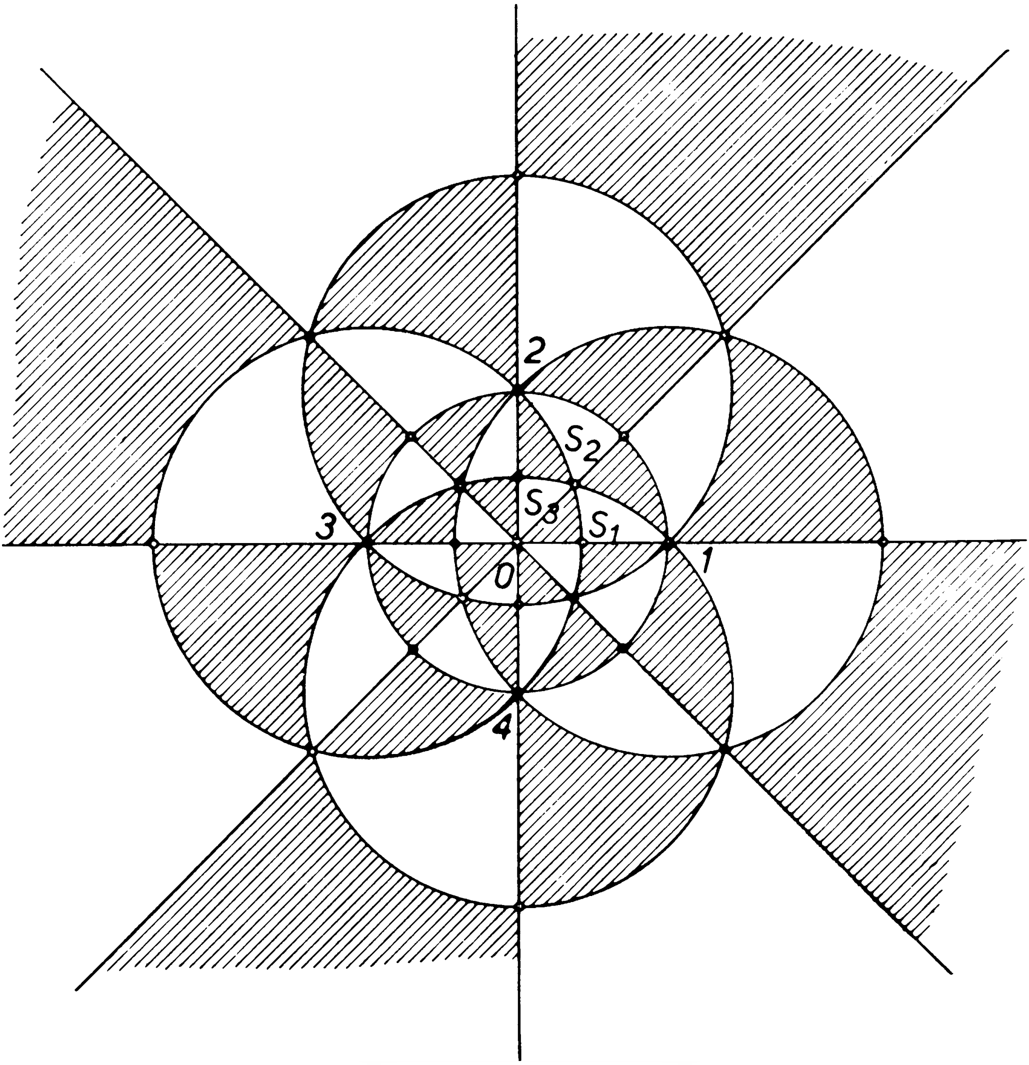


Fig. 14.2–12. Stereographic projection of the octahedral pattern

The permutations induced by A and B are even permutations of the vertices of the tetrahedron. Hence they generate a subgroup of the alternating group of four symbols. Since the order of the tetrahedral group is $12 = \frac{1}{2} \times 4!$ and only the identical transformation corresponds to the identical permutation we may conclude

The tetrahedral group is isomorphic to the alternating group of four symbols.

A3) The centres of a face of a cube are the vertices of a regular octahedron and, conversely, the centres of the faces of a octahedron are the vertices of a cube. We take the octahedron as in fig. 14.2–10 and denote the vertices by *0*, *0′*, *1*, *2*, *3*, *4*, where *0* and *0′* are the north pole and the south pole of the sphere and *1* is on the positive ξ-axis.

The faces of the octahedron can be divided into 3 shaded and 3 unshaded triangles, (fig. 14.2–11) as in the previous section. Projecting from *O* onto the sphere we obtain a pattern of 24 shaded and 24 unshaded triangles, each triangle having angles $\pi/2$, $\pi/3$, $\pi/4$. We are now in the situation A3) of section 14.2.2. By stereographic projection onto the s-plane we obtain the pattern as depicted in fig. 14.2–12. The same pattern is obtained by starting with a cube whose faces are divided into 4 shaded and 4 unshaded triangles, (fig. 14.2–13). Hence the group of automorphisms of the cube is the same as that of the octahedron.

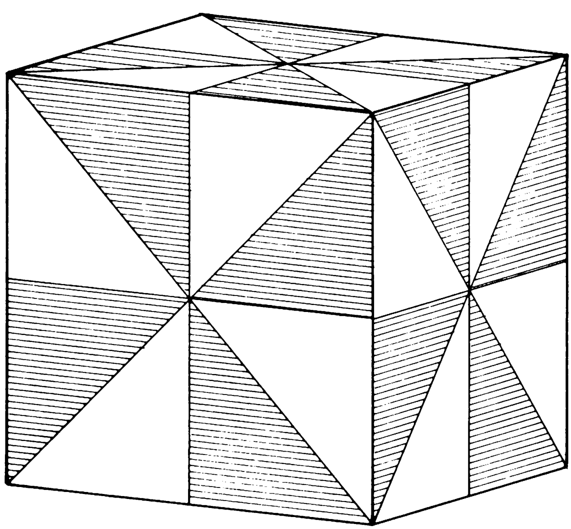


Fig. 14.2–13. The shaded and unshaded triangels on the faces of a cube

We turn to the description of this group. There are 2 rotations different from the identity through angles $2\pi/3$ and $4\pi/3$ about an axis connecting the centres of opposite faces of the octahedron which bring the polydron into coincidence with itself. Similarly there are 3 rotations about an axis connecting diametrical vertices through angles $2\pi/4$, $4\pi/4$, $6\pi/4$. There is one rotation through an angle π about an axis bisecting opposite edges. Including the identity we find at least $4\times2+3\times3+6+1 = 24$ rotations, carrying the octahedron into itself. On the other hand there are four rotations (including the identity) leaving a vertex invariant and, since each vertex can be carried into any other vertex by means of rotations of the sphere, we find at most $6\times4 = 24$ rotations.

The group of automorphisms of an octahedron is called the *octahedral group*. Its order is 24.

Let A denote the rotation through π about the axis bisecting the edge *01*. It induces the permutation

$$\begin{pmatrix} 0 & 0' & 1 & 2 & 3 & 4 \\ 1 & 3 & 0 & 4 & 0' & 2 \end{pmatrix}$$

of the vertices of the octahedron. The transformation B, being a rotation through $2\pi/3$ about the axis through the centre of the face *012*, induces the permutation

$$\begin{pmatrix} 0 & 0' & 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 0 & 4 & 0' \end{pmatrix}.$$

The transformation $C = B^{-1}A$ induces the permutation

$$\begin{pmatrix} 1 & 3 & 0 & 4 & 0' & 2 \\ 0 & 0' & 2 & 3 & 4 & 1 \end{pmatrix}\begin{pmatrix} 0 & 0' & 1 & 2 & 3 & 4 \\ 1 & 3 & 0 & 4 & 0' & 2 \end{pmatrix} = (2\ 3\ 4\ 1),$$

being effected by a rotation through $\pi/4$ about the axis connecting *0* and *0'*. Thus we have

The octahedral group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$A^2 = E, \qquad B^3 = E, \qquad C^4 = E, \qquad ABC = E. \qquad (14.2\text{–}12)$$

Let us denote by *I* the axis through the centres of *012* and *0'34*. We shall write

$$I = \{0\ 1\ 2\} = \{0'\ 3\ 4\}.$$

Similarly

$$II = \{0\ 2\ 3\} = \{0'\ 4\ 1\},$$
$$III = \{0\ 3\ 4\} = \{0'\ 1\ 2\},$$
$$IV = \{0\ 4\ 1\} = \{0'\ 2\ 3\}.$$

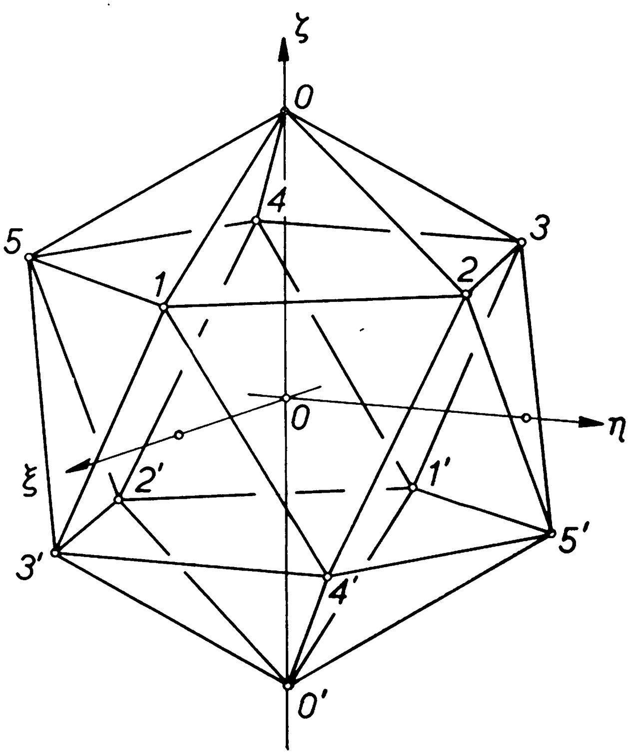


Fig. 14.2–14. The icosahedron

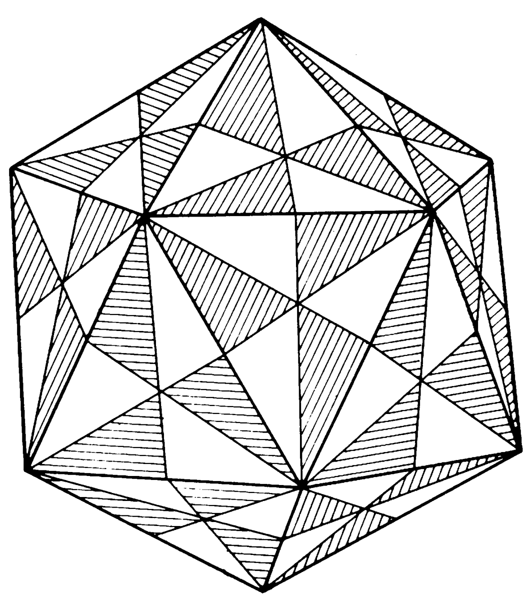


Fig. 14.2–15. The shaded and unshaded triangels on the faces of an icosahedron

Then A induces a permutation of these four elements, viz.,

$$\begin{pmatrix} I & II & III & IV \\ IV & II & III & I \end{pmatrix} = (I\ IV)$$

and B indices the permutation

$$\begin{pmatrix} I & II & III & IV \\ I & IV & II & III \end{pmatrix} = (II\ IV\ III).$$

These permutations generate a subgroup of the symmetric group of four symbols. Since the order of the octahedral group is 24 = 4! we have

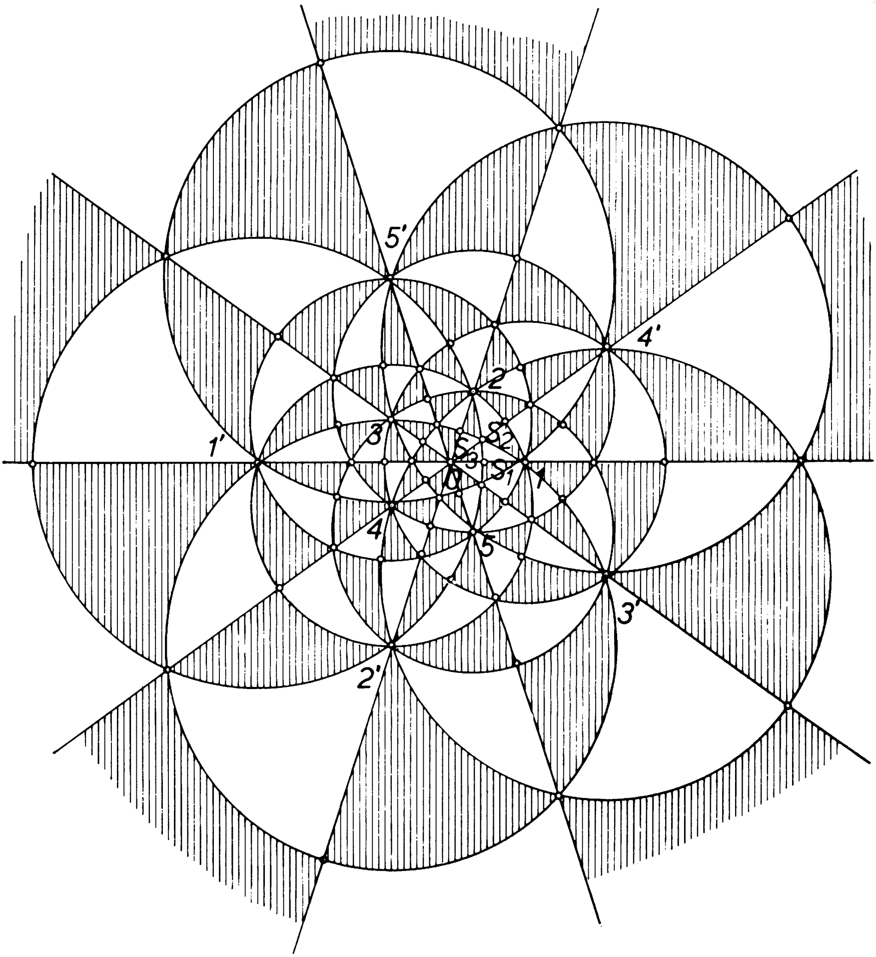


Fig. 14.2–16. Stereographic projection of the icosahedral pattern

The octahedral group is isomorphic to the symmetric group of four symbols.

A4) We consider a regular icosahedron inscribed in the unit sphere such that two vertices coincide with the north pole and the south pole. One of the vertices nearest the north pole is in the plane orthogonal to the η-axis (fig. 14.2–14). The vertices of the polyhedron will be denoted by 0, 1, 2, 3, 4, 5, $0'$, $1'$, $2'$, $3'$, $4'$, $5'$, such that the same symbols, primed and unprimed, denote diametral vertices.

The centres of the faces of a icosahedron is a dodecahedron and conversely. We prefer to study the icosahedron as we did in the case of the octahedron, for its faces are triangles.

The faces of an icosahedron can be divided in the usual way into 3 shaded and 3 unshaded triangles, (fig. 14.2–15). Projecting from O onto the sphere yields a pattern of 60 shaded and 60 unshaded triangles, each spherical triangle having angles $\pi/2$, $\pi/3$ and $\pi/5$. We have the case A4) of section 14.2.2. By stereographic projection onto the s-plane we obtain the pattern as depicted in fig. 14.2–16. The same pattern is obtained by starting with a dodecahedron whose faces are divided into 5 shaded and 5 unshaded triangles, (fig. 14.2–17). Hence the group related to an icosahedron is the same as the group related to a dodecahedron.

The description of the group proceeds along the same lines as in the previous sections. There are two rotations, different from the identity, through angles $2\pi/3$ and $4\pi/3$ about an axis connecting the centres of opposite faces of the octahedron which bring the polyhedron into coincidence with itself. Similarly there are 4 rotations about an axis connecting two opposite vertices through angles $2\pi/5$, $4\pi/5$, $6\pi/5$, $8\pi/5$. There is one rotation through the angle π about an axis bisecting opposite edges. Including the identity we have at least $10\times3+6\times4+5+1 = 60$ rotations, carrying the icosahedron into itself. On the other hand there are 5 rotations (including the identity) leaving a vertex invariant and since each vertex can be carried into any other vertex by means of a rotation of the sphere, we find at most $12\times5 = 60$ rotations. The group of automorphic rotations of an icosahedron is called the *icosahedral group*.

Let A denote the rotation through π about the axis bisecting the edge 01. It induces the permutation

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & 5 & 3' & 4' & 2 \end{pmatrix}$$

of the vertices, where the primed vertices in the upper line are omitted, because they are permuted in quite the same way. The transformation B, being a rotation through $2\pi/3$ about the axis connecting the centres of the

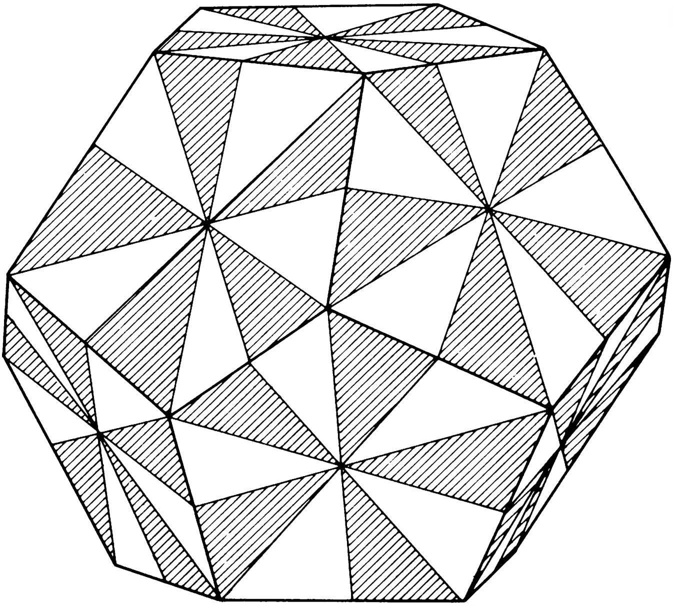


Fig. 14.2–17. The shaded and unshaded triangels on the faces of a dodecahedron

faces *0′2* and *0′1′2′* induces the permutation

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 0 & 5 & 3' & 4' \end{pmatrix},$$

Hence the transformation $\mathsf{C} = \mathsf{B}^{-1}\mathsf{A}$ induces the permutation

$$\begin{pmatrix} 1 & 0 & 5 & 3' & 4' & 2 \\ 0 & 2 & 3 & 4 & 5 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & 5 & 3' & 4' & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 2 & 3 & 4 & 5 & 1 \end{pmatrix} = (1\ 2\ 3\ 4\ 5),$$

i.e., a permutation being induced by a rotation about the vertical axis connecting *0* and *0′* through the angle $2\pi/5$. We may conclude

The octahedral group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$\mathsf{A}^2 = \mathsf{E}, \quad \mathsf{B}^3 = \mathsf{E}, \quad \mathsf{C}^5 = \mathsf{E}, \quad \mathsf{ABC} = \mathsf{E}. \tag{14.2–13}$$

It is easy to see that the plane through the edges *34* and *3′4′* is parallel to the edges *01* and *0′1′*. It bisects the edges *25′* and *2′5*. Thus it appears that the midpoint of the edges *01*, *0′1′*, *34*, *3′4′*, *25′*, *2′5* are the vertices of a regular octahedron, (fig. 14.2–18), which we shall denote by the symbol

$$I = \{01,\ 34,\ 25'\},$$

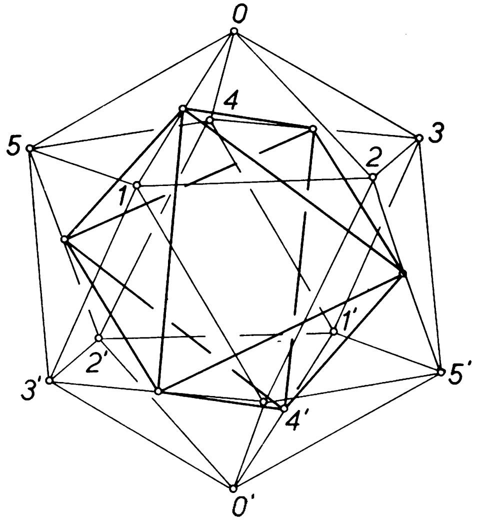


Fig. 14.2–18. Octahedron inscribed in an icosahedron

Similarly

$$\begin{aligned} II &= \{02,\ 45,\ 31'\}, \\ III &= \{03,\ 51,\ 42'\}, \\ IV &= \{04,\ 12,\ 53'\}, \\ V &= \{05,\ 23,\ 14'\}. \end{aligned}$$

A rotation of the polyhedron permutes these octahedrons and only the identity leaves one octahedron fixed. The permutation induced by A is

$$\begin{pmatrix} I & II & III & IV & V \\ I & III & II & V & IV \end{pmatrix} = (II\ III)(IV\ V);$$

the rotation B induces the permutation

$$\begin{pmatrix} I & II & III & IV & V \\ IV & I & III & II & V \end{pmatrix} = (I\ IV\ II).$$

These permutations are even and generate, therefore, a subgroup of the alternating group of five symbols. Since the order of the icosahedral group is $60 = \frac{1}{2} \times 5!$ we have

The icosahedral group is isomorphic to the alternating group of five symbols.

14.2.5 – The Schwarzian triangle of the second kind

A curvilinear triangle whose sum of the angles is equal to π can be transformed into a rectilinear triangle by means of a suitable fractional linear transformation. In this case it is possible to find a tessellation of the finite plane by congruent rectilinear triangles. This facilitates the discussion of the groups considerably.

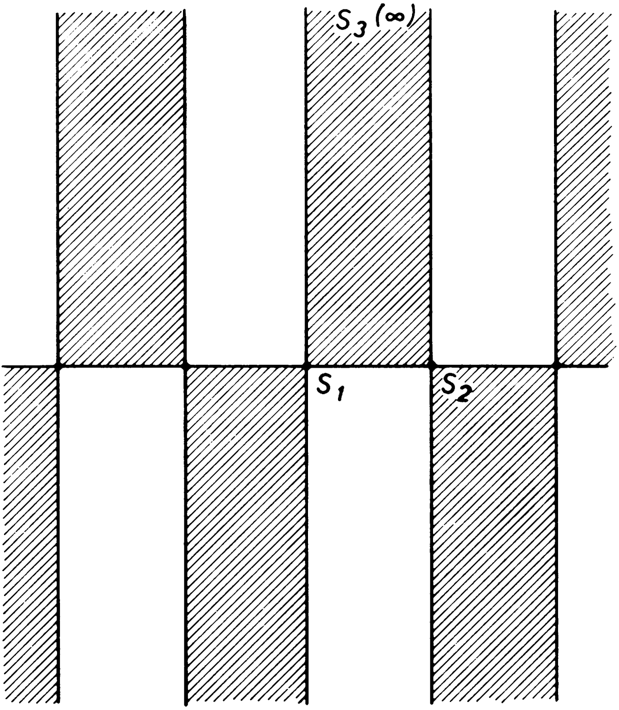


Fig. 14.2–19. Tessellation of the plane by triangles with $\gamma_1 = 2$, $\gamma_2 = 2$, $\gamma_3 = \infty$

B1) If two angles of the triangle are right angles, the third is equal to zero. The triangle is now a half strip and a tessellation of the plane is depicted in fig. 14.2–19. Let A and B denote a rotation about s_1 and s_2 respectively through the angle π. By a simple geometric consideration it is clear that C = BA is a translation. The third vertex s_3 is at infinity and parabolic; it is a limit point of the group. In this and also in the other cases the groups have precisely one limit point.

In the case B1) the group is isomorphic to a group generated by the *elements* A, B, C *and the defining relations*

$$A^2 = E, \qquad B^2 = E, \qquad ABC = E. \tag{14.2–14}$$

The subgroup of translations is isomorphic to the infinite group generated by

$$C = BA. \tag{14.2–15}$$

B2) The triangles with the angles $\pi/2$, $\pi/3$, $\pi/6$ can be arranged as in

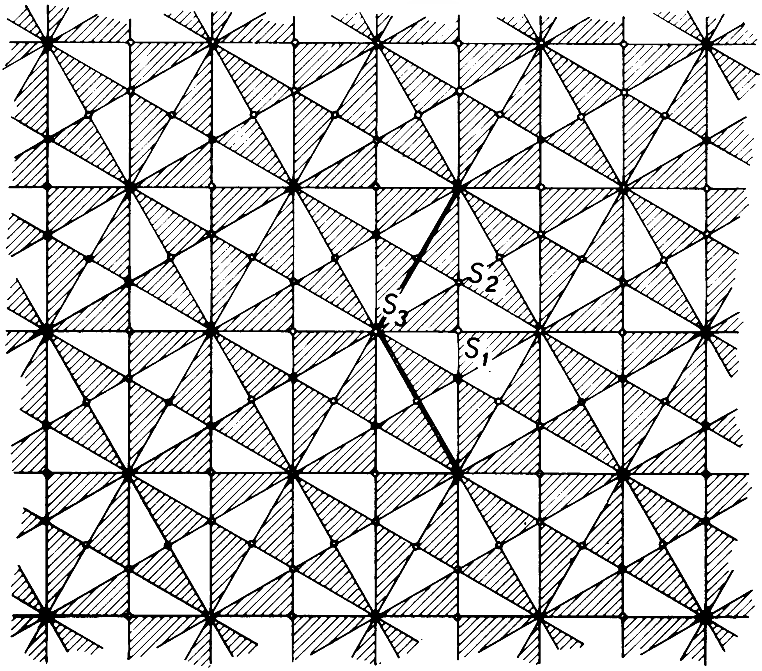


Fig. 14.2–20. Tessellation of the plane by triangles with $\gamma_1 = 2$, $\gamma_2 = 3$, $\gamma_3 = 6$

figure (14.2–20). Let A denote the rotation about s_1 through the angle $2\pi/2$, B the rotation about s_2 through the angle $2\pi/3$ and C the rotation about s_3 through the angle $2\pi/6$. Then

In the case B2) the group is isomorphic to a group generated by the elements ABC *and the defining relations*

$$A^2 = E, \qquad B^3 = E, \qquad C^6 = E, \qquad ABC = E. \tag{14.2–16}$$

It is easy to verify geometrically that the transformations $U = C^4BC^4$ and $V = B^2C^2$ are translations. Hence

The subgroup of translations is isomorphic to the infinite group generated *by the elements*

$$U = C^4BC^4, \qquad V = B^2C^2. \qquad (14.2\text{–}17)$$

B3) The triangles with the angles $\pi/2$, $\pi/4$ and $\pi/4$ constitute a pattern as shown in fig. 14.2–12. Let A denote the rotation about s_1 through the angle π, B and C rotations about s_2 and s_3 respectively through the angle $2\pi/4$. Thus

In the case B3) the group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$A^2 = E, \qquad B^4 = E, \qquad C^4 = E, \qquad ABC = E. \qquad (14.2\text{–}18)$$

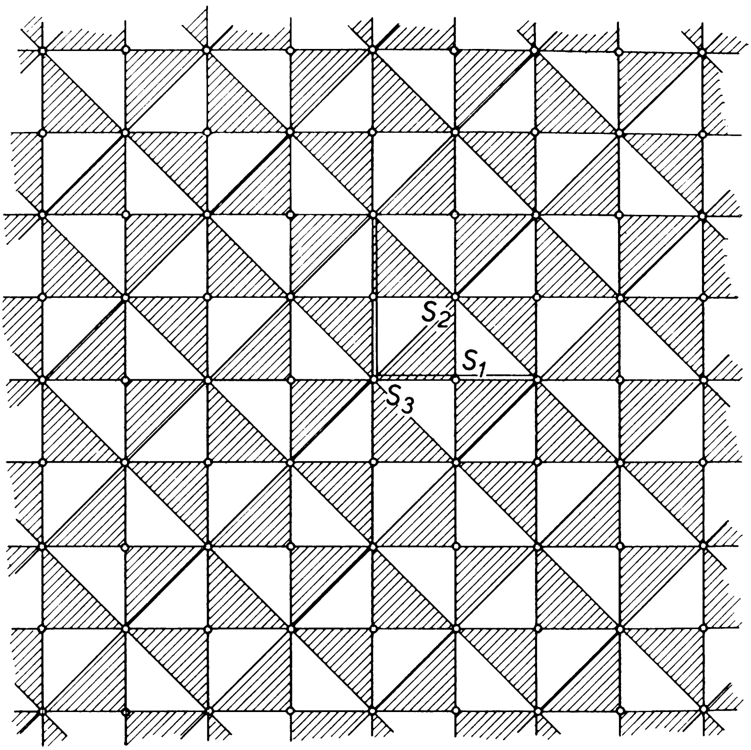


Fig. 14.2–21. Tessellation of the plane by triangles with $\gamma_1 = 2$, $\gamma_2 = 4$, $\gamma_3 = 4$

By a simple geometric consideration we find that the transformations $U = BC^3$ and $V = B^3C$ are translations. Hence

The subgroup of translations is isomorphic to the infinite group generated by the elements

$$U = BC^3, \qquad V = B^3C. \qquad (14.2\text{–}19)$$

B4) The isosceles triangles constitute a pattern as shown in fig. 14.2–22. Let A, B, C denote the rotation about s_1, s_2, s_3 respectively through the angle $2\pi/3$. Then

In the case B4) the group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$\mathsf{A}^3 = \mathsf{E}, \quad \mathsf{B}^3 = \mathsf{E}, \quad \mathsf{C}^3 = \mathsf{E}, \quad \mathsf{ABC} = \mathsf{E}. \tag{14.2–20}$$

The translations are the same as in the case B2), and, evidently, represented by $\mathsf{U} = \mathsf{C}^2\mathsf{BC}^2$, $\mathsf{V} = \mathsf{B}^2\mathsf{C}$.

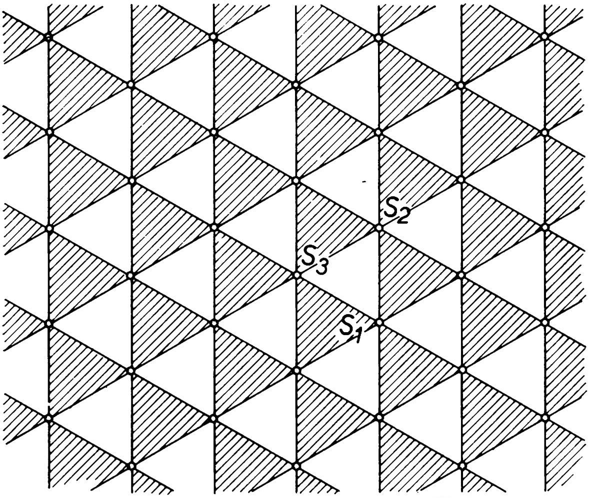


Fig. 14.2–22. Tessellation of the plane by triangles with $\gamma_1 = 3$, $\gamma_2 = 3$, $\gamma_3 = 3$

Hence

The subgroup of translations is isomorphic to the infinite group generated by the elements

$$\mathsf{U} = \mathsf{C}^2\mathsf{BC}^2, \qquad \mathsf{V} = \mathsf{B}^2\mathsf{C}. \tag{14.2–21}$$

It is clear that the group characterized by (14.2–20) is a subgroup with index 2 of the group characterized by (14.2–16). Indeed, eliminating A from (14.2–16) we find that the group of B2) is generated by B, C and the defining relations $\mathsf{B}^3 = \mathsf{E}$, $\mathsf{C}^6 = \mathsf{E}$. The group of B4) is the subgroup generated by the elements B and C^2.

14.2.6 – The Schwarzian triangles of the third kind

Let the sum of the angles of a Schwarzian triangle with vertices s_1, s_2 and s_3 be less than π. If $\alpha_3 \neq 0$ we can apply the same transformation as

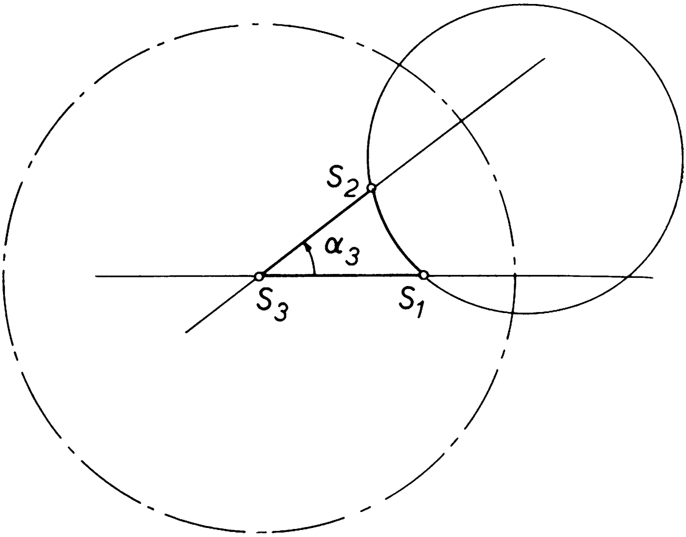


Fig. 14.2–23. Triangle of the third kind

in section 14.2.4. Since now the sum of the angles is less than π the third side of the transformed triangle is a circular arc convex towards $O = s_3$. As a consequence there is a circumference about O orthogonal to the sides of the triangle (fig. 14.2–23).

In the case that all angles are zero we can carry s_3 into infinity, (fig. 14.2–24) and the line connecting s_1 and s_2 may be considered as an (improper) circle orthogonal to all sides of the triangle.

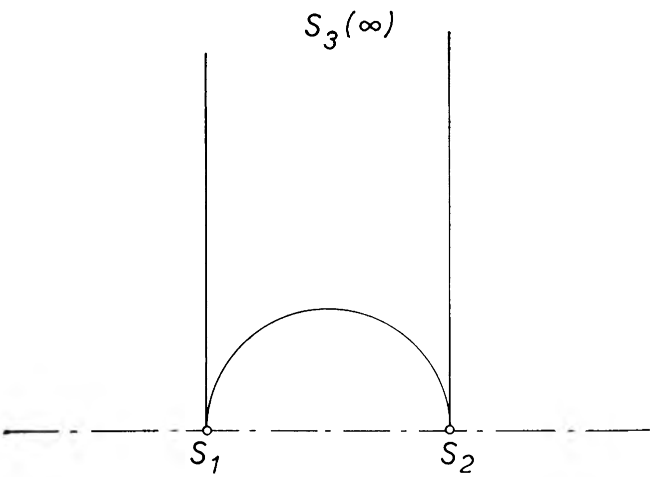


Fig. 14.2–24. Triangle with zero angles

Now we shall prove: *if the sides of a triangle in the interior of a circle are orthogonal to the circle, then the triangle obtained by reflecting the given triangle in a side is again in the interior on the circle.*

By a suitable transformation we can transform the triangle into a triangle $s_1 s_2 s_3$ which is situated on one side of a straight line. In fig. 14.2–25 we have taken this line as a horizontal line and the triangle in the upper half plane. If $s_3 = \infty$ the truth of the assertion is evident. Suppose, therefore, that s_1, s_2 and s_3 are finite. We reflect the triangle in the side $s_1 s_2$. Let O denote the centre of the circle on which the arc $s_1 s_2$ lies. In order to find the image of s_3 we draw a straight line from O

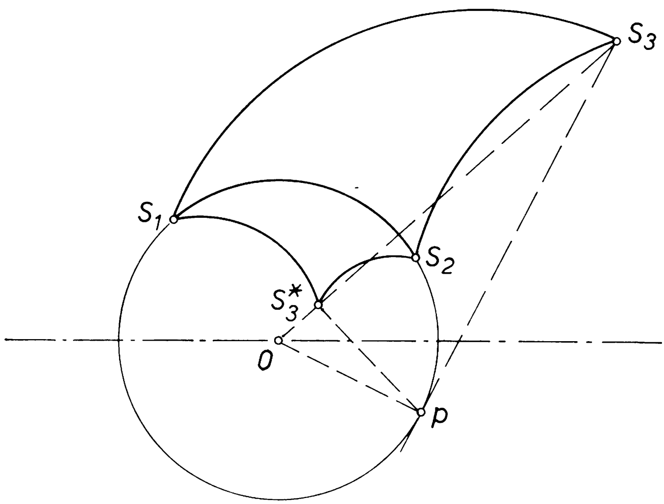


Fig. 14.2–25. Proof of the first theorem of section 14.2.6

to s_3. Let ps_3 be a tangent at p on the circle about O and s_3^* the point where the perpendicular through p on Os_3 meets this line. Then s_3^* is the desired image and it lies between O and s_3^*, i.e., also in the upper half-plane or on the horizontal line (if s_3 is on the horizontal line).

Hence

The covering transformations of a pattern of Schwarzian triangles of the third kind constitute a Fuchsian group of the first kind.

A fundamental domain consists of a shaded and an unshaded triangle. It follows from the considerations of section 13.3.2

The triangles obtained from a given Schwarzian triangle of the third kind inside a circle by a sequence of reflections in sides fill up without gaps or overlappings the whole interior of the circle. They cluster in infinite number about each point of the circumference.

The figure 14.2–26 gives an impression of the tessellation of the hyperbolic plane by means of Schwarzian triangles.

There is an endless number of possibilities of tessellations of the hyperbolic plane. The most remarkable cases are those in which

C1) $$\gamma_1 = 2, \quad \gamma_2 = \infty, \quad \gamma_3 = 3$$

and

C2) $$\gamma_1 = \infty, \quad \gamma_2 = \infty, \quad \gamma_3 = \infty$$

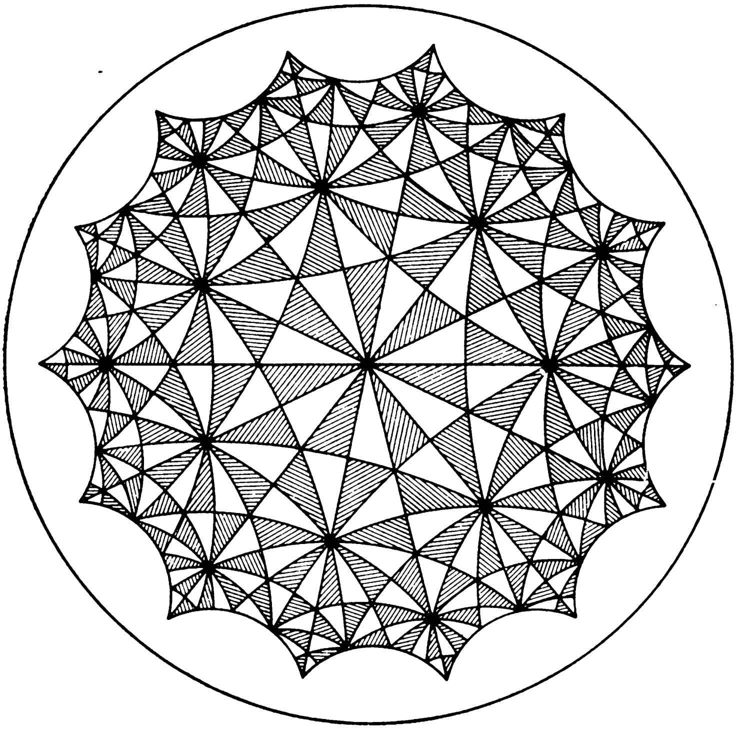


Fig. 14.2–26. Tessellation of the hyperbolic plane by triangles with $\gamma_1 = 2, \gamma_2 = 7, \gamma_2 = 3$

The corresponding groups are the *modular group* and the *congruence group* mod. 2.

We proceed to investigate these groups in more detail.

C1) As the region of discontinuity we take the upper half of the s-plane and we start with a triangle whose vertices are at $s_1 = i$, $s_2 = \infty$, $s_3 = \rho^2$, where $\rho = \frac{1}{2}+\frac{1}{2}i\sqrt{3}$.

Let A denote the rotation about s_1 through the angle $2\pi/2$, C the rotation about s_3 through the angle $2\pi/3$ and $B = AC^{-1}$. Then

The modular group is isomorphic to a group generated by the elements A, B, C *and the defining relations*

$$A^2 = E, \qquad C^3 = E, \qquad ABC = E. \tag{14.2–22}$$

B is the result of two reflections in vertical lines and thus the vertex s_3 is parabolic. Fig. 14.2–27 gives an impression of the pattern of the triangles.

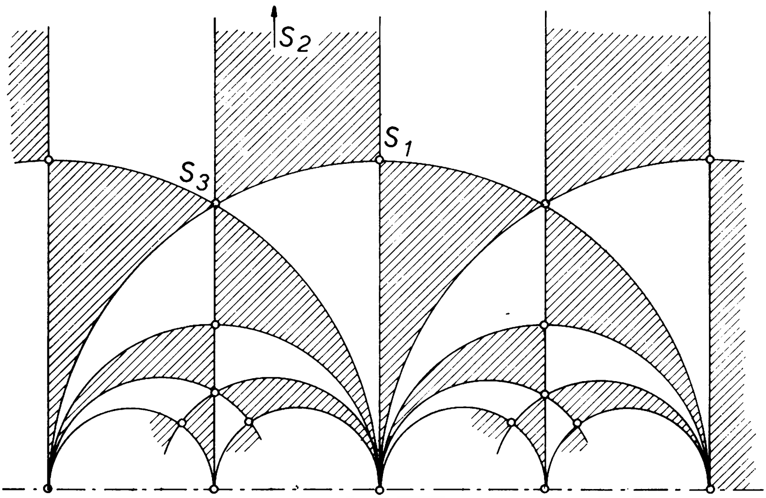


Fig. 14.2–27. Tessellation of the hyperbolic plane by triangles with $\gamma_1 = 2$, $\gamma_2 = \infty$, $\gamma_1 = 3$

The transformation A is elliptic of period 2. Since the fixed points are at $s_1 = i$ and $\bar{s}_1 = -i$ this transformation can be represented by

$$\frac{s' - i}{s' + i} = -\frac{s - i}{s + i},$$

or

$$s' = -\frac{1}{s}. \tag{14.2–23}$$

In symbolic form

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \tag{14.2–24}$$

The transformation C is elliptic of period 3. Its fixed points are

$s_3 = \rho^2 = -\frac{1}{2}+\frac{1}{2}i\sqrt{3}$ and $\bar{s}_3 = -\rho = -\frac{1}{2}-\frac{1}{2}i\sqrt{8}$. Since $\varrho^6 = 1$, we can represent the transformation as

$$\frac{s'-\rho^2}{s'+\rho} = \rho^2 \frac{s-\rho^2}{s+\rho}$$

or, since $\rho^3 = -1$,

$$s' = \frac{s+1}{-s}. \tag{14.2–25}$$

In symbolic form

$$\mathsf{C} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}. \tag{14.2–26}$$

The matrix associated with the transformation B is then

$$\mathsf{B} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{14.2–27}$$

i.e., B is the transformation

$$s' = s+1. \tag{14.2–28}$$

The modular group can be characterized in another way. We shall prove *The transformations of the modular group are the transformations*

$$s' = \frac{as+b}{cs+d}, \tag{14.2–29}$$

where a, b, c and d are real integers such that $ad-bc = 1$.

It is clear that

$$\mathsf{B}^k\mathsf{A} = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} k & -1 \\ 1 & 0 \end{bmatrix}.$$

If P denotes the matrix

$$\mathsf{P} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \tag{14.2–30}$$

then

$$\mathsf{PB}^{k_0}\mathsf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} k_0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} k_0 a+b & -a \\ k_0 c+d & -c \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}.$$

We can take k_0 such that $|k_0 a+b| < |a|$ i.e., $|a_1| < |b_1| = |a|$. Again

$$\mathsf{PB}^{k_0}\mathsf{AB}^{k_1}\mathsf{A} = \begin{bmatrix} k_1 a_1+b_1 & -a_1 \\ k_1 c_1+d_1 & -b_1 \end{bmatrix} = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}$$

and choosing k_1 suitably we have $|a_2| < |b_2| = |a_1|$. After a finite number of steps we have

$$\mathsf{PB}^{k_0}\mathsf{AB}^{k_1}\mathsf{A}\ldots\mathsf{B}^{k_r}\mathsf{A} = \begin{bmatrix} a_r & b_r \\ c_r & d_r \end{bmatrix},$$

with $a_r = 0$, $b_r c_r = -1$, e.g., $b_r = 1$, $c_r = -1$. Now

$$\begin{bmatrix} 0 & 1 \\ -1 & d_r \end{bmatrix}\begin{bmatrix} k_r & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -k_r+d_r & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

if we take $k_r = d_r$. As a consequence

$$\mathsf{PB}^{k_0}\mathsf{A}\ldots\mathsf{B}^{k_r}\mathsf{A} = \mathsf{E},$$

or, since $\mathsf{A}^{-1} = \mathsf{A}$,

$$\mathsf{P} = \mathsf{AB}^{-k_r}\ldots\mathsf{AB}^{-k_0}.$$

Thus we see that the group consisting of all elements of the type (14.2–30) is contained in the modular group. Conversely, A and B can be represented by transformations of the type (14.2–29). This concludes the proof of the theorem.

C2) We start with a triangle with vertices $s_1 = \infty$, $s_2 = 0$, $s_3 = 1$ and lying in the upper half plane. The angles of this triangle are all zero. Figure (14.2–28) gives an impression of the pattern associated with this triangle.

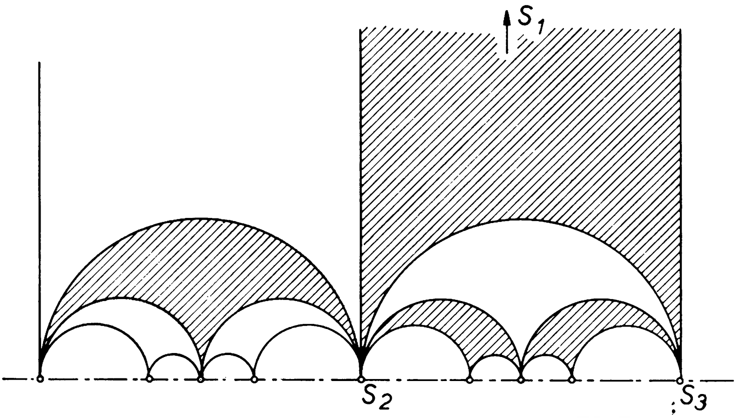


Fig. 14.2–28. Tessellation of the hyperbolic plane by triangles with $\gamma_1 = \gamma_2 = \gamma_3 = \infty$

By B we denote the parabolic transformation which is the result of two reflections in sides issuing from s_2. This transformation has a fixed point at $s = 0$ and it carries $s = \infty$ into $s = -\frac{1}{2}$. Hence this transformation must be

$$\frac{1}{s'} = \frac{1}{s} - 2$$

or

$$s' = \frac{s}{-2s+1}, \tag{14.2–31}$$

symbolically

$$\mathsf{B} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}. \tag{14.2–32}$$

Let C denote the parabolic transformation having its fixed point at $s = 1$ and carrying $s = \infty$ into $s = \frac{1}{2}$. This transformation is

$$\frac{1}{s'-1} = \frac{1}{s-1} - 2,$$

or

$$s' = \frac{s-2}{2s-3}. \tag{14.2–33}$$

Symbolically

$$\mathsf{C} = \begin{bmatrix} 1 & -2 \\ 2 & -3 \end{bmatrix}. \tag{14.2–34}$$

Now we define A by $\mathsf{A} = \mathsf{C}^{-1}\mathsf{B}^{-1}$,
i.e.,

$$\mathsf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \tag{14.2–35}$$

and it appears that A is the translation

$$s' = s+2, \tag{14.2–36}$$

which is also clear by a simple geometric consideration.

The group of covering transformations of the pattern of triangles with all angles equal to zero is generated by the parabolic transformations A, B, C *and the defining relation*

$$\mathsf{ABC} = \mathsf{E}. \tag{14.2–37}$$

This group is called the *congruence group modulo* 2 by the following reason

The congruence group mod 2 *consists of all transformations* (14.2–29) *with b and c even.*

Since $ad-bc = 1$, the numbers a and d are necessarily odd. The group can be represented symbolically by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (\text{mod } 2). \tag{14.2–38}$$

We shall prove that the group of elements (14.2–38) can be generated by the elements A and B, as given in (14.2–35) and (14.2–32). It is clear that

$$\mathsf{A}^k = \begin{bmatrix} 1 & 2k \\ 0 & 1 \end{bmatrix}, \qquad \mathsf{B}^{-k} = \begin{bmatrix} 1 & 0 \\ 2k & 1 \end{bmatrix}.$$

If P is again the matrix (14.2–28), then

$$\mathsf{PA}^{k_0} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 & 2k_0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & 2k_0 a + b \\ 0 & 2k_0 c + d \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}$$

and, taking k_0 suitably, $|b_1| < |a_1|$. Further

$$\mathsf{PA}^{k_0}\mathsf{B}^{-k'_0} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2k'_0 & 1 \end{bmatrix} = \begin{bmatrix} a_1 + 2k'_0 b_1 & b_1 \\ c_1 + 2k'_0 d_1 & d_1 \end{bmatrix} = \begin{bmatrix} a'_1 & b'_1 \\ c'_1 & d'_1 \end{bmatrix}$$

and, taking k'_0 suitably, we have $|a'_1| < |b'_1| = |b_1| < |a_1|$. After a finite number of steps, we arrive at

$$\mathsf{PA}^{k_0}\mathsf{B}^{-k'_0} \dots \mathsf{A}^{k_{r-1}}\mathsf{B}^{-k'_{r-1}} = \begin{bmatrix} 1 & b'_{r-1} \\ c'_{r-1} & d'_{r-1} \end{bmatrix}.$$

Now

$$\begin{bmatrix} 1 & b'_{r-1} \\ c'_{r-1} & d'_{r-1} \end{bmatrix} \begin{bmatrix} 1 & 2k_r \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2k_r + b'_{r-1} \\ c'_{r-1} & 2k_r c'_{r-1} + d'_{r-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ c_r & 1 \end{bmatrix},$$

if we take $k_r = -\frac{1}{2}b'_{r-1}$. This is possible, for b'_{r-1} is even. Moreover, $d_r = 1$, for the determinant of the matrix is unity. Finally

$$\begin{bmatrix} 1 & 0 \\ c_r & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2k'_r & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ c_r + 2k'_r & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

if we take $k'_r = -\frac{1}{2}c_r$. Thus

$$\mathsf{PA}^{k_0}\mathsf{B}^{-k'_0}\mathsf{A}^{k_r}\mathsf{B}^{-k'_r} = \mathsf{E},$$

or

$$\mathsf{P} = \mathsf{B}^{k'_r}\mathsf{A}^{-k_r} \dots \mathsf{B}^{k'_0}\mathsf{A}^{-k_0}.$$

Since A and B generate the group the proof of the theorem is complete.

With respect to the modulus 2 every transformation of the modular group can be represented as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$
$$\begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}, \tag{14.2–39}$$

i.e., we can divide the group into six disjunct classes. The first class consists of the congruence group mod. 2. Hence

The congruence group mod 2 is an invariant subgroup of the modular group with index six.

It is easy to see that mod 2 all transformations (14.2–39) are generated by the first two transformations following the identity. On the other hand the permutations (12) and (13) generate the symmetric group of three symbols. The correspondence

$$(1\quad 2) \leftrightarrow \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \quad (1\quad 3) \leftrightarrow \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

yields an isomorphism. Thus

The quotient group of the modular group with respect to the congruence group is isomorphic to the symmetric group of three symbols.

14.3 – Inverses of the Schwarzian triangle functions

14.3.1 – The polyhedral functions

The inverses of the Schwarzian functions of the first kind are called *polyhedral functions,* because they are automorphic with respect to the polyhedral groups. They can be found if we have solved the problem of the mapping of a half plane onto a curvilinear triangle with sum of angles $> \pi$. This requires the solution of the differential equation (14.1–5), with $\alpha_1 = 1/\gamma_1$, $\alpha_2 = 1/\gamma_2$, $\alpha_3 = 1/\gamma_3$ and the numbers $\gamma_1, \gamma_2, \gamma_3$ are those as listed in section 14.2.2 under A). In general the problems of solving an equation of this type is not easy.

In the case of a Schwarzian function of the first kind, however, it is not necessary to solve the equation, for the inverses of the desired solutions can be found by simple considerations. In this case the fact is that the inverses are rational functions.

This is geometrically clear. For let $z(s)$ denote the function which maps a triangle onto the upper half of the z-plane. This function is regular inside the triangle and continuous on the boundary. It can be continued throughout the extended s-plane by reflection in sides of triangles and the function is holomorphic in the extended plane, except possibly at the vertices of the triangle and the points congruent with these vertices with respect to the group associated with the triangle. Since the extended plane can be filled up by a finite number N of triangles, a value of $z(s)$ is taken in only a finite number of points. Hence it has no essential singular points and is, therefore, rational.

Let s_1, s_2, s_3 denote the vertices of a triangle with angles π/γ_1, π/γ_2, π/γ_3, corresponding to $z = 0, 1, \infty$ respectively.

A simple reasoning shows that the expansions of s in a neighbourhood of either of these points in the z-plane are of the type

$$s-s_1 = (c_0 z+c_1 z^2+ \dots)^{1/\gamma_1},$$
$$s-s_2 = (c_0(z-1)+c_1(z-1)^2+ \dots)^{1/\gamma_2},$$
$$s-s_3 = \left(\frac{c_0}{z}+\frac{c_1}{z^2}+\dots\right)^{1/\gamma_3}.$$

Inverting these expressions we obtain series of the type

$$\begin{aligned} z &= a_0(s-s_1)^{\gamma_1}+\dots, \\ z-1 &= a_0(s-s_2)^{\gamma_2}+\dots, \\ \frac{1}{z} &= a_0(s-s_3)^{\gamma_3}+\dots. \end{aligned} \tag{14.3–1}$$

Thus

The inverse $z(s)$ of the Schwarzian triangle function $s(1/\gamma_1, 1/\gamma_2, 1/\gamma_3; z)$ is a rational function which takes the value 0 with multiplicity γ_1, the value 1 with multiplicity γ_2 and which has a pole of order γ_3.

We recall that $s(z)$ is a multivalent function; if it takes the value s_0 it takes also all values obtained from s_0 by means of linear transformations belonging to the group associated with the triangle. Hence, what has been said about s_0, refers also to each point congruent to s_0, etc.

Now it is an easy matter to find the desired functions. Let $F_1(s)$ denote the polynomial of lowest degree whose zeros are the point s_1 and all congruent points different from s_1. The number of zeros of $F_1(s)$ is then N/γ_1 and this is also the degree of F_1. It may happen that the point ∞ is congruent to s_1. Then the degree of $F_1(s)$ diminished by one and we consider ∞ as a zero of F_1. Similarly, let $F_2(s)$ denote the polynomial of lowest degree whose zeros are the point s_2 and all congruent points different from s_2. The degree of $F_2(s)$ is N/γ_2. Finally, let $F_3(s)$ denote the polynomial of lowest degree whose zeros aie s_3 and the congruent points different from s_3. The degree of $F_3(s)$ is N/γ_3.

Taking into account the above theorem we may state

$$z(s) = A\,\frac{F_1^{\gamma_1}(s)}{F_3^{\gamma_3}(s)}, \qquad 1-z(s) = B\,\frac{F_2^{\gamma_2}(s)}{F_3^{\gamma_3}(s)} \tag{14.3–2}$$

wheie A and B are constants. They are determined by the identity in s

$$AF_1^{\gamma_1}+BF_2^{\gamma_2} = F_3^{\gamma_3}. \tag{14.3–3}$$

The various exceptional points, being the zeros of F_1, F_2, F_3 correspond to points on the Riemann sphere which are obtained by projection from the centre on to the sphere of the midpoints of the edges, the centres of the faces and the vertices of a regular polyhedron, inscribed in the sphere. This is also true for the dihedron if we take $\gamma_2 = n$, $\gamma_3 = 2$.

Consider a point on the sphere with spherical coordinates φ, θ, $0 \leqq \varphi \leqq 2\pi$, $-\pi \leqq \theta \leqq \pi$. Then the stereographic projection, (fig. 14.3–1) of this point onto the s-plane is given by

$$s = e^{i\varphi} \tan \tfrac{1}{2}\theta. \tag{14.3–4}$$

Let s_1 and s_2 be points corresponding to diametral points on the sphere. If $s_1 = e^{i\varphi_1} \tan \frac{1}{2}\theta_1$ then $s_2 = e^{i(\varphi_1+\pi)} \tan \frac{1}{2}(\theta_1+\pi) = -e^{i\varphi_1} \operatorname{ctn} \theta_1$, whence

$$s_1 \bar{s}_2 = -1, \tag{14.3–5}$$

in accordance with (9.3–28).

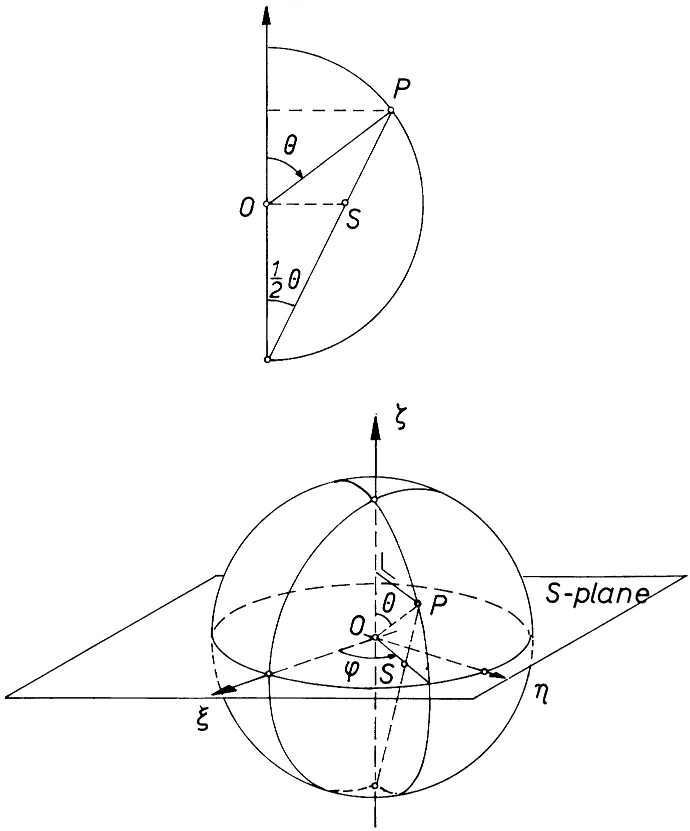


Fig. 14.3–1. Stereographic projection

It should be noticed that various forms of the functions $z(s)$ can be obtained, depending on the situation of the polyhedron in the sphere. These forms can be obtained by any of them by applying a fractional linear transformation corresponding to a rotation of the sphere.

It is clear that the functions thus obtained are automorphic and that every rational expression in terms of either function is again automorphic with respect to the same polyhedral group. In a certain sense the functions obtained by the above construction are the most simple ones.

A1) We consider first the *dihedral function*, being a function automorphic with respect to the dihedral group. It is uniquely determined if we agree that the dihedron shall be placed as pictured in fig. 14.2–3.

Let s_0 denote the point in the s-plane corresponding to the midpoint of the side *12*. It is evident that $s_0 = e^{\pi i/n}$, whence $s_0^n = e^{\pi i} = -1$. The other midpoints are represented by

$$s_0 e^{2\pi i/n}, \ldots, s_0 e^{2\pi i(n-1)/n}$$

and these points together with s_0 are the zeros of $s^n - s_0^n$, i.e.,

$$F_1(s) = s^n + 1. \qquad (14.3\text{–}6)$$

It is clear that the vertices are the points corresponding to the zeros of

$$F_3(s) = s^n - 1. \qquad (14.3\text{–}7)$$

The north pole and the south pole of the sphere correspond to $s = 0$ and $s = \infty$ respectively. They can be considered as the zeros of the polynomial

$$F_3(s) = s \qquad (14.3\text{–}8)$$

of virtual degree two.

In accordance with (14.3–3) we have to evaluate the constants A and B from the identity

$$A(s^n+1)^2 + Bs^n = (s^n-1)^2.$$

Taking $s = 0$ we find $A = 1$; taking $s^n = 1$ we find $B = -4$. Hence

$$\boxed{z(s) = \frac{(s^n+1)^2}{(s^n-1)^2}, \qquad z(s)-1 = \frac{4s^n}{(s^n-1)^2}.} \qquad (14.3\text{–}9)$$

Except for an additive constant the reciprocal of $z-1$ is the same as the function considered in section 10.2.4.

A2) Any function automorphic with respet to the tetrahedral group is called a *tetrahedral function*. Suppose that a tetrahedron *1234* is placed as in fig. 14.2–7. Let s_0 denote the point in the s-plane corresponding to the vertex *1*. It is easy to see that the other vertices correspond to $-s_0$, $1/s_0$ and $-1/s_0$ and thus

$$F_3(s) = (s^2-s_0^2)\left(s^2-\frac{1}{s_0^2}\right) = s^4-\left(s_0^2+\frac{1}{s_0^2}\right)s^2+1.$$

Now

$$s_0 = e^{\pi i/4}\tan\tfrac{1}{2}\theta,$$

whence

$$s_0^2 = i\tan^2\tfrac{1}{2}\theta$$

and

$$\frac{1}{s_0^2} = -i\operatorname{ctn}^2\tfrac{1}{2}\theta.$$

Observing that

$$\tan\theta = \frac{2\tan\frac{1}{2}\theta}{1-\tan^2\frac{1}{2}\theta}$$

and $\tan\theta = \sqrt{2}$, we see that $\tan\frac{1}{2}\theta$ and $-\operatorname{ctn}\frac{1}{2}\theta$ are the roots t_1 and t_2 of the equation

$$t^2+t\sqrt{2}-1 = 0.$$

We have

$$t_1^2-t_2^2 = (t_1+t_2)(t_1-t_2) = -\sqrt{2}\times\sqrt{6} = -2\sqrt{3}$$

and so

$$F_3(s) = s^4-2i\sqrt{3}\,s^2+1. \qquad (14.3\text{–}10)$$

The zeros of $F_2(s)$ correspond to the vertices of the twin tetraeder. The point corresponding to *1'* is $\bar{s}_0$ and we have immediately

$$F_2(s) = s^4+2i\sqrt{3}\,s^2+1. \qquad (14.3\text{–}11)$$

Finally the midpoints of the edges correspond to 1, i, -1, $-i$ and ∞, whence

$$F_1(s) = s(s^4-1). \qquad (14.3\text{–}12)$$

The constants A and B occurring in (14.3–2) are determined by the identity

$$As^2(s^4-1)^2+B(s^4-2i\sqrt{3}\,s^2+1)^3 = (s^4+2i\sqrt{3}\,s^2+1)^3.$$

Taking $s = 0$ we find $B = 1$. Taking $s^2 = i$ we find $4iA+24\sqrt{3} = -24\sqrt{3}$, whence $A = 12i\sqrt{3}$.

Thus

$$z(s) = 12i\sqrt{3}\,\frac{s^2(s^4-1)^2}{(s^4+2i\sqrt{3}\,s^2+1)^3}, \qquad z(s)-1 = \frac{(s^4-2i\sqrt{3}\,s^2+1)^3}{(s^4+2i\sqrt{3}\,s^2+1)^3}.$$

(14.3–13)

A3) A function automorphic with respect to the octahedral group is called an *octahedral function*. We obtain a special type if we place the octahedron as in fig. 14.2–10. The points corresponding to the vertices are evidently the zeros of

$$F_3(s) = s(s^4-1). \qquad (14.3\text{–}14)$$

The midpoints on the edges in the horizontal plane correspond to the zeros of s^4+1. The midpoint on the edge *01* corresponds to $s_0 = \tan \frac{1}{2}\theta$, and since $\tan \theta = 1$, it is a root of the equation

$$t^2+2t-1 = 0. \qquad (14.3\text{–}15)$$

The midpoints of the edges issuing from *0* are evidently the zeros of $s^4-s_0^4$ and those of the edges issuing from *0′* are the zeros of s^4-1/s_0^4. Hence

$$F_1(s) = (s^4+1)(s^4-s_0^4)\left(s^4-\frac{1}{s_0^4}\right) = (s^4+1)(s^8-\sigma_4 s^4+1),$$

where $\sigma_4 = s_0^4+1/s_0^4 = t_1^4+t_2^4$; t_1 and t_2 being the roots of (14.3–15). Now

$$\sigma_2 = t_1^2+t_2^2 = \sigma_1^2-2t_1t_2 = 4+2 = 6,$$
$$\sigma_4 = \sigma_2^2-2t_1^2t_2^2 = 36-2 = 34,$$

whence

$$F_1(s) = (s^4+1)(s^8-34s^4+1) = s^{12}-33s^8-33s^4+1. \qquad (14.3\text{–}16)$$

Finally we observe that the centres of the faces of the octahedron correspond to the vertices of the two desmic tetrahedra considered in A2). They correspond to the zeros of

$$(s^4-2i\sqrt{3}\,s^2+1)(s^4+2i\sqrt{3}\,s^2+1) = (s^4+1)^2+12s^4,$$

or

$$F_2(s) = s^8+14s^4+1. \qquad (14.3\text{–}17)$$

The constants A and B are obtained from the identity

$$A(s^{12}-33s^8-33s^4+1)^2+B(s^8+14s^4+1)^3 = s^4(s^4-1)^4.$$

Taking $s = 0$ we obtain $A+B = 0$. Comparing the coefficients of s^{20} we find

$$-2\times 33A+3\times 14B = -108A = 1$$

whence

$$A = -\tfrac{1}{108}, \qquad B = \tfrac{1}{108}$$

and so

$$z(s) = -\tfrac{1}{108}\frac{(s^{12}-33s^8-33s^4+1)^2}{s^4(s^4-1)^4}, \quad z(s)-1 = -\tfrac{1}{108}\frac{(s^8+14s^4+1)^3}{s^4(s^4-1)^4}. \tag{14.3–18}$$

A5) Finally we wish to construct an *icosahedral function*, which is a function automorphic with respect to the icosahedral group. We refer to fig. 14.2–14. In this case the numerical location of the critical points, i.e., those corresponding to $z = 0, 1, \infty$ is not so easy as in the previous cases.

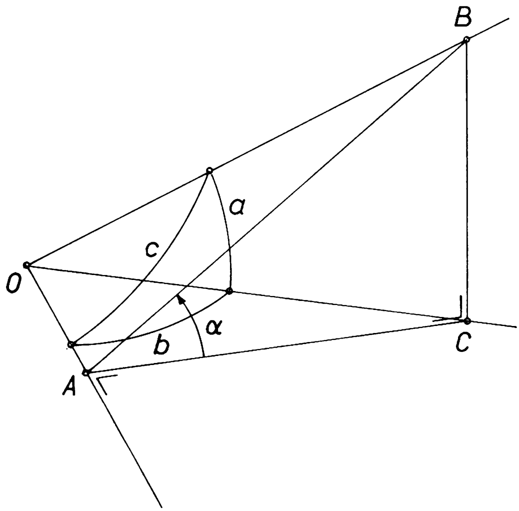


Fig. 14.3–2. Rectangular spherical triangle

First we derive a simple formula for a rectangular spherical triangle. We consider a trihedron with vertex at the centre of a sphere and a plane through a point A on one edge of the trihedron, perpendicular to this edge, (fig. 14.3–2). It cuts the other edges in B and C. The edges cut the sphere in the vertices of a spherical triangle with sides a, b and c and angles α, β and γ. Assuming that $\gamma = \frac{1}{2}\pi$ we have

$$\cos\alpha = \frac{AC}{AB} = \frac{OA\tan b}{OA\tan c},$$

whence

$$\tan b = \tan c\cos\alpha. \tag{14.3–19}$$

Next we consider an isosceles triangle with angle α and (spherical) side a. Let ρ denote the spherical radius of the circumscribed circle and

r that of the inscribed circle, (fig. 14.3–3). Applying (14.3–19) on a triangle which is part of the isosceles triangle we find

$$\tan \tfrac{1}{2}a = \tan \rho \cos \tfrac{1}{2}\alpha,$$

or

$$\tan \rho = \frac{\tan \frac{1}{2}a}{\cos \frac{1}{2}\alpha}. \tag{14.3–20}$$

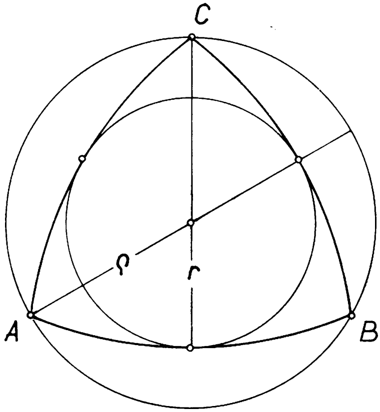


Fig. 14.3–3. The radii of the inscribed and the exscribed circle of an isosceles spherical triangle

In the same triangle

$$\tan r = \tan \rho \cos \tfrac{1}{3}\pi,$$

whence

$$\tan r = \tfrac{1}{2} \tan \rho. \tag{14.3–21}$$

If a is known we can find α by

$$\cos \alpha = \frac{\tan \frac{1}{2}a}{\tan a}. \tag{14.3–22}$$

We need also a formula for $\tan(\rho + r)$, viz.,

$$\tan(\rho + r) = \tan a \cos \tfrac{1}{2}\alpha. \tag{14.3–23}$$

If we project a face of the inscribed icosahedron centrally from O onto the sphere we obtain an isosceles spherical triangle whose side shall again be denoted by a. By some geometric considerations it is possible to find the value of $\tan a$.

First we wish to prove a remarkable relation between the sides of a regular pentagon and a regular dodecagon inscribed in a circle of radius

R about a centre M. Let, as in fig. 14.3–4 $AB = s_{10}$ denote the side of a regular decagon and $AC = s_5$ the side of a regular pentagon, such that B and C are on the same side of MA. Let P denote the intersection of AC and the line bisecting the angle AMB. Now it is easy to prove that the triangles APB and ABC are similar, as are the triangles CPM and CMA.

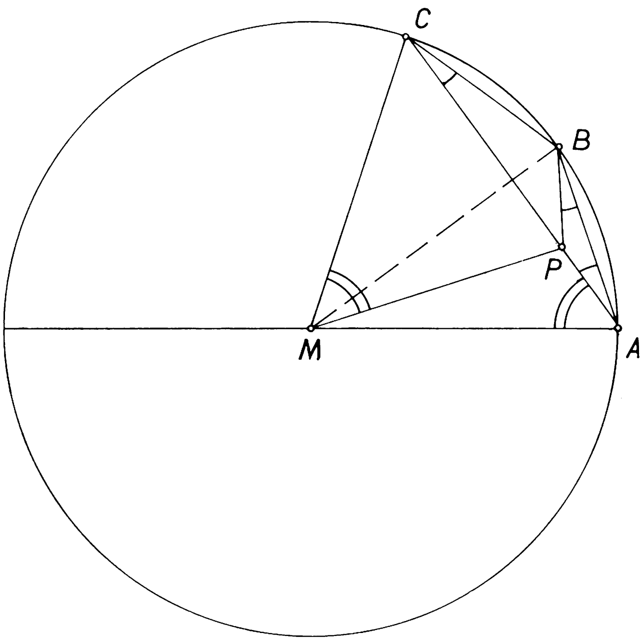


Fig. 14.3–4. Proof of the relation (14.3–24)

It follows that

$$AP : AB = AB : AC$$

and

$$MP : CM = CM : CA.$$

Hence

$$AB^2 + CM^2 = AP \cdot AC + MP \cdot CA = AC^2,$$

or

$$s_5^2 = s_{10}^2 + R^2, \tag{14.3–24}$$

the desired relation.

Let R denote the radius of the circumscribed circle of the upper pentagon *12345* of the iscosahedron. It is also the radius of the circumscribed circle of the lower pentagon *1'2'3'4'5'*. Projecting the upper pentagon orthogonally onto the plane of the lower pentagon the projections of the vertices of the upper pentagon constitute with the vertices of the lower pentagon the vertices of a decagon inscribed in the circle of radius R, (fig. 14.3–5). It is now readily seen that the distance of the point 1 to the plane of the lower pentagon is a side of a rectangular triangle whose other side is s_{10} and whose hypotenusa is s_5. In view of (14.3–24) we conclude that this distance is equal to R and it is now an easy matter to obtain the result

$$\tan a = 2. \qquad (14.3\text{–}25)$$

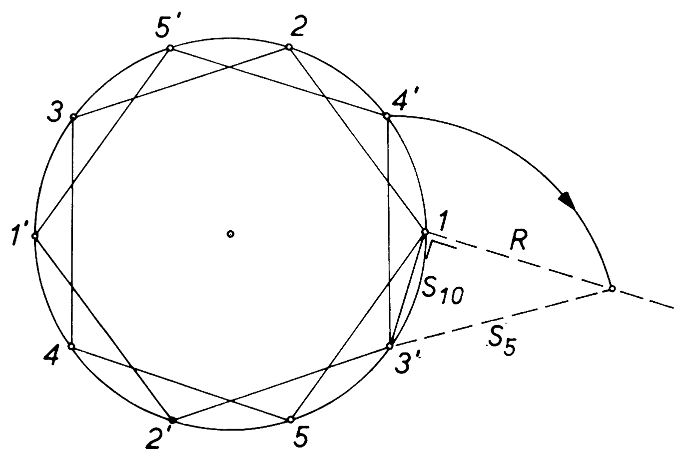


Fig. 14.3–5. Construction of the distance of the point *1* to the plane of the lower pentagon

Now $\tan \frac{1}{2}a$ is the positive root of the equation

$$t^2+t-1 = 0, \qquad (14.3\text{–}26)$$

whence

$$\tan \tfrac{1}{2}a = \tfrac{1}{2}(-1+\sqrt{5}) \qquad (14.3\text{–}27)$$

and from (14.3–22) we find

$$\cos \alpha = \tfrac{1}{4}(-1+\sqrt{5}). \qquad (14.3\text{–}28)$$

Further

$$\cos^2 \tfrac{1}{2}\alpha = \tfrac{1}{2}(1+\cos \alpha) = \tfrac{1}{8}(3+\sqrt{5}) = \tfrac{1}{16}(6+2\sqrt{5})$$

whence

$$\cos \tfrac{1}{2}\alpha = \tfrac{1}{4}(1+\sqrt{5}). \qquad (14.3\text{–}29)$$

In view of (14.3–20) we then have

$$\tan \rho = 3-\sqrt{5} \tag{14.3–30}$$

and, referring to (14.3–23)

$$\tan (\rho+r) = \tfrac{1}{2}(1+\sqrt{5}). \tag{14.3–31}$$

Finally we need the value of $\tan(\rho+2r) = \tan((\rho+r)+r)$. A simple calculation yields

$$\tan (\rho+2r) = 3+\sqrt{5}. \tag{14.3–32}$$

Now we are sufficiently prepared to find the polynomials $F_1(s)$, $F_2(s)$ and $F_3(s)$. In order to facilitate the calculations we shall firste evaluate the number

$$\sigma_5 = t_1^5+t_2^5,$$

where t_1 and t_2 are the roots of the equation

$$t^2+\xi t-1 = 0.$$

It is clear that

$$\sigma_2\sigma_3 = \sigma_5+t_1^2t_2^2\sigma_1 = \sigma_5+\sigma_1,$$

whence

$$\sigma_5 = \sigma_2\sigma_3+\xi.$$

Further

$$\sigma_2 = \sigma_1^2-2t_1t_2 = \sigma_1^2+2 = \xi^2+2$$

and

$$\sigma_1\sigma_2 = \sigma_3+t_1t_2\sigma_1 = \sigma_3+\xi,$$

or

$$\sigma_3 = \sigma_1\sigma_2-\xi = -\xi(\xi^2+3).$$

Hence

$$\sigma_5 = -\xi((\xi^2+2)(\xi^2+3)-1). \tag{14.3–33}$$

Let $t_1 = s_0$ denote the point in the s-plane corresponding to the vertex *1*. It is clear that the points in the s-plane corresponding to the vertices of the upper horizontal pentagon are the zeros of $s^5-s_0^5 = s^5-t_1^5$. The vertex *1′* corresponds to $t_2 = -1/s_0$ and the points in the s-plane corresponding to the vertices of the lower pentagon are the zeros of $s^5+1/s_0^5 = s^5-t_2^5$. The numbers t_1 and t_2 are the roots of the equation (14.3–26). The points in the s-plane corresponding to the vertices of the icosahedron are the zeros of

$$F_3(s) = s(s^{10}-\sigma_5 s^5-1).$$

The virtual degree of this polynomial is 12, in accordance with the fact that $s = \infty$ corresponds to a vertex of the icosahedron. Inserting $\xi = 1$ into (14.3–33) we find

$$F_3(s) = s(s^{10}+11s^5-1). \tag{14.3–34}$$

Let now s_0 denote the point in the s-plane corresponding to the centre of the face *012*. Its spherical coordinates are $\varphi = \pi/5$, $\theta = \rho$. Hence the points corresponding to the centres of all faces having *0* as vertex are the zeros of the polynomial $s^5+t_1^5$, where t_1 is the positive root of the equation

$$t^2+\frac{2}{\tan\rho}t-1 = t^2+\tfrac{1}{2}(3+\sqrt{5})t-1 = 0. \tag{14.3–35}$$

The points in the s-plane corresponding to the centres of the faces diametrically opposite to the faces considered above are evidently the zeros of $s^5+t_2^5$, where t_2 is the other root of the equation (14.3–35). The ten centres of the highest and the lowest faces correspond, therefore, to the zeros of

$$(s^5+t_1^5)(s^5+t_2^5) = s^{10}+\sigma_5 t^5-1.$$

The points corresponding to the centres of the remaining faces are obtained in quite the same way. We have only to replace ρ by $\rho+2r$ and they are the zeros of

$$s^{10}+\sigma_5^* s^5-1,$$

where σ_5^* is obtained from σ_5 by replacing $\sqrt{5}$ by $-\sqrt{5}$, This is at once clear if we compare (14.3–30) and (14.3–32). Hence the centres of the faces correspond to the zeros of

$$(s^{10}+\sigma_5 s^5-1)(s^{10}+\sigma_5^* s^5-1)$$
$$= s^{20}+(\sigma_5+\sigma_5^*)s^{15}+(\sigma_5\sigma_5^*-2)s^{10}-(\sigma_5+\sigma_5^*)s^5+1.$$

Inserting $\xi = \frac{1}{2}(3+\sqrt{5})$ into (14.3–33) we find

$$\sigma_5 = -114-50\sqrt{5}, \qquad \sigma_5^* = -114+50\sqrt{5}.$$

Thus

$$F_2(s) = s^{20}-228s^{15}+494s^{10}-228s^5+1. \tag{14.3–36}$$

The projection of the midpoint of the edge *01* from *0′* onto the sphere has the coordinates $\varphi = 0$, $\theta = \frac{1}{2}a$. Hence the midpoints of the edges issuing from *0* and *0′* correspond to the zeros of

$$(s^5-t_1^5)(s^5-t_2^5) = s^{10}-\sigma_5 s^5-1,$$

where t_1 and t_2 are the roots of the equation

$$t^2+\frac{2}{\tan\frac{1}{2}a}t-1 = t^2+(1+\sqrt{5})t-1 = 0.$$

The projection onto the sphere of the midpoint of the edge *12* has the coordinates $\varphi = r/5$, $\theta = \rho + r$. Hence the midpoints on the edges of the two horizontal pentagons correspond to the zeros of

$$s^{10} + \sigma_5^* s^5 - 1,$$

where σ_5^* is obtained from σ_5 by replacing $1+\sqrt{5}$ by $-1+\sqrt{5}$. Finally, it is easy to see that the midpoints of the remaining edges (which are in the horizontal coordinate plane) correspond to the zeros of $s^{10}+1$. Inserting $\xi = 1+\sqrt{5}$ into (14.3–33), we find

$$\sigma_5 = -261+125\sqrt{5}, \qquad \sigma_5^* = 261+125\sqrt{5}$$

and we get

$$\begin{aligned} F_1(s) &= (s^{10}+1)(s^{20}+522s^{15}-10006s^{10}-522s^5+1) \\ &= s^{30}+522s^{25}-10005s^{20}-10005s^{10}-522s^5+1. \end{aligned} \tag{14.3–37}$$

We have to evaluate A and B from

$$AF_1^2+BF_2^3 = F_3^5.$$

Taking $s = 0$ we find $A+B = 0$. Comparing the coefficients of s^{55} we find

$$2\times 522A - 3\times 228B = 1728A = 1,$$

whence

$$A = \tfrac{1}{1728}, \qquad B = -\tfrac{1}{1728}$$

and so

$$\boxed{\begin{aligned} z(s) &= \frac{(s^{30}+522s^{25}-10005s^{20}-10005s^{10}-522s^4+1)^2}{1728s^5(s^{10}+11s^5-1)^5}, \\ z(s)-1 &= \frac{(s^{20}-228s^{15}+494s^{10}+228s^5+1)^2}{1728s^5(s^{10}+11s^5-1)^5}. \end{aligned}} \tag{14.3–38}$$

It is easy to find the polynomial in s whose zeros correspond to the vertices of one of the regular octahedra considered at the end of section 14.2.4. Let us denote these polynomials by $f_k(s)$, $k = 0, 1, 2, 3, 4$, corresponding to the octahedra *I, II, III, IV, V* respectively. It is evident that $f_0(s)$ is the polynomial

$$\begin{aligned} &(s^2+1)\left(s^2+\frac{2s}{\tan\frac{1}{2}a}-1\right)\left(s^2-\frac{2s}{\tan 2(\rho+r)}-1\right) \\ &= (s^2+1)(s^2+(1+\sqrt{5})s-1)(s^2+(1-\sqrt{5})s-1) \end{aligned}$$

or

$$f_0(s) = s^6+2s^5-5s^4-5s^2-2s+1.$$

The other polynomials are obtained by means of the transformations

$$s \to C^k s = \varepsilon^k s, \qquad k = 1, 2, 3, 4,$$

where

$$\varepsilon = e^{2\pi i/5}.$$

Any transformation of the icosahedral group effects a permutation among the polynomials $f_k(s)$ and leave the elementary symmetric combinations of these polynomials invariant. The polynomials are the roots of the equation

$$w^5 + a_4 w^4 + a_3 w^3 + a_2 w^2 + a_1 w + a_0 = 0, \qquad (14.3\text{–}39)$$

where the coefficients are polynomials in s and are invariant with respect to the icosahedral group. Any polynomial of the lowest degree, being invariant with respect to the icosahedral group, is of degree at least. 12 Now if $a_4 \neq 0$ it is of degree 6, whence $a_4 = 0$. The polynomial a_3 is of degree 12 and can vanish only at the points corresponding to the vertices of the icosahedron. Hence $a_3 = aF_3$, where a is a constant. Since if $a_2 \neq 0$ it is of degree 18, it is also zero identically. The polynomial a_1 is of degree 24 and is, therefore, bF_3^2 (where b is again a constant), for it can only vanish at the vertices. Finally a_0 is of degree 30 and it is, apart from sign, equal to the product of the polynomials $f_k(s)$. Hence $a_0 = cF_1$ and comparing the coefficients of s^{30} we find $c = -1$.

The coefficients a and b may be found as follows. The recursive relations (12.3–1) yield

$$\sigma_2 + 2aF_3 = 0, \qquad \sigma_4 + (4b - 2a^2)F_3^2 = 0. \qquad (14.3\text{–}40)$$

Now

$$f_k(s) = \varepsilon^{3k}s^6 + 2\varepsilon^{2k}s^5 - 5\varepsilon^k s^4 + \ldots,$$
$$f_k^2(s) = \varepsilon^k s^{12} + 4s^{11} - 6\varepsilon^{4k}s^{10} + \ldots,$$
$$f_k^4(s) = \varepsilon^{2k}s^{24} + 8\varepsilon^k s^{23} + 4s^{22} + \ldots,$$

where $k = 0, 1, 2, 3, 4$. Inserting these polynomials into (14.3–40) and comparing the coefficients of s^{11}, s^{22} and s^{30} we find

$$20 + 2a = 0, \qquad 20 + 4b - 2a^2 = 0,$$

whence

$$a = -10, \qquad b = 45.$$

Thus we have

The octahedral polynomials f_0, f_1, f_2, f_3, f_4 are the roots of the equation

$$\boxed{w^5 - 10F_3(s)w^4 + 45F_3^2(s)w - F_3(s) = 0.} \qquad (14.3\text{–}41)$$

This result is due to A. Brioschi.

14.3.2 – THE INVERSES OF THE SCHWARZIAN FUNCTIONS OF THE SECOND KIND

The problem of finding a Schwarzian function of the second kind is not very difficult. We may suppose that the triangles are rectilinear and it is natural to apply the Schwarz-Christoffel integral (10.3–10). It takes the form

$$s\left(\frac{1}{\gamma_1}, \frac{1}{\gamma_2}, \frac{1}{\gamma_3}; z\right) = s_1 + c\int_0^z \frac{dt}{t^{1-1/\gamma_1}(1-t)^{1-1/\gamma_2}}, \tag{14.3–42}$$

where s_1 is the vertex of the triangle at which the interior angle is π/γ_1 and c is a constant. It is understood that z varies in the domain $\operatorname{Im} z \geqq 0$.

The argument of the integrand may be determined as follows. Let $\theta_0 = \arg t$, $\theta_1 = \arg(t-1)$. We agree that $\arg(1-t) = \theta_1 - \pi$ and that the argument of the denominator is

$$(1-1/\gamma_1)\theta_0 + (1-1/\gamma_2)\theta_1 - (1-1/\gamma_2)\pi.$$

Hence the denominator is positive for t between 0 and 1.

Now we shall discuss the various cases.

B1) Let $\gamma_1 = \gamma_2 = 2$. We take $s_1 = 0$. Then

$$s = c\int_0^z \frac{dt}{t^{\frac{1}{2}}(1-t)^{\frac{1}{2}}}. \tag{14.3–43}$$

We introduce a new variable u by $1-t = u^2$, ($u > 0$ if $0 < t < 1$). We find

$$s = -2c\int_1^{\sqrt{1-z}} \frac{du}{\sqrt{1-u^2}} = 2c\int_{\sqrt{1-z}}^1 \frac{du}{\sqrt{1-u^2}}.$$

The integral is uniquely defined if we agree that $\arg\sqrt{1-u^2} = \frac{1}{2}\arg t = \frac{1}{2}\theta_0$. Taking $c = \frac{1}{2}$ we get the function

$$s = \int_{\sqrt{1-z}}^1 \frac{du}{\sqrt{1-u^2}} \tag{14.3–44}$$

which maps the upper half of the z-plane onto a vertical half strip with angles $\frac{1}{2}\pi$, $\frac{1}{2}\pi$ and width

$$s_2 - s_1 = \int_0^1 \frac{du}{\sqrt{1-u^2}} = \tfrac{1}{2}\pi.$$

This strip is a triangle of the pattern as depicted in fig. 14.2–19.

Inverting the integral (14.3–44) we obtain

$$\boxed{1-z(s) = \cos^2 s, \qquad z(s) = \sin^2 s.} \tag{14.3–45}$$

The function $z(s)$ is a simply periodic function of the second order. A period strip consists of two shaded and two unshaded triangles. Since the integral (14.3–43) is divergent as $z \to \infty$ the third vertex s_3 is at infinity.
B2) In the case $\gamma_1 = 2, \gamma_2 = 3$ we have to consider the integral

$$s = s_1 + c \int_0^z \frac{dt}{t^{\frac{1}{2}}(1-t)^{\frac{2}{3}}}. \tag{14.3–46}$$

We introduce the variable u by $1-t = u^3$ and we get

$$s = s_1 - 3c \int_1^{\sqrt[3]{1-z}} \frac{du}{\sqrt{1-u^3}} = s_1 + 6c \int_{\sqrt[3]{1-z}}^1 \frac{du}{\sqrt{4-4u^3}}.$$

If we agree that $\arg\sqrt{1-u^3} = \frac{1}{2}\arg t = \frac{1}{2}\theta_0$ and $\arg\sqrt{u^3-1} = \frac{1}{2}\arg(-t)$ $= \frac{1}{2}\theta_0 - \frac{1}{2}\pi$, then

$$\sqrt{1-u^3} = i\sqrt{u^3-1}$$

and taking $c = i/6$ the function $s(z)$ appears as

$$s = s_1 + \int_{\sqrt[3]{1-z}}^1 \frac{du}{\sqrt{4u^3-4}}.$$

Next we take

$$s_1 = \int_1^\infty \frac{du}{\sqrt{4u^3-4}} \tag{14.3–47}$$

and we see that

$$s = \int_{\sqrt[3]{1-z}}^\infty \frac{du}{\sqrt{4u^3-4}} \tag{14.3–48}$$

is a function which maps the upper half of the z-plane onto a triangle having the angles $\pi/2, \pi/3, \pi/6$. It is a triangle of the pattern as pictured in fig. 14.2–20.

The integral (14.3–48) is of the Weierstrassian type as considered in section 5.13.1. Inverting it we find

$$\boxed{1 - z(s) = \wp^3(s;0,4), \qquad z(s) = 1 - \wp^3(s;0,4) = \tfrac{1}{4}\wp'^2(s;0,4).} \tag{14.3–49}$$

The function $z(s)$ is doubly periodic of order six. A rhombus having one vertex at the origin and containing six shaded and six unshaded triangles is a period parallellogram. The periods can be found from the numbers $\tilde{\omega}$ and $\tilde{\omega}'$, introduced in section 5.12.3. The first formula (5.13–10) gives

$$\tilde{\omega} = \int_1^\infty \frac{du}{\sqrt{4u^3-4}} = \frac{1}{6}\int_0^1 \frac{dt}{t^{5/6}(1-t)^{1/2}},$$

performing the substitution $u^3 = 1/t$. Hence

$$\tilde{\omega} = \tfrac{1}{6}\frac{\Gamma(\frac{1}{6})\Gamma(\frac{1}{2})}{\Gamma(\frac{2}{3})} = \frac{\sqrt{3}}{12\sqrt{\pi}}\Gamma(\tfrac{1}{3})\Gamma(\tfrac{1}{6}) = \frac{1}{4\pi\sqrt[3]{2}}\Gamma^3(\tfrac{1}{3}), \qquad (14.3\text{–}50)$$

by (10.3–13). It is geometrically clear that

$$\tilde{\omega}'/i = \tilde{\omega}\sqrt{3}. \qquad (14.3\text{–}51)$$

This follows also from

$$\tilde{\omega}'/i = 3|s_2 - s_1| = \tfrac{1}{2}\int_0^1 \frac{dt}{t^{\frac{1}{2}}(1-t)^{\frac{2}{3}}} = \tfrac{1}{2}\frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{3})}{\Gamma(\frac{5}{6})} = \frac{1}{4\sqrt{\pi}}\Gamma(\tfrac{1}{3})\Gamma(\tfrac{1}{6}).$$

In view of (5.6–22) we may express the function $z(s)$ in terms of the sigma functions of Weierstrass

$$z(s) = -\frac{\sigma_1^2(s)\sigma_2^2(s)\sigma_3^2(s)}{\sigma^6(s)}, \qquad (14.3\text{–}52)$$

a formula which exhibits the zeros and the poles.

B3) In the case $\gamma_1 = 2$, $\gamma_2 = 4$ we have to consider the integral

$$s = s_1 + c\int_0^z \frac{dt}{t^{\frac{1}{2}}(1-t)^{\frac{3}{4}}}. \qquad (14.3\text{–}53)$$

We introduce the variable u by $1-t = u^2$ and we get

$$s = s_1 - 4c\int_1^{\sqrt{1-z}} \frac{du}{\sqrt{4u-4u^3}} = s_1 + 4c\int_{\sqrt{1-z}}^1 \frac{du}{\sqrt{4u-4u^3}}.$$

If we agree that $\arg\sqrt{u-u^3} = \frac{1}{2}\arg t + \frac{1}{2}\arg(1-t) = \frac{1}{2}\theta_0 + \frac{1}{2}\theta_1 - \frac{1}{2}\pi$, $\arg\sqrt{4u^3-4u} = \frac{1}{2}\arg(-t) + \frac{1}{2}\arg(1-t) = \frac{1}{2}\theta_0 + \frac{1}{2}\theta_1 - \pi$, we have

$$\sqrt{u-u^3} = i\sqrt{u^3-u}$$

and taking $c = i/4$ the function $s(z)$ appears as

$$s = s_1 + \int_{\sqrt{1-z}}^1 \frac{du}{\sqrt{4u^3-4u}}.$$

Next we take

$$s_1 = \int_1^\infty \frac{du}{\sqrt{4u^3-4u}} \qquad (14.3\text{–}54)$$

and we see that

$$s = \int_{\sqrt{1-z}}^\infty \frac{du}{\sqrt{4u^3-4u}} \qquad (14.3\text{–}55)$$

is a function which maps the upper half of the z-plane onto a triangle of the pattern as pictured in fig. 14.2–21.

Inverting the integral (14.3–55) we find

$$\boxed{1-z(s) = \wp^2(s; 4, 0), \qquad z(s) = 1-\wp^2(s; 4, 0).} \tag{14.3–56}$$

This result does not differ essentially from (10.2–62).

The function $z(s)$ is doubly periodic and of order four. A square, having one vertex at the origin and containing four shaded triangles, is a period parallelogram. The real period is, according to (5.13–9),

$$2\omega = 2\int_1^\infty \frac{du}{\sqrt{4u^3-4u}} = \tfrac{1}{2}\int_0^1 \frac{dt}{t^{\frac{3}{4}}(1-t)^{\frac{1}{2}}},$$

where we have performed the substitution $u^2 = 1/t$.

Hence

$$2\omega = \tfrac{1}{2}\frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{4})}{\Gamma(\frac{3}{4})} = \frac{1}{2\sqrt{2\pi}}\Gamma^2(\tfrac{1}{4}), \tag{14.3–57}$$

by (10.2–59). The other period is determined by

$$2\omega'/i = \tfrac{1}{2}\int_0^1 \frac{dt}{t^{\frac{1}{2}}(1-t)^{\frac{3}{4}}} = 2\omega. \tag{14.3–58}$$

In view of (5.14–19) we may express the functions $z(s)$ and $1-z(s)$ in terms of the Jacobian elliptic function with modulus $k = 1/\sqrt{2}$. Since now $e_1 = 1$, $e_2 = 0$, $e_3 = -1$, $g = \sqrt{2}$, we have

$$1-z(s) = 4\frac{dn^4(s\sqrt{2})}{sn^4(s\sqrt{2})}, \qquad z(s) = -4\frac{cn^2(s\sqrt{2})}{sn^4(s\sqrt{2})}. \tag{14.3–59}$$

The functions under consideration are of the lemniscate type.

B4) A function automorphic with respect to the covering group of the pattern depicted in fig. 14.2–22 can be readily obtained by a simple geometric method, utilizing the results of part B2) of this section. On applying Schwarz's reflection principle we see that

$$z(s) = 1-\wp^3(s; 0, 4)$$

maps the z-plane slit along the positive axis onto an isosceles triangle. Hence $z(s) = \wp^3(s)-1$ maps the z-plane slit along the negative axis onto the triangle and

$$z(s) = \sqrt{\wp^3(s; 0, 4)-1} = -\tfrac{1}{2}\wp'(s; 0, 4)$$

(see e.g., 5.7–9) is the function which maps the right half of the z-plane onto the isosceles triangle under consideration. Multiplying by i, this half plane is rotated through a right angle about the origin. Hence the desired function is

$$\boxed{z(s) = -\tfrac{1}{2}i\wp'(s; 0, 4),} \tag{14.3–60}$$

being a doubly periodic function of the third order. The period parallelograms are the same as those mentioned in B2).

In view of (5.6–22) we may also write

$$z(s) = i\,\frac{\sigma_1(s)\sigma_2(s)\sigma_3(s)}{\sigma^3(s)}, \tag{14.3–61}$$

an expression exhibiting clearly the zeros and the poles of $z(s)$.

14.3.3 – The inverses of the Schwarzian functions of the third kind

There are an endless number of the Schwarzian functions corresponding to the solutions of the inequality (14.2–4). Their inverses are automorphic with respect to a Fuchsian group of the first kind whose region of discontinuity is the interior of the circle (or the half of the s-plane). The boundary of this region is a natural boundary (section 8.2.3); it is impossible to continue analytically a function of this kind across the boundary. This follows from the fact that the vertices of the triangles of the pattern obtained by any one by performing a transformation of the group cluster towards the points on the boundary.

The most remarkable Schwarzian functions of the third kind are those corresponding to $\gamma_1 = 2$, $\gamma_2 = \infty$, $\gamma_3 = 3$ and to $\gamma_1 = \gamma_2 = \gamma_3 = \infty$. Their inverses are automorphic with respect to the modular group and the congruence group, discussed in section 14.2.6. They are called *modular functions* because they are closely related to the modulus of Legendre's complete elliptic integrals to be considered in more detail in subsequent sections.

By rather simple arguments we can obtain a lot of information of the above mentioned functions. We start with a triangle in the upper half of the s-plane with vertices $s_1 = \infty$, $s_2 = 0$, $s_3 = 1$. Together with an adjacent triangle along the imaginary axis it constitutes a fundamental domain for the congruence group, (fig. 14.3–6).

The function $s(0, 0, 0; z)$ which maps the above triangle onto the upper half of the z-plane such that s_1, s_2, s_3 correspond to 0, 1, ∞, respectively will be denoted by

$$\tau(z). \tag{14.3–62}$$

This function is unique and we have

$$\tau(0) = \infty, \quad \tau(1) = 0, \quad \tau(\infty) = 1. \tag{14.3–63}$$

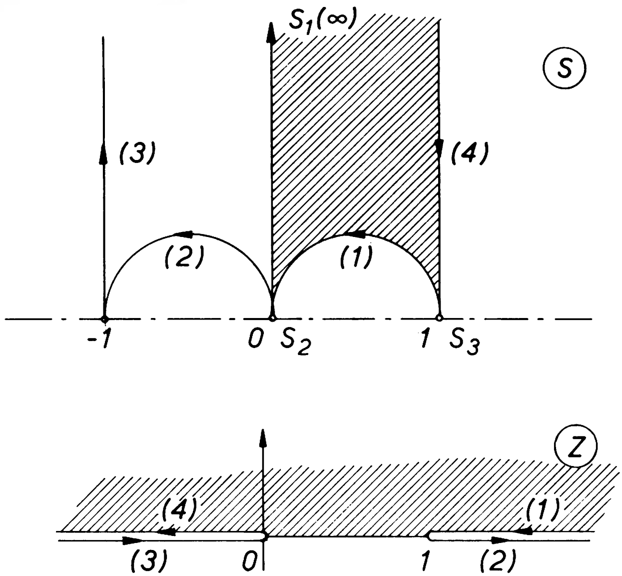


Fig. 14.3–6. The fundamental domain of the congruence group mod 2. There are 3 parabolic cycles, viz. $s = 0$, $(s = -1)$, $s = \infty$

The function $\tau(z)$ satisfies the Schwarzian differential equation

$$[\tau]_z = \frac{z^2-z+1}{2z^2(z-1)^2}, \tag{14.3–64}$$

as appears if we take $\alpha_1 = \alpha_2 = \alpha_3 = 0$ in (14.2–3). The function on the right remains unchanged if we replace z by $1-z$. On the other hand $[\tau]_z = [\tau]_{1-z}$, as follows from (14.3–64). The second theorem of section 13.4.5 implies that

$$\tau(1-z) = \frac{A\tau(z)+B}{C\tau(z)+D},$$

where A, B, C and D are constants. They are determined by the following considerations.

If z varies throughout the upper half of the z-plane, then $1-z$ varies throughout the lower half. Reflecting in the segment $0 < z < 1$ we see that $\tau(z)$ maps the interior of the fundamental region onto the z-plane slit along the real axis from 1 to ∞ and from 0 to $-\infty$. The upper border of the right cut corresponds to the circular arc between $s = 1$ and $s = 0$, the lower border to the circular arc between $s = 0$ and $s = -1$. The upper border of the left cut corresponds to the right half ray $\operatorname{Re} s = 1$, $\operatorname{Im} s > 0$ and the lower border of the left cut corresponds to the left half ray $\operatorname{Re} s = -1$, $\operatorname{Im} s > 0$.

If z moves along the upper border of the right cut to $+\infty$, then $1-z$ moves along the lower border of the left cut to $-\infty$. Then $\tau(z)$ moves along the right circular arc to $+1$ and $\tau(1-z)$ along the left circular arc to -1. Thus

$$-1 = \frac{A+B}{C+D}.$$

Taking $z = 0$ we find $A = 0$ and taking $z = 1$ we find $D = 0$. It follows that

$$\tau(1-z) = \frac{-1}{\tau(z)}. \tag{14.3–65}$$

Referring again to (14.3–64) a simple calculation affirms that this equation remains unchanged if we replace z by $1/z$, whence

$$\tau\left(\frac{1}{z}\right) = \frac{A\tau(z)+B}{C\tau(z)+D}.$$

If z moves along the upper border of the right cut to infinity, then $\tau(z)$ tends to 1 and $\tau(1/z)$ to ∞, since $1/z$ tends to zero. It follows that $C+D=0$. If z tends to 0 along the upper border then $1/z$ tends to ∞ along the lower border and $\tau(1/z)$ tends to -1, whence $A = -C$. If, finally, z tends to 1 so does $1/z$ and, as a consequence, $B = 0$.

Thus we also have

$$\tau\left(\frac{1}{z}\right) = \frac{\tau(z)}{-\tau(z)+1}. \tag{14.3–66}$$

The inverse function of $\tau(z)$ will be denoted by

$$\lambda(s). \tag{14.3–67}$$

This function is a simple automorphic function with respect to the congruence group mod 2 and it takes its values once at each point of the fundamental domain.

The equations (14.3–65) and (14.3–66) can be rewritten as

$$\lambda\left(\frac{s}{-s+1}\right) = \frac{1}{\lambda(s)} \tag{14.3–68}$$

and

$$\lambda\left(-\frac{1}{s}\right) = 1-\lambda(s) \tag{14.3–69}$$

The transformations performed on s are represented by the matrices

$$\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{14.3–70}$$

and they are the first two transformations different from the identity, listed in (14.2–39). They generate a group isomorphic to the quotient group of the modular group with respect to the congruence group and the elements of this group induce certain transformations of λ, namely

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} : \lambda, \quad \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} : \frac{1}{\lambda}, \quad \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} : 1-\lambda,$$
$$(14.3\text{–}71)$$
$$\begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix} : \frac{1}{1-\lambda}, \quad \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} : \frac{\lambda}{\lambda-1}, \quad \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} : \frac{\lambda-1}{\lambda}.$$

The matrices characterize the transformations of s. The expressions following the matrices are the results of the transformation of λ. The various transforms of λ are the six values of the cross ratio of four points on a projective line. By this reason the group interchanging these values is called *the group of the cross ratios.* By means of the construction as described in section 13.2.5 we obtain the fundamental domain depicted in fig. 14.3–7.

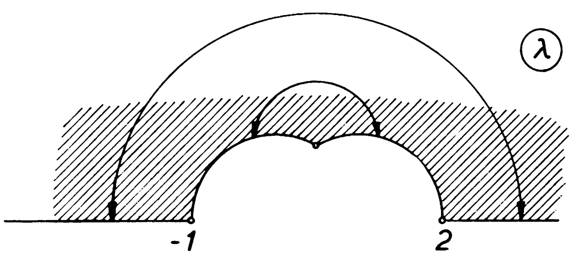


Fig. 14.3–7. Fundamental domain of the anharmonic ratio group. The rectilinear parts are equivalent by $\lambda \to 1-\lambda$; the circular paths are equivalent by $\lambda \to (\lambda-1)/\lambda$.

Some particular values of $\lambda(s)$ can easily be found. A fixed point of the transformation

$$s \to -\frac{1}{s}$$

is $s = i$. Inserting this into (14.3–69) we find

$$\lambda(i) = \tfrac{1}{2}. \tag{14.3–72}$$

A fixed point of the transformation

$$s \to \frac{s-1}{s}$$

is $\rho = \frac{1}{2}+\frac{1}{2}i\sqrt{3}$. From (14.3–71) we have

$$\lambda\left(\frac{s-1}{s}\right) = 1-\frac{1}{\lambda(s)},$$

whence

$$\lambda(\rho) = 1-\frac{1}{\lambda(\rho)}.$$

Since ρ is in the shaded triangle, $\operatorname{Im}\lambda(\rho) > 0$, and so

$$\lambda(\rho) = \rho. \tag{14.3–73}$$

Now we turn our attention towards the modular group. By means of the

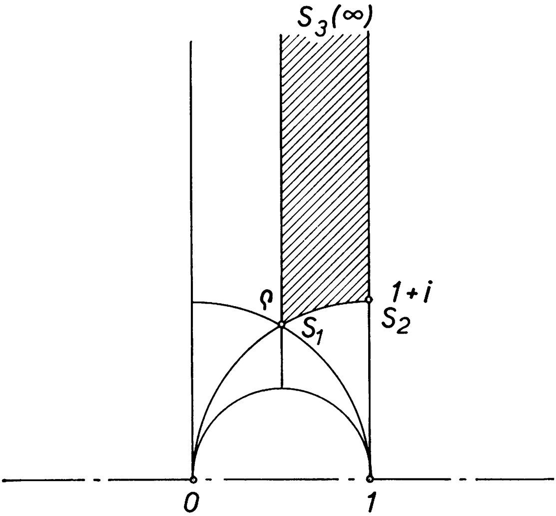


Fig. 14.3–8. The division of a triangle with zero angles into six congruent triangles

altitudes we can divide each triangle with zero angles into six congruent triangles with angles $\pi/2$, $\pi/3$, 0, (fig. 14.3–8). The pattern of these triangles remains invariant under the transformations of the modular group. We consider in particular the triangle with vertices $s_1 = \rho$, $s_2 = 1+i$, $s_3 = \infty$. Let

$$J(s) \tag{14.3–74}$$

denote the function which maps this triangle onto the upper half of the z-plane such that s_1, s_2 and s_3 correspond to $z = 0, 1, \infty$ respectively. The function $J(s)$ is a simple automorphic function of the first order and it takes its values six times in the fundamental region (fig. 14.3–9) of the congruence group. Since $J(s)$ is also automorphic with respect to the congruence group it must be a rational function of λ of order six.

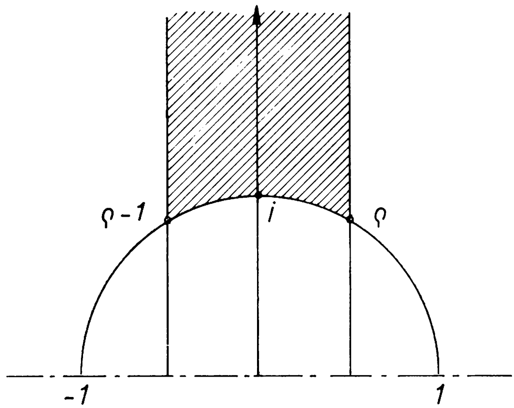


Fig. 14.3–9. Fundamental domain of the modular group

Now it is easy to find a function of this kind, viz.

$$f(s) = (1+\lambda)\left(1+\frac{1}{\lambda}\right)(1+1-\lambda)\left(1+\frac{1}{1-\lambda}\right)\left(1+\frac{\lambda}{\lambda-1}\right)\left(1+\frac{\lambda-1}{\lambda}\right),$$

for a transformation of the modular group merely interchanges the factors on the right. An easy calculation yields

$$f(s) = -\frac{(\lambda+1)^2(2-\lambda)^2(2\lambda-1)^2}{\lambda^2(\lambda-1)^2}. \tag{14.3–75}$$

Since $\lambda(\rho) = \rho$, and $\rho^2-\rho+1 = 0$, we readily find $f(\rho) = 27$. Hence the function

$$F(s) = 1-\tfrac{1}{27}f(s) \tag{14.3–76}$$

has a zero at $s = \rho$. Taking $s = i$, i.e., $\lambda(i) = \frac{1}{2}$, we find $F(i) = 1$ and from $F(s+1) = F(s)$ we deduce that also $F(i+1) = 1$. Finally $F(s) = \infty$ at $s = \infty$. The functions $F(s)$ and $J(s)$ are automorphic with respect to the same group and both are simple automorphic functions of the first order. It follows that either function is a rational function of the other. Since they coincide at the vertices of the triangle we have, evidently,

$$F(s) = J(s). \tag{14.3–77}$$

We notice that $F(\rho) = 0$ and that the numerator of $F(s)$ is a polynomial of the sixth degree in λ. There can be no other zeros in the fundamental domain and we conclude that

$$F(s) = c\,\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(\lambda-1)^3},$$

where the constant c is determined by the condition $F(i) = 1$. It follows $c = 4/27$ and so

$$\boxed{J(s) = \tfrac{4}{27}\,\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(\lambda-1)^2}.} \tag{14.3–78}$$

It is clear that this expression could be obtained by straight forward calculation from (14.3–75) inserting the expression for $f(s)$ in terms of λ.

In addition we have

$$\boxed{J(s)-1 = \tfrac{1}{27}\,\frac{(\lambda+1)^2(2-\lambda)^2(2\lambda-1)^2}{\lambda^2(\lambda-1)^2}.} \tag{14.3–79}$$

14.4 – **Picard's theorem and related theorems**

14.4.1 – Picard's first theorem

In section 9.9.3 we proved Picard's theorem for integral functions by means of a method due to Landau. A very simple proof is the original proof given by Picard himself based on the function $\tau(z)$ introduced in section 14.3.3.

If the half plane $\operatorname{Im} s > 0$ is filled by an infinity of successive reflections of the original triangle with vertices at 0, 1 and ∞, the z-plane is covered by an infinity of upper and lower half-planes which are the conformal images of the reflected triangles as given by $s = \tau(z)$. Each half plane has three adjacent half-planes which are connected with it along the segment $0 < z < 1$ and the rays $-\infty < z < 0$ and $1 < z < \infty$, respectively. The totality of half planes which are connected with each other in the manner indicated is known as the *modular surface*. There are no points of this surface above $z = 0, 1, \infty$, for these are logarithmic branch points. The modular surface is the Riemann surface of the analytic function obtained from $\tau(z)$ by analytic continuation along paths which avoid the points $z = 0$ and $z = 1$. This analytic function may again be denoted by $\tau(z)$.

Let $f(z)$ denote a non-constant integral function omitting the values 0 and 1. Consider a single-valued branch of the analytic function $\tau(z)$

in a sufficiently small neighbourhood of $f(z_0)$. We shall denote this branch by $\tau(z)$. Then

$$g(z) = \tau(f(z)) \tag{14.4–1}$$

is a single-valued function element defined in a neighbourhood of $z = z_0$. Since $f(z)$ avoids the values 0, 1 and is everywhere regular it can be continued analytically throughout the entire (finite)z-plane giving rise to a single-valued function, by virtue of the monodromy theorem of section 12.2.3. But $g(z)$ takes only values in the upper half of the complex plane, i.e., Im $g(z) > 0$. Hence the modulus of exp $ig(z)$ is less than 1 and by the Cauchy–Liouville theorem (section 2.12.1) it must be constant. It follows that $f(z)$ is constant, contrary to our assumption.

An alternative statement of Picard's theorem is

If $f(z)$ is meromorphic throughout the whole plane and does not take three different values a, b and c, then it is a constant.

This follows from the fact that the function

$$\frac{c-a}{c-b}\,\frac{f(z)-b}{f(z)-a}$$

is an integral function which does not take the values 0 and 1.

Picard's theorem for integral functions will be referred to as *Picard's first theorem.* A similar theorem is concerned with the behaviour of a function in a neighbourhood of any essential singular point. It will be discussed in section 14.4.4.

14.4.2 – Landau's theorem

In this section we will establish another proof of Landau's theorem, stated in section 9.9.4 and we shall obtain an explicit expression for the *Landau radius* $R(a_0, a_1)$.

Let $f(z)$ denote a function holomorphic in $|z| < R$ which takes neither the value 0, nor the value 1. As in the preceding section we consider a function element $\tau(f(z))$ in a neighbourhood of $z = 0$. This can be continued analytically throughout the disc $|z| < R$ and yields a holomorphic function $g(z)$ with Im $g(z) > 0$.

Now we make the following remark: *If* Im $z > 0$, Im $a > 0$ *and*

$$w = \frac{z-a}{z-\bar{a}}, \tag{14.4–2}$$

then $|w| < 1$.

In fact, this transformation carries the real axis into a circumference through $w = 1$ corresponding to $z = \infty$. Since a and $\bar{a}$ are symmetric with respect to the real axis, their images, $w = 0$ and $w = \infty$, are symmetric with respect to the circumference.

If we put

$$h(z) = \frac{g(z)-g(0)}{g(z)-\overline{g(0)}} \tag{14.4-3}$$

we have

$$|h(z)| < 1, \qquad |z| < R.$$

By Schwarz's lemma (section 2.21.2)

$$|h(z)|R \leqq |z|,$$

whence

$$|h'(0)|R \leqq 1,$$

or

$$R \leqq \frac{1}{|h'(0)|}. \tag{14.4-4}$$

Now

$$h'(0) = \lim_{z\to 0} \frac{g(z)-g(0)}{z} \cdot \frac{1}{g(z)-\overline{g(0)}} = \lim_{z\to 0} \frac{\tau(f(z))-\tau(f(0))}{z} \cdot \frac{1}{g(z)-\overline{g(0)}}$$

$$= \tau'(a_0)a_1 \frac{1}{\tau(a_0)-\overline{\tau(a_0)}},$$

with

$$a_0 = f(0), \qquad a_1 = f'(0).$$

It follows that

$$R \leqq R(a_0, a_1)$$

with

$$R(a_0, a_1) = \frac{|\tau(a_0)-\overline{\tau(a_0)}|}{|a_1||\tau'(a_0)|} = \frac{2 \operatorname{Im} \tau(a_0)}{|a_1||\tau'(a_0)|}. \tag{14.4-5}$$

This expression has been obtained by Carathéodory.

The result is sharp. In order to prove this we consider the function

$$f(z) = \lambda\left(i\frac{1+z}{1-z}\right)$$

which is certainly holomorphic if $|z| < 1$ (in accordance with 10.2–3). This function does not take the values 0 and 1. Now $a_0 = f(0) = \lambda(i)$, whence $\tau(a_0) = i$. Further $a_1 = f'(0) = \lambda'(i)\times 2i = 2i/\tau'(a_0)$ and so

$$R(a_0, a_1) = \frac{2}{\dfrac{2}{|\tau'(a_0)|}|\tau'(a_0)|} = 1.$$

This concludes the proof.

14.4.3 – Schottky's theorem

We adopt the same assumptions as in section 9.9.5. The functions

$$g(z) = \tau(f(z))$$

can be defined as a single-valued holomorphic function in the disc $|z| < 1$. Introducing a hyperbolic metric in this disc we deduce from (9.5–22) that the hyperbolic radius of the circumference $|z| = \vartheta$ about the origin is

$$\log \frac{1+\vartheta}{1-\vartheta}. \tag{14.4–6}$$

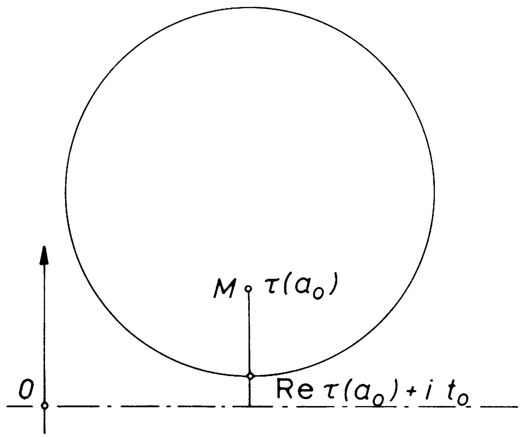


Fig. 14.4–1. Determination of t_0 in Schottky's theorem

The geometric form of Schwarz's lemma (section 9.8.2) states that the values of $g(z)$ in the s-plane are contained in a circular disc with centre $\tau(f(0)) = \tau(a_0)$ and the same hyperbolic radius. In the Poincaré model of the hyperbolic plane let Re $\tau(a_0)+it_0$ denote a point on the circumference right under the centre, (fig. 14.4–1). Then, in view of (9.5–31) the radius of the circle in the s-plane is

$$\int_{t_0}^{\operatorname{Im} \tau(a_0)} \frac{dy}{y} = \log \frac{\operatorname{Im} \tau(a_0)}{t_0}$$

and equating this to (14.4–6) we get

$$t_0 = \frac{1-\vartheta}{1+\vartheta} \operatorname{Im} \tau(a_0). \tag{14.4–7}$$

The modular function $\lambda(s)$ is periodic with period 2. A maximum $\mu(t_0)$

of the modulus of this function on the line $\operatorname{Im} s = t_0$ is attained on the segment $0 \leqq \operatorname{Re} s \leqq 2$, $\operatorname{Im} s = t_0$. By the maximum principle the maximum of $\lambda(s)$ in the infinite strip $0 \leqq \operatorname{Re} s \leqq 2$, $\operatorname{Im} s \geqq t_0$, is attained on the boundary. On the lines $\operatorname{Re} s = 0$ and $\operatorname{Re} s = 2$ the function $\lambda(s)$ is real and tends monotonously to zeros as $\operatorname{Im} s \to \infty$. Hence

$$|\lambda(s)| < \mu(t_0), \qquad \operatorname{Im} s > t_0.$$

As a consequence the moduli of the values of $f(z) = \lambda(g(z))$ in the disc $|z| \leqq \vartheta$ do not exceed the number

$$\mu(t_0) = \mu\left(\frac{1-\vartheta}{1+\vartheta} \operatorname{Im} \tau(a_0)\right) = \varphi(\vartheta, a_0), \qquad (14.4\text{–}8)$$

as asserted in Schottky's theorem.

We shall make an additional remark which will be useful in the next section. Suppose that $|a_0| < \frac{1}{2}$. Then we contend that the upper bound in Schottky's theorem depends only on ϑ. For let $\tau(z)$ denote the branch which maps the plane slit along the half rays $z > 1$ and $z < 0$ onto the fundamental domain consisting of two triangles with vertex at infinity (fig. 14.4–2). Since $\tau(z) \to \infty$ as $z \to 0$ the disc $|z| \leqq \frac{1}{2}$, slit along a radius to the left, corresponds to a closed point set within the strip of the fundamental domain. If z tends to a point of the slit then $\tau(z)$ tends to a point on one

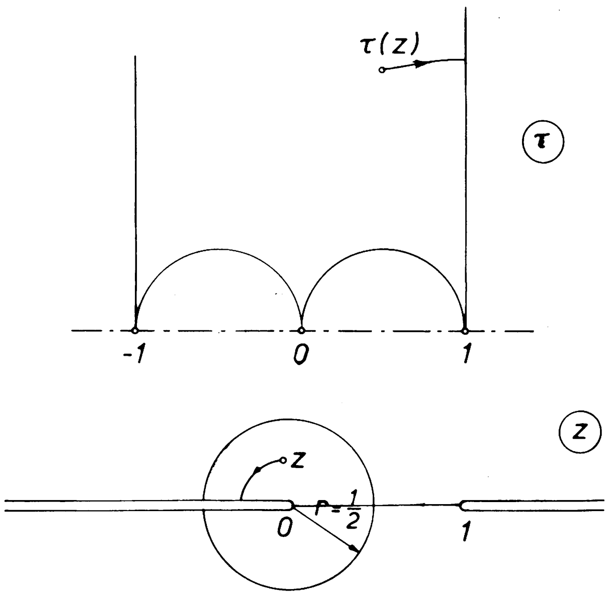


Fig. 14.4–2. Additional remark to Schottky's theorem

of the vertical bounding lines. This closed set has a positive distance to the real axis in the s-plane, for it cannot penetrate into a neighbourhood of $s = 0$ or $s = 1$, points corresponding to $z = 1$ and $z = \infty$ respectively. Hence $\operatorname{Im} \tau(a_0) > \eta$, where η is a positive constant, not depending on a_0, as long as $|a_0| < \frac{1}{2}$. In Schottky's theorem we now may take

$$t_0 = \frac{1-\vartheta}{1+\vartheta}\eta.$$

14.4.4 – Picard's second theorem

Picard's second theorem is a statement of the behaviour of a function in the vicinity of an isolated essential singular point and it completes the Casorati-Weierstrass theorem of section 3.2.1.

Let $f(z)$ be holomorphic in a region $0 < |z| < 1$ and omit the values 0 *and* 1. *Then $z = 0$ is not an essential singular point.*

From the Casorati-Weierstrass theorem follows that we can find a sequence $a_1, a_2, \ldots$ such that

$$e^{-4\pi} > |a_1| > |a_2| > \ldots, \qquad |a_n| \to 0, \quad \text{as} \quad n \to \infty \tag{14.4–9}$$

and

$$|f(a_n)-\tfrac{1}{4}| < \tfrac{1}{4}. \tag{14.4–10}$$

To every number a_n we assign a number b_n such that

$$e^{b_n} = a_n, \qquad -\pi \leqq \operatorname{Im} b_n \leqq \pi. \tag{14.4–11}$$

Next we consider the function

$$g(w) = f(e^w) = f(z), \qquad z = e^w, \tag{14.4–12}$$

which is holomorphic for $\operatorname{Re} w < 0$ and omits the values 0 and 1. The function is periodic with period $2\pi i$. The inequalities (14.4–9) are equivalent to

$$-4\pi > \operatorname{Re} b_1 > \operatorname{Re} b_2 > \ldots, \qquad \operatorname{Re} b_n \to -\infty. \tag{14.4–13}$$

The image of the segment $-\pi \leqq \operatorname{Im} w \leqq \pi$ in the w-plane is the circumference $|z| = |a_n|$ in the z-plane.

Let us now consider the function

$$h(w) = g(b_n+4\pi w). \tag{14.4–14}$$

It is holomorphic in the disc $|w| < 1$, for

$$\operatorname{Re}(b_n+4\pi w) \leqq \operatorname{Re} b_n+4\pi|w| < -4\pi+4\pi = 0.$$

It omits the values 0 and 1. Finally, in view of (14.4–10)

$$|h(0)-\tfrac{1}{4}| = |g(b_n)-\tfrac{1}{4}| = |f(a_n)-\tfrac{1}{4}| < \tfrac{1}{4},$$

whence

$$|h(0)| < \tfrac{1}{2}. \tag{14.4–15}$$

Now we apply Schottky's theorem, taking $\vartheta = \tfrac{1}{2}$. We find that in the disc the inequality

$$|h(w)| \leqq \beta$$

holds, where β is an absolute constant for according to the remark at the end of the last section β does not depend on $a_0 = h(0)$.

The vertical segment $-\pi \leqq \operatorname{Im} w \leqq \pi$ through b_n corresponds to the circumference $|z| = |a_n|$ in the z-plane. This segment belongs to the disc $|w-b_n| \leqq 2\pi$ and according to (14.4–15) the function does not exceed in absolute value the constant β. This means, however, that on each circumference $|z| = |a_n|$, $n = 1, 2, \ldots$, we also have

$$|f(z)| \leqq \beta$$

and this inequality remains true within the annulus between two such circumferences. Since each z satisfying $0 \leqq |z| \leqq |a_1|$ is in at least one such annulus, we also have

$$|f(z)| \leqq \beta, \qquad |z| \leqq |a_1|,$$

in contradiction with the fact that $z = 0$ is an essential singular point.

A somewhat more general statement of Picard's second theorem is

If $f(z)$ is holomorphic in the region $0 < |z-z_0| < R$ and omits the values a and b, $a \neq b$, then z_0 is not an essential singular point.

Indeed, we may apply the previous theorem to the function

$$\frac{f(z_0+Rz)-a}{b-a}.$$

We conclude that in every neighbourhood of an isolated essential singular point the function omits at most one value.

Picard's first theorem may be obtained from the above theorem by the substitution $z \to 1/(z-a)$.

14.4.5 – Normal families

An astonishing simple criterium for normality of a family of functions has been given by Montel. It states that a family is certainly normal in a region if the functions omit two values, say 0 and 1.

In order to prove this statement we will employ an elegant test for normality due to Ahlfors.

By the *chordal derivative* of a function $f(z)$ is understood

$$\chi(f) = \lim_{h\to 0} \frac{\chi(f(z+h), f(z))}{|h|}, \tag{14.4–16}$$

where $\chi(a, b)$ denotes the chordal distance of the points a and b, (section 1.1.5). Assuming that $f(z)$ is differentiable at the point z, we find from (1.1–15) and

$$\chi(f(z+h), f(z)) = \frac{2|f(z+h)-f(z)|}{\sqrt{1+|f(z+h)|^2}\sqrt{1+|f(z)|^2}}$$

that

$$\boxed{\chi(f) = \frac{2|f'(z)|}{1+|f(z)|^2}.} \tag{14.4–17}$$

Ahlfors's theorem states

A family of holomorphic functions in a region $\mathfrak{R}$ is normal if and only if on every closed and bounded subset of $\mathfrak{R}$ the chordal derivatives of the functions are uniformly bounded.

Suppose first that this condition is not fulfilled. Then we can find a closed and bounded subset $\mathfrak{C}$ of $\mathfrak{R}$ and a sequence $f_1, f_2, \ldots$ such that $\max \chi(f_n)$ on $\mathfrak{C}$ tends to ∞. If there would be a subsequence $f_{n_1}, f_{n_2}, \ldots$ being uniformly convergent on any closed and bounded subset of $\mathfrak{R}$, then this subsequence would tend to a holomorphic function g by Weierstrass's theorem of section 2.20.1. Then $\chi(f_{n_k})$ would be bounded on $\mathfrak{C}$, contrary to the assumption. If the subsequence would tend uniformly to ∞, then the sequence of the reciprocals $1/f_{n_1}, 1/f_{n_2}, \ldots$ would tend to zero, uniformly. Observing that

$$\chi(f) = \chi\left(\frac{1}{f}\right), \tag{14.4–18}$$

it would follow $\chi(f_{n_k}) \to 0$ which is again in contradiction with the assumption.

The proof of the sufficienency of the condition is based on the following statement: Every point z_0 in $\mathfrak{R}$ has a neighbourhood $\mathfrak{U}$ such that, if $|f(z_0)| < A$ for all functions of the family, then $|f(z)| < 2A$ for all z in $\mathfrak{U}$ and, if $|f(z_0)| > B$ for all functions of the family, then $|f(z)| > \frac{1}{2}B$ for all z in $\mathfrak{U}$.

Consider a closed disc $|z-z_0| \leqq \rho$ within $\mathfrak{R}$. Then, by hypothesis, $\chi(f) < M$ for a fixed M and all z in this disc. Let $|f(z_0)| < A$ for all f in the family. Assume that $|f(z)| < 2A$ does not hold for some function f and $|z-z_0| < \rho$. Then there is a point z_1 closest to z_0 at which $|f(z_1)| = 2A$ and $|f(z)| < 2A$ on the segment connecting z_0 and z_1. Integrating along this segment we have by Darboux's inequality (section 2.4.3)

$$|f(z_1)-f(z_0)| = \left|\int_{z_0}^{z_1} f'(\zeta)d\zeta\right| \leqq \rho \max |f'(z)| < \rho M(1+4A^2).$$

On the other hand we have

$$2A = |f(z_1)| \leqq |f(z_1)-f(z_0)|+|f(z_0)| < |f(z_1)-f(z_0)|+A,$$

whence

$$A < |f(z_1)-f(z_0)|.$$

Since ρ can be taken as small as desired this leads to a contradiction. Thus $|f(z)| < 2A$ if $|z-z_0| < \rho$, provided that ρ is sufficiently small. If $|f(z_0) > B$ for every function of the family, then $|1/f(z_0)| < 1/B$ and by the previous result $|1/f(z)| < 2/B$ in a suitable neighbourhood of z_0. This concludes the proof of the assertion.

It follows that the set of points of $\mathfrak{R}$ at which $f(z)$ is bounded for all functions of the family is open. Also the set of points of $\mathfrak{R}$ at which the $f(z)$ are unbounded is open. Since $\mathfrak{R}$ is connected it cannot be decomposed into disjunct open subsets (section 9.1.1) and thus we find that the functions f are either bounded at all points of $\mathfrak{R}$ or unbounded at all points.

Let $f_1, f_2, \ldots$ be any sequence of functions of the family. Then either the sequence is bounded at every point of $\mathfrak{R}$, or the sequence is unbounded on every point of $\mathfrak{R}$.

Consider the second case. Let z_0 denote a given point and B an arbitrary positive number. Then $|f_n(z_0)| > B$ if $n > n_0$ and hence $|f_n(z)| > \frac{1}{2}B$ in a sufficiently small neighbourhood of z_0. If $\mathfrak{C}$ is a closed and bounded subset of $\mathfrak{R}$, then $\mathfrak{C}$ can be covered by a finite number of discs of the considered kind and we may conclude that $|f_n(z)| > \frac{1}{2}B$ for all z in $\mathfrak{C}$ and n sufficiently large. Hence the sequence $f_1, f_2, \ldots$ tends to infinity uniformly on $\mathfrak{C}$.

By the same reasoning we may infer that the sequence is uniformly bounded on every closed and bounded subset of $\mathfrak{R}$, if the sequence is bounded at every point of $\mathfrak{R}$. Then there is a subsequence which converges uniformly on every closed and bounded subset of $\mathfrak{R}$ and thus we see that Ahlfors's condition implies the normality of the family.

The following theorem can be proved by means of Ahlfors's theorem.

Assume that a family of holomorphic functions in a region is such that on every closed and bounded subset $\operatorname{Im} f(z) > 0$. *Then the family is normal.*

By straight-forward computation it may be verified that

$$\chi(g(z)) = \chi(f(z)),$$

if

$$g(z) = \frac{f(z)-i}{f(z)+i}.$$

From (10.2–3) we deduce that $|g(z)| < 1$ if $\operatorname{Im} f(z) > 0$. Hence the family of the functions $g(z)$, being uniformly bounded throughout $\mathfrak{R}$, is normal and from Ahlfors's criterium follows that also the family of the functions $f(z)$ is normal.

A family of functions regular at a point z_0 is said to be *normal at this* point, if there exists a neighbourhood of z_0 in which the family is normal.

A family of holomorphic functions is normal in a region $\mathfrak{R}$ *if and only if the family is normal at every point of* $\mathfrak{R}$.

The necessity of the condition is trivial. Let $\mathfrak{C}$ denote a closed and bounded subset of $\mathfrak{R}$. About every point of $\mathfrak{C}$ there is a neighbourhood $\mathfrak{U}$ such that $\chi(f)$ is uniformly bounded on a closed disc within $\mathfrak{U}$ about the point. Since we can cover $\mathfrak{C}$ by a finite number of neighbourhoods of this kind, if follows that $\chi(f)$ is also uniformly bounded on $\mathfrak{C}$.

Now we are sufficiently prepared to prove *Montel's theorem*

If the functions of a family are holomorphic in a region $\mathfrak{R}$ *and omit the values 0 and 1, then the family is normal in* $\mathfrak{R}$.

Let z_0 denote any point of $\mathfrak{R}$ and $\mathfrak{K}$ an open disc centred about z_0 and included in $\mathfrak{R}$. Within a sufficiently small neighbourhood of z_0 we can find a function element $\tau(f(z))$, where f is a function of the family. Among the various values of $\tau(f(z))$ we take the value that is in the fundamental domain consisting of two adjacent triangles with vertex in infinity.

Since $f(z)$ avoids the values 0 and 1 the continuation of $\tau(f(z))$ throughout $\mathfrak{K}$ is possible and yields a single-valued function $g(z)$ holomorphic throughout $\mathfrak{K}$. Moreover $\operatorname{Im} g(z) > 0$. From the second theorem of this section follows that the family of the functions g is normal.

Consider now a sequence of functions f of the given family. The sequence of the corresponding functions g contains a subsequence of functions $g_n(z) = \tau(f_n(z))$, $n = 1, 2, \ldots$, which converges either to a function $g(z)$ holomorphic in $\mathfrak{K}$, or to infinity. By virtue of the theorem of section 3.12.1 the function $g(z)$ does not take values on the real axis, for $\operatorname{Im} g(z) > 0$, unless it is a real constant. Thus either $\operatorname{Im} g(z) > 0$ or $\operatorname{Im} g(z) = 0$ for all z in $\mathfrak{K}$. Since the $g_n(z)$ are chosen in the above mentioned fundamental domain the limit $g(z_0)$ is in this domain. We can take the domain such that the only real finite boundary points are 0 and ± 1.

Suppose that $g(z) = 0$ identically. In a closed disc about z_0 within $\mathfrak{K}$ we have

$$|g_n(z)| < \varepsilon,$$

uniformly in this disc, provided that n is sufficiently large. Now

$$f_n(z) = \lambda(g_n(z)),$$

where λ is the function (14.3–67) and if $g_n(z) \to 0$ then $f_n(z) \to 1$, uniformly in the closed disc.

If $g(z) = +1$ or -1, then $g_n(z) \to +1$ or -1 and $f_n(z) \to \infty$, uniformly in the closed disc.

It remains to consider the case that $g(z)$ does not have one of these three exceptional constant values. Since $g(z)$ is holomorphic in $\mathfrak{K}$ the image of this disc, as given by g, is an open set which does not meet the real axis. The values of $g(z)$ at points of a closed disc within $\mathfrak{K}$ constitute a bounded set which has a positive distance from the real axis. On this set the functions $g_n(z)$ tend uniformly to $g(z)$. Hence

$$f_n(z) = \lambda(g_n(z)) \to \lambda(g(z)) = f(z),$$

uniformly on the closed disc within $\mathfrak{K}$.

Thus we proved that the given family is normal at the point z_0. Since z_0 is an arbitrary point of $\mathfrak{R}$ we may conclude that the family is normal throughout $\mathfrak{R}$.

14.4.6 – An alternative proof of Picard's second theorem

The following theorem enables us to give a very short proof of Picard's second theorem.

If $z = 0$ is an essential singularity of the function $f(z)$ then the sequence of functions

$$f_n(z) = f(2^{-n}z) \tag{14.4–19}$$

is in no punctured disc about $z = 0$ a normal family.

By a punctured disc we understand a disc from which the centre is omitted.

Suppose that the function $f(z)$ is holomorphic in the punctured disc $0 < |z| < 1$. Consider the sequence (14.4–19) in the annulus

$$\mathfrak{A}_0 \colon 2^{-2} < |z| < 3 \times 2^{-2}.$$

The values that $f_n(z)$ takes in $\mathfrak{A}_0$ coincide with the values of $f(z)$ in the annulus

$$\mathfrak{A}_n \colon 2^{-2-n} < |z| < 3 \times 2^{2-n}.$$

Since $\mathfrak{A}_n$ and $\mathfrak{A}_{n+1}$ overlap, each value taken by $f(z)$ in the disc $0 < |z| < 3 \times 2^{-2}$ is taken by at least one of the functions $f_n(z)$ in $\mathfrak{A}_0$. Supposing that the sequence (14.4–19) is a normal family, there is a subsequence $f_{n_1}, f_{n_2}, \ldots$ which converges either to a holomorphic function $f_0(z)$ or tends to ∞.

In the first case $f_0(z)$ is holomorphic in $\mathfrak{A}_0$ and, consequently, bounded on the circumference $|z| = \frac{1}{2}$. Since the convergence of the subsequence is uniform on the circumference the functions $f_{n_k}(z)$ are uniformly bounded on $|z| = \frac{1}{2}$, i.e., we can find a constant M such that

$$|f_{n_k}(\tfrac{1}{2}e^{i\theta})| \leqq M$$

and

$$|f(2^{-1-n_k}e^{i\theta})| \leqq M, \qquad k = 1, 2, \ldots.$$

This asserts the existence of a sequence of concentric circles shrinking to the origin on which $|f(z)|$ is bounded. Since $f(z)$ is holomorphic in each annulus bounded by two consecutive circles, the maximum principle (section 2.5.3) shows that $f(z)$ is locally bounded at $z = 0$. By Riemann's theorem (section 2.8.3) the function $f(z)$ can be extended to a function holomorphic throughout the disc about $z = 0$, contrary to the hypothesis.

In the second case the subsequence f_{n_k}, $k = 1, 2, \ldots$ tends to ∞ uniformly on $|z| = \frac{1}{2}$. Hence the sequence $1/f_{n_k}$ tends to zero, uniformly on this circumference. As above we may conclude that $1/f(z)$ is locally bounded at $z = 0$ and can be extended to a regular function having a zero at $z = 0$. It follows that $f(z)$ possesses a pole at $z = 0$ (section 3.2.1), contrary to the assumption that $z = 0$ is an essential singularity. This concludes the proof of the theorem.

A direct consequence is Picard's theorem. Without destroying the generality we may assume that $f(z)$ is holomorphic in the punctured disc $0 < |z| < 1$ and omits the values 0 and 1. Then the functions $f_n(z) = f(2^{-n}z)$ of the above theorem constitute a normal family, by Montel's theorem. This is a contradiction.

The theorem of this section permits a stronger conclusion. Since the family $f_n(z)$ is not normal in the annulus $\mathfrak{A}_0$ there is at least one point z_0 of $\mathfrak{A}_0$ and a disc $\mathfrak{K}_0$: $|z-z_0| < \varepsilon$ in $\mathfrak{A}_0$, such that the sequence is not normal in $\mathfrak{K}_0$. We consider the homothetic discs

$$\mathfrak{K}_n : |z-2^{-n}z_0| < 2^{-n}\varepsilon, n = 1, 2, \ldots.$$

The values taken on by $f(z)$ in $\mathfrak{K}_0$ are the values taken on by $f(z)$ in $\mathfrak{K}_n$. Let now a and b be two different numbers. Suppose it is not true that at least one of the equations $f(z) = a$, $f(z) = b$ has a root in infinitely many of the discs $\mathfrak{K}_n$. Then there is an integer n_0 such that $f(z) \neq a$, $f(z) \neq b$ with z in $\mathfrak{K}_n$, $n \geqq n_0$. It follows that the sequence $f_n(z)$ is normal in $\mathfrak{K}_0$ and we have a contradiction. This result is due to Julia and may be stated as follows

If $z = 0$ is an essential singular point of the function $f(z)$, then there exists a sequence of homothetic discs in which $f(z)$ assumes every value with at most one exception.

This theorem may also be stated in the form

If $z = 0$ is an essential singular point of $f(z)$, then there exists at least one ray $\arg z = \theta$ *such that $f(z)$ assumes every value with at most one exception in each angular region* $\theta-\alpha < \arg z < \theta+\alpha$, *no matter how small the positive number α is.*

A ray of this kind is known as *direction of Julia.* A simple example provides the function $\exp(1/z)$. The Julia directions are given by $\arg z = \pm\frac{1}{2}\pi$.

14.5 – The elliptic modular function

14.5.1 – The mapping problem for the trebly asymptotic triangle

The Schwarzian function $\tau(z)$, introduced in section 14.3.3, maps the upper half of the z-plane onto a triangle with vertices at $s_1 = \infty$, $s_2 = 0$, $s_3 = 1$ and angles equal to zero. It is a solution of Schwarz's differential equation

$$[s]_z = \frac{z^2-z+1}{2z^2(z-1)^2}. \tag{14.5–1}$$

This is an equation of the third order. The solution is much facilitated by considering the linear differential equation of the second order

$$w''+\tfrac{1}{4}\,\frac{z^2-z+1}{z^2(z-1)^2}\,w = 0. \tag{14.5–2}$$

As was pointed out in section 13.4.5 a quotient of two linearly independent solutions of this equation is a solution of the equation (14.5–1). The equation takes a more manageable form if we perform the substitution

$$w = uz^{\frac{1}{2}}(1-z)^{\frac{1}{2}}.$$

This has no consequences for the quotient of two solutions. Inserting this into (14.5–2) and replacing afterwards u by w again we get the differential equation

$$\boxed{z(1-z)w''+(1-2z)w'-\tfrac{1}{4}w = 0.} \tag{14.5–3}$$

This is an equation of the so-called *Fuchsian type.* A general theory of these equations will be dealt with in the next chapters.

The theory asserts the existence of at least one solution which is regular at $z = 0$. Such a solution can be obtained by inserting the power series

$$w(z) = \sum_{\nu=0}^{\infty} c_\nu z^\nu \tag{14.5–4}$$

into the left member of the differential equation. Supposing that this series has a positive radius of convergence we may differentiate it term by term. Equating the sum of the coefficients of equal powers of z to zero we obtain the recursive relations

$$n(n+1)c_{n+1}-n(n-1)c_n+(n+1)c_{n+1}-2nc_n-\tfrac{1}{4}c_n = 0, \qquad n = 0, 1, \ldots,$$

or

$$\frac{c_{n+1}}{c_n} = \left(\frac{n+\frac{1}{2}}{n+1}\right)^2. \tag{14.5–5}$$

Since $c_{n+1}/c_n \to 1$ as $n \to \infty$ we find that the radius of convergence of the series (14.5–5) is unity.

Solving the equations (14.5–5) we find

$$c_n = c_0 \left(\frac{(n-\frac{1}{2})(n-\frac{3}{2})\ldots\frac{1}{2}}{n(n-1)\ldots 1}\right)^2 = \frac{c_0}{\pi}\frac{\Gamma^2(n+\frac{1}{2})}{\Gamma^2(n+1)} \tag{14.5–6}$$

and the desired solution is

$$w_0(z) = \frac{c_0}{\pi}\sum_{\nu=0}^{\infty}\frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)}z^\nu. \tag{14.5–7}$$

It follows that all solutions regular at $z = 0$ differ only by a multiplicative constant.

14.5.2 – Legendre's complete elliptic integrals

Although we are now in possession of an analytic expression for a solution of the differential equation we need a second one which is not regulai at $z = 0$. This can be found in a very easy way, for we may indentify (14.5–7) by a simple integral, namely *Legendre's complete elliptic integral of the first kind*

$$\boxed{\mathrm{K}(z) = \int_0^{\frac{1}{2}\pi}\frac{d\theta}{\sqrt{1-z\sin^2\theta}},} \tag{14.5–8}$$

where the square root takes the value 1 at $z = 0$. This is the same integral as (5.14–4) occurring in the theory of the Jacobian elliptic functions. There $z = k^2$ is the square of the modulus of these functions.

It is our next aim to study the integral as a function of z. It is not difficult to show that this function is holomorphic in the z-plane cut along the real axis from 1 to ∞.

If $|z| < 1$ the integrand may be expanded in a series of powers of z, the series being uniformly convergent with respect to θ (since $|\sin\theta| \leq 1$). Hence we may integrate term by term and we get

$$\mathsf{K}(z) = \sum_{\nu=0}^{\infty} (-1)^\nu A_\nu \binom{-\frac{1}{2}}{\nu} z^\nu, \tag{14.5–9}$$

with

$$A_n = \int_0^{\frac{1}{2}\pi} \sin^{2n}\theta \, d\theta, \qquad n = 0, 1, 2, \ldots. \tag{14.5–10}$$

By virtue of (4.7–39) we have

$$A_n = \tfrac{1}{2}\mathsf{B}(n+\tfrac{1}{2}, \tfrac{1}{2}) = \tfrac{1}{2}\frac{\Gamma(n+\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(n+1)}.$$

From (2.16–19) we deduce

$$\binom{-\frac{1}{2}}{n} = \frac{-\frac{1}{2}(-\frac{1}{2}-1)\ldots(-\frac{1}{2}-n+1)}{n!} = (-1)^n \frac{(n-\frac{1}{2})\ldots\frac{1}{2}}{n!}$$

$$= (-1)^n \frac{1}{\Gamma(\frac{1}{2})} \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+1)}.$$

Thus

$$\boxed{\mathsf{K}(z) = \tfrac{1}{2}\sum_{\nu=0}^{\infty} \frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)} z^\nu,} \qquad |z| < 1 \tag{14.5–11}$$

and this is the same series as (14.5–7) if we take $c_0 = \frac{1}{2}\pi$. It follows that

Legendre's complete elliptic integral of the first kind is a solution of the differential equation (14.5–3), *being regular at* $z = 0$.

It is easily verified that the differential equation remains unchanged if we introduce the variable $z' = 1-z$. As a consequence the equation is also satisfied by the function

$$\boxed{\mathsf{K}'(z) = \mathsf{K}(z') = \mathsf{K}(1-z),} \tag{14.5–12}$$

the *complementary integral of* $\mathsf{K}(z)$. Here the primes do not indicate differentiation.

The function (14.5–12) is regular at $z = 1$; it is holomorphic in the entire z-plane cut along the real axis from 0 to $-\infty$. From the representation of $\mathsf{K}(z)$ as an integral follows that $\mathsf{K}(z) \to \infty$ as $z \to 1$. Hence $\mathsf{K}'(z) \to \infty$ as $z \to 0$. Thus

The integrals $\mathsf{K}(z)$ *and* $\mathsf{K}'(z)$ *represent two linearly independent solutions on the differential equation* (14.5–3).

There exists an important relation between these two integrals which is a special case of Abel's identity (15.1–19) in the theory of homogeneous linear differential equations of the second order. Since the functions are solutions of (14.5–3) we have

$$z(1-z)\frac{d^2\mathsf{K}}{dz^2}+(1-2z)\frac{d\mathsf{K}}{dz}-\tfrac{1}{4}\mathsf{K}=0$$

and

$$z(1-z)\frac{d^2\mathsf{K}'}{dz^2}+(1-2z)\frac{d\mathsf{K}'}{dz}-\tfrac{1}{4}\mathsf{K}'=0.$$

Multiplying the second equation by K and substracting from the result thus obtained the first equation multiplied by K′ we readily find

$$\frac{d}{dz}\,z(1-z)\left(\mathsf{K}\frac{d\mathsf{K}'}{dz}-\mathsf{K}'\frac{d\mathsf{K}}{dz}\right)=0,$$

whence

$$z(1-z)\left(\mathsf{K}\frac{d\mathsf{K}'}{dz}-\mathsf{K}'\frac{d\mathsf{K}}{dz}\right)=c \tag{14.5–13}$$

where c is a constant. This is *Abel's identity.*

We may also write, observing that $\mathsf{K}(0)=\tfrac{1}{2}\pi$,

$$\frac{d}{dz}\frac{\mathsf{K}'}{\mathsf{K}}=\frac{c}{z(1-z)\mathsf{K}^2(z)}=\frac{4c}{\pi^2}\frac{1}{z}+\psi(z),$$

where $\psi(z)$ denotes an ordinary power series in z. Integrating we get

$$\mathsf{K}'(z)=\frac{4c}{\pi^2}\mathsf{K}(z)\log z+\varphi(z), \tag{14.5–14}$$

where $\varphi(z)$ is regular at $z=0$. It appears that $\mathsf{K}'(z)$ has a logarithmic singularity at $z=0$. Notice that the logarithm denotes the principal branch.

The evaluation of the constant c in Abel's identity is not easy and requires some preparatory work.

Finally we remark that the results obtained in this section are special instances of more general considerations which are the subject matter of the chapters 15 and 16.

14.5.3 – Legendre's relations

By *Legendre's complete elliptic integral* of the second kind we understand the function

$$\mathsf{E}(z)=\int_0^{\frac{1}{2}\pi}\sqrt{1-z\sin^2\theta}\,d\theta, \tag{14.5–15}$$

where the square root is positive at $z=0$. This function is holomorphic in the z-plane cut along the real axis from 1 to ∞, as is $\mathsf{K}(z)$. We encountered this integral in section 5.16.5.

If $|z| < 1$ we have the expansion

$$\mathsf{E}(z) = \sum_{\nu=0}^{\infty} (-1)^\nu A_\nu \binom{\frac{1}{2}}{\nu} z^\nu$$

where A_n is again the integral (14.5–10). Since

$$\binom{\frac{1}{2}}{n} = \frac{-1}{2n-1}\binom{-\frac{1}{2}}{n}$$

we now have

$$\mathsf{E}(z) = -\tfrac{1}{2}\sum_{\nu=0}^{\infty} \frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)} \frac{z^\nu}{2\nu-1}, \qquad |z| < 1. \qquad (14.5\text{–}16)$$

Replacing $\sin\theta$ by the variable t Legendre's integrals appear in the form

$$\mathsf{K}(z) = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-zt^2)}} \qquad (14.5\text{–}17)$$

and

$$\mathsf{E}(z) = \int_0^1 \frac{\sqrt{1-zt^2}}{\sqrt{1-t^2}}\, dt. \qquad (14.5\text{–}18)$$

Performing the substitution $u = t^{-2}$ we get

$$\mathsf{K}(z) = \tfrac{1}{2}\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\, du \qquad (14.5\text{–}19)$$

and

$$\mathsf{E}(z) = \tfrac{1}{2}\int_1^\infty u^{-\frac{3}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{\frac{1}{2}}\, du. \qquad (14.5\text{–}20)$$

The derivatives are

$$\frac{d\mathsf{K}}{dz} = \tfrac{1}{4}\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{3}{2}}\, du \qquad (14.5\text{–}21)$$

and

$$\frac{d\mathsf{E}}{dz} = -\tfrac{1}{4}\int_1^\infty u^{-\frac{3}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\, du. \qquad (14.5\text{–}22)$$

Legendre's famous relations are expressions of these derivatives in terms of K and E.

In the integral (14.5–20) we may split off the factor $u-z$ by writing

$$(u-z)^{\frac{1}{2}} = u(u-z)^{-\frac{1}{2}} - z(u-z)^{-\frac{1}{2}}$$

and we obtain the identity

$$\int_1^\infty u^{-\frac{3}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{\frac{1}{2}}\,du$$
$$= \int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du - z\int_1^\infty u^{-\frac{3}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du. \quad (14.5\text{–}23)$$

In view of (14.5–22), (14.5–20) and (14.5–19) this is equivalent to

$$\mathsf{E}(z) = \mathsf{K}(z)+2z\,\frac{d\mathsf{E}(z)}{dz},$$

or

$$\boxed{2z\,\frac{d\mathsf{E}(z)}{dz} = \mathsf{E}(z)-\mathsf{K}(z).} \quad (14.5\text{–}24)$$

The identity

$$\int_1^\infty \frac{d}{du}\left(u^{-\frac{1}{2}}(u-1)^{\frac{1}{2}}(u-z)^{-\frac{1}{2}}\right)du = 0$$

yields

$$\int_1^\infty u^{-\frac{3}{2}}(u-1)^{\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du$$
$$= \int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du - \int_1^\infty u^{-\frac{1}{2}}(u-1)^{\frac{1}{2}}(u-z)^{-\frac{3}{2}}\,du. \quad (14.5\text{–}25)$$

Writing

$$(u-z)^{-\frac{1}{2}} = (u-z)^{-\frac{3}{2}}(u-z) = (u-z)^{-\frac{3}{2}}(u-1)+(u-z)^{-\frac{3}{2}}(1-z),$$

the first integral on the right appears as

$$\int_1^\infty u^{-\frac{1}{2}}(u-1)^{\frac{1}{2}}(u-z)^{-\frac{3}{2}}\,du+(1-z)\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{3}{2}}\,du,$$

and so the equation (14.5–25) becomes

$$\int_1^\infty u^{-\frac{3}{2}}(u-1)^{\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du = (1-z)\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{3}{2}}\,du. \quad (14.5\text{–}26)$$

In the integral on the left we split off the factor $u-1$ by writing

$$(u-1)^{\frac{1}{2}} = u(u-1)^{-\frac{1}{2}}-(u-1)^{-\frac{1}{2}}$$

and we get

$$\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du - \int_1^\infty u^{-\frac{3}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{1}{2}}\,du$$
$$= (1-z)\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}(u-z)^{-\frac{3}{2}}\,du.$$

By virtue of (14.5–19), (14.5–21) and (14.5–22) this turns out to be equivalent to

$$\mathsf{K}(z)+2\,\frac{d\mathsf{E}(z)}{dz} = 2(1-z)\,\frac{d\mathsf{K}(z)}{dz}. \tag{14.5–27}$$

Inserting the expression for $d\mathsf{E}/dz$ from (14.5–24) we finally have

$$\boxed{2z(1-z)\,\frac{d\mathsf{K}(z)}{dz} = \mathsf{E}(z)-(1-z)\mathsf{K}(z).} \tag{14.5–28}$$

It is clear that similar relations must exist for $\mathsf{K}'(z)$ and $\mathsf{E}'(z)$, where $\mathsf{E}'(z)$ is the function

$$\boxed{\mathsf{E}'(z) = \mathsf{E}(z') = \mathsf{E}(1-z).} \tag{14.5–29}$$

These relations are evidently

$$\boxed{2(1-z)\,\frac{d\mathsf{E}'(z)}{dz} = \mathsf{K}'(z)-\mathsf{E}'(z)} \tag{14.5–30}$$

and

$$\boxed{2z(1-z)\,\frac{d\mathsf{K}'(z)}{dz} = z\mathsf{K}'(z)-\mathsf{E}'(z).} \tag{14.5–31}$$

It is now an easy matter to evaluate the constant c in Abel's relation (14.5–13). First we observe that

$$\lim_{z\to 0} z\mathsf{K}'(z) = 0, \tag{14.5–32}$$

as follows immediately from (14.5–14). Secondly

$$\lim_{z\to 0} \mathsf{E}'(z) = \int_0^{\frac{1}{2}\pi} \cos\theta\, d\theta = 1. \tag{14.5–33}$$

By making $z \to 0$ in (14.5–13) we now find, taking into account (14.5–31),

$$c = \lim_{z\to 0} z(1-z)\mathsf{K}(z)\,\frac{d\mathsf{K}'(z)}{dz} = \tfrac{1}{2}\lim_{z\to 0} \mathsf{K}(z)(z\mathsf{K}'(z)-\mathsf{E}'(z)) = -\tfrac{1}{4}\pi$$

and Abel's identity appears as

$$\boxed{\mathsf{K}(z)\,\frac{d\mathsf{K}'(z)}{dz} - \mathsf{K}'(z)\,\frac{d\mathsf{K}(z)}{dz} = -\frac{\pi}{4z(1-z)}.} \tag{14.5–34}$$

Accordingly the relation (14.5–14) becomes

$$\mathsf{K}'(z) = -\frac{1}{\pi}\mathsf{K}(z)\log z+\varphi(z). \tag{14.5–35}$$

The functions $\varphi(z)$ will be determined in section 14.5.11.

Replacing in (14.5–34) $d\mathsf{K}'/dz$ and $d\mathsf{K}/dz$ by their expressions (14.5–31) and (14.5–28) we obtain

$$\boxed{\mathsf{K}'\mathsf{E}+\mathsf{K}\mathsf{E}'-\mathsf{K}\mathsf{K}' = \tfrac{1}{2}\pi,} \tag{14.5–36}$$

which is Legendre's relation (5.17–8), but now proved without the theory of the Jacobian elliptic functions.

14.5.4 – Solution of the mapping problem

According to the general theory the function

$$\boxed{\tau(z) = \frac{i\mathsf{K}'(z)}{\mathsf{K}(z)}} \tag{14.5–37}$$

is a solution of Schwarz's differential equation (14.5–1) and maps, therefore, the upper half of the z-plane onto a triangle in the s-plane with zero angles. It is clear that $\tau(0) = \infty$, for $\mathsf{K}'(0) = \infty$ and that $\tau(1) = 0$, for $\mathsf{K}(1) = \infty$. Further, since $\mathsf{K}'(\frac{1}{2}) = \mathsf{K}(\frac{1}{2})$ we also have $\tau(\frac{1}{2}) = i$. Hence the function (14.5–37) is the same as the function $\tau(z)$ considered in section 14.3.3.

It follows immediately from (14.5–37) that

$$\tau(1-z) = \frac{i\mathsf{K}(z)}{\mathsf{K}'(z)} = -\frac{1}{\tau(z)},$$

which is the relation (14.3–65).

It is clear that $\tau(z)$ as defined by (14.5–37) is holomorphic in the z-plane cut along the real axis from 1 to ∞ and from 0 to $-\infty$. If we continue $\tau(z)$ analytically along a small circumference about the origin in the anti-clockwise sense the function $\mathsf{K}(z)$ retakes its original value, but $\mathsf{K}'(z)$ is transformed into $\mathsf{K}'(z)-2i\mathsf{K}(z)$, since $\log z$ becomes $\log z+2\pi i$. Hence $\tau(z)$ undergoes the transformation

$$\tau(z) \to \tau(z)+2. \tag{14.5–38}$$

Instead of (14.5–35) we may write

$$\mathsf{K}'(z) = -\frac{1}{\pi}\mathsf{K}(z')\log z'+\varphi(z'),\ z'=1-z,$$

If z is continued along a small circumference about $z = 1$ in the clockwise sense, then z' moves along a circle about $z' = 0$ in the anti-clockwise sense and $\mathsf{K}(z)$ is transformed into $\mathsf{K}(z)-2i\mathsf{K}'(z)$. Hence $\tau(z)$ undergoes the transformation

$$\tau(z) \to \frac{\tau(z)}{-2\tau(z)+1}. \tag{14.5–39}$$

These transformations are characterized by the matrices (14.2–35) and (14.2–32) respectively. Hence, repeating the above process in all possible ways we see that the various values of the analytic function generated from $\tau(z)$ by analytic continuation throughout the z-plane are obtained from one value by means of the linear transformations of the congruence group mod 2.

Since Abel's identity (14.5–34) is equivalent to

$$\frac{d\tau(z)}{dz} = -\frac{\pi i}{4z(1-z)\mathsf{K}^2(z)}, \tag{14.5–40}$$

we may represent $\tau(z)$ by the integral

$$\tau(z) = \frac{\pi}{4i}\int_1^z \frac{d\zeta}{\zeta(1-\zeta)\mathsf{K}^2(\zeta)}. \tag{14.5–41}$$

This integral is convergent at $z = 1$, for $\mathsf{K}(z)$ behaves like the logarithm of $1-z$ at this point.

14.3.5 – A COVERING THEOREM

An interesting application of the theory of the function $\tau(z)$ is the following covering theorem

If a function

$$f(z) = z+\sum_{\nu=2}^{\infty} a_\nu z^\nu \tag{14.5–42}$$

is holomorphic throughout the open disc $|z| < 1$ then the image of this disc as given by this function covers an open segment of arbitrarily given direction which contains the origin and whose length is not less than

$$l_0 = \frac{4\pi^2}{\Gamma^4(\frac{1}{4})} = 0{,}228\ldots \tag{14.5–43}$$

This result is sharp.

Consider an open segment containing the origin whose end points are not covered by the image of the disc $|z| < 1$. Let α denote the end point with the largest distance to the origin. Then $f(z)$ does not take the values α and $-\alpha$. Hence the function

$$h(z) = \tfrac{1}{2}(1+\alpha^{-1}f(z))$$

omits the values 0 and 1.

Among the possible values of $\tau(\frac{1}{2})$ we may take

$$\tau(\tfrac{1}{2}) = i,$$

whence

$$\tau(h(0)) = i.$$

It follows that the function $\tau(h(z))$ is regular at $z = 0$ and since $h(z)$ omits the values 0 and 1 if $|z| < 1$, it can be continued analytically throughout the open disc. By the monodromy theorem this function is single-valued.

Since Im $\tau(h(z)) > 0$, the values of

$$g(z) = \frac{\tau(h(z))-i}{\tau(h(z))+i}$$

are for $|z| < 1$ within the unit circle (see (10.2–3)) and it follows from Schwarz's lemma that

$$|g'(0)| \leqq 1.$$

Now

$$g'(0) = \lim_{z\to 0} \frac{\tau(h(z))-i}{z} \frac{1}{\tau(h(z))+i} = \tau'(\tfrac{1}{2})\frac{h'(0)}{2i} = \frac{1}{4\alpha i}\tau'(\tfrac{1}{2}).$$

Here, of course, the prime denotes differentiation. It follows that

$$|\alpha| \geqq \tfrac{1}{4}|\tau'(\tfrac{1}{2})|. \tag{14.5–44}$$

According to (10.3–33) we have

$$\mathrm{K}(\tfrac{1}{2}) = \frac{1}{4\sqrt{\pi}}\Gamma^2(\tfrac{1}{4}) \tag{14.5–45}$$

and from (14.5–40) we get

$$\tau'(\tfrac{1}{2}) = \frac{\pi}{i\mathrm{K}^2(\frac{1}{2})} = \frac{16\pi^2}{i\Gamma^4(\frac{1}{4})}$$

and (14.5–44) appears as

$$|\alpha| \geqq \frac{4\pi^2}{\Gamma^4(\frac{1}{4})}.$$

This proves the first part of the theorem.

If $\lambda(s)$ is the inverse of $\tau(z)$ we readily see that the function

$$f_0(z) = 2\lambda\left(i\frac{1+z}{1-z}\right) - 1$$

is holomorphic throughout the open disc $|z| < 1$. It does not take the values $2\times 0-1 = -1$ and $2\times 1-1 = 1$. Further

$$f_0'(0) = 4i\lambda'(i) = \frac{4i}{\tau'(\frac{1}{2})} = -\frac{\Gamma^4(\frac{1}{4})}{4\pi^2}.$$

Hence the function $f_0(z)/f_0'(0)$ omits the values $\pm 4\pi^2/\Gamma^4(\frac{1}{4})$. This concludes the proof of the theorem.

In quite the same way we may prove

If the odd function

$$f(z) = z + \sum_{\nu=1}^{\infty} a_\nu z^{2\nu+1}$$

is holomorphic throughout the open disc $|z| < 1$ *then the image of this disc covers the disc* $|w| < l_0$, *where* l_0 *is the number* (14.5–43).

14.5.6 – The periods of Weierstrass's pe function

A wholly different approach to the theory of the modular functions is possible by the study of the Weierstrass pe function, considered as a function of its periods. We shall see that the functions $\lambda(\tau)$ and $J(\tau)$ are closely related to the elliptic functions.

Let 2ω, $2\omega'$ be a pair of primitive periods of a Weierstrass pe function. We can express all other periods by adding integral multiples of the two periods. The numbers

$$\begin{aligned} 2\tilde{\omega}' &= a2\omega' + b2\omega, \\ 2\tilde{\omega} &= c2\omega' + d2\omega, \end{aligned} \tag{14.5–46}$$

a, b, c and d being integers, are again primitive periods if and only if 2ω, $2\omega'$ can be expressed as sums of integral mutiples of $2\tilde{\omega}$, $2\tilde{\omega}'$. This condition is necessary, but also sufficient, for any period is then expressible in integral multiples of $2\tilde{\omega}$, $2\tilde{\omega}'$.

Solving (14.5–46) we have, writing D for $ad-bc$,

$$\begin{aligned} 2\omega' &= \frac{1}{D}(d2\tilde{\omega}' - b2\tilde{\omega}), \\ 2\omega &= \frac{1}{D}(-c2\tilde{\omega}' + a2\tilde{\omega}). \end{aligned} \tag{14.5–47}$$

The determinant of this transformation is

$$\frac{D}{D^2} = \frac{1}{D}.$$

Since the coefficients in (14.5–47) are integers it follows that $D = \pm 1$.

We introduce the ratio of the periods

$$\tau = \frac{\omega'}{\omega} \qquad (14.5\text{–}48)$$

and we assume that ω and ω' are such that

$$\operatorname{Im} \tau > 0, \qquad (14.5\text{–}49)$$

as in section 5.8.1. The ratio of the new periods is then

$$\tilde{\tau} = \frac{a\tau+b}{c\tau+d} \qquad (14.5\text{–}50)$$

and Im $\tilde{\tau} > 0$ if and only if $D = 1$. Hence

A change of primitive periods of the Weierstrass pe function induces a transformation of the ratio τ by an element of the modular group.

14.5.7 – The absolute invariant

It is possible to find an automorphic function of the modular group by means of the invariants g_2 and g_3 of Weierstrass's pe function. In (5.11–17) we derived expressions for these invariants in terms of q and ω, where as in (5.8–8)

$$q = \exp \pi i\tau. \qquad (14.5\text{–}51)$$

We have

$$g_2 = \left(\frac{\pi}{\omega}\right)^4 (\tfrac{1}{12}+20q^2+\ldots) \qquad (14.5\text{–}52)$$

and

$$g_3 = \left(\frac{\pi}{\omega}\right)^6 (\tfrac{1}{216}-\tfrac{7}{3}q^2+\ldots). \qquad (14.5\text{–}53)$$

An easy calculation yields

$$\Delta = g_2^3-27g_3^2 = \left(\frac{\pi}{\omega}\right)^{12} (q^2+\ldots). \qquad (14.5\text{–}54)$$

Now we combine these expressions to form a function from which the variable ω cancels and which is, therefore, a function of the ratio τ alone, namely

$$\boxed{J(\tau) = \frac{g_2^3}{\Delta}.} \qquad (14.5\text{–}55)$$

This is the so-called *absolute invariant* of the pe function.

Inserting the expansion (14.5–52) and (14.5–54) we get

$$J(\tau) = \tfrac{1}{1728}\left(\frac{1}{q^2} + c_0 + c_1 q^2 + \ldots\right) \tag{14.5–56}$$

and introducing the variable $t = q^2 = \exp 2\pi i\tau$ we see that the function $J(\tau)$, when expressed in terms of t, has a pole of the first order at $t = 0$. This corresponds to $\tau = i\infty$.

Let 2ω. $2\omega'$ of given ratio $\tau = \omega'/\omega$ generate the set of lattice points representing the periods of the pe function. Then the periods $2\tilde{\omega} = a2\omega' + b2\omega$, $2\tilde{\omega}' = c2\omega' + d2\omega$, where a, b, c and d are integers and $ad - bc = 1$ generate the same pattern of lattice points and the associated pe function and the invariants g_2, g_3, Δ are the same. Hence $J(\tau)$ is unaltered if τ undergoes a transformation of the modular group. A fundamental domain is e.g. the figure consisting of two triangles in the upper half of the τ-plane with vertices at $\rho - 1$, i, ρ and ∞ (fig. 14.3–9), with $\rho = \exp(2\pi i/3)$. Since $J(\tau)$ has a pole at a parabolic point we may state

The function $J(\tau)$ is a simple automorphic function with respect to the modular group.

The series (14.5–52) and (14.5–53) are uniformly convergent if $|q^2| \leq r < 1$. It follows

The function $J(\tau)$ is holomorphic in the upper half of the τ-plane.

According to the convention of section 13.4.3. the function $J(\tau)$ has a single pole of the first order in the fundamental domain. Hence $J(\tau)$ is of the first degree. Otherwise stated

The function $J(\tau)$ takes on each value once and only once in the fundamental domain.

It is clear that the function $J(\tau)$ introduced in this section coincides with the function considered in section 14.3.3 if we identify the variables s and τ. For in section 5.12.1 we found that $g_3 = 0$ if $\tau = i$, and $g_2 = 0$ if $\tau = \rho = \exp(2\pi i/3)$. Hence

$$J(i) = 1, \qquad J(\rho) = 0.$$

Since two simple automorphic functions belonging to the same group and being both of the first order can only differ in a multiplicative constant, the identity is established.

Squaring the expressions (5.10–2) we easily deduce

$$\begin{aligned} e_1 &= \frac{\pi^2}{12\omega^2}\left(\vartheta_3^4(0) + \vartheta_4^4(0)\right), \\ e_2 &= \frac{\pi^2}{12\omega^2}\left(\vartheta_2^4(0) - \vartheta_4^4(0)\right), \\ e_3 &= -\frac{\pi^2}{12\omega^2}\left(\vartheta_2^4(0) + \vartheta_3^4(0)\right). \end{aligned} \tag{14.5–57}$$

In section 5.10.2 we already obtained

$$\Delta = 16\left(\frac{\pi}{2\omega}\right)^{12} \vartheta_2^8(0)\vartheta_3^8(0)\vartheta_4^8(0) \tag{14.5–58}$$

and from (14.5–57) we easily deduce

$$g_2 = \tfrac{2}{3}\left(\frac{\pi}{2\omega}\right)^4 (\vartheta_2^8(0)+\vartheta_3^8(0)+\vartheta_4^8(0)). \tag{14.5–59}$$

Thus we may bring $J(\tau)$ into the elegant form

$$\boxed{J(\tau) = \frac{1}{54}\frac{(\vartheta_2^8(0)+\vartheta_3^8(0)+\vartheta_4^8(0))^3}{\vartheta_2^8(0)\vartheta_3^8(0)\vartheta_4^8(0)}.} \tag{14.5–60}$$

The mapping properties of the function may be summarized as follows

The function $J(\tau)$ maps the fundamental domain $\mathfrak{R}_0$, consisting of two triangles with vertices at ∞, i, ρ, $\rho-1$ and angles, 0, $\frac{1}{2}\pi$, $\frac{1}{3}\pi$ and adjacent along the imaginary axis in the τ-plane from i to $i\infty$ onto the z-plane, the mapping being one-to-one. The interior corresponds to the z-plane cut along the real axis from $-\infty$ to 1. If we complete this interior with a vertical side on the left of the boundary and a circular arc from $\rho-1$ to i, then the points of the upper border of the slit must be added to the region in the z-plane. The left half of the fundamental region corresponds to the upper half of the z-plane.

14.5.8 – Existence of a function with prescribed invariants

A consequence of the last theorem of the previous section is that the equation

$$J(\tau) = z, \tag{14.5–61}$$

where z is any given complex number, has always a solution τ in the fundamental region. This enables us to solve the following problem:

Given two numbers g_2, g_3 we ask whether there exist two numbers ω, ω' such that

$$g_2(\omega, \omega') = g_2, \qquad g_3(\omega, \omega') = g_3, \tag{14.5–62}$$

where $g_2(\omega, \omega')$ and $g_3(\omega, \omega')$ are the expressions (5.4–2) with $w = 2n\omega+2n'\omega'$. The answer is affirmative.

It is always possible to choose the periods 2ω, $2\omega'$ of a pe function in such a way that its invariants $g_2(\omega, \omega')$ and $g_3(\omega, \omega')$ take prescribed values g_2, g_3, provided that

$$g_2^3-27g_3^2 \neq 0. \tag{14.5–63}$$

In section 5.13.4 we solved this problem under the restriction that g_2 and g_3 are real numbers. Our general method covers this case.

1) If $g_2 = 0$ we take $\tau = \rho = \exp(2\pi i/3)$. Then $\omega' = \omega\rho$ and from (5.12–4) follows that $g_2(\omega, \omega') = 0$. If we determine ω from

$$(2\omega)^6 = \frac{140}{g_3}\sum{}' \frac{1}{(n+n'\rho)^6}$$

the equations (14.5–62) are satisfied ($\sum'$ denotes summation excluding $n = n' = 0$).

2) If $g_3 = 0$ we take $\tau = i$, hence $\omega' = \omega i$. Then, in view of (5.12–3), $g_2(\omega, \omega') = 0$ and ω may be determined from

$$(2\omega)^4 = \frac{60}{g_2}\sum{}' \frac{1}{(n+n'i)^4}.$$

3) There remains the case $g_2 \neq 0$, $g_3 \neq 0$. The equations (14.5–62) are satisfied if and only if

$$\frac{g_2(\omega, \omega')}{g_3(\omega, \omega')} = \frac{g_2}{g_3}, \quad \frac{g_2^3(\omega, \omega')}{g_2^3(\omega, \omega') - 27g_3^2(\omega, \omega')} = \frac{g_2^3}{g_2^3 - 27g_3^2}.$$

Now we solve the equation

$$J(\tau) = \frac{g_2^3}{g_2^3 - 27g_3^2}$$

and determine ω from

$$\omega^2 = \frac{g_2}{g_3}\,\frac{140\sum' \dfrac{1}{(n+n'\tau)^6}}{60\sum' \dfrac{1}{(n+n'\tau)^4}}.$$

This concludes the proof of the assertion.

An alternative statement is

The differential equation

$$w'^2 = 4w^3 - g_2 w - g_3 \qquad (14.5\text{–}64)$$

can be solved by a Weierstrassian pe function

$$w = \wp(z|\omega, \omega'),$$

provided that the inequality (14.5–63) *holds.*

14.5.9 – The function $\lambda(\tau)$

If a pe function is given, the numbers

$$e_1 = \wp(\omega), \quad e_2 = \wp(\omega+\omega'), \quad e_3 = \wp(\omega') \qquad (14.5\text{–}65)$$

are uniquely determined. In view of (5.15–4) it is natural to consider the function

$$\boxed{\lambda(\tau) = \frac{e_2-e_3}{e_1-e_3}.} \tag{14.5–66}$$

It may be seen from (5.11–19) that λ depends on τ only.

The function $\lambda(\tau)$ takes the values 0 and 1 nowhere in the upper half of the τ-plane.

This is a direct consequence of the fact that e_1, e_2 and e_3 are unequal.

Let us now perform a transformation (14.5–46) of the periods. For the new periods we have

$$\begin{aligned}\tilde{e}_1 &= \wp(c\omega'+d\omega),\\ \tilde{e}_2 &= \wp(c\omega'+d\omega+a\omega'+b\omega),\\ \tilde{e}_3 &= \wp(a\omega'+b\omega).\end{aligned} \tag{14.5–67}$$

Since we obtain the same pe function, the numbers $\tilde{e}_1, \tilde{e}_2, \tilde{e}_3$ must be the same as e_1, e_2, e_3, but possibly arranged in different order. In general the function $\lambda(\tau)$ is altered. If, however, $c\omega'+d\omega$ and ω differ by a period, that is, if $c\omega'+(d-1)\omega$ is a period, we have $\tilde{e}_1 = e_1$. This occurs if and only if c is even and d is odd. Similarly, if $a\omega'+b\omega$ and ω' differ by a period, that is, if $(a-1)\omega'+b\omega$ is a period, we have $\tilde{e}_3 = e_3$. This occurs if a is odd and b is even. If both situations occur then, of course, $\tilde{e}_2 = e_2$ and λ is unchanged. If, conversely, b and c are even, then a and d are odd, as a consequence of $ad-bc = 1$. Thus

The function $\lambda(\tau)$ is unaltered by the transformations of the congruence group mod 2.

As we pointed out in section 14.3.3 every transformation of the modular group is contained in one of the six cosets of the congruence group. It is clear that the transformations of the same coset perform the same permutation on e_1, e_2, e_3. We may list the various permutations as follows, (the matrices characterizing the cosets):

$$\begin{aligned}
&\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} : \tilde{e}_1 = e_1, \quad \tilde{e}_2 = e_2, \quad \tilde{e}_3 = e_3 : \lambda,\\
&\begin{bmatrix}1 & 0\\ -1 & 1\end{bmatrix} : \tilde{e}_1 = e_2, \quad \tilde{e}_2 = e_1, \quad \tilde{e}_3 = e_3 : \frac{1}{\lambda},\\
&\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix} : \tilde{e}_1 = e_3, \quad \tilde{e}_2 = e_2, \quad \tilde{e}_3 = e_1 : 1-\lambda,\\
&\begin{bmatrix}0 & -1\\ 1 & 1\end{bmatrix} : \tilde{e}_1 = e_2, \quad \tilde{e}_2 = e_3, \quad \tilde{e}_3 = e_1 : \frac{1}{1-\lambda},\\
&\begin{bmatrix}1 & -1\\ 0 & 1\end{bmatrix} : \tilde{e}_1 = e_1, \quad \tilde{e}_2 = e_3, \quad \tilde{e}_3 = e_2 : \frac{\lambda}{\lambda-1},\\
&\begin{bmatrix}1 & -1\\ 1 & 0\end{bmatrix} : \tilde{e}_1 = e_3, \quad \tilde{e}_2 = e_1, \quad \tilde{e}_3 = e_2 : \frac{\lambda-1}{\lambda}.
\end{aligned} \tag{14.5–68}$$

In the last column are listed the corresponding transforms of λ.

Thus we established

The quotient group of the modular group with respect to the congruence group mod 2 *is isomorphic to the symmetric group of three symbols and to the group of the cross ratios.*

It is easy to verify that $J(\tau)$ can be expressed rationally in terms of $\lambda(\tau)$. From

$$\lambda = \frac{e_2-e_3}{e_1-e_3}, \qquad 1-\lambda = \frac{e_1-e_2}{e_1-e_3}$$

follows

$$1-\lambda(1-\lambda) = \frac{(e_1-e_3)^2-(e_2-e_3)(e_1-e_2)}{(e_1-e_3)^2}$$

$$= \frac{e_1^2+e_2^2+e_3^2-e_1e_2-e_1e_3-e_2e_3}{(e_1-e_3)^2},$$

whence in view of (5.4–4) and (5.4–5)

$$1-\lambda(1-\lambda) = \tfrac{3}{4}\frac{g_2}{(e_1-e_3)^2}.$$

Taking into account (5.4–11) we now have

$$J(\tau) = \frac{g_2^3}{\Delta} = \frac{g_2^3}{16(e_1-e_2)^2(e_2-e_3)^2(e_3-e_1)^2} = \frac{\dfrac{g_2^3}{(e_1-e_3)^6}}{16\left(\dfrac{e_2-e_3}{e_1-e_3}\right)^2\left(\dfrac{e_1-e_2}{e_1-e_3}\right)^2}$$

$$= \tfrac{4}{27}\frac{(1-\lambda(1-\lambda))^3}{\lambda^2(1-\lambda)^2},$$

or finally

$$J(\tau) = \tfrac{4}{27}\frac{(\lambda^2-\lambda+1)^3}{\lambda^2(1-\lambda)^2}, \tag{14.5–69}$$

being the same expression as (14.3–78). As a consequence the function $\lambda(\tau)$ can have no singularities other than ordinary poles. Hence

The function $\lambda(\tau)$ is a simple automorphic function with respect to the congruence group mod 2.

Let $\mathfrak{B}_0$ denote the fundamental domain of the group consisting of two triangles with zero angles and vertices at ∞, the other vertices being at -1, 0, $+1$. Some particular values of λ may be found as follows.

If we make the change of periods by means of

$$\tilde{\omega}' = \omega'+\omega,$$
$$\tilde{\omega} = \omega,$$

i.e.

$$\tilde{\tau} = \tau+1,$$

then λ changes into $\lambda/(\lambda-1)$. The transformation in the τ-plane has a fixed point at $\tau = \infty$, whence

$$\lambda(\infty) = \frac{\lambda(\infty)}{\lambda(\infty)-1},$$

or

$$\lambda(\infty) = 0.$$

Making the change of periods by means of

$$\tilde{\omega}' = -\omega,$$
$$\tilde{\omega} = \omega',$$

i.e.,

$$\tilde{\tau} = -\frac{1}{\tau},$$

then λ changes into $1-\lambda$. The point $\tau = i$ is a fixed point, whence

$$\lambda(i) = \tfrac{1}{2}.$$

Since $\tau = \infty$ is transformed into $\tilde{\tau} = 0$ we have

$$\lambda(0) = 1-\lambda(\infty),$$

whence

$$\lambda(0) = 1.$$

Performing the change

$$\tilde{\omega}' = \omega',$$
$$\tilde{\omega} = \omega+\omega',$$

i.e.,

$$\tilde{\tau} = \frac{\tau}{1+\tau},$$

then $\tau = \infty$ is transformed into $\tilde{\tau} = 1$ and λ into $1/\lambda$, whence

$$\lambda(1) = \frac{1}{\lambda(\infty)},$$

or

$$\lambda(1) = \infty.$$

If we introduce the variable $t = \exp \pi i \tau$, then (14.5–56) takes the form

$$J(\tau) = \tfrac{1}{1728}\left(\frac{1}{t^2}+c_0+c_1 t^2+\ldots\right) \tag{14.5–70}$$

and it appears that the function has a pole of the second order at $t = 0$ of the r-plane. From (14.5–69) we deduce

$$\lim_{\tau\to\infty} \lambda^2(\tau)J(\tau) = \tfrac{4}{27}. \tag{14.5–71}$$

Hence $\lambda(\tau)$ has a simple zero at $t = 0$, and thus we see that $\lambda(\tau)$ is of the first degree.

The function $\lambda(\tau)$ takes on each value once and only once in the fundamental domain. The function maps the right half of its fundamental domain onto the upper half of the τ-plane.

The last assertion follows from the fact that λ increases from 0 to 1 if τ moves along the imaginary axis from $i\infty$ to 0.

It shoud be noticed that the t of (14.5–56) is the auxiliar variable (13.4–6) for the transformation $\tilde{\tau} = \tau+1$ and the t of (14.5–70) that of the transformation $\tilde{\tau} = \tau+2$. The first is a transformation of the modular group which alters $\lambda(\tau)$.

It is now clear that the function $\lambda(\tau)$ introduced in this section is the same as the function $\lambda(s)$ of section 14.3.3 if we identify the variables s and τ.

We can now complete a remark made in section 5.15.1. To every value of τ with $\operatorname{Im} \tau > 0$ we can construct Jacobian functions, for k and k' may be evaluated, e.g., by (5.15–6).

Since the equation

$$\lambda(\tau) = k^2$$

has a solution in the upper half of the closed τ-plane, provided that $k^2 \neq 0$, $k^2 \neq 1$, we may state

To every value of k which is different from 0 and ± 1 there exist Jacobian functions with modulus k.

Otherwise stated

The differential equation

$$w'^2 = (1-w^2)(1-k^2w^2) \tag{14.5–72}$$

admits as a solution a Jacobian elliptic function with modulus k, provided that $k^2 \neq 0$, $k^2 \neq 1$.

Finally we wish to express $\lambda(\tau)$ in terms of the theta functions at $z = 0$. From (5.10–2) we deduce

$$\boxed{\lambda(\tau) = \frac{\vartheta_2^4(0)}{\vartheta_3^4(0)}} \tag{14.5–73}$$

and this yields, in view of (5.9–15), introducing the variable $q = \exp \pi i \tau$,

$$\boxed{\lambda(\tau) = 16q \frac{\prod_{\nu=1}^{\infty}(1+q^{2\nu})^8}{\prod_{\nu=1}^{\infty}(1+q^{2\nu-1})^8},} \qquad |q| < 1. \tag{14.5–74}$$

Denoting the expression on the right by $Q(q)$ it is clear that this function is holomorphic within the open disc $|q| < 1$ and does not assume the value 1. In particular

$$Q(0) = 0, \quad Q'(0) = 16. \tag{14.5–75}$$

This latter result has remarkable consequences.

14.5.10 – A covering theorem

With the help of the function $Q(q)$ introduced at the end of the previous section we can prove the following remarkable covering theorem

Let $f(z)$ be holomorphic in the unit disc $|z| < 1, f(0) = 0, f'(0) = 1$ and $f(z) \neq 0$ at all points of the disc punctured at $z = 0$. Then the conformal image of this disc as given by $f(z)$ completely covers the interior of a circle about the origin whose radius is $1/16$. This result is sharp.

Since $Q'(0) \neq 0$ the function $Q(q)$ is uniquely invertible in the vicinity of $q = 0$. Suppose that α is a value not taken by $f(z)$ in $|z| < 1$. Consider the function

$$g(z) = \check{Q}\left(\frac{f(z)}{\alpha}\right), \tag{14.5–76}$$

where $\check{Q}$ denotes the inverse of Q. The function has a meaning for z in a small neighbourhood of $z = 0$ and since $f(z)/\alpha$ omits the value 1 it can be continued analytically throughout $|z| < 1$ and by the monodromy theorem $g(z)$ turns out to be a single valued function of z. In particular $g(0) = 0$. Since $|g(z)| < 1$ we have by Schwarz's lemma

$$|g'(0)| \leqq 1.$$

Now

$$f(z) = \alpha Q(g(z)),$$

whence

$$1 = |f'(0)| = |\alpha| Q'(0) |g'(0)| \leqq 16\alpha$$

and so

$$|\alpha| \geqq \tfrac{1}{16}.$$

That the constant $1/16$ cannot be improved upon is shown by the function

$$f(z) = \tfrac{1}{16} Q(z) = z + \ldots$$

which leaves out the value $1/16$.

14.5.11 – The expansion of the complementary elliptic integral of the first kind

In section 14.5.3 we obtained the formula (14.5–35) for the function $\mathsf{K}'(z)$. No information was got for the function $\varphi(z)$ and we devote this section to the problem of finding the expansion of this function in a neighbourhood of $z = 0$.

Suppose that the function

$$-\frac{1}{\pi}\,\mathsf{K}(z)\log z+\psi(z),$$

where $\psi(z)$ is regular at $z = 0$, is also a solution of the differential equation (14.5–3). Then the difference of this function and the function (14.5–35) is a solution which is regular at $z = 0$.

Hence

$$\varphi(z)-\psi(z) = c\mathsf{K}(z).$$

If now $\varphi(0) = \psi(0)$ we conclude that $c = 0$ and so $\varphi(z) = \psi(z)$.

Our first task will be the evaluation of the constant $\varphi(0)$. Consider (with reference to (14.5–35) and (14.5–37)) the function

$$\tau(z) = \frac{1}{\pi i}\log z+\frac{i\varphi(z)}{\mathsf{K}(z)}. \tag{14.5–77}$$

Next we introduce

$$q(z) = \exp \pi i\tau(z) = z \exp\left(-\frac{\pi\varphi(z)}{\mathsf{K}(z)}\right). \tag{14.5–78}$$

Then

$$Q(q(z)) = \lambda(\tau(z)) = z$$

and so

$$\frac{Q(q(z))}{q(z)} = \frac{z}{q(z)}.$$

By making $z \to 0$ along the real axis from the positive side, the function $\tau(z) \to i\infty$ and $q(z) \to 0$.

Hence

$$16 = Q'(0) = \frac{1}{q'(0)},$$

or

$$q'(0) = \tfrac{1}{16}.$$

It follows from (14.5–78) by differentiating and taking afterwards $z = 0$

$$\tfrac{1}{16} = \exp\left(-\frac{\pi\varphi(0)}{\mathsf{K}(0)}\right) = \exp\left(-2\varphi(0)\right),$$

whence

$$\varphi(0) = \log 4. \tag{14.5–79}$$

In order to obtain an expansion for $\varphi(z)$ we try to find a solution of (14.5–3) having a logarithmic singularity at the origin. A simple trick, rather common in the theory of the linear differential equations, is the following: instead of the power series (14.5–4) we substitute the series

$$w(\rho, z) = z^\rho \sum_{\nu=0}^{\infty} c_\nu z^\nu = \sum_{\nu=0}^{\infty} c_\nu z^{\nu+\rho}, \tag{14.5–80}$$

where ρ is a real and positive number. The power z^ρ is the principal power. Inserting (14.5–80) into (14.5–3) we find

$$z(1-z)w''(\rho, z)+(1-2z)w'(\rho, z)-\tfrac{1}{4}w(\rho, z) = \rho^2 c_0 z^{\rho-1},$$

provided that

$$\frac{c_{n+1}}{c_n} = \frac{(n+\rho+\frac{1}{2})^2}{(n+\rho+1)^2}. \tag{14.5–81}$$

It follows that

$$c_n = \frac{\Gamma^2(n+\rho+\frac{1}{2})}{\Gamma^2(n+\rho+1)}\frac{c_0}{\pi}. \tag{14.5–82}$$

Assuming that c_0 is a constant the functions c_n, considered as functions of the complex variable ρ, are regular at $\rho = 0$. The series (14.5–80) is convergent in the disc $|z| < 1$ and uniformly convergent with respect to ρ for ρ near $\rho = 0$. Hence the sum $w(\rho, z)$ is also regular at $\rho = 0$ at each point $z \neq 0$. The derivatives with respect to ρ are obtained by differentiating term by term. Since the operations

$$\frac{d}{dz}, \frac{\partial}{\partial \rho}$$

are permutable, we get

$$z(1-z)\frac{\partial w''}{\partial \rho}+(1-2z)\frac{\partial w'}{\partial \rho}-\tfrac{1}{4}\frac{\partial w}{\partial \rho} = c_0 z^{\rho-1}(2\rho+\rho^2 \log z).$$

By making $\rho \to 0$ we see that

$$\left.\frac{\partial w(\rho, z)}{\partial \rho}\right|_{\rho=0} \tag{14.5–83}$$

is again a solution of (14.5–3).

Differentiating (14.5–82) logarithmically we find in view of (4.8–1)

$$\frac{\partial c_n/\partial \rho}{c_n} = 2(\Psi(n+\rho+\tfrac{1}{2})-\Psi(n+\rho+1))$$

and so

$$\left.\frac{\partial c_n}{\partial \rho}\right|_{\rho=0} = 2\frac{\Gamma^2(n+\frac{1}{2})}{\Gamma^2(n+1)}(\Psi(n+\tfrac{1}{2})-\Psi(n+1))\frac{c_0}{\pi}. \tag{14.5–84}$$

Since

$$\left.\frac{\partial w(z,\rho)}{\partial \rho}\right|_{\rho=0} = \sum_{\nu=0}^{\infty} c_\nu z^\nu \log z + \sum_{\nu=0}^{\infty} \left.\frac{\partial c_\nu}{\partial \rho}\right|_{\rho=0} z^\nu \tag{14.5–85}$$

we find that the part regular at the origin is

$$\sum_{\nu=0}^{\infty} \left.\frac{\partial c_\nu}{\partial \rho}\right|_{\rho=0} z^\nu.$$

The coefficient of z^0 is

$$\frac{2c_0}{\pi}\pi(\Psi(\tfrac{1}{2})-\Psi(1)). \tag{14.5–86}$$

Differentiating both members of (4.6–26) logarithmically yields

$$\Psi(z)+\Psi(z+\tfrac{1}{2}) = -2\log z+2\Psi(2z)$$

and, taking $z = \frac{1}{2}$, we get

$$\Psi(1)-\Psi(\tfrac{1}{2}) = 2\log 2. \tag{14.5–87}$$

If in (14.5–86) we take $c_0 = -\frac{1}{2}$ we find the value log 4. The part of (14.5–85) involving the logarithm becomes then

$$-\frac{1}{\pi}\mathsf{K}(z)\log z$$

and we may conclude that the function $\varphi(z)$ in (14.5–35) has the expansion

$$\varphi(z) = -\frac{1}{\pi}\sum_{\nu=0}^{\infty}\frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)}(\Psi(\nu+\tfrac{1}{2})-\Psi(\nu+1))z^\nu.$$

From (4.8–4) we deduce

$$\Psi(n+\tfrac{1}{2}) = \frac{1}{n-\frac{1}{2}}+\Psi(n-\tfrac{1}{2}) = \frac{2}{2n-1}+\frac{2}{2n-3}+\ldots+\frac{2}{1}+\Psi(\tfrac{1}{2})$$

and

$$\Psi(n+1) = \frac{1}{n}+\Psi(n) = \frac{2}{2n}+\frac{2}{2n-2}+\ldots+\frac{2}{2}+\Psi(1),$$

whence in view of (14.5–87)

$$\Psi(n+\tfrac{1}{2})-\Psi(n+1) = 2\left(\frac{1}{1}-\frac{1}{2}+\ldots-\frac{1}{2n}\right)-2\log 2,$$

where the expression between the brackets on the right is to be interpreted as 0 if $n = 0$. Thus we finally have

$$\mathsf{K}(z) = -\frac{1}{2\pi}\sum_{\nu=0}^{\infty}\frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)}z^\nu\left(\log\frac{z}{16}+4\left(\frac{1}{1}-\frac{1}{2}+\ldots-\frac{1}{2\nu}\right)\right), \tag{14.5–88}$$

the desired expansion. The expansion is valid for $|z| < 1$ and $|\arg z| < \pi$.

CHAPTER 15

LINEAR HOMOGENEOUS DIFFERENTIAL EQUATIONS

15.1 – General theory

15.1.1 – THE EXISTENCE THEOREM

In many problems we encounter analytic functions which are characterized by functional equations of a special type, viz., differential equations, rather than by their power series expansions. We shall restrict ourselves to an important class of differential equations, having simple properties, the *linear homogeneous equations* which are of the form

$$w^{(n)}+p_1(z)w^{(n-1)}+\ldots+p_{n-1}(z)w'+p_n(z)w = 0. \qquad (15.1\text{–}1)$$

The coefficients $p_1(z), \ldots, p_n(z)$, are supposed to be single-valued functions.

The main theorem of this section states:

If the coefficients in (15.1–1) *are regular at* $z = z_0$ *then there exists a unique solution of the equation such that this solution and its first* $n-1$ *derivatives assume arbitrarily assigned values, at* $z = z_0$.

Consider first the case $n = 1$. We have to solve an equation of the form

$$w'+p(z)w = 0. \qquad (15.1\text{–}2)$$

The solution which takes a prescribed value $w(z_0)$ at $z = z_0$ is evidently

$$w(z) = w(z_0)\exp\left(-\int_{z_0}^{z} p(\zeta)d\zeta\right), \qquad (15.1\text{–}3)$$

where the path of integration is included in a sufficiently small neighbourhood of $z = z_0$.

We wish to give another proof which may be generalized to the higher cases. Since $p(z)$ is supposed to be regular at $z = z_0$ we can expand this function in a power series

$$p(z) = \sum_{\nu=0}^{\infty} p_\nu(z-z_0)^\nu, \qquad (15.1\text{–}4)$$

having a positive radius of convergence. We try to find a solution

$$w(z) = \sum_{\nu=0}^{\infty} c_\nu(z-z_0)^\nu, \qquad (15.1\text{–}5)$$

where c_0 has a prescribed value. The other coefficients are so determined that the series satisfies the differential equation formally. Inserting (15.1–5) into (15.1–2) we find by collecting terms with equal powers of $z-z_0$ the recurrence relations

$$\begin{aligned} c_1+p_0c_0 &= 0,\\ 2c_2+p_0c_1+p_1c_0 &= 0,\\ &\cdots\\ nc_n+p_0c_{n-1}+p_1c_{n-2}+\ldots+p_{n-1}c_0 &= 0,\\ &\cdots \end{aligned} \qquad (15.1\text{–}6)$$

It is obvious that we can evaluate successively $c_1, c_2, \ldots$ in terms of c_0 and thus the uniqueness of the solution has been established. It remains to prove that the relations (15.1–6) lead to a power series (15.1–5) with a positive radius of convergence. It then follows that the formal operations leading to (15.1–6) are permissible and, consequently, the series (15.1–5) represents the desired solution with the prescribed initial value.

Let r_0 denote a positive number not exceeding the radius of convergence of the power series (15.1–4). If M_0 is not smaller than the maximum modulus of $p(z)$ on the circumference $|z-z_0| = r_0$, then by Cauchy's inequality (2.18–2)

$$|p_n| \leqq M_0 r_0^{-n}, \qquad n = 0, 1, 2, \ldots. \qquad (15.1\text{–}7)$$

Now we shall show that numbers M and r can be found such that

$$|c_n| \leqq Mr^{-n}. \qquad (15.1\text{–}8)$$

For $n = 0$ this is trivial. We may take $r = r_0$. Suppose the inequalities (15.1–8) to be valid for all subscripts $< n$, where now n is positive. From (15.1–6) we deduce

$$n|c_n| \leqq MM_0nr^{1-n} = nM(rM_0)r^{-n},$$

if $r = r_0$ and this remains true for all smaller r. If r is sufficiently small then $rM_0 \leqq 1$ and (15.1–8) follows. Hence the series (15.1–5) is convergent, provided that $|z-z_0| < r$.

Next consider the case of an equation of the second order. We shall write it in the form

$$w''+p(z)w'+q(z)w = 0. \qquad (15.1\text{–}9)$$

Expanded in series the coefficients are

$$\begin{aligned} p(z) &= \sum_{\nu=0}^{\infty} p_\nu(z-z_0)^\nu,\\ q(z) &= \sum_{\nu=0}^{\infty} q_\nu(z-z_0)^\nu, \end{aligned} \qquad (15.1\text{–}10)$$

both having a positive radius of convergence.

We try to find a solution of the form (15.1–5), where c_0 and c_1 have prescribed values. Inserting (15.1–5) into (15.1–9) we find by collecting terms with equal powers of $z-z_0$ the recurrence relations

$$\begin{aligned}
&2c_2+p_0c_1+q_0c_0 = 0,\\
&3\times 2c_3+2p_0c_2+p_1c_1+q_0c_1+q_1c_0 = 0,\\
&n(n-1)c_n+(n-1)p_0c_{n-1}+(n-2)p_1c_{n-2}+\ldots+p_{n-2}c_1+\\
&\quad +q_0c_{n-1}+q_1c_{n-3}+\ldots+q_{n-2}c_0 = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{15.1–11}$$

It is obvious that we can evaluate successively $c_2, c_3, \ldots$ in terms of c_0 and c_1 and thus the uniqueness of the solution is proved.

Next we wish to establish that the relations (15.1–11) lead to a power series (15.1–5) which has a positive radius of convergence.

Let again r_0 denote a positive number not exceeding the radii of convergence of the power series (15.1–10). If M_0 is not smaller than the maximum moduli of $p(z)$ and $q(z)$ respectively on the circumference $|z-z_0| = r_0$, then

$$|p_n| \leqq M_0 r_0^{-n}, \qquad |q_n| \leqq M_0 r_0^{-n}, \qquad n = 0, 1, 2, \ldots \tag{15.1–12}$$

It remains to show that we can find numbers M and r such that (15.1–8) holds. For $n = 0,1$ this is trivial; we may take $r = r_0$. Suppose the inequalities (15.1–8) to be valid for all subscripts $< n$. From (15.1–11) we deduce

$$\begin{aligned}
n(n-1)|c_n| &\leqq MM_0((n-1)+\ldots+1)r^{1-n}+(n-1)r^{2-n})\\
&= n(n-1)M\left(\tfrac{1}{2}r+\frac{r^2}{n}\right)M_0 r^{-n}
\end{aligned}$$

and these inequalities remain true for all smaller r. If r is sufficiently close to zero then $(\frac{1}{2}r+r^2/n)\, M_0 \leqq 1$ and (15.1–8) follows.

The above proofs show already the general features for the treatment of equations of higher order. In our subsequent work we need only the cases of the first and the second order.

Finally we wish to observe that the solution obtained is the only regular solution satisfying the prescribed initial conditions. It remains to investigate the question whether or not there may exist non-regular solutions which satisfy the initial conditions. Since $w(z_0)$ is finite a power series representing $w(z)$ cannot involve powers of $z-z_0$ with negative exponents. If the series should involve fractional powers of $z-z_0$ with positive exponents, then after a certain order the derivatives would become infinite at $z = z_0$. The derivatives are obtained from the differential equation by successive differentiation and are, therefore, necessarily finite at $z = z_0$.

We tacitly assumed that z_0 is finite, If, however, $z_0 = \infty$ we have to consider power series in terms of z^{-1}.

15.1.2. – The Wronskian

Two functions $w_0(z)$ and $w_1(z)$, holomorphic in a region $\mathfrak{R}$, are said to be *linearly independent* if the function

$$w(z) = c_0 w_0(z) + c_1 w_1(z), \tag{15.1–13}$$

where c_0 and c_1 are numerical constants, vanishes identically if and only if $c_0 = c_1 = 0$. In the contrary case the functions are called *linearly dependent*. That means: we can find constants c_0 and c_1, such that

$$c_0 w_0(z) + c_1 w_1(z) = 0 \tag{15.1–14}$$

identically, and at least one of the numbers c_0, c_1 is different from zero. Differentiation yields

$$c_0 w_0'(z) + c_1 w_1'(z) = 0 \tag{15.1–15}$$

identically. These two relations between c_0 and c_1 are only consistent if the *determinant of Wronski* (*Wronskian*)

$$W(w_0, w_1) = \begin{vmatrix} w_0(z) & w_1(z) \\ w_0'(z) & w_1'(z) \end{vmatrix} \tag{15.1–16}$$

vanishes identically. Hence

The non-vanishing of the Wronskian is sufficient for linear independence of the functions $w_0(z)$, $w_1(z)$.

Thus, for instance, the Wronskian of the functions $\cos z$ and $\sin z$ is $\cos^2 z + \sin^2 z = 1$. Hence these functions are linearly independent.

A converse of the above assertion will be stated in the following form

If $w_0(z)$ and $w_1(z)$ are solutions of the differential equation (15.1–9), *regular at $z = z_0$, and if their Wronskian vanishes at $z = z_0$, then the functions are linearly dependent and their Wronskian is identically zero.*

Since the Wronskian is zero at $z = z_0$ we can find numbers c_0, c_1 from which at least one is different from zero, such that

$$c_0 w_0(z_0) + c_1 w_1(z_0) = 0,$$
$$c_0 w_0'(z_0) + c_1 w_1'(z_0) = 0.$$

Hence the function (15.1–13), being also a solution of the differential equation (15.1–9), takes the value 0 at $z = z_0$ as does its derivative. From the theorem of the previous section follows that $w(z)$ is identically zero. This concludes the first part of the theorem. The second part is a consequence of the previous theorem.

This latter part can be established in another way. Multiplying the first of the equations

$$\begin{aligned} w_0''+p(z)w_0'+q(z)w_0 &= 0, \\ w_1''+p(z)w_1'+q(z)w_1 &= 0, \end{aligned} \tag{15.1–17}$$

by $-w_1$ and the second by w_0 we have after adding

$$w_0 w_1''-w_0''w_1+p(z)(w_0 w_1'-w_0' w_1) = 0. \tag{15.1–18}$$

Observing that the Wronskian has the derivative

$$W'(z) = w_0 w_1''-w_0''w_1,$$

we find

$$W'(z)+p(z)W(z) = 0,$$

whence

$$\boxed{W(z) = W(z_0)\exp\int_{z_0}^{z} -p(\zeta)d\zeta.} \tag{15.1–19}$$

This identity is referred to as *Abel's identity.* We encountered this in a particular instance in section 14.5.2.

It is clear that $W(z)$ vanishes identically if and only if $W(z_0) = 0$. It follows that *two linearly independent solutions cannot vanish simultaneously at a regular point.*

Another consequence is the following

Any solution is expressible as a linear combination

$$w(z) = c_0 w_0(z)+c_1 w_1(z) \tag{15.1–20}$$

of any two linearly independent solutions $w_0(z)$ *and* $w_1(z)$.

At the regular point z_0 we determine the constants c_0 and c_1 in such a fashion that

$$\begin{aligned} c_0 w_0(z_0)+c_1 w_1(z_0) &= w(z_0), \\ c_0 w_0'(z_0)+c_1 w_1'(z_0) &= w'(z_0). \end{aligned}$$

This is possible since $W(z_0) \neq 0$. Hence the solutions $w(z)$ and $c_0 w_0(z)+ + c_1 w_1(z)$ have the same initial values, as have their derivatives and, consequently, they coincide throughout a neighbourhood of z_0. It follows that we can write down the general solution as soon as we are in possession of two linearly independent solutions. Now there exist solutions $w_0(z)$, $w_1(z)$ with $w_0(z_0) = 1$, $w_0'(z_0) = 0$ and $w_1(z_0) = 0$, $w_1'(z_0) = 1$; they are evidently linearly independent since $W(z_0) = 1$. Thus

The differential equation (15.1–9) *possesses a system of two linearly independent solutions regular at* $z = z_0$, *where* z_0 *is a regular point of the coefficients* $p(z)$ *and* $q(z)$.

A set of linearly independent solutions will be called *a fundamental system at* $z = z_0$.

15.1.3 – The monodromy group

Let $w(z)$ denote a solution of (15.1–9) in a neighbourhood of $z = z_0$. Consider a closed path C in the open z-plane, beginning and ending at z_0 and not passing through any singularity of the coefficients. We can apply the process of analytic continuation simultaneously to the coeffients and the solution $w(z)$ under consideration along C. The continuation will not disturb the relation (15.1–9) because of the principle of permanence of functional equations (section 12.1.7).

After completing the circuit the coefficients of the differential equation resume their original forms as power series in $z - z_0$, being, by hypothesis, single-valued. But $w(z)$ need not resume its original form, though it will remain some solution $v(z)$, say. To trace this change the process of analytic continuation must be simultaneously applied to both solutions $w_0(z)$ and $w_1(z)$ of a fundamental system at $z = z_0$. They will assume the new form

$$\begin{aligned} v_0(z) &= a_{00} w_0(z) + a_{01} w_1(z), \\ v_1(z) &= a_{10} w_0(z) + a_{11} w_1(z). \end{aligned} \tag{15.1–21}$$

Between the Wronskians of v_0, v_1 and w_0, w_1 respectively exists the relation

$$W(v_0, v_1) = \begin{vmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{vmatrix} W(w_0, w_1). \tag{15.1–22}$$

If the determinant involving the coefficients a_{ij} in (15.1–22) were zero, then there would exist a linear relation between v_0 and v_1. But returning along the same path the linear relation would persist.

This can also be seen in the following way. If C is a closed path then by Abel's identity

$$W(v_0(z_0), v_1(z_0)) = W(w_0(z_0), w_1(z_0)) \exp \int_C -p(\zeta) d\zeta,$$

whence

$$\begin{vmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{vmatrix} = \exp \int_C -p(\zeta) d\zeta. \tag{15.1–23}$$

Let γ_1, γ_2 be two closed continuous paths starting and ending at the same point z_0. Continuing the solutions w_0, w_1 along γ_1 we obtain a set which may be represented symbolically by

$$\begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

on applying the rule for multiplication of matrices. Continuing afterwards along γ_2 we obtain the set

$$\begin{bmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{bmatrix} \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix},$$

where the elements of the matrix on the right are obtained by applying the rule for multiplication to the matrices on the left. This is the result of continuation along the path $\gamma_2 \gamma_1$.

Thus we see that the linear transformations belonging to closed paths, beginning and ending at z_0, form a group, the *monodromy group* of the equation. It might seem that this group depends on z_0. But this is not true, for any closed loop beginning and ending at a point z_1 can be transformed into a circuit beginning and ending at z_0 by adding a path connecting z_0 and z_1 which is percorsed in opposite directions. A relation between v_0, v_1 and w_0, w_1 remains unaltered by continuation along this path. It is readily seen that the monodromy group is a *representation* (i.e., a homomorphic image) of the homotopy group of the extended plane punctured at the singularities of the coefficients (section 12.7.5).

15.1.4 – The differential resolvent

From (15.1–21) follows that the quotient $v_1(z)/v_0(z)$ is obtained from the quotient

$$s(z) = \frac{w_1(z)}{w_0(z)} \tag{15.1–24}$$

by means of a linear fractional substitution

$$\frac{v_1}{v_0} = \frac{a_{11}s+a_{10}}{a_{01}s+a_{00}}. \tag{15.1–25}$$

Since w_0 and w_1 are linearly independent the function $s(z)$ has a meaning. It is now clear that the Schwarzian derivative (13.4–12) is a single-valued function

$$\boxed{[s]_z = R(z),} \tag{15.1–26}$$

called *the differential resolvent* of the equation (15.1–9); the expression on the right is called the *invariant* of this equation, for it does not depend on the particular choice of a fundamental system.

In order to evaluate $R(z)$ we substitute $w = u\varphi$ into (15.1–9), where φ is a function to be chosen suitably. The function $u(z)$ satisfies the differential equation

$$u'' + \left(2\frac{\varphi'}{\varphi} + p\right)u' + \left(\frac{\varphi''}{\varphi} + \frac{\varphi'}{\varphi}p + q\right)u = 0. \tag{15.1–27}$$

The middle term vanishes if

$$\frac{2\varphi'}{\varphi} = -p,$$

i.e., we may take

$$\varphi(z) = \exp\left(-\frac{1}{2}\int_{z_0}^{z} p(\zeta)d\zeta\right).$$

Then

$$\frac{\varphi''}{\varphi} - \left(\frac{\varphi'}{\varphi}\right)^2 = -\tfrac{1}{2}p',$$

or

$$\frac{\varphi''}{\varphi} = \tfrac{1}{4}p^2 - \tfrac{1}{2}p'$$

and (15.1–27) takes the form

$$\boxed{u'' + \tfrac{1}{2}R(z)u = 0,} \tag{15.1–28}$$

the so-called *reduced form* of (15.1-9). The function $R(z)$ stands for

$$R(z) = 2q(z) - \tfrac{1}{2}p^2(z) - p'(z). \tag{15.1–29}$$

Let u_0, u_1 denote a fundamental system of (15.1–28) at a regular point of $R(z)$. It is clear that

$$u_0 u_1'' - u_1 u_0'' = 0$$

identically. Introducing the function

$$s(z) = \frac{u_1(z)}{u_0(z)},$$

we have

$$s' = \frac{ds}{dz} = \frac{u_0 u_1' - u_0' u_1}{u_0^2}$$

and by logarithmic differentiation, taking the above identity between u_0 and u_1 into account,

$$\frac{d}{dz}\log s' = -2\frac{u_0'}{u_0}.$$

Differentiating again

$$\frac{d^2}{dz^2}\log s' = -2\frac{u_0''}{u_0} + 2\left(\frac{u_0'}{u_0}\right)^2 = R(z) + \frac{1}{2}\left(\frac{d}{dz}\log s'\right)^2$$

and this leads to (15.1–26).

Any Schwarzian equation (15.1–26) *is the differential resolvent of a linear homogeneous equation* (15.1–9), *where* $p(z)$ *can be chosen arbitrarily.*

Indeed, if $R(z)$ is given we can take $p(z)$ arbitrarily. Then $q(z)$ is determined by (15.1–29) and the equation (15.1–9) with these coefficients $p(z)$ and $q(z)$ has precisely the reduced form (15.1–28). The above considerations ensure that (15.1–26) is the corresponding resolvent.

This theorem finds an application in the theory of conformal mapping of circular polygons. For we may replace the Schwarzian equation by a linear equation of the second order and in many cases we obtain a manageable form by choosing appropriately $p(z)$.

As an illustrative example we consider the equation (14.5–1). Now

$$R(z) = \frac{z^2-z+1}{2z^2(z-1)^2}.$$

If we take

$$p(z) = \frac{2z-1}{z(z-1)},$$

then

$$\tfrac{1}{2}p^2+p' = \frac{-1}{2z^2(z-1)^2},$$

whence

$$2q = \frac{1}{2}\frac{z^2-z}{z^2(z-1)^2} = \frac{1}{2}\frac{1}{z(z-1)}.$$

This yields again (14.5–3).

15.1.5 – The characteristic numbers

Any solution $v(z)$ of (15.1–9) regular at $z = z_0$ can be expressed as a linear combination

$$v = c_0 w_0 + c_1 w_1$$

of the solutions of a fundamental system. We ask whether it is possible to select the constants c_0 and c_1 in such a way that, after continuing v along a closed path, it is reproduced apart from a multiplicative constant σ. It is clear that we exclude the trivial case $c_0 = c_1 = 0$.

After continuation along C the functions w_0, w_1 are carried into v_0, v_1 as represented by (15.1–21) and v becomes σv. As a consequence the relation

$$\sigma(c_0 w_0 + c_1 w_1) = c_0 v_0 + c_1 v_1 = (c_0 a_{00} + c_1 a_{10})w_0 + (c_0 a_{01} + c_1 a_{11})w_1$$

must hold identically. Hence

$$\begin{aligned} \sigma c_0 &= c_0 a_{00} + c_1 a_{10}, \\ \sigma c_1 &= c_0 a_{01} + c_1 a_{11}. \end{aligned} \tag{15.1–30}$$

Since the constants c_0 and c_1 do not vanish simultaneously, a relation (15.1–30) can subsist if and only if σ is a root of the *characteristic equation* associated with the path C

$$\begin{vmatrix} a_{00}-\sigma & a_{10} \\ a_{01} & a_{11}-\sigma \end{vmatrix} = 0. \tag{15.1–31}$$

It cannot have a zero root, for the determinant (15.1–23) is different from zero. To every root of (15.1–31) corresponds at least one system c_0, c_1 which affords a function v having the desired property along C .The roots of (15.1–31) are called the *characteristic numbers* of the transformation.

We contend that *the characteristic numbers are independent of the choice of the initial system of solutions* w_0, w_1. Let w_0^*, w_1^* denote another fundamental system. For the sake of brevity we shall write

$$\mathsf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}.$$

Then there is a square 2×2-matrix Q such that

$$\mathsf{w}^* = \mathsf{Qw}.$$

Analytic continuation along C yields

$$\mathsf{v}^* = \mathsf{Qv} = \mathsf{QAw}.$$

If

$$\mathsf{v}^* = \mathsf{Bw}^* = \mathsf{BQw},$$

we have evidently

$$\mathsf{QA} = \mathsf{BQ}$$

and also

$$\mathsf{Q}(\mathsf{A}-\sigma\mathsf{E}) = (\mathsf{B}-\sigma\mathsf{E})\mathsf{Q}.$$

Since the determinant of a product of matrices is the product of the determinants of the factors, we find, because $\det \mathsf{Q} \neq 0$,

$$\det(\mathsf{A}-\sigma\mathsf{E}) = \det(\mathsf{B}-\sigma\mathsf{E}),$$

which exhibits the *invariance of the characteristic equation and, consequently, of the characteristic numbers.*

Assume first that the roots σ_0 and σ_1 of (15.1–31) are different. The corresponding solutions, starting with a given fundamental system, are denoted by v_0, v_1. After the loop has been described once they become u_0 and u_1 respectively with

$$u_0 = \sigma_0 v_0, \qquad u_1 = \sigma_1 v_1. \tag{15.1–32}$$

The functions v_0, v_1 are linearly independent, for a relation $c_0 v_0 + c_1 v_1 = 0$ becomes, after performing the continuation along the circuit once, $c_0 \sigma_0 v_0 + c_1 \sigma_1 v_1 = 0$. Since

$$\begin{vmatrix} 1 & 1 \\ \sigma_0 & \sigma_1 \end{vmatrix} = \sigma_1 - \sigma_0 \neq 0,$$

by hypothesis, we conclude that $c_0 v_0 = c_1 v_1 = 0$ identically. Hence $c_0 = c_1 = 0$.

Secondly we suppose that $\sigma_0 = \sigma_1 = \sigma$. We can find a solution v_0 such that after performing the continuation along C it becomes $u_0 = \sigma v_0$. Let v_0, v_1 denote a fundamental system. After performing the continuation we must have, a being a constant,

$$\begin{aligned} u_0 &= \sigma v_0 \\ u_1 &= a v_0 + \sigma v_1 \end{aligned} \tag{15.1–33}$$

for the characteristic equation has only the root σ.

The transformations (15.1–32) and (15.1–33) are called *canonical* with respect to the given loop C. The system v_0, v_1 is also called canonical.

15.2 – The theory of Fuchs

15.2.1 – Solutions at a singular point

By a *singular point* of a homogeneous linear differential equation we understand a singular point of at least one of the coefficients. We consider only isolated singularities. At a singular point of the differential equation the behaviour of the solutions become more involved. It is, however, possible to state general rules in special cases, discovered by Fuchs.

Without loss of generality we may assume that a singular point is at the origin. First we consider an equation of the first order

$$w' + p(z)w = 0. \tag{15.2–1}$$

A fundamental system at a regular point consists of one function only. If $z = 0$ is a singular point of $p(z)$, this function may be expanded in a Laurent series

$$p(z) = \sum_{\nu=-\infty}^{\infty} c_\nu z^\nu. \tag{15.2–2}$$

Let z_0 denote a point in the vicinity of the origin. Integration of (15.2–2) along a path, issuing from z_0 and remaining in a neighbourhood of z_0, yields

$$-\int_{z_0}^{z} p(\zeta)d\zeta = \psi(z)+\rho \log z, \qquad \rho = c_{-1}.$$

Hence a solution of (15.2–1) regular at $z = z_0$ appears as

$$w(z) = z^{\rho}\varphi(z), \tag{15.2–3}$$

with $\varphi(z) = \exp \psi(z)$. It is determined up to a multiplicative constant. The function $\varphi(z)$ is single-valued but not necessarily regular at $z = 0$.

Next we focus our attention on the case of a second order equation. We assume again that $z = 0$ is a singular point. Let C denote a small circle about the origin such that within it or on it there is no other singular point of the equation. In a small neighbourhood of z_0 on C we can find a fundamental system of solutions regular at z_0 and we shall suppose that this system is canonical with respect to C.

Suppose first that the two characteristic numbers are different. Performing the analytic continuation along C the solutions w_0 and w_1 are carried into the solutions $\sigma_0 w_0$ and $\sigma_1 w_1$, where σ_0 and σ_1 are the characteristic numbers. The function z^{ρ} has a similar property, for continuation along C yields the function $e^{2\pi i\rho}z^{\rho}$. It is, therefore, natural to take the numbers ρ_0 and ρ_1 in such a way that

$$e^{2\pi i\rho_0} = \sigma_0, \qquad e^{2\pi i\rho_1} = \sigma_1. \tag{15.2–4}$$

The numbers ρ_0 and ρ_1 are determined up to an additive constant, being an integer. Since σ_0 and σ_1 are different the difference $\rho_0-\rho_1$ is not an integer.

It is now clear that the functions

$$\varphi_0(z) = z^{-\rho_0}w_0(z), \qquad \varphi_1(z) = z^{-\rho_1}w_1(z)$$

return to their initial values after continuation along C. Otherwise stated: the functions φ_0 and φ_1 are single-valued functions in a neighbourhood of the origin. Thus we have

If the characteristic roots at a singular point $z = 0$ are different then (15.1–9) *possesses two linearly independent solutions*

$$w_0(z) = z^{\rho_0}\varphi_0(z), \qquad w_1(z) = z^{\rho_1}\varphi_1(z) \tag{15.2–5}$$

where φ_0 and φ_1 are single-valued functions in a neighbourhood of $z = 0$.

It should be noticed that the characteristic numbers do not depend on the radius of C. Any closed path homotopic to C in a neighbourhood of $z = 0$ (from which the origin is deleted) gives rise to the same transformation of the fundamental system.

If the characteristic roots are equal the situation is more complicated. Let again C denote a small circumference about $z = 0$ and assume that w_0, w_1 is a canonical system regular at a point z_0 on C. Analytic continuation along C carries w_0 into σw_0 and w_1 into $aw_0+\sigma w_1$. Hence the quotient w_1/w_0 is transformed into $w_1/w_0+a/\sigma$. The function

$$\frac{a/\sigma}{2\pi i}\log z$$

increases by a/σ after continuing along C once. As a consequence the difference

$$\varphi(z) = \frac{w_1(z)}{w_0(z)} - \frac{a/\sigma}{2\pi i}\log z$$

returns to its initial value. As above we find that

$$w_0(z) = z^\rho\varphi_0(z),$$

with $e^{2\pi i\rho} = \sigma$, the function φ_0 being single-valued. Writing $\varphi_1(z) = \varphi(z)\varphi_0(z)$ we have in addition

$$w_1(z) = z^\rho(c\varphi(z)\log z+\varphi_1(z)).$$

The function φ_1 is again single-valued. Thus

If the characteristic roots at $z = 0$ are equal then (15.1–9) *possesses a fundamental system of solutions*

$$w_0(z) = z^\rho\varphi_0(z), \qquad w_1(z) = z^\rho(c\varphi_0(z)\log z+\varphi_1(z)), \tag{15.2–6}$$

where φ_0 and φ_1 are single-valued functions, c being a constant.

The functions φ_0 and φ_1 may be singular at $z = 0$. This is rather harmless if they have only a pole there, for we can increase the exponents ρ_0, ρ_1 or ρ by a suitable integer in order to cancel the pole. In the case of an essential singular point this is not possible. The solutions for which the functions φ have no essential singularity at the point under consideration will be called *semi-regular* and a singular point of the differential equation at which there are only semi-regular solutions will be referred to as a *regular-singular point.*

15.2.2 – Fuchs's criterion for a regular singularity

There is a very simple test which enables us to decide whether or not a singular point of a linear homogeneous differential equation is a regular singular point. This is the content of *Fuch's theorem* which states for an equation of the first order

The necessary and sufficient condition that an isolated singular point *of the equation* (15.1–2) *should be a regular singularity is that it should be at most a simple pole of the coefficient* $p(z)$.

Let $z = 0$ be the singular point. Suppose that the equation (15.1–2) has the solution

$$w(z) = z^{\rho}\varphi(z), \tag{15.2–7}$$

where $\varphi(z)$ is regular at $z = 0$ and $\varphi(0) \neq 0$. Then

$$-p(z) = \frac{w'(z)}{w(z)} = \frac{\rho}{z} + \frac{\varphi'(z)}{\varphi(z)}. \tag{15.2–8}$$

Since φ'/φ is also regular at $z = 0$ we see that $p(z)$ has a pole of at most the first order. This pole is absent, of course, if $\rho = 0$. Conversely, if $p(z)$ is of the form $p/z+\psi(z)$, the function $\psi(z)$ being regular at $z = 0$, we may determine φ from $\varphi'/\varphi = \psi$ and it is clear that (15.2–7) is a solution of the differential equation under consideration.

Now we turn our attention to the case of an equation of the second order. For this case Fuchs's theorem states

The necessary and sufficient conditions that an isolated singularity of the equation (15.1–9) *should be a regular-singular point are that it should be at most a simple pole of the coefficient* $p(z)$ *and at most a double pole of the coefficient* $q(z)$.

We suppose that w_0 and w_1 are semi-regular solutions at the singular point $z = 0$ and that w_0 has the form

$$w_0(z) = z^{\rho_0}\varphi_0(z). \tag{15.2–9}$$

According as the characteristic roots are unequal or equal, w_1 has the form

$$w_1(z) = \begin{cases} z^{\rho_1}\varphi_1(z), \\ z^{\rho}(c\varphi_0(z)\log z+\varphi_1(z)), \end{cases} \tag{15.2–10}$$

the functions φ_0 and φ_1 being regular at $z = 0$. The ratio of these solutions

$$s(z) = \frac{w_1(z)}{w_0(z)} = \begin{cases} z^{\rho_1-\rho_0}\dfrac{\varphi_1(z)}{\varphi_0(z)}, \\ c\log z+\dfrac{\varphi_1(z)}{\varphi_0(z)}, \end{cases} \tag{15.2–11}$$

is again semi-regular and the same can be stated for

$$s'(z) = \frac{ds}{dz} = \begin{cases} z^{\rho_1-\rho_0-1}\left((\rho_1-\rho_0)\dfrac{\varphi_1(z)}{\varphi_0(z)}+z\dfrac{d}{dz}\dfrac{\varphi_1(z)}{\varphi_0(z)}\right), \\ z^{-1}\left(c+z\dfrac{d}{dz}\dfrac{\varphi_1(z)}{\varphi_0(z)}\right). \end{cases} \tag{15.2–12}$$

In view of (15.1–18) we have

$$\frac{ds'}{dz} = -\frac{d}{dz}\left(\frac{w_0'(z)w_1(z)-w_0(z)w_1'(z)}{w_0^2(z)}\right) = -\frac{w_0''(z)w_1(z)-w_0(z)w_1''(z)}{w_0^2(z)} +$$

$$+2w_0'(z)\,\frac{w_0'(z)w_1(z)-w_0(z)w_1'(z)}{w_0^3(z)}$$

$$= \left(p(z)+2\,\frac{w_0'(z)}{w_0(z)}\right)\frac{w_0'(z)w_1(z)-w_0(z)w_1'(z)}{w_0^2(z)}$$

$$= -\left(p(z)+2\,\frac{w_0'(z)}{w_0(z)}\right)s'(z).$$

Hence $s'(z)$ is a solution of the first order equation

$$w'+\left(p(z)+2\,\frac{w_0'(z)}{w_0(z)}\right)w = 0. \qquad (15.2\text{–}13)$$

To this equation we apply Fuchs's theorem for equations of the first order. Since w_0'/w_0 has a pole of at most the first order we deduce that $p(z)$ has a pole of at most the first order at $z = 0$.

Since $w_0(z)$ is a solution of (15.1–9) we have

$$q(z) = -\left(\frac{w_0''(z)}{w_0(z)}+p(z)\,\frac{w_0'(z)}{w_0(z)}\right). \qquad (15.2\text{–}14)$$

Differentiating the coefficient of w in (15.2–13) we obtain

$$p'(z)+2\,\frac{w_0''(z)}{w_0(z)}-2\left(\frac{w_0'(z)}{w_0(z)}\right)^2,$$

a function which has at most a double pole at $z = 0$ and so has w_0''/w_0. It is now clear that $q(z)$ has at most a double pole at the origin. This establishes the necessity of Fuchs's conditions.

In order to prove the sufficiency we proceed as follows. First we assume provisionally the existence of a semi-regular solution $w_0(z)$. Then (15.2–13) satisfies Fuchs's condition and it possesses a semi-regular solution

$$s'(z) = z^\rho(a_0+a_1 z+\ldots),$$

from which follows

$$s(z) = \frac{a_0}{\rho+1}\,z^{\rho+1}+\frac{a_1}{\rho+2}\,z^{\rho+2}+\ldots+c\log z,$$

where the logarithmic term occurs only if ρ is a negative integer; then the corresponding term in the series is missing. As a consequence also $w_1(z) = s(z)w_0(z)$ is semi-regular.

It remains to prove the existence of at least one semi-regular solution. We proceed as in section 15.1.1. Let

$$p(z) = \frac{p_{-1}}{z} + p_0 + p_1 z + \ldots,$$
$$q(z) = \frac{q_{-2}}{z^2} + \frac{q_{-1}}{z} + q_0 + q_1 z + \ldots \tag{15.2–15}$$

be the Laurent expansions of $p(z)$ and $q(z)$ about $z = 0$. First we consider a special case: we assume that $q(z)$ has only a simple pole. i.e., $q_{-2} = 0$ and, moreover, that p_{-1} is not zero or a negative integer. As we shall see, in this case there is even a regular solution. If we suppose the existence of a solution

$$w(z) = c_0 + c_1 z + \ldots, \tag{15.2–16}$$

formal substitution yields

$$\begin{aligned} &z(2\times 1c_2 + 3\times 2c_3 z + \ldots + n(n-1)c_n z^{n-2} + \ldots) + \\ &\quad +(p_{-1} + p_0 z + p_1 z^2 + \ldots)(c_1 + 2c_2 z + 3c_3 z^2 + \ldots + nc_n z^{n-1} + \ldots) \\ &\quad +(q_{-1} + q_0 z + q_1 z^2 + \ldots)(c_0 + c_1 z + c_2 z^2 + \ldots + c_n z^n + \ldots) = 0. \end{aligned}$$

Combining coefficients of equal powers of z we get the recurrence relations

$$\begin{aligned} &p_{-1}c_1 + q_{-1}c_0 &&= 0, \\ &2(1+p_{-1})c_2 + p_0 c_1 + q_{-1}c_1 + q_0 c_0 &&= 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &n(n-1+p_{-1})c_n + (n-1)p_0 c_{n-1} + \ldots + \\ &+p_{n-2}c_1 + q_{-2}c_{n-1} + q_0 c_{n-2} + \ldots + q_{n-2}c_0 &&= 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{15.2–17}$$

The number c_0 can be chosen arbitrarily; then the other coefficients of the series (15.2–16) can be evaluated successively in a unique way.

The convergence of the series (15.2–16) whose coefficients are evaluated from (15.2–17) can be proved by almost the same method as applied in section 15.1.1. We have the same inequalities as in (15.1–12), but the estimate for $|c_n|$ is now

$$n|n-1+p_{-1}||c_n| \leq n(n-1)M\left(\tfrac{1}{2}r + \frac{r^2}{n}\right)M_0 r^{-n} + |q_{-1}|rMr^{-n}.$$

Then (15.1–8) follows if we take r sufficiently small.

The general case (i.e., the case without the restrictive conditions for p_{-1} and q_{-2}) may be reduced to this special one. We contend that there is always a solution of the type

$$w(z) = z^\rho \varphi(z), \tag{15.2–18}$$

where $\varphi(z)$ is regular at the origin and ρ is chosen appropiately. In fact, substituting (15.2–18) into (15.1–9) we find that $\varphi(z)$ must satisfy the differential equation

$$\varphi''+\left(p(z)+\frac{2\rho}{z}\right)\varphi'+\left(q(z)+\frac{\rho}{z}p(z)+\frac{\rho(\rho-1)}{z^2}\right)\varphi=0. \qquad (15.2\text{–}19)$$

The coefficient of φ has only a pole of the first order if ρ satisfies the *indicial equation*

$$\boxed{\rho(\rho-1)+p_{-1}\rho+q_{-2}=0.} \qquad (15.2\text{–}20)$$

This same equation is obtained if we substitute the series

$$c_0z^\rho+c_1z^{\rho+1}+c_2z^{\rho+2}+\ldots \qquad (15.2\text{–}21)$$

into (15.1–9) and collect coefficients of equal powers of z.

We shall denote the roots of the equation (15.2–20) by ρ_0 and ρ_1 and we shall assume that $\rho_0\geqq\rho_1$ if ρ_0 and ρ_1 are real. It follows from

$$\rho_0-\rho_1=2\rho_0-(\rho_0+\rho_1)=2\rho_0+p_{-1}-1 \qquad (15.2\text{–}22)$$

that $2\rho_0+p_{-1}$ cannot be zero or a negative integer. Hence our preliminary considerations allow us to conclude that to ρ_0 corresponds a regular solution φ_0 and thus a solution

$$w_0(z)=z^{\rho_0}\varphi_0(z) \qquad (15.2\text{–}23)$$

is asserted. If the difference $\rho_0-\rho_1$ is not an integer then $2\rho_1+\rho_0$ is not zero, nor it is a negative integer and we find a second solution

$$w_1(z)=z^{\rho_1}\varphi_1(z), \qquad (15.2\text{–}24)$$

such that w_0 and w_1 are linearly independent.

If the roots differ by an integer an independent solution with respect to w_0 can be found as follows. We employ the function $s=w_1/w_0$, where w_0 is the solution (15.2–23) and w_1 has to be determined from s. Its derivative satisfies the differential equation (15.2–13), that is

$$w'+\left(\frac{2\rho_0+p_{-1}}{z}+r(z)\right)w=0 \qquad (15.2\text{–}25)$$

where $r(z)$ is regular at $z=0$. By assumption $\rho_0-\rho_1=m$ is a non-negative integer and from (15.2–22) follows that $2\rho_0+p_{-1}=m+1$. Hence

$$\frac{w'(z)}{w(z)}=-\frac{m+1}{z}-r(z)$$

and

$$s'(z) = z^{-(m+1)}(b_0+b_1z+\ldots),$$

whence

$$s(z) = z^{-m}\varphi(z)+c\log z,$$

φ being regular at $z = 0$. The constant c is not necessarily different from zero. Finally we have

$$w_1(z) = s(z)w_0(z) = z^{\rho_0-m}\varphi_0(z)\varphi(z)+cz^{\rho_0}\varphi_0(z)\log z$$
$$= cz^{\rho_0}\varphi_0(z)\log z+z^{\rho_1}\varphi_1(z),$$

in accordance with (15.2–6). Summing up we have

If $z = 0$ is a regular singular point of the differential equation (15.1–9) *and if ρ_0 and ρ_1 are the roots of the indicial equation* (15.2–20), *there exist linearly independent solutions*

$$w_0(z) = z^{\rho_0}\varphi_0(z), \qquad w_1(z) = z^{\rho_1}\varphi_1(z), \tag{15.2–26}$$

where φ_0 and φ_1 are regular at $z = 0$, corresponding to the roots of the indicial equation, provided that the difference of the roots is not an integer. If, however, the difference $\rho_0-\rho_1$ is a non-negative integer. there exist linearly independent solutions

$$w_0(z) = z^{\rho_0}\varphi_0(z), \qquad w_1(z) = z^{\rho_0}c\,\varphi_0(z)\log z+z^{\rho_1}\varphi_1(z). \tag{15.2–27}$$

For the sake of illustration we consider again the differential equation (14.5–3). Now $p_{-1}=1$, $q_{-2}=0$ and the indicial equation is $\rho^2=0$. Hence there is a regular solution at $z = 0$. By taking the constant appropriately we get the function $\mathsf{K}(z)$. The other solution $\mathsf{K}'(z)$ is related to $\mathsf{K}(z)$ as may be seen from (14.5–14). This is in accordance with the general theory.

15.3 – Bessel functions

15.3.1 – The differential equation of Bessel

A famous example of a differential equation with a regular singular point is *Bessel's equation*

$$\boxed{z^2w''+zw'+(z^2-\kappa^2)w = 0,} \tag{15.3–1}$$

where κ is an arbitrary constant. The origin $z=0$ is a regular singularity. Replacing z by $1/z$ the equation becomes

$$z^2w''+zw'+\left(\frac{1}{z^2}-\kappa^2\right)w = 0 \tag{15.3–2}$$

In this case Fuchs's condition is not satisfied at $z = 0$. Otherwise stated: *The point $z = \infty$ is not a regular singular point of Bessel's equation* (15.3–1). Else the equation has no singular points.

The indicial equation of (15.3–1) is

$$\rho(\rho-1)+\rho-\kappa = \rho^2-\kappa^2 = 0 \tag{15.3–3}$$

whose roots are $\rho = \pm\kappa$; we may assume that Re $\kappa \geqq 0$.

In accordance with the general theory there must be a solution at $z=0$

$$z^\kappa \varphi(z) \tag{15.3–4}$$

where φ is regular at $z = 0$. Inserting the series

$$\sum_{\nu=0}^{\infty} c_\nu z^{\kappa+\nu} \tag{15.3–5}$$

into (15.3–1) we find that the coefficients must satisfy the recursive relations

$$\begin{aligned} &(2\kappa+1)c_1 = 0, \\ &n(2\kappa+n)c_n + c_{n-2} = 0, \quad n = 2, 3, \ldots, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{15.3–6}$$

The fact that the equation (15.3–1) remains the same by changing z into $-z$ suggests that $c_{2n+1} = 0$, $n = 0, 1, 2, \ldots$. The relations (15.3–6) are satisfied if we take the remaining coefficients such that

$$\begin{aligned} c_{2n} &= -\frac{c_{2n-2}}{4n(\kappa+n)} = \ldots = (-1)^n \frac{c_0}{2^{2n} n!(\kappa+1)\ldots(\kappa+n)} \\ &= (-1)^n c_0 \frac{\Gamma(\kappa+1)}{2^{2n} n!\Gamma(\kappa+n+1)}, \qquad n = 0, 1, 2, \ldots, \end{aligned}$$

provided that κ is not a negative integer. It is common use to take

$$c_0 = \frac{1}{2^\kappa \Gamma(\kappa+1)},$$

2^κ being exp (κ log 2). Thus we obtain the function

$$\boxed{J_\kappa(z) = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(\kappa+\nu+1)} \left(\frac{z}{2}\right)^{\kappa+2\nu},} \tag{15.3–7}$$

a solution of Bessel's differential equation. It is easy to verify that $z^{-\kappa}J_\kappa(z)$ is an integral function. Hence term-by-term differentiation is permissible and the formal process for finding a solution is justified.

The function (15.3–7) is called *a Bessel function of the first kind of order* κ. It is assumed that z^κ represents its principal value in the region $z+|z| \neq 0$. Recapitulating:

If κ is not an integer the functions $J_\kappa(z)$ and $J_{-\kappa}(z)$ constitute a fundamental system of solutions of Bessel's differential equation.

The expression (15.3–7) makes sense if κ is a negative integer $-k$. Then the first k terms in the expansion of $J_{-k}(z)$ vanish and we have

$$J_{-k}(z) = \sum_{\nu=k}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(-k+\nu+1)} \left(\frac{z}{2}\right)^{-k+2\nu}$$

$$= \sum_{\nu=0}^{\infty} \frac{(-1)^{k+\nu}}{(k+\nu)!\nu!} \left(\frac{z}{2}\right)^{k+2\nu} = (-1)^k \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(\kappa+\nu+1)} \left(\frac{z}{2}\right)^{k+2\nu},$$

whence

$$\boxed{J_{-k}(z) = (-1)^k J_k(z),} \tag{15.3–8}$$

Thus it appears that $J_\kappa(z)$ and $J_{-\kappa}(z)$ are linearly dependent if κ is an integer.

15.3.2 – The recurrence formulae

There exist simple relations between Bessel functions whose orders differ by an integral number.

Differentiating $z^\kappa J_\kappa(z)$ we get

$$\kappa z^{\kappa-1} J_\kappa(z) + z^\kappa \frac{dJ_\kappa}{dz} = \frac{d}{dz} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(\kappa+\nu+1)} \frac{z^{2\kappa+2\nu}}{2^{\kappa+2\nu}}$$

$$= \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{z^\kappa}{\Gamma(\kappa+\nu)} \left(\frac{z}{2}\right)^{\kappa+2\nu-1} = z^\kappa J_{\kappa-1}(z),$$

or

$$\frac{dJ_\kappa}{dz} = -\frac{\kappa}{z} J_\kappa(z) + J_{\kappa-1}(z). \tag{15.3–9}$$

Differentiating $z^{-\kappa} J_\kappa(z)$ we get

$$-\kappa z^{-\kappa-1} J_\kappa(z) + z^{-\kappa} \frac{dJ_\kappa}{dz} = \frac{d}{dz} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(\kappa+\nu+1)} \frac{z^{2\nu}}{2^{\kappa+2\nu}}$$

$$= \sum_{\nu=1}^{\infty} \frac{(-1)^\nu}{(\nu-1)!} \frac{z^{-\kappa}}{\Gamma(\kappa+\nu+1)} \left(\frac{z}{2}\right)^{\kappa-1+2\nu}$$

$$= -\sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{z^{-\kappa}}{\Gamma(\kappa+1+\nu+1)} \left(\frac{z}{2}\right)^{\kappa+1+2\nu} = -z^{-\kappa} J_{\kappa+1}(z),$$

or

$$\frac{dJ_\kappa}{dz} = \frac{\kappa}{z} J_\kappa(z) - J_{\kappa+1}(z). \tag{15.3–10}$$

Adding corresponding members of (15.3–9) and (15.3–10) we obtain

$$\boxed{2\frac{dJ_\kappa}{dz} = J_{\kappa-1}(z)-J_{\kappa+1}(z)} \tag{15.3–11}$$

and subtracting corresponding members

$$\boxed{J_{\kappa+1}(z)+J_{\kappa-1}(z) = \frac{2\kappa}{z}J_\kappa(z).} \tag{15.3–12}$$

The relation (15.3–12) may serve to prove that *the Bessel functions of the first kind whose order is half an odd integer are elementary functions.*

It is sufficient to verify this for $\kappa = -\frac{1}{2}$ and $\kappa = \frac{1}{2}$. According to (15.3–7) we have the expansion

$$J_{-\frac{1}{2}}(z) = \left(\frac{z}{2}\right)^{-\frac{1}{2}} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!\Gamma(\frac{1}{2}+\nu)} \left(\frac{z}{2}\right)^{2\nu}.$$

From

$$\Gamma(n+\tfrac{1}{2}) = \frac{(2n)!}{2^{2n}n!}\Gamma(\tfrac{1}{2})$$

and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ it follows that

$$J_{-\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi z}} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu z^{2\nu}}{(2\nu)!},$$

or

$$\boxed{J_{-\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi z}}\cos z.} \tag{15.3–13}$$

If $\kappa = \frac{1}{2}$ we have the expansion

$$J_{\frac{1}{2}}(z) = \left(\frac{z}{2}\right)^{\frac{1}{2}} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(\frac{3}{2}+\nu)} \left(\frac{z}{2}\right)^{2\nu}.$$

From

$$\Gamma(n+\tfrac{3}{2}) = (n+\tfrac{1}{2})\Gamma(n+\tfrac{1}{2})$$

and the above results we find that

$$J_{\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi z}} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu z^{2\nu+1}}{(2\nu+1)!}$$

or

$$\boxed{J_{\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi z}}\sin z.} \tag{15.3–14}$$

A simple consequence of these formulas is that *Fresnel's integrals* (8.13–31) *may be expressed in terms of Bessel functions.* Performing the substitution $t = \frac{1}{2}\pi u^2$ we find

$$C(z) = \int_0^z \cos \tfrac{1}{2}\pi u^2\,du = \frac{1}{2}\int_0^{\frac{1}{2}\pi z^2} \sqrt{\frac{2}{\pi t}}\cos t\,dt,$$

or

$$C(z) = \frac{1}{2}\int_0^{\frac{1}{2}\pi z^2} J_{-\frac{1}{2}}(t)\,dt \tag{15.3–15}$$

and similarly

$$S(z) = \frac{1}{2}\int_0^{\frac{1}{2}\pi z^2} J_{\frac{1}{2}}(t)\,dt. \tag{15.3–16}$$

15.3.3 – Schläfli's integral

The representation (4.7–35) of the reciprocal of the gamma function is the key to obtain a representation of Bessel functions of the first kind as a contour integral. The expansion (15.3–7) may be written as

$$J_\kappa(z) = \frac{1}{2\pi i}\sum_{\nu=0}^{\infty}\frac{(-1)^\nu}{\nu!}\left(\frac{z}{2}\right)^{\kappa+2\nu}\int_{L(a)} e^t t^{-\kappa-\nu-1}\,dt.$$

Inverting summation and integration, and observing that

$$\exp\left(-\frac{z^2}{4t}\right) = \sum_{\nu=0}^{\infty}\frac{(-1)^\nu}{\nu!}\left(\frac{z}{2}\right)^{2\nu}\frac{1}{t^\nu}$$

we readily find

$$\boxed{J_\kappa(z) = \frac{1}{2\pi i}\left(\frac{z}{2}\right)^\kappa \int_{L(a)} t^{-(\kappa+1)}\exp\left(t - \frac{z^2}{4t}\right)dt.} \tag{15.3–17}$$

If we suppose that $a \geqq 1$, we have on the rectilinear part of the Hankel contour $|t| \geqq 1$ and hence for $|z| \leqq h$

$$\left|\exp\left(\frac{-z^2}{4t}\right)\right| \leqq \exp\left|\frac{z^2}{4t}\right| \leqq e^{h^2}.$$

As a consequence the integrand in (15.3–17) is uniformly bounded on a closed disc about the origin and by arguments similar to those employed in section 2.20.5 it follows that the integral represents a holomorphic function of z which may be differentiated within the sign of integration. Expanding it in a Taylor series we obtain the series for $J_\kappa(z)$.

The integral (15.3–17) is due to Schläfli. It will be convenient to modify it slightly. Let $t = \frac{1}{2}uz$ and assume $z > 0$. Then (15.3–17) transforms into

$$\boxed{J_\kappa(z) = \frac{1}{2\pi i}\int_{L(a)} u^{-(\kappa+1)} \exp \tfrac{1}{2}z\left(u - \frac{1}{u}\right) du.} \tag{15.3–18}$$

The same Hankel contour may be taken, for the transformation yields a contour of the same form as the original contour.

The integral on the right is a holomorphic function of z as long as $\mathrm{Re}(zu) < 0$ on the rectilinear part of the contour, i.e., if $\mathrm{Re}\, z > 0$. Hence (15.3–18) is valid for $\mathrm{Re}\, z > 0$ by the identity principle.

Let us now take $a = 1$. Then $L(a)$ is a contour consisting of half rays from $-\infty$ to -1 and a circumference C of radius unity about the origin. On this circumference we have $u = e^{i\theta}$, $-\pi \leqq \theta \leqq \pi$ and on the straight tails $u = ve^{\mp\pi i}$, $v > 0$. Inserting the variable into (15.3–18) we get

$$J_\kappa(z) = \frac{1}{2\pi i}\left(e^{(\kappa+1)\pi i} - e^{-(\kappa+1)\pi i}\right)\int_1^\infty v^{-(\kappa+1)} \exp \tfrac{1}{2}z\left(-v + \frac{1}{v}\right) dv +$$
$$+ \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp(-\kappa i\theta + iz\sin\theta)d\theta. \tag{15.3–19}$$

In the last integral we bisect the range of integration and write $-\theta$ instead of θ in the lower part. This integral becomes

$$\frac{1}{2\pi}\int_0^\pi \exp(\kappa i\theta - iz\sin\theta)d\theta + \frac{1}{2\pi}\int_0^\pi \exp(-\kappa i\theta + iz\sin\theta)d\theta.$$

In the first integral of (15.3–19) we write $v = e^\theta$. It follows that

$$J_\kappa(z) = \frac{\sin(\kappa+1)\pi}{\pi}\int_0^\infty \exp(-\kappa\theta - z\sinh\theta)d\theta +$$
$$+ \frac{1}{\pi}\int_0^\pi \cos(\kappa\theta - z\sin\theta)d\theta. \tag{15.3–20}$$

This expression simplifies considerably if $\kappa = k$ is an integer. We then have

$$\boxed{J_k(z) = \frac{1}{\pi}\int_0^\pi \cos(k\theta - z\sin\theta)d\theta.} \tag{15.3–21}$$

At the end of section 3.14.4 we encountered the particular case

$$J_0(z) = \frac{1}{\pi}\int_0^\pi \cos(z \sin\theta)d\theta. \tag{15.3–22}$$

15.3.4 – Bessel's solution of Kepler's equation

Bessel discovered his functions by attempting to solve Kepler's equation (3.13–23), viz.,

$$u - \varepsilon \sin u \equiv t, \qquad 0 < \varepsilon < 1. \tag{15.3–23}$$

This equation defines t as a strictly increasing function of u, since

$$\frac{dt}{du} = 1 - \varepsilon \cos u > 0.$$

Hence u can be considered as a strictly increasing function $u(t)$ of t.

It is clear that

$$u(t_0) - \varepsilon \sin u(t_0) = t_0$$

and

$$u(t_0+2\pi) - \varepsilon \sin u(t_0+2\pi) = t_0+2\pi.$$

But also

$$u(t_0)+2\pi - \varepsilon \sin(u(t_0)+2\pi) = t_0+2\pi.$$

Since the solution of the equation (15.3–23) is uniquely determined it follows that

$$u(t_0+2\pi) = u(t_0)+2\pi$$

and, consequently, $\sin u(t)$ is a periodic function of t, the period being 2π. By similar arguments it is readily seen that $\sin u(t)$ is an odd function of t. Since $u(\pi) = \pi$, we have $\sin u(\pi) = 0$.

The theory of Fourier series asserts that $\varepsilon \sin u(t)$ can be expanded in a uniformly convergent sine series. Thus we have

$$\varepsilon \sin u(t) = \sum_{\nu=1}^{\infty} A_\nu \sin \nu t, \tag{15.3–24}$$

with

$$A_n = \frac{2}{\pi}\int_0^\pi \varepsilon \sin u(t) \sin nt\, dt, \qquad n = 1, 2, \ldots.$$

Integration by parts yields

$$A_n = -\left.\frac{2\varepsilon \sin u(t) \cos nt}{n\pi}\right|_0^\pi + \frac{2}{n\pi}\int_0^\pi \cos nt \frac{d(\varepsilon \sin u(t))}{dt}\, dt,$$

whence, in view of (15.3–23)

$$A_n = \frac{2}{n\pi}\int_0^\pi \left(\frac{du}{dt}-1\right)\cos nt\,dt = \frac{2}{n\pi}\int_0^\pi \frac{du}{dt}\cos nt\,dt - \frac{2}{n\pi}\int_0^\pi \cos nt\,dt.$$

The last integral vanishes. In the preceding integral we replace t by $u-\varepsilon \sin u$ and we find in view of (15.3–21)

$$A_n = \frac{2}{n\pi}\int_0^\pi \cos n(u-\varepsilon\sin u)du = \frac{2}{n}J_n(n\varepsilon)$$

and the equation (15.3–23) is solved by the series

$$u(t) = t+\sum_{\nu=1}^{\infty}\frac{2}{\nu}J_\nu(\nu\varepsilon)\sin \nu t.$$

15.3.5 – Bessel's functions of integral order

If $\kappa = k$ is an integer, then the Hankel contour in (15.3–18) may be replaced by a circumference about the origin, for $u^{-(k+1)}$ returns to its initial value if we percorse this contour once. In this case

$$J_k(z) = \frac{1}{2\pi i}\int_C u^{-(k+1)}\exp \tfrac{1}{2}z\left(u-\frac{1}{u}\right)du. \tag{15.3–25}$$

This formula may be the starting point for obtaining elementary properties of the Bessel functions of integral order. Thus e.g., the formula (15.3–8) is readily obtained if we replace u by $-1/u$.

If we expand $\exp \frac{1}{2}z(u-1/u)$, considered as a function of u in a Laurent series near the origin, we deduce at once from (2.23–8)

$$\boxed{\exp \tfrac{1}{2}z\left(u-\frac{1}{u}\right) = \sum_{\nu=-\infty}^{\infty} u^\nu J_\nu(z).} \tag{15.3–26}$$

The function on the left is *a generating function* for the Bessel coefficients on the right.

Performing the substitution $u = e^{i\theta}$ yields

$$\exp(iz\sin\theta) = \sum_{\nu=-\infty}^{\infty} J_\nu(z)e^{i\nu\theta}.$$

Adding and subtracting corresponding members of this equation and the equation obtained by replacing θ by $-\theta$ we get the Fourier series

$$\begin{aligned}\cos(z\sin\theta) &= J_0(z)+2\sum_{\nu=1}^{\infty} J_{2\nu}(z)\cos 2\nu\theta,\\ \sin(z\sin\theta) &= 2\sum_{\nu=1}^{\infty} J_{2\nu+1}(z)\sin(2\nu+1)\theta,\end{aligned} \tag{15.3–27}$$

due to Jacobi. In particular, if $\theta = \frac{1}{2}\pi$,

$$\begin{aligned}\cos z &= J_0(z)-2J_2(z)+2J_4(z)+\ldots,\\ \sin z &= 2J_1(z)-2J_3(z)+\ldots.\end{aligned} \tag{15.3–28}$$

If $\theta = 0$ we have the identity

$$1 = J_0(z)+2\sum_{\nu=1}^{\infty} J_{2\nu}(z). \tag{15.3–29}$$

Another interesting application of (15.3–18) is the finding of an expansion of a polynomial in a series involving Bessel functions. We start with the following simple remark. According to (3.3–4) the residue of

$$t^{-(m+1)}e^{zt},$$

m being a positive integer, at the point $t = 0$ in the t-plane is

$$\frac{1}{m!}\,\frac{d^m}{dt^m}\,e^{zt}\Big|_{t=0} = \frac{z^m}{m!}.$$

In view of the definition of residue we then have

$$z^m = \frac{m!}{2\pi i}\int_C t^{-(m+1)}e^{zt}\,dt, \tag{15.3–30}$$

where C is any contour encircling the origin in the t-plane once in the anti-clockwise sense. The transformation

$$t = \frac{1}{2}\left(u-\frac{1}{u}\right)$$

carries a small circle C about the origin in the u-plane into an ellipse in the t-plane percorsed in opposite sense. If $-C'$ is the image of C, where C is percorsed in the anti-clockwise sense, then it follows from (15.3–30)

$$\begin{aligned} z^m &= \frac{(m-1)!}{2\pi i}\int_{-C'} e^{zt}d(t^{-m})\\ &= \frac{2^m(m-1)!}{2\pi i}\int_C \exp\tfrac{1}{2}z\left(u-\frac{1}{u}\right)\frac{d}{du}\left(u-\frac{1}{u}\right)^{-m}du.\end{aligned}$$

Expanding by means of the binomial theorem (2.16–20) we have

$$\begin{aligned}\left(u-\frac{1}{u}\right)^{-m} &= (-1)^m u^m(1-u^2)^{-m} = (-1)^m u^m\sum_{\nu=0}^{\infty}(-1)^\nu\binom{-m}{\nu}u^{2\nu}\\ &= (-1)^m u^m\sum_{\nu=0}^{\infty}\frac{(m+\nu-1)!}{(m-1)!\,\nu!}\,u^{2\nu}.\end{aligned}$$

Within the unit circle term-by-term differentiation of the binomial series is permissible and if C is within the unit circle we may also integrate term-by-term. We find

$$\left(\frac{z}{2}\right)^m = \frac{(-1)^m}{2\pi i}\sum_{\nu=0}^{\infty}\frac{(2\nu+m)(m+\nu-1)!}{\nu!}\int_C u^{2\nu+m-1}\exp\tfrac{1}{2}z\left(u-\frac{1}{u}\right)du.$$

If κ is an integer the Hankel contour in (15.3–18) may be replaced by a circle about the origin and the above result appears in the form

$$(-1)^m\sum_{\nu=0}^{\infty}\frac{(2\nu+m)(m+\nu-1)!}{\nu!}J_{-(2\nu+m)}(z).$$

Finally, taking account of (15.3–8)

$$\left(\frac{z}{2}\right)^m = \sum_{\nu=0}^{\infty}\frac{(2\nu+m)(m+\nu-1)!}{\nu!}J_{2\nu+m}(z). \qquad (15.3\text{–}31)$$

This formula has been derived under the assumption $m > 0$. It does not hold for $m = 0$ as is clear from (15.3–29). We notice the particular case

$$\frac{z}{2} = J_1(z)+3J_3(z)+5J_5(z)+\ \ldots. \qquad (15.3\text{–}32)$$

15.3.6 – Sommerfeld's integral

Bessel's function $J_\kappa(z)$ can be written as

$$J_\kappa(z) = z^\kappa\varphi(z), \qquad (15.3\text{–}33)$$

where φ is regular at $z = 0$. Starting from (15.3–1) a simple computation shows that $\varphi(z)$ satisfies the differential equation

$$z^2\varphi''+(2\kappa+1)z\varphi'+z^2\varphi = 0. \qquad (15.3\text{–}34)$$

A method of frequent use in the theory of linear differential equations is the application of Laplace's integral

$$\int_C e^{-st}f(t)dt, \qquad (15.3\text{–}35)$$

where s is a complex variable and C a suitably chosen path. We shall apply this method to the equation (15.3–34). It is convenient to introduce the variable $s = -z^2$. Then (15.3–34) takes the form

$$s\frac{d^2\varphi}{ds^2}+(\kappa+1)\frac{d\varphi}{ds}-\tfrac{1}{4}\varphi = 0. \qquad (15.3\text{–}36)$$

Replacing φ by (15.3–35) yields

$$\int_C(st^2-(\kappa+1)t-\tfrac{1}{4})e^{-st}f(t)dt = 0.$$

This is trivially satisfied if $f(t) = 0$ identically. Another solution of (15.3–36) can be obtained by performing integration by parts on

$$\int_C e^{-st} st^2 f(t)dt = -\int_C t^2 f(t) de^{-st}$$
$$= -t^2 f(t)e^{-st}\Big|_C + \int_C e^{-st}(t^2 f'(t) + 2tf(t))dt,$$

where the first term in the last member of the equation denotes the difference of the values of the function before the bar at the endpoints of C. The equation (15.3–36) is satisfied if

$$t^2 f'(t) - (\kappa - 1)tf(t) - \tfrac{1}{4}f(t) = 0 \qquad (15.3\text{–}37)$$

together with

$$t^2 f(t)e^{-st}\Big|_C = 0. \qquad (15.3\text{–}38)$$

It is easy to verify that

$$f(t) = t^{\kappa-1}e^{-1/4t}$$

is a solution of (15.3–38). Hence

$$\varphi(z) = \int_C t^{\kappa-1} \exp\left(z^2 t - \frac{1}{4t}\right) dt \qquad (15.3\text{–}39)$$

is a solution of (15.3–34), provided that

$$t^{\kappa+1} \exp\left(z^2 t - \frac{1}{4t}\right)\Big|_C = 0. \qquad (15.3\text{–}40)$$

Now we perform the transformation $t = -1/4t'$ and omit afterwards the prime. The image of C shall again be denoted by C. Omitting a multiplicative constant we find

$$\varphi(z) = \int_C t^{-(\kappa+1)} \exp\left(t - \frac{z^2}{4t}\right) dt,$$

with the additional condition

$$t^{-(\kappa+1)} \exp\left(t - \frac{z^2}{4t}\right)\Big|_C = 0. \qquad (15.3\text{–}41)$$

Adjusting constants we may say that

$$\frac{1}{2\pi i}\left(\frac{z}{2}\right)^{\kappa} \int_C t^{-(\kappa+1)} \exp\left(t - \frac{z^2}{4t}\right) dt \qquad (15.3\text{–}42)$$

satisfies Bessel's differential equation, provided that (15.3–41) holds. By taking C suitably we can find a variety of solutions represented by an integral. If we identify C with a Hankel contour $L(a)$ then (15.3–41) is true and we obtain again the integral (15.3–17) for $J_\kappa(z)$.

Applying the substitution $t = \frac{1}{2}z \exp -iu$ the integral (15.3–42) transforms into *Sommerfeld's integral*

$$\boxed{\frac{-1}{2\pi}\int_C \exp i(\kappa u - z \sin u)du} \tag{15.3–43}$$

which represents a solution of Bessel's differential equation, provided that

$$\exp i(\kappa u - z \sin u)|_C = 0. \tag{15.3–44}$$

Here again C is the image of the path occurring in (15.3–42). In particular, the Hankel contour $L(1)$ transforms into a figure L consisting of two half-rays from $\pi + i\infty$ to π and from $-\pi$ to $-\pi + i\infty$, connected by a segment from π to $-\pi$, (fig. 15.3–1). This is Sommerfeld's representation of $J_\kappa(z)$.

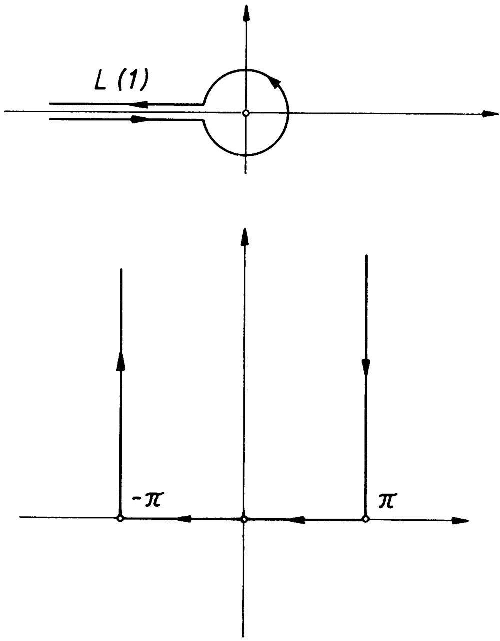


Fig. 15.3–1. Hankel's contour and Sommerfeld's contour

15.3.7 – The Hankel functions

As already remarked we can find other solutions of Bessel's differential equation by modifying the path C. Consider the function

$$F(z) = \int_{\alpha-i\infty}^{\beta+i\infty} \exp i(\kappa u - z \sin u)du, \tag{15.3–45}$$

the path of integration being a half ray from $\alpha - i\infty$ to α, a segment from α to β and finally a half ray from β to $\beta + i\infty$, α and β being real and the half rays being vertical, (fig. 15.3–2).

Put $z = re^{i\theta}$. Consider first the lower part of the path of integration and let $t = \alpha - i\eta$, $\eta > 0$. An easy computation yields

$$\begin{aligned}\text{Re}\,(-iz \sin t) &= r \sin \alpha \sin \theta \cosh \eta - r \cos \alpha \cos \theta \sinh \eta \\ &= \tfrac{1}{2}re^{-\eta} \cos (\theta - \alpha) - \tfrac{1}{2}re^{\eta} \cos (\theta + \alpha) \leqq \tfrac{1}{2}re^{-\eta} - \tfrac{1}{2}re^{\eta} \cos (\theta + \alpha).\end{aligned}$$

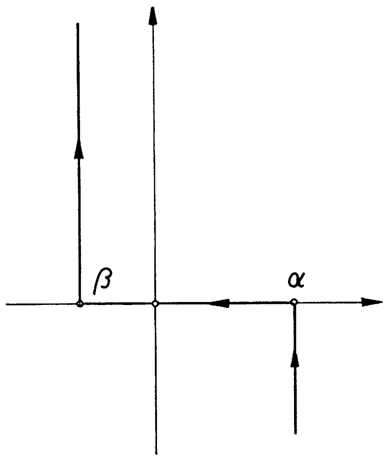


Fig. 15.3–2. The contour C of (15.3–45)

This tends to $-\infty$ as $\eta \to \infty$, provided that $\cos (\theta + \alpha) > 0$. This is the case if

$$-\tfrac{1}{2}\pi + \varepsilon < \theta + \alpha < \tfrac{1}{2}\pi - \varepsilon, \qquad 0 < \varepsilon < \tfrac{1}{2}\pi, \tag{15.3–46}$$

for then $\cos (\theta + \alpha) \geqq \sin \varepsilon$.

Since

$$-\tfrac{1}{2}\pi - \alpha + \varepsilon \leqq \theta \leqq \tfrac{1}{2}\pi - \alpha - \varepsilon$$

the point z must be restricted to the closed domain

$$-\tfrac{1}{2}\pi - \alpha + \varepsilon \leqq \arg z \leqq \tfrac{1}{2}\pi - \alpha - \varepsilon \tag{15.3–47}$$

which is in a half plane bounded by the line $\arg t = \frac{1}{2}\pi - \alpha$, (fig. 15.3–3).

In a similar way we can handle the upper part of the path of integration.

Writing now $t = \beta + i\eta$, $\eta > 0$, we have

$$\begin{aligned} \operatorname{Re}(-iz \sin t) &= \tfrac{1}{2}re^{\eta}\cos(\theta-\beta) - \tfrac{1}{2}re^{-\eta}\cos(\theta+\beta) \\ &\leq \tfrac{1}{2}re^{\eta}\cos(\theta-\beta) + \tfrac{1}{2}re^{-\eta} \end{aligned}$$

and this tends to $-\infty$ as $\eta \to \infty$, provided that $\cos(\theta-\beta) < 0$. This is the case if

$$+\tfrac{1}{2}\pi + \varepsilon \leqq \theta - \beta \leqq \tfrac{3}{2}\pi - \varepsilon$$

or

$$+\tfrac{1}{2}\pi + \varepsilon + \beta \leqq \theta \leqq \tfrac{3}{2}\pi - \varepsilon + \beta. \qquad (15.3\text{–}48)$$

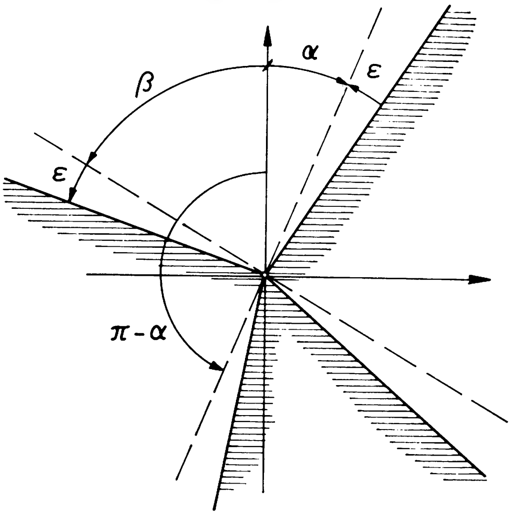


Fig. 15.3–3. The domains (15.3–47) and (15.3–48)

This range coincides with (15.3–47) if we take $\beta = -\pi-\alpha$ or $\beta = \pi-\alpha$. Because e^{η} tends more quickly to ∞ as η it is now clear that the integrals

$$\int_{\alpha-i\infty}^{\mp\pi-\alpha+i\infty} \exp i(\kappa u - z \sin u)du \qquad (15.3\text{–}49)$$

are uniformly convergent in any bounded and closed set of the z-plane included in the domain (15.3–47). In addition we find that the condition (15.3–44) is also satisfied.

The following remark is now of utmost importance. Sommerfeld's path can be considered as the sum of two paths L_1 and L_2, (fig. 15.3–4), whereby L_1 consists of the imaginary axis from $-i\infty$ to 0, the segment from 0 to $-\pi$ and the half ray from $-\pi$ to $-\pi+i\infty$. It is the path considered above with $\alpha = 0$, $\beta = -\pi$. Similarly L_2 consists of the half ray

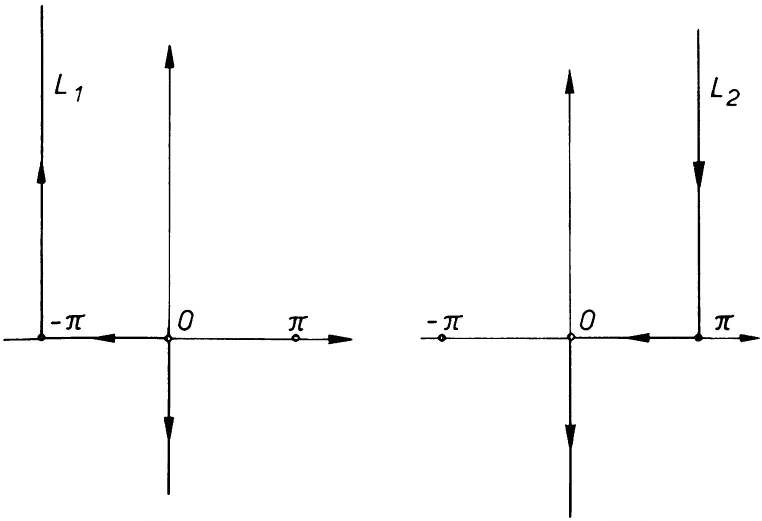


Fig. 15.3–4. The paths of integration for the Hankel functions

from $\pi+i\infty$ to π, the segment from π to 0 and the negative imaginary axis from 0 to $-i\infty$. It is the above path with $\alpha = 0$, $\beta = \pi$ percorsed in the opposite sense.

The functions

$$\boxed{\begin{aligned} H_\kappa^1(z) &= -\frac{1}{\pi}\int_{L_1} \exp i(\kappa u - z \sin u)du, \\ H_\kappa^2(z) &= -\frac{1}{\pi}\int_{L_2} \exp i(\kappa u - z \sin u)du, \end{aligned}} \tag{15.3–50}$$

are solutions of Bessel's differential equation. They are referred to as *Hankel's functions* and from the above considerations follows that they are holomorphic throughout the right half plane. It is clear that they are also holomorphic functions of the parameter κ.

In view of the fact that Sommerfeld's contour is $L = L_1+L_2$ we conclude that

$$\boxed{J_\kappa(z) = \tfrac{1}{2}(H_\kappa^1(z)+H_\kappa^2(z)).} \tag{15.3–51}$$

The integral (15.3–49) provides an analytic continuation of Hankel's functions throughout the z-plane. We take $\alpha = -\frac{1}{2}\pi$. The integral

$$\frac{1}{\pi}\int_{-\frac{1}{2}\pi-i\infty}^{-\frac{1}{2}\pi+i\infty} \exp i(\kappa u - z \sin u)du$$

represents a solution of Bessel's differential equation throughout the region $0 < \arg z < \pi$ i.e., the upper half plane. The path of the last integral is obtained from L_1 by shifting it to the left over a distance $\frac{1}{2}\pi$.

Let now z denote a point in the first quadrant. Integrating along the perimeter of the rectangle with vertices $-\pi$, $-\pi+i\eta$, $-\frac{3}{2}\pi+i\eta$, $-\frac{3}{2}\pi$, $\eta > 0$, (fig. 15.3–5) we obtain the value zero. Since the integrand tends to zero as $\eta \to \infty$ the integral taken along the segment from $-\pi+i\eta$ to $-\frac{3}{2}\pi+i\eta$ tends to zero as $\eta \to \infty$ and we may infer that

$$\int_{-\pi}^{-\pi+i\infty} = \int_{-\pi}^{-\frac{3}{2}\pi} + \int_{-\frac{3}{2}\pi}^{-\frac{3}{2}\pi+i\infty} .$$

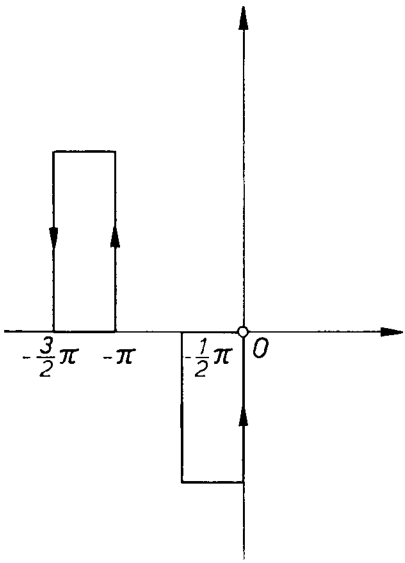


Fig. 15.3–5. Analytic continuation of the Hankel functions

By a similar reasoning we find

$$\int_{-i\infty}^{0} + \int_{0}^{-\frac{1}{2}\pi} = \int_{-\frac{1}{2}\pi-i\infty}^{-\frac{1}{2}\pi} .$$

As a consequence

$$\int_{-i\infty}^{0} + \int_{0}^{-\pi} + \int_{-\pi}^{-\pi+i\infty} = \int_{-i\infty}^{0} + \int_{0}^{-\frac{1}{2}\pi} + \int_{-\frac{1}{2}\pi}^{-\pi} + \int_{-\pi}^{-\frac{3}{2}\pi} + \int_{-\frac{3}{2}\pi}^{-\frac{3}{2}\pi+i\infty}$$

$$= \int_{-\frac{1}{2}\pi-i\infty}^{-\frac{1}{2}\pi} + \int_{-\frac{1}{2}\pi}^{-\frac{3}{2}\pi} + \int_{-\frac{3}{2}\pi}^{-\frac{3}{2}\pi+i\infty} .$$

Thus the integrals along the two paths yield the same value for all z in the first quadrant. By shifting again to the left we can repeat the process of continuation and thus we find that $H^1_\kappa(z)$ can be continued through-out the whole z-plane (avoiding the origin). The reasoning for $H^2_\kappa(z)$ is quite the same.

15.3.8 – Neumann's functions

The function $H^1_{-\kappa}(z)$ is represented by the integral

$$H^1_{-\kappa}(z) = -\frac{1}{\pi}\int_{L_1} \exp i(-\kappa u - z \sin u)du.$$

Replacing the variable u by $u-\pi$ effects the shifting of the path of integration to the right through the distance π. Denoting this new path symbolically by $L_1+\pi$ we get

$$H^1_{-\kappa}(z) = -\frac{e^{\kappa\pi i}}{\pi}\int_{L_1+\pi} \exp i(-\kappa u + z \sin u)du.$$

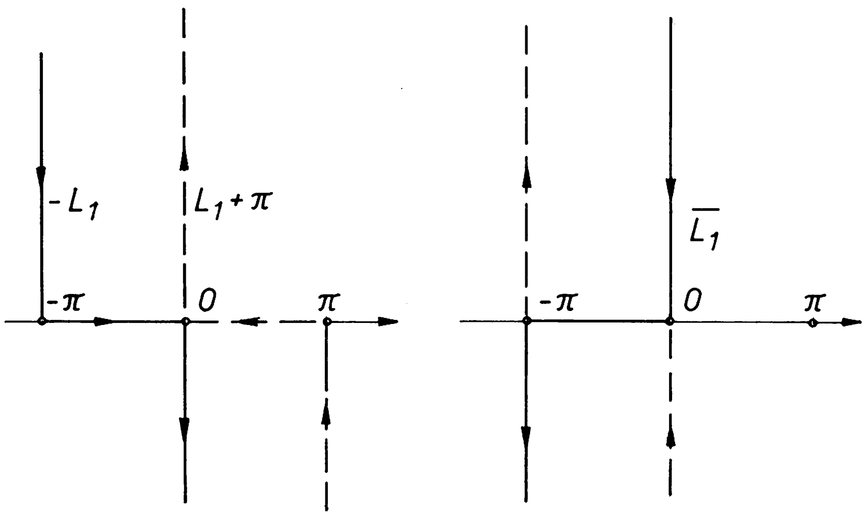


Fig. 15.3–6. Modification of the paths for the Hankel functions

Reflecting the path $L_1+\pi$ in the origin we obtain L_1, percorsed in the reverse sense (fig. 15.3–6). Hence replacing u by $-u$ we get

$$H^1_{-\kappa}(z) = -\frac{e^{\kappa\pi i}}{\pi}\int_{L_1} \exp i(\kappa u - z \sin u)du,$$

the integral on the right being that for $H^1_\kappa(z)$. A similar reasoning is applicable in the case of $H^2_{-\kappa}(z)$. Thus we arrive at

$$\boxed{\begin{aligned} H^1_{-\kappa}(z) &= e^{\kappa\pi i}H^1_\kappa(z), \\ H^2_{-\kappa}(z) &= e^{-\kappa\pi i}H^2_\kappa(z). \end{aligned}} \tag{15.3–52}$$

The complex conjugate of $H^1_\kappa(z)$ is

$$\overline{H^1_\kappa(z)} = -\frac{1}{\pi}\int_{\bar{L}_1} \exp -i(\bar{\kappa}u-\bar{z}\sin u)du,$$

where $\bar{L}_1$ is obtained by reflecting L_1 in the real axis. Replacing u by $-u$ the path $\bar{L}_1$ becomes $-L_2$, (fig. 15.3–6). Hence

$$\overline{H^1_\kappa(z)} = -\frac{1}{\pi}\int_{L_2} \exp i(\bar{\kappa}u-\bar{z}\sin u)du,$$

or

$$\boxed{\overline{H^1_\kappa(z)} = H^2_{\bar{\kappa}}(\bar{z}).} \tag{15.3–53}$$

It follows that *if κ and z are real the Hankel functions are conjugate complex.*

The functions

$$\boxed{N_\kappa(z) = \frac{1}{2i}(H^1_\kappa(z)-H^2_\kappa(z))} \tag{15.3–54}$$

are called *Neumann's functions*. They are real if κ and z are real.

Solving $H^1_\kappa(z)$ and $H^2_\kappa(z)$ from (15.3–51) and (51.3–54) we obtain

$$\boxed{\begin{aligned} H^1_\kappa(z) &= J_\kappa(z)+iN_\kappa(z),\\ H^2_\kappa(z) &= J_\kappa(z)-iN_\kappa(z).\end{aligned}} \tag{15.3–55}$$

We notice that from (15.3–52) and (15.3–51) follows

$$\boxed{J_{-\kappa}(z) = \tfrac{1}{2}e^{\kappa\pi i}H^1_\kappa(z)+\tfrac{1}{2}e^{-\kappa\pi i}H^2_\kappa(z)} \tag{15.3–56}$$

and this yield (15.3–8) if κ is an integer.

Finally we wish to express Neumann's function of order κ in terms of Bessel functions. The equation (15.3–56) may be written as

$$\begin{aligned} J_{-\kappa}(z) &= \tfrac{1}{2}(H^1_\kappa(z)+H^2_\kappa(z))\cos\kappa\pi+\tfrac{1}{2}i(H^1_\kappa(z)-H^2_\kappa(z))\sin\kappa\pi\\ &= J_\kappa(z)\cos\kappa\pi-N_\kappa(z)\sin\kappa\pi,\end{aligned}$$

whence

$$\boxed{N_\kappa(z) = \frac{J_\kappa(z)\cos\kappa\pi-J_{-\kappa}(z)}{\sin\kappa\pi}.} \tag{15.3–57}$$

15.3.9 – Bessel's functions of the second kind

Neumann's functions multiplied by π

$$Y_\kappa(z) = \pi N_\kappa(z) \tag{15.3–58}$$

are called *Bessel's functions of the second kind.* They are of particular importance if $\kappa = k$ is a non-negative integer, for then we have

If κ is a non-negative integer k then the functions $J_\kappa(z)$ and $N_k(z)$ constitute a fundamental system of solutions of Bessel's differential equation. It is clear from Sommerfeld's integral (15.3–43), that $J_\kappa(z)$ is holomorphic with respect to κ, (section 2.9.1). Applying de l'Hôpital's rule we find from (15.3–57)

$$\begin{aligned} Y_k(z) &= \lim_{\kappa\to k} \pi N_\kappa(z) \\ &= \lim_{\kappa\to k} -\pi J_\kappa(z) \tan \kappa\pi + \lim_{\kappa\to k} \left(\frac{\partial J_\kappa(z)}{\partial \kappa} - \frac{1}{\cos \kappa\pi} \frac{\partial J_{-\kappa}(z)}{\partial \kappa}\right), \end{aligned}$$

or

$$Y_k(z) = \lim_{\kappa\to k} \left(\frac{\partial J_\kappa(z)}{\partial \kappa} - (-1)^k \frac{\partial J_{-\kappa}(z)}{\partial \kappa}\right). \tag{15.3–59}$$

Differentiating the series (15.3–7) with respect to κ we get

$$\frac{\partial J_\kappa(z)}{\partial k} = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(\kappa+\nu+1)} \left(\frac{z}{2}\right)^{\kappa+2\nu} \left(\log \frac{z}{2} - \Psi(\kappa+\nu+1)\right),$$

where Ψ is the function (4.8–1). Hence

$$\lim_{\kappa\to k} \frac{\partial J_\kappa(z)}{\partial \kappa} = J_k(z) \log \frac{z}{2} - \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} \frac{1}{(k+\nu)!} \left(\frac{z}{2}\right)^{k+2\nu} \Psi(k+\nu+1).$$

The evaluation of the limit of $\partial J_{-\kappa}(z)/\partial \kappa$ requires the separate consideration of the sum of the first k terms in the expansion of $J_{-\kappa}(z)$. Differentiating this sum we get

$$-\sum_{\nu=0}^{k-1} \frac{(-1)^\nu}{\nu!} \frac{1}{\Gamma(-\kappa+\nu+1)} \left(\frac{z}{2}\right)^{-\kappa+2\nu} \log \frac{z}{2} + \\ + \sum_{\nu=0}^{k-1} \frac{(-1)^\nu}{\nu!} \frac{\Psi(-\kappa+\nu+1)}{\Gamma(-\kappa+\nu+1)} \left(\frac{z}{2}\right)^{-\kappa+2\nu}.$$

Now we observe that $\Gamma(t)$ and $\Psi(t)$ have poles at $t = 0, -1, -2, \ldots$. If $-m \leq 0$, m being an integer, we have in view of (4.6–12) and the first remark in section 4.8.1

$$\lim_{t\to -m} \frac{\Psi(t)}{\Gamma(t)} = \lim_{t\to -m} \frac{(t+m)\Psi(t)}{(t+m)\Gamma(t)} = \frac{-m!}{(-1)^m} = (-1)^{m+1} m!.$$

Making $\kappa \to k$ the second sum tends to

$$(-1)\sum_{\nu=0}^{k-1} \frac{(k-\nu-1)!}{\nu!}\left(\frac{z}{2}\right)^{-k+2\nu}.$$

Differentiating the remainder of the series we get

$$-\sum_{\nu=k}^{\infty} \frac{(-1)^{\nu}}{\nu!} \frac{1}{\Gamma(-\kappa+\nu+1)}\left(\frac{z}{2}\right)^{-\kappa+2\nu}\left(\log\frac{z}{2} - \Psi(-\kappa+\nu+1)\right)$$

and making $\kappa \to k$ this tends to

$$-\sum_{\nu=k}^{\infty} \frac{(-1)^{\nu}}{\nu!} \frac{1}{\Gamma(-k+\nu+1)}\left(\frac{z}{2}\right)^{-k+2\nu}\left(\log\frac{z}{2} - \Psi(-k+\nu+1)\right)$$

$$= -(-1)^k \sum_{\nu=0}^{\infty} \frac{(-1)^{\nu}}{\nu!} \frac{1}{\Gamma(k+\nu+1)}\left(\frac{z}{2}\right)^{k+2\nu}\left(\log\frac{z}{2} - \Psi(\nu+1)\right)$$

$$= -(-1)^k J_k(z)\log\frac{z}{2} + (-1)^k \sum_{\nu=0}^{\infty} \frac{(-1)^{\nu}}{\nu!(k+\nu)!}\left(\frac{z}{2}\right)^{k+2\nu}\Psi(\nu+1).$$

Hence (15.3–59) becomes

$$Y_k(z) = 2J_k(z)\log\frac{z}{2} - \sum_{\nu=0}^{\infty} \frac{(-1)^{\nu}}{\nu!(k+\nu)!}\left(\frac{z}{2}\right)^{k+2\nu}(\Psi(\nu+1)+\Psi(k+\nu+1)+$$

$$-\sum_{\nu=0}^{k-1} \frac{(k-\nu-1)!}{\nu!}\left(\frac{z}{2}\right)^{-k+2\nu}. \qquad (15.3\text{–}60)$$

If $k = 0$ the last sum must be omitted. The function $Y_k(z)$ is clearly linearly independent of $J_k(z)$. The coefficients occurring in the second sum on the right are "known" numbers, for by virtue of (4.8–3) and (4.8–6) we have

$$\Psi(1) = -\gamma, \qquad \Psi(m+1) = 1+\tfrac{1}{2}+\ldots+\frac{1}{m}-\gamma, \qquad (15.3\text{–}61)$$

m being a positive integer.

Since $Y_k(z)$ is essentially Neumann's function and the Hankel functions can be expressed in terms of Bessel's and Neumann's functions, we may conclude that

The Hankel functions $H^1_\kappa(z)$ and $H^2_\kappa(z)$ constitute a fundamental system of solutions of Bessel's differential equation for every value of the parameter κ.

15.4 – Legendre's functions

15.4.1 – Solution of Legendre's differential equation

In section 3.14.1 we obtained the result that the polynomials of Legendre satisfy the differential equation

$$\boxed{(z^2-1)w''+2zw'-n(n+1)w = 0,} \tag{15.4–1}$$

where n is a non-negative integer. This equation will be referred to as *Legendre's differential equation.*

It is easily verified that the equation has regular singular points at $z = \pm 1$. At these points the indicial equation is

$$\rho(\rho-1)+\rho = 0, \tag{15.4–2}$$

having the double root $\rho = 0$.

Replacing z by $1/z$ the equation (15.4–1) becomes

$$(1-z^2)w''-2zw'-\frac{n(n+1)}{z^2}w = 0 \tag{15.4–3}$$

and it follows that $z = \infty$ is a regular singular point of (15.4–1). The indicial equation at this point is

$$\rho(\rho-1)-n(n+1) = 0, \tag{15.4–4}$$

with roots

$$\rho_0 = n+1, \qquad \rho_1 = -n.$$

The most manageable forms of the solutions are those which proceed in descending powers of z and are, therefore, appropriate to the singularity ad infinity. Substituting the series corresponding to

$$\sum_{\nu=0}^{\infty} c_\nu z^{n-\nu} \tag{15.4–5}$$

into (15.4–1) we obtain the recursion formulas

$$(n-m+2)(n-m+1)c_{m-2}+m(2n-m+1)c_m = 0 \tag{15.4–6}$$

where $m \geqq 2$. We may take $c_0 = 1$ and it follows from the fact that the equation (15.4–1) remains unaltered if we change z into $-z$ that all coefficients with odd subscripts are zero. Hence a solution is

$$w_1(z) = z^n\left(1-\frac{n(n-1)}{2(2n-1)}z^{-2}+\frac{n(n-1)(n-2)(n-3)}{2\cdot 4(2n-1)(2n-3)}z^{-4}+\dots\right). \tag{15.4–7}$$

Actually the series terminates after a finite number of terms and is, therefore, a polynomial of degree n.

The solution corresponding to the root $\rho_0 = n+1$ may be found by replacing n by $-(n+1)$ in the recursion formulas (15.4–6). Now we obtain the solution

$$w_0(z) = z^{-(n+1)}\left(1+\frac{(n+1)(n+2)}{2(2n+3)}z^{-2}+\right.$$
$$\left.+\frac{(n+1)(n+2)(n+3)(n+4)}{2\cdot 4(2n+3)(2n+5)}z^{-4}+\ldots\right). \qquad (15.4\text{–}8)$$

By applying the ratio test (1.6–12) it is found that this series is convergent if $|z| > 1$. Hence $w_0(z)$ is regular at $z = \infty$.

Thus we see that $w_0(z)$ and $w_1(z)$ constitute a fundamental system at $z = \infty$. Notwithstanding the fact that the difference of ρ_0 and ρ_1 is an integer the solution $w_1(z)$ does not contain a logarithmic term. The equation (15.4–1) cannot have two independent solutions which are polynomials, for in the contrary case all solutions would be polynomials.

Since Legendre's polynomial (3.13–27) is a solution, it must be equal to $w_1(z)$ apart from a multiplicative constant. It follows from Rodrigues's formula (3.13–27) that the leading term in $P_n(z)$ is

$$\frac{1}{2^n n!}\frac{d^n}{dz^n}z^{2n} = \frac{2n(2n-1)\ldots n}{2^n n!}z^n = \frac{(2n)!}{2^n(n!)^2}z^n.$$

Hence

$$\boxed{\begin{aligned}P_n(z) = \frac{(2n)!}{2^n(n!)^2}\left(z^n - \frac{n(n-1)}{2(2n-1)}z^{n-2}+\right.\\ \left.+\frac{n(n-1)(n-2)(n-3)}{2\cdot 4(2n-1)(2n-3)}z^{n-4}+\ldots\right).\end{aligned}} \qquad (15.4\text{-}9)$$

The first polynomials are

$$\begin{aligned}
P_0(z) &= 1,\\
P_1(z) &= z,\\
P_2(z) &= \tfrac{1}{2}(3z^2-1),\\
P_3(z) &= \tfrac{1}{2}z(5z^2-3)\\
P_4(z) &= \tfrac{1}{8}(35z^4-30z^2+3)\\
P_5(z) &= \tfrac{1}{8}z(63z^4-70z^2+15), \text{ etc.}
\end{aligned}$$

15.4.2. – Legendre's functions of the second kind

A second solution of (14.5–1), besides the polynomial $P_n(z)$, is the power series (15.4–8) for $w_0(z)$. In order to obtain a convenient multiplier we make use of Abel's identity (15.1–19). The Wronskian of $w_0(z)$ and $w_1(z)$ is

$$w_0 w_1' - w_0' w_1 = \frac{c}{1-z^2}, \tag{15.4–10}$$

where c is a constant. Since $z^2\,(w_0 w_1' - w_0' w_1)$ tends to $2n+1$ as $z \to \infty$ the value of this constant is $-(2n+1)$. If we now define

$$Q_n(z) = \frac{2^n(n!)^2}{(2n+1)!}\, w_0(z), \tag{15.4–11}$$

then evidently

$$\boxed{Q_n'(z)P_n(z) - Q_n(z)P_n'(z) = \frac{1}{1-z^2}.} \tag{15.4–12}$$

The function

$$\boxed{\begin{aligned} Q_n(z) = \frac{2^n(n!)^2}{(2n+1)!}\Bigg(& z^{-(n+1)} + \frac{(n+1)(n+2)}{2(2n+3)}\, z^{-(n+3)} \\ &+ \frac{(n+1)(n+2)(n+3)(n+4)}{2\cdot 4(2n+3)(2n+5)}\, z^{-(n+5)} + \ldots\Bigg) \end{aligned}} \tag{15.4–13}$$

is called *Legendre's function of the second kind* (and of order n). An alternative expression for this function may be obtained from (15.4–12) which is equivalent with

$$\frac{d}{dz}\left(\frac{Q_n(z)}{P_n(z)}\right) = \frac{1}{(1-z^2)P_n^2(z)}. \tag{15.4–14}$$

We cut the z-plane along a line-segment connecting the points -1 and $+1$. If z is in the remaining region we have evidently

$$\boxed{Q_n(z) = P_n(z)\int_\infty^z \frac{dt}{(1-t^2)P_n^2(t)}.} \tag{15.4–15}$$

In fact, if $z \to \infty$ then $P_n(z)/z^n$ tends to a constant and $Q_n(z)/z^n$ tends to zero as does the integral.

This integral yields interesting information about $Q_n(z)$. At the end of section 3.14.1 we proved that $P_n(z)$ has exactly n simple roots between -1 and $+1$. We denote them by $a_1, \ldots, a_n$. Expanding the integrand of (15.4–15) in partial fractions we get

$$\frac{1}{(1-t^2)P_n^2(t)} = \frac{A}{1-t} + \frac{B}{1+t} + \sum_{\nu=1}^{n} \frac{A_\nu}{(t-a_\nu)^2} + \sum_{\nu=1}^{n} \frac{B_\nu}{t-a_\nu}.$$

Multiplying both members by $1-t^2$ and taking $t = +1$ or $t = -1$ we see at once that $A = B = \frac{1}{2}$.
In order to find the contribution of the first two terms by integrating between z_0 and z we put

$$z+1 = re^{i\theta}, \qquad z-1 = r'e^{i\theta'},$$

where θ and θ' are between $-\pi$ and π. Hence

$$\int_{z_0}^{z} \frac{dt}{1+t} = \int_{r_0}^{r} \frac{d\rho}{\rho} + i\int_{\theta_0}^{\theta} d\theta = \log\frac{r}{r_0} + i(\theta-\theta_0)$$

and

$$\int_{z_0}^{z} \frac{dt}{1-t} = -\int_{r_0}^{r'} \frac{d\rho}{\rho} - i\int_{\theta'_0}^{\theta'} d\theta = -\log\frac{r'}{r'_0} - i(\theta'-\theta'_0),$$

whence

$$2\int_{z_0}^{z} \frac{dt}{1-t^2} = \log\frac{r}{r'} - \log\frac{r_0}{r'_0} + i(\theta-\theta') - i(\theta_0-\theta'_0).$$

If we let $z_0 \to \infty$ then $r_0/r'_0 \to 1$ and $\theta_0-\theta'_0 \to 0$. It follows that

$$\int_{\infty}^{z} \frac{dt}{1-t^2} = \frac{1}{2}\left(\log\frac{r}{r'} + i(\theta-\theta')\right) = \frac{1}{2}\log\frac{z+1}{z-1}.$$

Next we observe that the numbers $B_1, \ldots, B_n$ are all equal to zero. Indeed, if not, integration would yield logarithmic terms and this is in contradiction with the fact that the zeros of $P_n(z)$ are regular points for the differential equation and hence also for $Q_n(z)$. Thus we may infer that

$$\boxed{Q_n(z) = \frac{1}{2}P_n(z)\log\frac{z+1}{z-1} - W_{n-1}(z)} \tag{15.4–16}$$

provided that z is not on the segment $-1 \leqq x \leqq 1$. The last term is

$$W_{n-1}(z) = P_n(z)\sum_{\nu=1}^{n} \frac{A_\nu}{z-a_\nu} \tag{15.4–17}$$

a polynomial of degree $n-1$. If $n = 0$ this polynomial is absent. In the next section we shall describe a method for evaluating $W_{n-1}(z)$. The formula (15.4–16) shows that $Q_n(z)$ has a logarithmic singularity at the points $z = \pm 1$.

It is possible to define $Q_n(x)$ for $-1 < x < 1$. If z tends to x from the upper half of the z-plane then $\theta \to 0$, $\theta' \to \pi$, $r \to 1+x$ and $r' \to 1-x$. Hence $Q_n(z)$ has a limiting value which we shall denote by $Q_n(x+0i)$.

Thus

$$Q_n(x+0i) = \tfrac{1}{2}P_n(x)\log\frac{1+x}{1-x} - \tfrac{1}{2}\pi i P_n(x) - W_{n-1}(x).$$

By a similar reasoning we find, denoting by $Q_n(x-0i)$ the limiting value of $Q_n(z)$ as $z \to x$ from below,

$$Q_n(x-0i) = \tfrac{1}{2}P_n(x)\log\frac{1+x}{1-x} + \tfrac{1}{2}\pi i P_n(x) - W_{n-1}(x).$$

Hence

$$Q_n(x+0i) - Q_n(x-0i) = -\pi i P_n(x). \tag{15.4–18}$$

It appears that $Q_n(z)$ is discontinuous at every point between -1 and $+1$, the jump being $-\pi i P_n(x)$, if the segment is crossed from above to below.

If $-1 < x < 1$ we define

$$Q_n(x) = \tfrac{1}{2}(Q_n(x+0i) + Q_n(x-0i)) \tag{15.4–19}$$

and it follows that

$$Q_n(x) = \tfrac{1}{2}P_n(x)\log\frac{1+x}{1-x} - W_{n-1}(x). \tag{15.4–20}$$

15.4.3 – Neuman's and Schläfli's integral

The first member on the right of (15.4–16) may be written as

$$\tfrac{1}{2}P_n(z)\log\frac{z+1}{z-1} = \frac{1}{2}\int_{-1}^{1}\frac{P_n(z)}{z-t}\,dt. \tag{15.4–21}$$

It is natural to consider the function

$$f_n(z) = \frac{1}{2}\int_{-1}^{+1}\frac{P_n(t)}{z-t}\,dt,$$

where z is in the z-plane cut along the segment connecting -1 and $+1$. By straightforward computation it may be verified that this function is a solution of Legendre's equation (15.4–1).

Indeed, the derivatives of this function are

$$f_n'(z) = \int_{-1}^{+1} P_n(t)\frac{d}{dz}\frac{1}{z-t}\,dt = -\int_{-1}^{+1} P_n(t)\frac{d}{dt}\frac{1}{z-t}\,dt$$

$$= -\left.\frac{P_n(t)}{z-t}\right|_{-1}^{+1} + \int_{-1}^{+1}\frac{P_n'(t)}{z-t}\,dt$$

and

$$f_n''(z) = \left(\frac{P_n(t)}{(z-t)^2} - \frac{P_n'(t)}{z-t}\right)\Bigg|_{-1}^{+1} + \int_{-1}^{+1} \frac{P_n''(t)}{z-t}\, dt.$$

Hence

$$\begin{aligned}
&(z^2-1)f_n''(z)+2zf_n'(z)-n(n+1)f_n(z)\\
&= \int_{-1}^{+1} \frac{(z^2-1)P_n''(t)+2zP_n'(t)-n(n+1)P_n(t)}{z-t}\, dt +\\
&+ \left(-\frac{z^2-2zt+1}{(z-t)^2}P_n(t) - \frac{z^2-1}{z-t}P_n'(t)\right)\Bigg|_{-1}^{+1}\\
&= \int_{-1}^{+1} \frac{(t^2-1)P_n''(t)+2tP_n'(t)-n(n+1)P_n(t)}{z-t}\, dt +\\
&+ \int_{-1}^{+1}((z+t)P_n''(t)+2P_n'(t))dt - P_n(1)+P_n(-1)-(z+1)P_n'(1)+\\
&+(z-1)P_n'(-1) = (z+t)P_n'(t)\Bigg|_{-1}^{+1} + \int_{-1}^{+1} P_n'(t)dt - P_n(t)\Bigg|_{-1}^{+1} +\\
&-(z+t)P_n'(t)\Bigg|_{-1}^{+1} = 0.
\end{aligned}$$

This proves the assertion. Taking into account (15.4–16) and (15.4–21) we may conclude that

$$Q_n(z) - \frac{1}{2}\int_{-1}^{+1} \frac{P_n(t)}{z-t}\, dt = \frac{1}{2}\int_{-1}^{+1} \frac{P_n(z)-P_n(t)}{z-t}\, dt - W_{n-1}(z).$$

The expression on the left is a solution of (15.4–1). The right-hand member is clearly a polynomial of degree not exceeding $n-1$. Since only polynomials of exactly the degree n satisfy Legendre's equation (15.4–1) this right hand member is identically zero. Thus we have derived *Neumann's integral*

$$\boxed{Q_n(z) = \frac{1}{2}\int_{-1}^{+1} \frac{P_n(t)}{z-t}\, dt} \tag{15.4–22}$$

and

$$\boxed{W_{n-1}(z) = \frac{1}{2}\int_{-1}^{+1} \frac{P_n(z)-P_n(t)}{z-t}\, dt.} \tag{15.4–23}$$

The formula (15.4–22) holds also for z between -1 and $+1$. It must then be interpreted as a Cauchy principal value.

From Neumann's formula (15.4–22) and (3.14–11) we deduce at once the relation

$$\boxed{(n+1)Q_{n+1}(z)-(2n+1)zQ_n(z)+(n-1)Q_{n-1}(z) = 0.} \tag{15.4–24}$$

Inserting (15.4–9) into (15.4–23) we find by straight-forward computation

$$\boxed{\begin{aligned} W_{n-1}(z) = \frac{(2n)!}{2^n(n!)^2}\Bigg(z^{n-1}+z^{n-3}\left(\frac{1}{3}-\frac{n(n-1)}{2(2n-1)}\right)+ \\ +z^{n-5}\left(\frac{1}{5}-\frac{1}{3}\frac{n(n-1)}{2(2n-1)}+\frac{n(n-1)(n-2)(n-3)}{2\times 4(2n-1)(2n-3)}\right)+\ldots..\end{aligned}}$$

(15.4–25)

The first polynomials $W_{n-1}(z)$ are

$$\begin{aligned} W_{-1}(z) &= 0, \qquad \text{for} \quad P_0(z) = 1 \text{ identically;} \\ W_0(z) &= 1, \\ W_1(z) &= \tfrac{3}{2}z, \\ W_2(z) &= \tfrac{1}{6}(15z^2-4), \\ W_3(z) &= \tfrac{5}{24}z(21z^2-11), \\ W_4(z) &= \tfrac{1}{120}(945z^4-735z^2+64), \text{ etc.} \end{aligned}$$

It is not difficult to obtain a formula of the type (3.14–17) for $Q_n(z)$. Taking into account Rodrigues's formula (3.13–27) we may write (15.4–22) as

$$Q_n(z) = \frac{1}{2^{n+1}n!}\int_{-1}^{+1}(z-t)^{-1}\frac{d^n}{dt^n}(t^2-1)^n\,dt.$$

Integrating by parts we get in succession

$$\begin{aligned} Q_n(z) &= \frac{-1}{2^{n+1}n!}\int_{-1}^{+1}(z-t)^{-2}\frac{d^{n-1}}{dt^{n-1}}(t^2-1)^n\,dt \\ &= \frac{(-1)^2 2}{2^{n+1}n!}\int_{-1}^{1}(z-t)^{-3}\frac{d^{n-2}}{dt^{n-2}}(t^2-1)^n\,dt = \ldots, \end{aligned}$$

etc. After n steps we arrive at

$$\boxed{Q_n(z) = \frac{1}{2^{n+1}}\int_{-1}^{1}\frac{(1-t^2)^n}{(z-t)^{n+1}}\,dt,} \tag{15.4–26}$$

being *Schläfli's integral* for the Legendre function of the second kind.

15.4.4 – Legendre's functions for general parameters

The equation (15.4–1) is a special case of the equation

$$(z^2-1)w''+2zw'-\kappa(\kappa+1)w = 0, \tag{15.4–27}$$

where κ is any complex number. In view of the representation of Legendre's polynomials by means of a Schäfli integral (3.14–17) it is natural to investigate under what circumstances an integral of the type

$$f_\kappa(z) = \int_C \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\,dt \tag{15.4–28}$$

represents a solution of (15.4–27). It is clear that if z is in some appropiately chosen simply connected region the function $f_\kappa(z)$ is holomorphic.

Inserting (15.4–28) into the left hand member of (15.4–27) we get

$$\begin{aligned}(z^2-1)f_\kappa''(z)&+2zf_\kappa'(z)-\kappa(\kappa+1)f_\kappa(z)\\ &= (\kappa+1)\int_C \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+3}}((\kappa+2)(z^2-1)+2z(t-z)-\kappa(t-z)^2)dt\\ &= (\kappa+1)\int_C \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+3}}((\kappa+2)(t^2-1)-2(\kappa+1)t(t-z))dt\\ &= -(\kappa+1)\int_C \frac{d}{dt}\frac{(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+2}}\,dt = -(\kappa+1)\left.\frac{(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+2}}\right|_C,\end{aligned}$$

where the last expression denotes the difference of the values of the function

$$V(t) = -(\kappa+1)\frac{(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+2}} \tag{15.4–29}$$

obtained by analytic continuation along C. Thus we see that

The integral (15.4–28) *is a solution of Legendre's equation* (15.4–27) *if the function* $V(t)$ *as defined in* (15.4–29) *returns to its initial value if* t *percorses the path* C.

By taking C appropriately we can find generalizations of the Legendre functions considered in the previous section.

It should be noticed that the theorem of section 2.9.1 applied to (15.4–28) asserts that $f_\kappa(z)$ is also holomorphic with respect to κ.

15.4.5 – Legendre's functions of the first kind

The points $t = \pm 1$ and $t = z$ are in general ramification points of the function $V(t)$. We cut the complex plane along the half ray from $-\infty$ to -1 and we take C as a simple closed circuit encircling the points $t = 1$ and $t = z$ once in the anti-clockwise sense, whereby C remains in the cut plane, (fig. 15.4–1). It is clear that in this case $V(t)$ returns to its initial value if t describes the path C. If κ is a non-negative integer then $t=1$ is an ordinary point and C may be taken as a small circle $t = z$. Then the integral (15.4–28) reduces to Schläfli's integral (3.14–17), apart from the multiplier. Hence it is natural to define with Pochhammer's notation

$$\boxed{P_\kappa(z) = \frac{1}{2\pi i}\int^{(1+,z+)} \frac{(t^2-1)^\kappa}{2^\kappa(t-z)^{\kappa+1}}\,dt,} \tag{15.4–30}$$

as *Legendre's function of the first kind of order* κ. The integrand will be specified precisely by taking $\arg(t+1) = \arg(t-1) = 0$, if t is a point of C on the real axis to the right of $t = 1$. Further we shall take $-\pi \leq \arg(t-z) \leq \pi$ at this point. Then $P_\kappa(z)$ is well-defined as a single-valued holomorphic function in the cut plane.

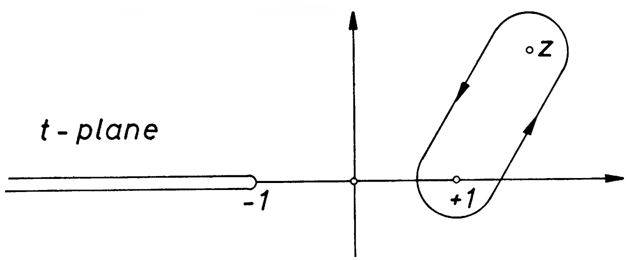


Fig. 15.4–1. The path of integration $(1+, z+)$ for Legendre's function of the first kind

We wish to evaluate $P_\kappa(z)$ at $z = 1$. Then C may be taken as a small circle about $t = 1$ and we have

$$P_\kappa(1) = \frac{1}{2\pi i}\int_C \frac{(t^2-1)^\kappa}{2^\kappa(t-1)^{\kappa+1}}\,dt = \frac{1}{2\pi i}\int_C \frac{(t+1)^\kappa}{2^\kappa(t-1)}\,dt = \operatorname*{Res}_{z=1} \frac{(z+1)^\kappa}{2^\kappa(z-1)}$$

and it follows from (3.3–5) that

$$P_\kappa(1) = 1 \tag{15.4–31}$$

irrespective of the value of κ.

15.4.6 – Legendre's functions of the second kind

By taking another contour we may expect another solution of the equation (15.4–27). Let z not be a real number between -1 and 1. The contour will be taken as a figure-of-eight encircling the point $t = 1$ once in the clockwise sense and the point $t = -1$ once in the anti-clockwise sense and such that z is outside either of the loops. We may take the path such that it crosses itself at $t = 0$, (fig. 15.4–2). Also in this case the function returns to its initial value if t returns to its starting point on C. Let $-\pi < \arg z \leqq \pi$ and let $\arg(z-t) \to \arg z$ as $t \to 0$ along C. Further we shall suppose that $\arg(t+1) = \arg(t-1) = 0$ if t is the point on C on the real axis to the right of $t = 1$.

We proceed to prove that, after introducing an appropriate multiplier the function $f_\kappa(z)$ represents $Q_n(z)$ if $\kappa = n$, a non-negative integer.

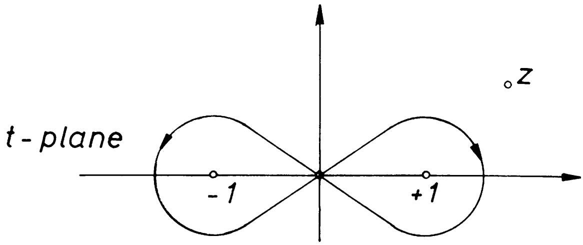


Fig. 15.4–2. The path of integration $(-1+, +1-)$ for Legendre's function of the second kind

Assuming that $\mathrm{Re}\,\kappa > -1$ we have

$$\int^{(-1+,+1-)} \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\,dt = \int_0^{-1} + e^{2\kappa\pi i}\int_{-1}^{0} + e^{2\kappa\pi i}\int_0^1 + \int_1^0$$

$$= (e^{2\kappa\pi i}-1)\int_{-1}^{+1} \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\,dt = e^{\kappa\pi i} 2i \sin\kappa\pi \int_{-1}^{+1} \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\,dt.$$

Interchanging t and z in the denominator, on agreeing that

$$z-t = (t-z)e^{\pi i},$$
$$1-t = (t-1)e^{\pi i},$$

we may conclude that

$$\frac{1}{2i\sin\kappa\pi}\int^{(-1+,+1-)} \frac{(t^2-1)^\kappa}{2^{\kappa+1}(z-t)^{\kappa+1}}\,dt = e^{\kappa\pi i}\int_{-1}^{+1} \frac{(t^2-1)^{\kappa+1}}{2^{\kappa+1}(z-t)^{\kappa+1}}\,dt$$

$$= \int_{-1}^{1} \frac{(1-t^2)^\kappa}{2^{\kappa+1}(z-t)^{\kappa+1}}\,dt.$$

For $\kappa = n$ this last integral represents $Q_n(z)$ as is clear from (15.4–26). By this reason we define

$$\boxed{Q_\kappa(z) = \frac{1}{2i\sin\kappa\pi}\int^{(-1+,+1-)} \frac{(t^2-1)^\kappa}{2^{\kappa+1}(z-t)^{\kappa+1}}\,dt} \qquad (15.4\text{–}32)$$

as *Legendre's function of the second kind* of order κ. The function is single-valued and holomorphic in the plane cut along the real axis from $-\infty$ to $+1$.

15.4.7 – Asymptotic behaviour of Legendre's functions for large values of $|z|$

As we have seen in the previous section the function $Q_\kappa(z)$ may be represented by

$$Q_\kappa(z) = \frac{1}{2^{\kappa+1}} \int_{-1}^{+1} \frac{(1-t^2)^\kappa}{(z-t)^{\kappa+1}}\, dt \tag{15.4–33}$$

provided that $\operatorname{Re}\kappa > -1$. We may write this as

$$z^{\kappa+1} Q_\kappa(z) = \frac{1}{2^{\kappa+1}} \int_{-1}^{+1} \frac{(1-t^2)^\kappa}{(1-t/z)^{\kappa+1}}\, dt.$$

If we let $z \to \infty$ under the condition $|\arg z| < \pi$ we get

$$\lim_{z\to\infty} z^{\kappa+1} Q_\kappa(z) = \frac{1}{2^\kappa} \int_0^1 (1-t^2)^\kappa\, dt$$

$$= \frac{1}{2^{\kappa+1}} \int_0^1 (1-u)^\kappa u^{-\frac{1}{2}}\, du = \frac{1}{2^{\kappa+1}} \frac{\Gamma(\kappa+1)\Gamma(\frac{1}{2})}{\Gamma(\kappa+\frac{3}{2})},$$

or

$$\boxed{\lim_{z\to\infty} z^{\kappa+1} Q_\kappa(z) = \frac{\sqrt{\pi}}{2^{\kappa+1}} \frac{\Gamma(\kappa+1)}{\Gamma(\kappa+\frac{3}{2})},} \qquad \operatorname{Re}\kappa > -1. \tag{15.4–34}$$

The similar problem for $P_\kappa(z)$ requires more attention, because the path of integration includes the point z. In

$$\int^{(1+,z+)} \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\, dt$$

we perform the substitutions

$$\begin{aligned} t-1 &= (z-1)u, \\ t+1 &= (z-1)u+2, \\ t-z &= (z-1)(u-1). \end{aligned}$$

We get

$$(z-1)^\kappa \int^{(1+,z+)} u^\kappa (u-1)^{-(\kappa+1)} \left(u + \frac{2}{z-1}\right)^\kappa du.$$

If we let $z \to \infty$ as above we find, taking into account (15.4–30),

$$\lim_{z\to\infty} z^{-\kappa} P_\kappa(z) = \frac{1}{2^{\kappa+1}\pi i} \int^{(1+,0+)} u^{2\kappa} (u-1)^{-(\kappa+1)}\, du$$

$$= \frac{-e^{-\kappa\pi i}}{2^{\kappa+1}\pi i} \int^{(1+,0+)} u^{2\kappa} (1-u)^{-(\kappa+1)}\, du.$$

If we now suppose that $-\frac{1}{2} < \operatorname{Re}\kappa < 0$ we have

$$\int^{(1+,0+)} = (e^{2\kappa\pi i}-1)\int_0^1 u^{2\kappa}(1-u)^{-(\kappa+1)}\,du,$$

whence

$$\lim_{z\to\infty} z^{-\kappa}P_\kappa(z) = \frac{-\sin\kappa\pi}{2^\kappa\pi}\int_0^1 u^{2\kappa}(1-u)^{-(\kappa+1)}\,du.$$

We evaluate the last integral by employing (4.6–26) and (4.6–13). We have

$$\int_0^1 u^{2\kappa}(1-u)^{-(\kappa+1)}\,du = \frac{\Gamma(2\kappa+1)\Gamma(-\kappa)}{\Gamma(\kappa+1)} = 2\kappa\frac{\Gamma(2\kappa)\Gamma(-\kappa)}{\Gamma(\kappa+1)}$$

$$= \frac{2^{2\kappa}}{\sqrt{\pi}}\Gamma(\kappa+\tfrac{1}{2})\Gamma(-\kappa) = -\frac{2^{2\kappa}\sqrt{\pi}}{\sin\kappa\pi}\frac{\Gamma(\kappa+\frac{1}{2})}{\Gamma(\kappa+1)}.$$

Thus we obtain

$$\boxed{\lim_{z\to\infty} z^{-\kappa}P_\kappa(z) = \frac{2^\kappa}{\sqrt{\pi}}\frac{\Gamma(\kappa+\frac{1}{2})}{\Gamma(\kappa+1)},} \quad -\tfrac{1}{2} < \operatorname{Re}\kappa < 0. \qquad (15.4\text{–}35)$$

15.4.8 – Legendre's functions as infinite series

Supposing that $\operatorname{Re}\kappa > -1$ Legendre's function of the second kind may be represented by (15.4–33), or

$$Q_\kappa(z) = \frac{1}{2^{\kappa+1}z^{\kappa+1}}\int_{-1}^{+1}\frac{(1-t^2)^\kappa}{(1-t/z)^{\kappa+1}}\,dt. \qquad (15.4\text{–}36)$$

If $|z| > 1$ the integrand can be expanded in a series uniformly convergent with respect to t, so that

$$Q_\kappa(z) = cz^{-(\kappa+1)}F_{-(\kappa+1)}(z),$$

where $F_{-(\kappa+1)}(z)$ is an ordinary power series in terms of z^{-1} with leading coefficient equal to unity. Inserting $Q_\kappa(z)$ into the left hand member of (15.4–27) we obtain the recursive relations (15.4–6) for the coefficients of $F_{-(\kappa+1)}(z)$, by replacing n by $-(\kappa+1)$. Hence

$$F_{-(\kappa+1)}(z) = 1+\frac{(\kappa+1)(\kappa+2)}{2(2\kappa+3)}z^{-2}$$

$$+\frac{(\kappa+1)(\kappa+2)(\kappa+3)(\kappa+4)}{2\times 4(2\kappa+3)(2\kappa+5)}z^{-4}+\ldots \qquad (15.4\text{–}37)$$

and in view of (15.4–34) we have evidently

$$\boxed{Q_\kappa(z) = \frac{\sqrt{\pi}}{2^{\kappa+1}} \frac{\Gamma(\kappa+1)}{\Gamma(\kappa+\frac{3}{2})} \frac{1}{z^{\kappa+1}} F_{-(\kappa+1)}(z).} \tag{15.4–38}$$

By the identity principle this result is true for unrestricted values of κ, negative integral values excluded, and for all values of $|z|>1$, $|\arg z|<\pi$.

It is clear from (15.4–36) that $Q_\kappa(z)$ is singular at $z = 1$. On the other hand $P_\kappa(z)$ and $P_{-(\kappa+1)}(z)$ are regular at $z = 1$, taking the value 1 there. Since these latter functions are solutions of the differential equation (15.4–27) (this equation remaining invariant if we replace κ by $-(\kappa+1)$) and cannot constitute a fundamental system, we necessarily have

$$\boxed{P_\kappa(z) = P_{-(\kappa+1)}(z),} \tag{15.4–39}$$

this result being true for unrestricted values of κ on account of the identity principle.

If κ is not an integer the functions $Q_\kappa(z)$ and $Q_{-(\kappa+1)}(z)$ constitute a fundamental system. Taking into account (15.4–39) we may conclude that

$$P_\kappa(z) = A_\kappa Q_\kappa(z) + A_{-(\kappa+1)} Q_{-(\kappa+1)}(z), \tag{15.4–40}$$

where A_κ and $A_{-(\kappa+1)}$ do not depend on z. Multiplying by $z^{-\kappa}$ and making $z \to \infty$ we find under the assumption $-\frac{1}{2} < \operatorname{Re} z < 0$ from (15.4–35) and (15.4–34)

$$\frac{2^\kappa}{\sqrt{\pi}} \frac{\Gamma(\kappa+\frac{1}{2})}{\Gamma(\kappa+1)} = A_{-(\kappa+1)} 2^\kappa \sqrt{\pi} \frac{\Gamma(-\kappa)}{\Gamma(\frac{1}{2}-\kappa)},$$

whence

$$A_{-(\kappa+1)} = \frac{1}{\pi} \frac{\Gamma(\frac{1}{2}+\kappa)\Gamma(\frac{1}{2}-\kappa)}{\Gamma(\kappa+1)\Gamma(-\kappa)}.$$

From (4.6–13) follows

$$\Gamma(\tfrac{1}{2}+\kappa)\Gamma(\tfrac{1}{2}-\kappa) = \frac{\pi}{\cos \kappa\pi},$$

$$\Gamma(1+\kappa)\Gamma(-\kappa) = \frac{-\pi}{\sin \kappa\pi},$$

whence

$$A_{-(\kappa+1)} = -\frac{1}{\pi} \tan \kappa\pi$$

and

$$A_\kappa = \frac{1}{\pi} \tan \kappa\pi.$$

Thus we have proved

$$P_\kappa(z) = \frac{1}{\pi} (Q_\kappa(z) - Q_{-(\kappa+1)}(z)) \tan \kappa\pi. \tag{15.4–41}$$

By the identity principle this is true for all values of κ which are not halves of an integer.

Replacing $\tan \kappa\pi$ by

$$\frac{\Gamma(\frac{1}{2}-\kappa)\Gamma(\frac{1}{2}+\kappa)}{\Gamma(1-\kappa)\Gamma(\kappa)},$$

we find by some elementary manipulations

$$P_\kappa(z) = \frac{2^\kappa}{\sqrt{\pi}}\frac{\Gamma(\kappa+\frac{1}{2})}{\Gamma(\kappa+1)} z^\kappa F_\kappa(z) + \frac{1}{2^{\kappa+1}\sqrt{\pi}}\frac{\Gamma(-\kappa-\frac{1}{2})}{\Gamma(-\kappa)}\frac{1}{z^{\kappa+1}} F_{-(\kappa+1)}(z), \tag{15.4–42}$$

where

$$F_\kappa(z) = 1 - \frac{\kappa(\kappa-1)}{2(\kappa-1)} z^{-2} + \frac{\kappa(\kappa-1)(\kappa-2)(\kappa-3)}{2\cdot 4(2\kappa-1)(2\kappa-3)} z^{-4} + \ldots,$$

provided that κ is not the half of an odd integer. If κ is a non-negative integer the second term vanishes and it is easily shown that the first term reduces to (15.4–9).

15.4.9 – Recurrent relations

The relation between Legendre's polynomials derived in section 3.14.3 remains true for the general Legendre functions $P_\kappa(z)$ and $Q_\kappa(z)$. It is sufficient to prove them for the function $f_\kappa(z)$, defined in (15.4–28).

Observing that

$$\frac{d}{dt}\frac{(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+1}} = \frac{2(\kappa+1)t(t^2-1)^\kappa}{(t-z)^{\kappa+1}} - \frac{(\kappa+1)(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+2}},$$

we find on integrating along any path C

$$0 = 2\int_C \frac{t(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\, dt - \int_C \frac{(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+2}}\, dt,$$

whence

$$\int_C \frac{(t^2-1)^\kappa}{(t-z)^\kappa}\, dt = \frac{1}{2}\int_C \frac{(t^2-1)^{\kappa+1}}{(t-z)^{\kappa+2}}\, dt - z\int_C \frac{(t^2-1)^\kappa}{(t^2-z)^{\kappa+1}},$$

or

$$\int_C \frac{(t^2-1)^\kappa}{(t-z)^\kappa}\, dt = \tfrac{1}{2} f_{\kappa+1}(z) - z f_\kappa(z). \tag{15.4–43}$$

Expanding the equation

$$\int_C \frac{d}{dt}\frac{t(t^2-1)^\kappa}{(t-z)^\kappa}\, dt = 0$$

we get

$$\int_C \frac{(t^2-1)^\kappa}{(t-z)^\kappa}\,dt+2\kappa\int_C \frac{t^2(t^2-1)^{\kappa-1}}{(t-z)^\kappa}\,dt-\kappa\int_C \frac{t(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\,dt = 0.$$

Writing $(t^2-1)+1$ for t^2 and $(t-z)+z$ for t we find that

$$(\kappa+1)\int_C \frac{(t^2-1)^\kappa}{(t-z)^\kappa}\,dt+2\kappa\int_C \frac{(t^2-1)^{\kappa-1}}{(t-z)^\kappa}\,dt-\kappa z\int_C \frac{(t^2-1)^\kappa}{(t-z)^{\kappa+1}}\,dt = 0$$

and using (15.4–43) we get

$$\tfrac{1}{2}(\kappa+1)f_{\kappa+1}(z)-(2\kappa+1)f_\kappa(z)+2\kappa f_{\kappa+1}(z) = 0. \qquad (15.4\text{–}44)$$

By adjusting constants we obtain relations for $P_\kappa(z)$ and $Q_\kappa(z)$ which are generalizations of (3.4–11).

Differentiating (15.4–43) we get

$$\tfrac{1}{2}f_{\kappa+1}(z)-zf_\kappa'(z) = (\kappa+1)f_\kappa(z) \qquad (15.4\text{–}45)$$

and by adjusting constants we obtain a generalization of (3.14–15).

By simple algebraic manipulations it is possible to derive formulas which are generalizations of the remaining recurrent relations listed in section 3.14.3.

15.5 – Fuchsian equations

15.5.1 – Fundamental relations

Linear differential equations having no other singularities than regular singular points are of particular interest. They were investigated thoroughly by Fuchs and are called after him *Fuchsian equations.*

At first sight it is not clear which conditions $p(z)$ and $q(z)$ of the second order equation (15.1–9) must satisfy in order that $z = \infty$ be a regular point or a regular singular point. To study the behaviour at $z = \infty$ we have to perform the substitution $z' = 1/z$. Omitting primes from z the transformation leads to

$$w''+\left(\frac{2}{z}-\frac{1}{z^2}\,p\left(\frac{1}{z}\right)\right)w'+\frac{1}{z^4}\,q\left(\frac{1}{z}\right)w = 0. \qquad (15.5\text{–}1)$$

The behaviour of the solutions of (15.1–9) at $z = \infty$ is determined by that of the solutions of (15.5–1) at $z = 0$. Hence $z = \infty$ is a regular point of (15.1–9) if the functions

$$\frac{2}{z}-\frac{1}{z^2}\,p\left(\frac{1}{z}\right), \qquad \frac{1}{z^4}\,q\left(\frac{1}{z}\right) \qquad (15.5\text{–}2)$$

are regular at $z = 0$, or, which amounts to the same, if the functions

$$2z - z^2 p(z), \qquad z^4 q(z) \tag{15.5–3}$$

are regular at $z = \infty$.

The point $z = \infty$ is a regular singular point if the functions (15.5–2) have at most a simple pole and a double pole at $z = 0$. The first condition asserts that

$$\frac{1}{z}\, p\left(\frac{1}{z}\right)$$

remains finite as $z \to 0$, or that $zp(z)$ remains finite as $z \to \infty$. The second condition asserts that

$$\frac{1}{z^2}\, q\left(\frac{1}{z}\right)$$

remains finite if $z \to 0$ or that $z^2 q(z)$ remains finite as $z \to \infty$. This may be stated in the following form

The point $z = \infty$ is a regular singularity if $p(z)$ has at least a simple zero at $z = \infty$ and $q(z)$ at least a double zero.

As an example we may consider Legendre's equations (15.4–27). Here we have

$$p(z) = \frac{2z}{z^2-1}, \qquad q(z) = -\frac{\kappa(\kappa+1)}{z^2-1}$$

and the conditions of a regular singular point at $z = \infty$ are fulfilled.

In the case of Bessel's equation (15.3.–1) we have

$$p(z) = \frac{1}{z}, \qquad q(z) = \frac{z^2-\kappa^2}{z^2}$$

and it appears that $q(z)$ has no zero at $z = \infty$. Hence $z = \infty$ is not a regular singular point.

Let $a_1, \ldots, a_n$ denote the singular points of the equation (15.1–9) in the open plane. Supposing that they are regular singularities the functions $p(z)$ and $q(z)$ have the form

$$p(z) = \frac{P(z)}{\psi(z)}, \qquad q(z) = \frac{Q(z)}{\psi^2(z)}, \tag{15.5–4}$$

where

$$\psi(z) = (z-a_1)\ldots(z-a_n) \tag{15.5–5}$$

and $P(z)$, $Q(z)$ are polynomials of degree $\leq n-1$ and $\leq 2n-2$ respectively. If $z = \infty$ is a regular point, then the degree of $Q(z)$ does not exceed $2n-4$.

The decompositions of $p(z)$ and $q(z)$ in partial fractions are

$$p(z) = \sum_{\nu=1}^{n} \frac{A_\nu}{z-a_\nu}, \qquad q(z) = \sum_{\nu=1}^{n} \frac{B_\nu}{(z-a_\nu)^2} + \sum_{\nu=1}^{n} \frac{C_\nu}{z-a_\nu}. \tag{15.5–6}$$

The condition for the degree of $Q(z)$ leads at once to the relation

$$\sum_{\nu=1}^{n} C_\nu = 0. \tag{15.5–7}$$

If $z = \infty$ is an ordinary point we have additional relations. For large $|z|$ we have

$$2z - z^2 p(z) = 2z - \sum_{\nu=1}^{n} A_\nu \left(z + a_\nu + \frac{a_\nu^2}{z} + \dots\right)$$

and this remains finite as $z \to \infty$ if and only if

$$\sum_{\nu=1}^{n} A_\nu = 2. \tag{15.5–8}$$

Further

$$z^4 q(z) = \sum_{\nu=1}^{n} B_\nu (z^2 + 2a_\nu z + \dots) + \sum_{\nu=1}^{n} C_\nu (z^3 + a_\nu z^2 + a_\nu^2 z + \dots).$$

Taking into account (15.5–7) we conclude that $z^4 q(z)$ remains finite as $z \to \infty$ if and only if

$$\sum_{\nu=1}^{n} (B_\nu + C_\nu a_\nu) = 0 \tag{15.5–9}$$

and

$$\sum_{\nu=1}^{n} (2B_\nu a_\nu + C_\nu a_\nu^2) = 0. \tag{15.5–10}$$

These are the fundamental relations for Fuchs' equation if $z = \infty$ is a regular point.

15.5.2 – The resolvent

We consider in more detail the resolvent (15.1–29) of the equation (15.1–9). The behaviour of this function at $z = \infty$ is described in the theorem

If the point at infinity is a regular singular point it is a zero of multiplicity at least two for $R(z)$. It is of multiplicity at least four if the point is a regular point for the equation (15.1–9).

If $z = \infty$ is a regular singular point then $p(z)$ has at least a simple zero at $z = \infty$ and hence $p'(z)$ has a double zero there. Since also $q(z)$ has a double zero the assertion is at once clear.

Suppose $z = \infty$ is a regular point. Then we have an expansion

$$2z-z^2p(z) = c_0+\frac{c_1}{z}+\dots,$$

or

$$p(z) = \frac{2}{z}-\frac{c_0}{z^2}-\frac{c_1}{z^3}-\dots$$

and

$$p'(z) = -\frac{2}{z^2}+\frac{2c_0}{z^3}+\dots.$$

It is now clear that in $\frac{1}{2}p^2(z)+p'(z)$ the terms with z^{-2} and z^{-3} are absent. Since also $q(z)$ has a zero at least of order four the truth of the assertion follows.

At the (finite) singular points $a_1, \dots, a_n$ the function $R(z)$ has double poles. Let us introduce the numbers $\alpha_1, \dots, \alpha_{n+1}$ by means of

$$\begin{aligned} \tfrac{1}{2}(1-\alpha_k^2) &= \lim_{z\to a_k}(z-a_k^2)R(z), \qquad k = 1,\dots, n, \\ \tfrac{1}{2}(1-\alpha_{n+1}^2) &= \lim_{z\to\infty} z^2R(z). \end{aligned} \tag{15.5–11}$$

Since

$$\begin{aligned} \lim_{z\to a_k}(z-a_k)^2q(z) &= B_k, \\ \lim_{z\to a_k}(z-a_k)^2p^2(z) &= A_k^2, \\ \lim_{z\to a_k}(z-a_k)^2p'(z) &= -A_k \end{aligned}$$

we have for $k = 1, \dots, n$,

$$\tfrac{1}{2}(1-\alpha_k^2) = 2B_k-\tfrac{1}{2}A_k^2+A_k = 2B_k-\tfrac{1}{2}A_k(A_k-2). \tag{15.5–12}$$

Again

$$\begin{aligned} \lim_{z\to\infty} z^2q(z) &= \sum_{\nu=1}^{n}(B_\nu+C_\nu a_\nu), \\ \lim_{z\to\infty} z^2p^2(z) &= (\sum_{\nu=1}^{n}A_\nu)^2, \\ \lim_{z\to\infty} z^2p'(z) &= -\sum_{\nu=1}^{n}A_\nu. \end{aligned}$$

Hence

$$\tfrac{1}{2}(1-\alpha_{n+1}^2) = 2\sum_{\nu=1}^{n}(B_\nu+C_\nu a_\nu)-\tfrac{1}{2}(\sum_{\nu=1}^{n}A_\nu)(\sum_{\nu=1}^{n}A_\nu-2). \tag{15.5–13}$$

Assuming that $z = \infty$ is a regular point the relations (15.5–9) and (15.5–10) appear in the form (14.1–8) and 14.1–9) respectively, and in addition we have $\alpha_{n+1}^2 = 1$. It should be noticed, however, that the relations in section 14.1.2 are obtained under more restrictive assumptions, namely that $R(z)$ is related to a problem of conformal mapping.

Finally we may state the following theorem

The resolvent of a Fuchsian equation of the second order may be written in the form

$$\boxed{R(z) = \frac{1}{2}\sum_{\nu=1}^{n}\frac{(1-\alpha_\nu^2)}{(z-a_\nu)^2} + \frac{\beta_0+\beta_1 z+\ldots+\beta_{n-2}z^{n-2}}{\psi(z)},} \tag{15.5–14}$$

where $\psi(z) = (z-a_1)\ldots(z-a_n)$.

In fact, the numerator in the second member on the right is of degree $\leq n-2$, since $\lim_{z\to\infty} z^2R(z)$ is finite. From (15.5–11) follows

$$\tfrac{1}{2}\sum_{\nu=1}^{n}(1-\alpha_\nu^2) + \beta_{n-2} = \tfrac{1}{2}(1-\alpha_{n+1}^2). \tag{15.5–15}$$

15.5.3 – The relation of Fuchs

At the finite singular points the indicial equations are

$$\rho(\rho-1)+A_k\rho+B_k = 0, \qquad k = 1,\ldots, n. \tag{15.5–16}$$

It is readily seen that at $z = \infty$ the indicial equation is

$$\rho(\rho-1)+(2-\sum_{\nu=1}^{n}A_\nu)\rho+\sum_{\nu=1}^{n}(B_\nu+C_\nu a_\nu) = 0. \tag{15.5–17}$$

Let $\rho_0^{(k)}$, $\rho_1^{(k)}$ denote the roots of the equations (15.5–16) and $\rho_0^{(n+1)}$, $\rho_1^{(n+1)}$ the roots of (15.5–17). It is clear that we have the following relations

$$\rho_0^{(k)}+\rho_1^{(k)} = 1-A_k, \qquad k = 1,\ldots, n, \tag{15.5–18}$$

and

$$\rho_0^{(n+1)}+\rho_1^{(n+1)} = \sum_{\nu=1}^{n}A_\nu-1. \tag{15.5–19}$$

As a consequence

$$\boxed{\sum_{\nu=1}^{n+1}(\rho_0^{(\nu)}+\rho_1^{(\nu)}) = n-1,} \tag{15.5–20}$$

the *relation of Fuchs* between the roots of the indicial equations. If $z = \infty$ is an ordinary point the roots of (15.5–17) are in view of (15.5–8) and (15.5–9) equal to 1 and 0 and (15.5–20) remains true if we replace n by $n-1$.

Given the singular points $a_1, \ldots, a_n$ and the roots of the indicial equations, it is possible to evaluate the coefficients A_k and B_k, $k=1, \ldots, n$. In fact, the A_k follow from (15.5–18) and the coefficients B_k are determined by

$$\rho_0^{(k)}\rho_1^{(k)} = B_k, \qquad k = 1, \ldots, n. \tag{15.5–21}$$

The coefficients C_k satisfy the relation (15.5–7) and

$$\rho_0^{(n+1)}\rho_1^{(n+1)} = \sum_{\nu=1}^{n} (B_\nu + C_\nu a_\nu). \tag{15.5–22}$$

Hence $n-2$ of them remain undetermined. They are called the *accessory parameters* (see also section 14.1.2).

The number of accessory parameters does not exceed the number of singular points minus three.

Finally we wish to deduce a relation between the roots of the indicial equation and the numbers α_k, introduced in the previous section. If $k = 1, \ldots, n$ we find from (15.5–12), (15.5–18) and (15.5–21)

$$\begin{aligned} \alpha_k^2 &= -4B_k + A_k(A_k-2)+1 = -4B_k+(A_k-1)^2 \\ &= -4\rho_0^{(k)}\rho_1^{(k)} + (\rho_0^{(k)}+\rho_1^{(k)})^2 = (\rho_0^{(k)}-\rho_1^{(k)})^2. \end{aligned}$$

If $k = n+1$ we have from (15.5–13), (15.5–19) and (15.5–22)

$$\begin{aligned} \alpha_{n+1}^2 &= -4\sum_{\nu=1}^{n} (B_\nu + C_\nu a_\nu) + \Big(\sum_{\nu=1}^{n} A_\nu\Big)\Big(\sum_{\nu=1}^{n} A_\nu - 2\Big)+1 \\ &= -4\sum_{\nu=1}^{n} (B_\nu + C_\nu a_\nu) + \Big(\sum_{\nu=1}^{n} A_\nu - 1\Big)^2 \\ &= -4\rho_0^{(n+1)}\rho_1^{(n+1)} + (\rho_0^{(n+1)}+\rho_1^{(n+1)})^2 = (\rho_0^{(n+1)}-\rho_1^{(n+1)})^2. \end{aligned}$$

Since the sign of the α_k has not been fixed we may put

$$\boxed{\alpha_k = \rho_0^{(k)} - \rho_1^{(k)},} \qquad k = 1, \ldots, n+1. \tag{15.5–23}$$

15.5.4 – Fuchsian equations with one or two singular points

It is clear that a Fuchsian equation with no singular points has no other solutions than constant functions, for every solution is regular thoughout the extended plane.

The roots of an indicial equation determine the linear transformation which undergoes a fundamental system of solutions, if we perform analytic continuation about the singular point to which the indicial equation is associated. Hence these roots are not affected by a fractional linear transformation of the independent variable.

First we donsider an equation having only one singular point. According to the above remark we may suppose without loss of generality that this point is placed at the origin. By virtue of (15.5–8), expressing that the sum of the residues of $p(z)$ at the finite singular points is equal to two, we may infer that

$$p(z) = \frac{2}{z}. \tag{15.5–24}$$

Since $z = \infty$ is a regular point the function $q(z)$ has a zero at least of order four at $z = \infty$. On the other hand the multiplicity of its pole at $z = 0$ does not exceed two. This is only possible if $q(z)$ is identically zero and the desired equation is of the form

$$zw'' + 2w' = 0. \tag{15.5–25}$$

The indicial equation is

$$\rho(\rho-1) + 2\rho = 0$$

having the roots $\rho_0 = 0$, $\rho_1 = -1$. A fundamental system of solutions is

$$w_0(z) = 1, \qquad w_1(z) = z^{-1}.$$

In the case that there are two singularities we can place them at $z = 0$ and $z = \infty$. Then, since there is only one finite singularity

$$p(z) = \frac{A}{z}, \qquad q(z) = \frac{B}{z^2} + \frac{C}{z},$$

with $C = 0$, in accordance with (15.5–7). The desired equation is

$$z^2w'' + Azw' + Bw = 0, \tag{15.5–26}$$

a linear differential equation of the Eulerian type. The indicial equation at $z = 0$ is

$$\rho(\rho-1) + A\rho + B = 0$$

and if the roots ρ_0 and ρ_1 are different it is easily verified that

$$w_0(z) = z^{\rho_0}, \qquad w_1(z) = z^{\rho_1}$$

is a fundamental system of solutions. If both roots have the value ρ then a linearly independent system of solutions is

$$w_0(z) = z^{\rho}, \qquad w_1(z) = z^{\rho} \log z.$$

Thus we see that the Fuchsian differential equations of the second order with one or two singular points can be integrated in terms of elementary functions.

The next type in order of complexity is that with three regular singular points. It gives rise to non-elementary solutions and will be amply studied in the subsequent paragraph.

15.6 – Riemann's equation

15.6.1 – Papperitz's form of the equation

The most celebrated equation in the Fuchsian class is that with three singular points. It has been studied thoroughly by Riemann and, therefore, named after him. Riemann did not derive an explicit form of the equation if the singular points and the roots of the indicial equations are given. For he made the fundamental discovery that a lot of information can be obtained without knowledge of the coefficients of the equation. An elegant and symmetric form of the equation has been obtained for the first time by E. Papperitz. Knowing the explicit form of the equation it will be possible to write down the solutions.

For the sake of convenience we assume that the singular points a_1, a_2 and a_3 are finite. Since there are no accessory parameters, the coefficients A_k, B_k and C_k occurring in (15.5–6) can be expressed in terms of the affixes of the singular points and the roots of the indicial equations. These roots are usually called the *exponents of the equation.*

The evaluation of the numbers A_k and B_k offers no difficulty as we saw in section 15.5.3. There remains the evaluation of the C_k. In order to avoid unnecessary calculations we proceed as follows. We may write

$$q(z) = \frac{Q(z)}{\psi^2(z)}, \tag{15.6–1}$$

where $Q(z)$ is a polynomial and $\psi(z) = (z-a_1)(z-a_2)(z-a_3)$. By hypothesis $z = \infty$ is a regular point of the equation and, therefore, $q(z)$ has a zero of multiplicity at least four at $z=\infty$. Hence the degree of $Q(z)$ cannot exceed the number two. Thus $q(z)$ may be written as

$$q(z) = \frac{1}{\psi(z)} \sum_{\nu=1}^{3} \frac{D_\nu}{z-a_\nu}, \tag{15.6–2}$$

where the constants D_1, D_2, D_3 are determined by

$$B_k = \lim_{z\to a_k} (z-a_k)^2 q(z) = \lim_{z\to a_k} D_k \frac{z-a_k}{\psi(z)} = \frac{D_k}{\psi'(a_k)},$$

with $k = 1, 2, 3$. Hence

$$D_1 = \rho_0^{(1)}\rho_1^{(1)}(a_1-a_2)(a_1-a_3) \tag{15.6–3}$$

and cyclically.

It is common use to denote the singular points by a, b and c respectively and the corresponding exponents by α, α'; β, β'; γ, γ'. These exponents satisfy Fuchs's relation (15.5–20) which now takes the form

$$\boxed{\alpha+\alpha'+\beta+\beta'+\gamma+\gamma' = 1.} \tag{15.6–4}$$

In view of (15.5–18), (15.6–2) and (15.6–3) we have

A Fuchsian equation with three singular points is

$$\begin{aligned} w''+\left(\frac{1-\alpha-\alpha'}{z-a}+\frac{1-\beta-\beta'}{z-b}+\frac{1-\gamma-\gamma'}{z-c}\right)w'+ \\ +\left(\frac{\alpha\alpha'(a-b)(a-c)}{z-a}+\frac{\beta\beta'(b-c)(b-a)}{z-b}+\frac{\gamma\gamma'(c-a)(c-b)}{z-c}\right)\frac{w}{\psi(z)} = 0. \end{aligned} \tag{15.6–5}$$

with

$$\psi(z) = (z-a)(z-b)(z-c). \tag{15.6–6}$$

This result is due to Papperitz.

15.6.2 – Solution of Riemann's equation by means of integrals

One of the pleasant features of Riemann's equation is the fact that it can be solved in a closed form, i.e., we can find explicit expressions for the solutions. The derivation of these expressions requires some heavy computation but the result is worth the trouble.

First we make the substitution

$$w = (z-a)^{\alpha}(z-b)^{\beta}(z-c)^{\gamma}v = f(z)v, \tag{15.6–7}$$

in order to obtain a solution with exponent α at $z = a$, etc.

Differentiation yields

$$w' = \Sigma\frac{\alpha}{z-a}f(z)v+f(z)v'$$

and

$$w'' = \left(\Sigma\frac{\alpha(\alpha-1)}{(z-a)^2}+2\Sigma\frac{\alpha\beta}{(z-a)(z-b)}\right)f(z)v+2\Sigma\frac{\alpha}{z-a}f(z)v'+f(z)v'',$$

where the summation must be performed about all terms which are obtained by cyclic permutation from the term written behind the sigma sign.

Inserting these expressions into Riemann's equation we may omit the common factor $f(z)$. The coefficients of v' becomes

$$\Sigma \frac{1-\alpha-\alpha'}{z-a} + 2\Sigma \frac{\alpha}{z-a} = \Sigma \frac{1+\alpha-\alpha'}{z-a}. \qquad (15.6\text{–}8)$$

The coefficient of v becomes

$$\Sigma \frac{\alpha(\alpha-1)}{(z-a)^2} + 2\Sigma \frac{\alpha\beta}{(z-a)(z-b)} + \left(\Sigma \frac{\alpha}{z-a}\right)\left(\Sigma \frac{1-\alpha-\alpha'}{z-a}\right) +$$
$$+\Sigma \frac{\alpha\alpha'(a-b)(a-c)}{(z-a)\psi(z)}$$
$$= -\Sigma \frac{\alpha\alpha'}{(z-a)^2} + \Sigma \frac{\alpha(1-\beta-\beta')+\beta(1-\alpha-\alpha')+2\alpha\beta}{(z-a)(z-b)} + \Sigma \frac{\alpha\alpha'(a-b)(a-c)}{(z-a)\psi(z)}$$
$$= -\Sigma \frac{((\alpha\alpha'(z-b)(z-c)-(a-b)(a-c))}{(z-a)\psi(z)} + \Sigma \frac{(\alpha+\beta-\alpha\beta'-\alpha'\beta)(z-c)}{\psi(z)}$$
$$= \Sigma \frac{-\alpha\alpha'(z+a-b-c)+(\alpha+\beta-\alpha\beta'-\alpha'\beta)(z-c)}{\psi(z)}.$$

Taking into account (15.6–4) we find that the coefficient of z in the numerator is

$$-\Sigma\alpha\alpha' + 2\Sigma\alpha - \Sigma\alpha\beta' - \Sigma\alpha'\beta = 2\Sigma\alpha - (\Sigma\alpha)(\Sigma\alpha')$$
$$= (\Sigma\alpha)(2-\Sigma\alpha') = (\Sigma\alpha)(1+\Sigma\alpha).$$

The constant term in the numerator is

$$-\Sigma\alpha\alpha'(a-b-c) - \Sigma(\alpha+\beta-\alpha\beta'-\alpha'\beta)c$$
$$= \Sigma a(\alpha(\alpha+\beta+\beta'+\gamma+\gamma'-1)+\beta\beta'+\gamma\gamma'+\beta\gamma'+\beta'\gamma-\beta-\gamma)$$
$$= \Sigma a((\alpha+\beta+\gamma)(\alpha+\beta'+\gamma')-(\alpha+\beta+\gamma)) = (\Sigma\alpha)(\Sigma a(\alpha+\beta'+\gamma'-1)$$
$$= -(\Sigma\alpha)\Sigma a(\alpha'+\beta+\gamma).$$

Hence the coefficient of v turns out to be

$$\frac{(\Sigma\alpha)((1+\Sigma\alpha)z-\Sigma a(\alpha'+\beta+\gamma))}{\psi(z)}. \qquad (15.6\text{–}9)$$

Now it is easy to verify that the new differential equation may be written as

$$\psi(z)v'' - (\mu\psi'(z)+\chi(z))v' + \left(\tfrac{1}{2}\mu(\mu+1)\psi''(z)+(\mu+1)\chi'(z)\right)v = 0, \quad (15.6\text{–}10)$$

with

$$\begin{aligned} \mu+1 &= -(\alpha+\beta+\gamma), \\ \psi(z) &= (z-a)(z-b)(z-c), \\ \chi(z) &= \Sigma(\alpha'+\beta+\gamma)(z-b)(z-c). \end{aligned} \qquad (15.6\text{–}11)$$

Indeed

$$\begin{aligned}\mu\psi'+\chi &= -(1+\Sigma\alpha)\Sigma(z-b)(z-c)+\Sigma(\alpha'+\beta+\gamma)(z-b)(z-c)\\ &= -(1+\Sigma\alpha)\Sigma(z-b)(z-c)+(\Sigma\alpha)\Sigma(z-b)(z-c)+\\ &\quad +\Sigma(\alpha'-\alpha)(z-b)(z-c) = \Sigma(-1+\alpha'-\alpha)(z-b)(z-c).\end{aligned}$$

Differentiation yields

$$\begin{aligned}\mu\psi''+\chi' &= \Sigma(-1+\alpha'-\alpha)(z-c)+\Sigma(-1+\alpha'-\alpha)(z-b)\\ &= \Sigma(-2+\beta'-\beta+\gamma'-\gamma)(z-a) = \Sigma(-1-\Sigma\alpha-1+\alpha+\beta'+\gamma')(z-a)\\ &= -(1+\Sigma\alpha)\Sigma(z-a)-(\alpha'+\beta+\gamma)(z-a).\end{aligned}$$

Since

$$-\tfrac{1}{2}\mu\psi'' = (1+\Sigma\alpha)\Sigma(z-a),$$

we get

$$\begin{aligned}\tfrac{1}{2}\mu\psi''+\chi' &= -\Sigma(\alpha'+\beta+\gamma)(z-a) = -\Sigma(\alpha'+\beta+\gamma)z+\\ &\quad +\Sigma(\alpha'+\beta+\gamma)a = -(1+\Sigma\alpha)z+\Sigma(\alpha'+\beta+\gamma)a.\end{aligned}$$

The next step is the task to find a function $U(t)$ such that any integral of the type

$$v = \int_C U(t)(t-z)^{\mu+1}dt \qquad (15.6\text{–}12)$$

is a solution of the equation (15.6–10), provided that the path C is suitably chosen. The exponent $\mu+1$ has been chosen in order to ensure that the function $(z-a)^\alpha(z-b)^\beta(z-c)^\gamma\, v$ should be regular at $z = \infty$.

First we observe that differentiation within the sign of integration is legitimate if C does not depend on z, nor passes through one of the singular points a, b and c. We suppose that C remains in the open plane. Inserting (15.6–12) into (15.6–10) the condition that the equation should be satisfied becomes

$$\begin{aligned}\int_C (t-z)^{\mu-1}(\mu(\psi(z)+(t-z)\psi'(z)+\tfrac{1}{2}(t-z)^2\psi''(z))+\\ +(t-z)(\chi(z)+(t-z)\chi'(z)))U(t)dt = 0.\end{aligned}$$

Since $\psi(z)$ and $\chi(z)$ are polynomials of the third and the second degree respectively we have (observing that $\psi''(z) = 6$)

$$\psi(z)+(t-z)\psi'(z)+\tfrac{1}{2}(t-z)^2\psi''(z) = \psi(t)-(t-z)^3$$

and

$$\chi(z)+(t-z)\chi'(z) = \chi(t)-\tfrac{1}{2}(t-z)^2\chi''(z).$$

But

$$\chi''(z) = 2\Sigma(\alpha'+\beta+\gamma) = 2(1+\alpha+\beta+\gamma) = -2\mu.$$

Hence the condition becomes

$$\int_C (t-z)^{\mu-1}(\mu\psi(t)+(t-z)\chi(t))U(t)dt = 0. \qquad (15.6\text{–}13)$$

At this stage we introduce an auxiliary function

$$V(t) = (t-z)^{\mu}\psi(t)U(t). \qquad (15.6\text{–}14)$$

Its derivative with respect to t is

$$\frac{dV}{dt} = (t-z)^{\mu}\frac{d(\psi U)}{dt} + \mu(t-z)^{\mu-1}\psi U.$$

This is identical with the integrand in (15.6–13) if $U(t)$ satisfies the differential equation

$$\frac{d(\psi U)}{dt} = \chi(t)U,$$

or

$$\frac{d(\psi U)/dt}{\psi U} = \frac{\chi(t)}{\psi(t)} = \Sigma\frac{\alpha'+\beta+\gamma}{t-a}.$$

A solution is

$$U(t) = (t-a)^{\alpha'+\beta+\gamma-1}(t-b)^{\alpha+\beta'+\gamma-1}(t-c)^{\alpha+\beta+\gamma'-1}.$$

Summing up we have established the following theorem

A function of the type

$$\boxed{(z-a)^{\alpha}(z-b)^{\beta}(z-c)^{\gamma}\int_C (t-a)^{\mu_a-1}(t-b)^{\mu_b-1}(t-c)^{\mu_c-1}(t-z)^{\mu+1}\,dt,} \qquad (15.6\text{–}15)$$

where the exponents occurring in the integrand are

$$\begin{aligned}\mu_a &= \alpha'+\beta+\gamma,\\ \mu_b &= \alpha+\beta'+\gamma,\\ \mu_c &= \alpha+\beta+\gamma',\\ \mu &= -\alpha-\beta-\gamma-1,\end{aligned} \qquad (15.6\text{–}16)$$

is a solution of Riemann's equation provided that C is a path in the t-plane such that

$$\int_C dV = 0, \qquad (15.6\text{–}17)$$

V being the function

$$V(t) = (t-a)^{\mu_a}(t-b)^{\mu_b}(t-c)^{\mu_c}(t-z)^{\mu}. \qquad (15.6\text{–}18)$$

From Fuch's relation (15.6–4) we deduce that

$$\mu_a+\mu_b+\mu_c+\mu = 0. \tag{15.6–19}$$

The condition (15.6–17) is certainly fulfilled if C is a double-contour of the Jordan-Pochhammer type as considered in section 12.7.9. We recall that (c_+, b_+, c_-, b_-) indicates a path encircling c in the positive direction then b in the positive direction, then again c in the negative direction and finally b in the negative direction, remaining outside a small circle about z. The integral (15.6–15) along this path is denoted by

$$(z-a)^\alpha(z-b)^\beta(z-c)^\gamma \int^{(c+,b+,c-,b-)} U(t)(t-z)^{\mu+1}\,dt. \tag{15.6–20}$$

Let t_0 denote any point on the contour under consideration. We shall prove that the integral (15.6–20) may be written as the sum of two ordinary loop integrals. Denoting by $\int^{(c+)}$ an integral taken along a simple loop starting and ending at t_0 and surrounding c once in a positive direction, such that the loop does not surround other singular points, we have in accordance with (12.7–35), observing that the integrand possesses the multipliers $e^{2\pi i\mu_b}$ and $e^{2\pi i\mu_c}$ at b and c respectively

$$\int^{(c+,b+,c-,b-)} = (1-e^{2\pi i\mu_b})\int^{(c+)} -(1-e^{2\pi i\mu_c})\int^{(b+)}. \tag{15.6–21}$$

If we make the additional assumptions that Re $\mu_b > -1$, Re $\mu_c > -1$ then the integrals $\int_{t_0}^{b}$, $\int_{t_0}^{c}$, taken along a simple path from t_0 to b and t_0 to c respectively, exist and we have, in accordance with (12.7–36),

$$\int^{(c+,b+,c-,b-)} = (1-e^{2\pi i\mu_b})(1-e^{2\pi i\mu_c})\int_b^c. \tag{15.6–22}$$

In order to make our considerations complete we shall prove that (15.6–20) provides a solution which is semi-regular at $z = a$ and has the exponent α there. Take z in a small circle about a which does not include either the points b or c. If necessary the double-loop C about b and c may be deformed so as to be wholly outside the circle. Then for all points t on C we have $|z-a| < |t-a|$. Let $|\arg(z-a)| < \pi$. Also let $\arg(a-b)$, $\arg(a-c)$ have their principal values and let $\arg(t-a)$, $\arg(t-b)$, $\arg(t-c)$ be similarly made definite when t is at t_0 on C. Then, if $\arg(z-b)$, $\arg(z-c)$, $\arg(t-z)$ are defined such that they reduce resp. to $\arg(a-b)$, $\arg(a-c)$, $\arg(t-a)$ as $z \to a$, we have

$$(z-b)^\beta = (a-b)^\beta\left(1+\beta\frac{z-a}{a-b}+\ldots\right),$$

$$(z-c)^{\gamma} = (a-c)^{\gamma}\left(1+\gamma\,\frac{z-a}{a-c}+\ldots\right),$$

$$(t-z)^{\mu+1} = (t-a)^{\mu+1}\left(1+(\mu+1)\,\frac{a-z}{t-a}+\ldots\right)$$

and the series on the right converge near $z = a$ and for all t on C. Now it is clear that

$$P_{\alpha}(z) = (z-a)^{\alpha}\varphi_{\alpha}(z), \tag{15.6–23}$$

where $\varphi_{\alpha}(z)$ stands for

$$\varphi_{\alpha}(z) = \left(\frac{z-b}{a-b}\right)^{\beta}\left(\frac{z-c}{a-c}\right)^{\gamma}\frac{\displaystyle\int^{(c+,b+,c-,b-)} U(t)(t-z)^{\mu+1}\,dt}{\displaystyle\int^{(c+,b+,c-,b-)} U(t)(t-a)^{\mu+1}\,dt}, \tag{15.6–24}$$

is a solution with $\varphi_{\alpha}(a) = 1$. The function φ_{α} is regular at $z = a$. The value of the integral in the denominator may be found from (12.7–42).

In a similar way other solutions of this kind, corresponding to other exponents and other singular points, being semi-regular there, may be obtained. The exceptional cases in which one or more of the differences $\alpha-\alpha'$, $\beta-\beta'$, $\gamma-\gamma'$ are integers require a special treatment which we shall omit.

15.6.3 – Riemann's theorem about contiguous functions

Let $P(z)$ denote a solution of Riemann's equation, semi-regular at a given singular point. A function obtained from it by augmenting one of the exponents by unity and diminishing at the same time another exponent by the same number, and which is a solution of the modified Riemann equation, is called a *contiguous function* of $P(z)$. Since these processes may be effected for any two of the exponents, these are $6\times 5 = 30$ contiguous functions. A famous theorem due to Riemann states

The function $P(z)$ and any two of its contiguous functions are connected by a linear relation, the coefficients in which are polynomials in z.

We represent $P(z)$ by the integral (15.6–15), where C is a double-contour about c and b. Hence $P(z)$ is a multiple of $P_{\alpha}(z)$ defined in (15.6–23). It is clear that

$$\int_C \frac{d}{dt}\left((t-a)^{\mu_a}(t-b)^{\mu_b-1}(t-c)^{\mu_c-1}(t-z)^{\mu+1}\right)dt = 0, \tag{15.6–25}$$

for the function to be differentiated resumes its initial value if z has described C. This leads to

$$\mu_a P+(\mu_b-1)P_{\alpha'+1,\beta'-1}+(\mu_c-1)P_{\alpha'+1,\gamma'-1}+\frac{\mu+1}{z-b}P_{\beta+1,\gamma'-1} = 0, \tag{15.6–26}$$

where $P_{\alpha'+1,\beta'-1}$ arises from P by replacing α' by $\alpha'+1$ and β' by $\beta'-1$, etc. Since (15.6–15) is symmetric with respect to b and c we also have

$$\mu_a P+(\mu_b-1)P_{\alpha'+1,\beta'-1}+(\mu_c-1)P_{\alpha'+1,\gamma'-1}+\frac{\mu+1}{z-c}P_{\gamma+1,\beta'-1} = 0. \tag{15.6–27}$$

Interchanging cyclically (a, α, α'), (b, β, β') and (c, γ, γ') we obtain six linear relations connecting P with twelve contiguous functions:

$$\begin{gathered} P_{\alpha+1,\beta'-1},\ P_{\beta+1,\gamma'-1},\ P_{\gamma+1,\alpha'-1},\ P_{\alpha+1,\gamma'-1},\ P_{\beta+1,\alpha'-1},\ P_{\gamma+1,\beta'-1} \\ P_{\alpha'+1,\beta'-1},\ P_{\beta'+1,\gamma'-1},\ P_{\gamma'+1,\alpha'-1},\ P_{\alpha'+1,\gamma'-1},\ P_{\beta'+1,\alpha'-1},\ P_{\gamma'+1,\beta'-1}. \end{gathered} \tag{15.6–28}$$

Writing $(t-a)^{\mu_a}$ as $(t-a)^{\mu_a-1}((t-b)+(b-a))$ we find from the expression for P, if $Q_{\alpha'-1}$ is obtained from P by replacing α' by $\alpha'-1$:

$$P = P_{\alpha'-1,\beta'+1}+(b-a)Q_{\alpha'-1}$$

and similarly

$$P = P_{\alpha'-1,\gamma'+1}+(c-a)Q_{\alpha'-1}.$$

Eliminating $Q_{\alpha'-1}$ from these equations we get

$$(c-b)P+(a-c)P_{\alpha'-1,\beta'+1}+(b-a)P_{\alpha'-1,\gamma'+1} = 0 \tag{15.6–29}$$

and by cyclic permutation we obtain two more relations of this type.

Writing $(t-z)^{\mu+1}$ as $(t-z)^{\mu}((t-a)-(z-a))$ we obtain

$$P = \frac{1}{z-b}P_{\beta+1,\gamma'-1}-(z-a)^{\alpha+1}(z-b)^{\beta}(z-c)^{\gamma}\times \\ \times\int_C (t-a)^{\mu_a-1}(t-b)^{\mu_b-1}(t-c)^{\mu_c-1}(t-z)^{\mu}\,dt,$$

whence

$$\frac{1}{z-a}\left(P-\frac{P_{\beta+1,\gamma'-1}}{z-b}\right) = \frac{1}{z-b}\left(P-\frac{P_{\gamma+1,\alpha'-1}}{z-c}\right) = \frac{1}{z-c}\left(P-\frac{P_{\alpha+1,\beta'-1}}{z-a}\right). \tag{15.6–30}$$

There are two more linear relations between P and the contiguous functions (15.6–28). We are now in possession of eleven linear relations between P and two of the above twelve contiguous functions, the coefficients in these relations being rational functions of z. As a consequence each of these functions can be expressed in terms of P and some selected one of them. Hence between P and any two of the functions (15.6–28) exists a linear relation with rational coefficients. Multiplying throughout by a common multiple of the denominators these coefficients become polynomials.

Starting with relations which are obtained from (15.6–25) by cyclic permutation we can extend the result to all contiguous functions. This concludes the proof of Riemann's theorem.

An illustrative example is provided by the relation (3.14–11) between Legendre's polynomials which holds also for the general Legendre functions as we pointed out in section 15.4.9.

Other examples are the Gaussian relations between contiguous hypergeometric functions to be dealt with in section 16.3.2.

15.6.4 – The monodromy group

The representation of the solutions of Riemann's equation by means of double loop integrals affords a means for obtaining the monodromy group, as has been shown by C. Jordan.

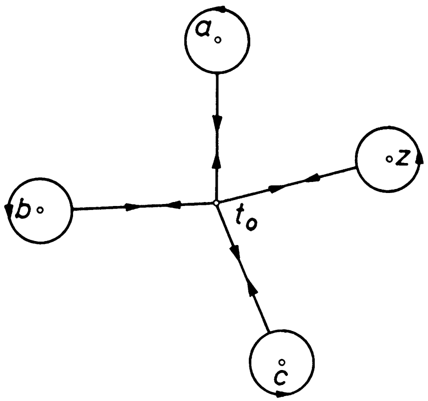


Fig. 15.6–1. Elementary loops in the t-plane punctured at $t = a, b, c, z$

Let t_0 be any point in the complex t plane different from the singular points and a given point z. By $\lambda_a, \lambda_b, \lambda_c, \lambda$ we denote simple loops starting and ending at t_0 and encircling once the points a, b, c, z respectively in the positive direction, with the understanding that λ_a encircles the point a but neither of the other points, etc. (fig. 15.6–1). In most cases we can take them as a simple curve described in two directions and closed by a small circumference.

We consider the functions

$$\begin{aligned} w_a(z) &= \int^{(a+,z+,a-,z-)} U(t)(t-z)^{\mu+1}\,dt, \\ w_b(z) &= \int^{(b+,z+,b-,z-)} U(t)(t-z)^{\mu+1}\,dt, \\ w_c(z) &= \int^{(c+,z+,c-,z-)} U(t)(t-z)^{\mu+1}\,dt, \end{aligned} \tag{15.6–31}$$

the integrals being defined unambiguously if we assign any function element of the integrand at t_0. The integrand possesses the multipliers

$$\begin{aligned} \xi_a &= e^{2\pi i\mu_a} \quad \text{at} \quad t = a, \\ \xi_b &= e^{2\pi i\mu_b} \quad \text{at} \quad t = b, \\ \xi_c &= e^{2\pi i\mu_c} \quad \text{at} \quad t = c, \\ \xi &= e^{2\pi i\mu} \quad \text{at} \quad t = z \end{aligned} \tag{15.6–32}$$

and is arbitrarily continuable in the extended t-plane punctured at $t = a, b, c, z$. It follows from (12.7–35) that

$$\begin{aligned} \int^{(b+,a+,b-,a-)} &= (1-\xi_a)\int_{\lambda_b} -(1-\xi_b)\int_{\lambda_a}, \\ w_a = \int^{(a+,z+,a-,z-)} &= (1-\xi)\int_{\lambda_a} -(1-\xi_a)\int_{\lambda}, \\ w_b = \int^{(b+,z+,b-,z-)} &= (1-\xi)\int_{\lambda_b} -(1-\xi_b)\int_{\lambda}. \end{aligned} \tag{15.6–33}$$

Eliminating the integrals on the right we find

$$(1-\xi)\int^{(b+,a+,b-,a-)} + (1-\xi_b)w_a - (1-\xi_a)w_b = 0.$$

Hence the solutions of Riemann's equation involving integrals along Jordan-Pochhammer contours about two of the three singular points may be expressed linearly in terms of the functions w_a, w_b, w_c multiplied by $(z-a)^\alpha(z-b)^\beta(z-c)^\gamma$, provided that μ is not an integer (i.e., $\xi \neq 1$). It is clear that under this latter assumption the pair w_a, w_b forms a fundamental system.

First we ask what happens if w_a, w_b are continued along an elementary loop encircling the point $t = a$. It is clear that we may assume that z is in the vicinity of $t = a$, for it is always possible to connect $t = z$ with a point near $t = a$ by means of a simple path (e.g. a polygon) percorsed back and

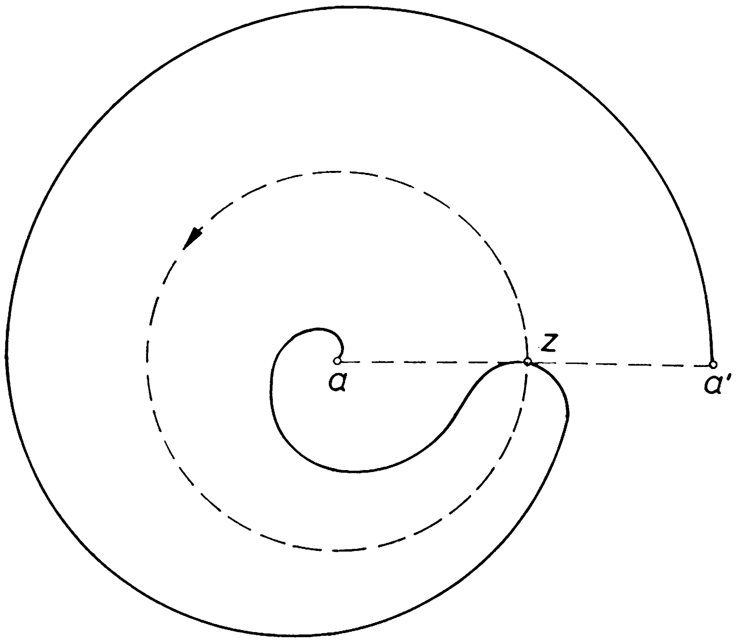


Fig. 15.6–2. The deformation of the t-plane in a neighbourhood of $t = a$ as given by
$$t = a+2ru \exp (8\pi\, iu(1-u)v).$$
The curve connecting the points a and a', where a' is such that $z = \frac{1}{2}(a+a')$ is the deformation of the rectilinear segment connecting these points. The point z describes a full circle

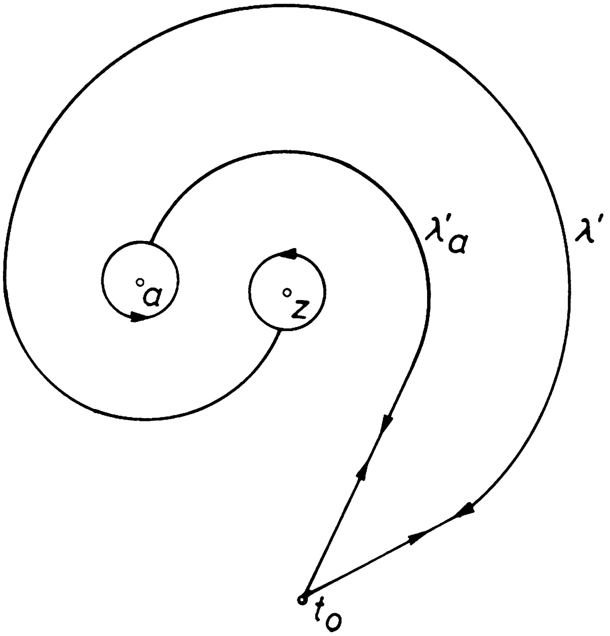


Fig. 15.6–3. Deformation of the loops λ and λ_a into λ' and λ'_a

forth and closed by a small circumference surrounding $t = a$. Let r denote the distance between $t = z$ and $t = a$. In view of the above remark we may assume, without loss of generality, that the points t_0, b and c are outside the circumference $|t-a| = 2r$, as well as the loops λ_b and λ_c.

An encircling of $t = a$ by z may be effected by means of a deformation in a neighbourhood of $t = a$ which leaves all points outside $|t-a| = 2r$ invariant, for instance, the deformation

$$t-a = 2ru \exp(8\pi iu(1-u)v), \qquad 0 \leq u \leq 1, \qquad 0 \leq v \leq 1.$$

In particular, t describes the full circle $t-a = \exp 2\pi iv$, $0 \leq v \leq 1$, if $u = \frac{1}{2}$. This deformation is defined throughout the disc $|t-a| \leq 2r$, (fig. 15.6–2). If z has returned to its initial position the loops λ_a and λ are deformed into loops λ_a' and λ', (fig. 15.6–3), whereas λ_b and λ_c remain fixed. This deformation takes place in the region $\mathfrak{R}'$, being the extended plane punctured at $t = a, b, c$. But λ_a' and λ_a are not homotopic in the region $\mathfrak{R}$, being the region $\mathfrak{R}'$ punctured at $t = z$, nor are λ' and λ homotopic in $\mathfrak{R}$.

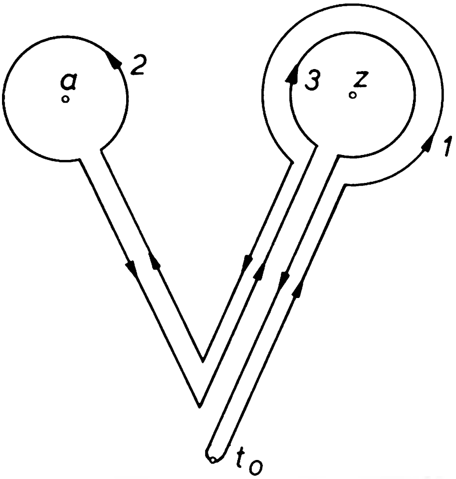


Fig. 15.6–4. Further deformation of the path λ_a'

Consider first the path λ_a', (fig. 15.6.4). By means of a suitable deformation in $\mathfrak{R}$ we may carry it into the product of three loops, two encircling $t = z$ and $t = a$ in this order in a positive sense and one encircling $t = z$ in a negative sense.

Similarly we see that λ' (fig. 15.6.5), may be deformed into a product of five loops, three encircling $t = z$, $t = a$, $t = z$ in this order in a positive sense and two encircling $t = a$ and $t = z$ in a negative sense.

Thus we obtain the relations of homotopy

$$\lambda'_a \approx \lambda^{-1}\lambda_a\lambda$$

and

$$\lambda' \approx \lambda^{-1}\lambda_a^{-1}\lambda\lambda_a\lambda.$$

Now it is clear that

$$\lambda'_a \approx \lambda_a\lambda_a^{-1}\lambda^{-1}\lambda_a\lambda \approx \lambda_a(\lambda^{-1}\lambda_a^{-1}\lambda\lambda_a)^{-1}.$$

Since the path $\lambda^{-1}\,\lambda_a^{-1}\lambda\lambda_a$ is closed with respect to analytic continuation of the integrand occurring in (15.6–31) (considered as a function of t),

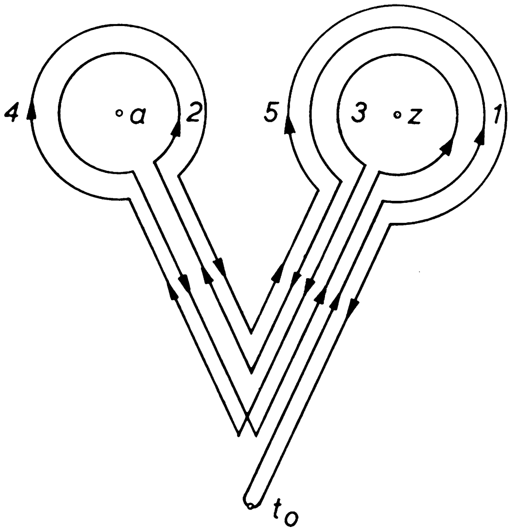


Fig. 15.6–5. Further deformation of the path λ'

we may apply (12.7–33) and (12.7–31). We find, in view of the first expression (15.6–31),

$$\int_{\lambda'_a} = -w_a + \int_{\lambda_a}. \tag{15.6–34}$$

Since

$$\lambda' \approx (\lambda^{-1}\lambda_a^{-1}\lambda\lambda_a)\lambda$$

we find, by virtue of (12.7–34),

$$\int_{\lambda'} = \xi w_a + \int_{\lambda}. \tag{15.6–35}$$

Taking into account (15.6–33) we see that the function w_a is continued analytically to the function

$$(1-\xi)\int_{\lambda'_a} -(1-\xi_a)\int_{\lambda'}$$

and the function w_b to the function

$$(1-\xi)\int_{\lambda_b} -(1-\xi_b)\int_{\lambda'}.$$

Using the equation (15.6–34) and (15.6–35) we find by simple calculation that w_a is transformed into $\xi\xi_a w_a$ and w_b is transformed into

$$\xi(\xi_b-1)w_a+w_b.$$

The monodromy group is known if we know the effect of an analytic continuation about the points a, b and c on the quotient w_b/w_a for this is equal to the quotient of two fundamental solutions of Riemann's equation. It is readily seen that analytic continuation along a loop about the point $t = a$ induces a linear fractional transformation of the quotient w_b/w_a, characterized by the matrix

$$\mathsf{S}_a = \begin{bmatrix} 1 & \xi(\xi_b-1) \\ 0 & \xi\xi_a \end{bmatrix} = \begin{bmatrix} 1 & \xi_b-1 \\ 0 & \xi_a \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \xi \end{bmatrix}.$$

Interchanging the roles of a and b, we find in a similar way

$$\mathsf{S}_b = \begin{bmatrix} \xi\xi_b & 0 \\ \xi(\xi_a-1) & 1 \end{bmatrix} = \begin{bmatrix} \xi_b & 0 \\ \xi_a-1 & 1 \end{bmatrix} \begin{bmatrix} \xi & 0 \\ 0 & 1 \end{bmatrix}.$$

Analytic continuation about $t = c$ induces a transformation S_c, but it is not easy to find this directly from the above results. Reasoning, however, as at the end of section 12.7.8 we see that the product of three elementary loops encircling the points a, b and c respectively is in the extended plane homotopic to unity. Since outside a large circle containing a, b and c every function element of the integrand occurring in (15.6–31) is reproduced by analytic continuation along a closed curve, we infer that

$$\mathsf{S}_a\mathsf{S}_b\mathsf{S}_c = \mathsf{E}.$$

Summing up we have in view of (15.6–32)

The monodromy group of Riemann's equation may be generated by the transformations

$$\mathsf{S}_a = \begin{bmatrix} 1 & e^{2\pi i\mu_b}-1 \\ 0 & e^{2\pi i\mu_a} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & e^{2\pi i\mu} \end{bmatrix} \tag{15.6–36}$$

and

$$\mathsf{S}_b = \begin{bmatrix} e^{2\pi i\mu_b} & 0 \\ e^{2\pi i\mu_a}-1 & 1 \end{bmatrix} \begin{bmatrix} e^{2\pi i\mu} & 0 \\ 0 & 1 \end{bmatrix}. \tag{15.6–37}$$

Notice that another fundamental system gives rise to a conjugate group, that is essentially the same group if we do not distinguish between isomorphic groups.

15.6.5 – Riemann's method for obtaining the monodromy group

An alternative method for finding the monodromy group goes back to Riemann. The interesting feature of this method is that it does not need the explicit form of the solutions of Riemann's differential equation. It has an algebraic character and it employs only general principles.

According to the general theory of Fuchs there exists a system of three pairs of solutions, each pair constituting a fundamental system

$$P_\alpha, P_{\alpha'}; P_\beta, P_{\beta'}; P_\gamma, P_{\gamma'}, \tag{15.6–38}$$

such that $(z-a)^{-\alpha}P_\alpha(z)$, $(z-a)^{-\alpha'}P_{\alpha'}(z)$ are regular at $z=a$, etc., provided that none of the differences $\alpha-\alpha'$, $\beta-\beta'$, $\gamma-\gamma'$ is an integer. In the course of this section we adopt this assumption.

Within the simply connected region bounded by the circumference through the singular points the solutions can be defined as single-valued holomorphic functions. The same is true for the region outside this circumference. It is assumed that the circumference is percorsed in the positive sense if we pass from a to b to c to a. In the case that these points are on a straight line we consider the region on the left of this line as the interior region.

It is clear that in the interior region exist relations of the form

$$\begin{aligned} P_\alpha &= a_\beta P_\beta + a_{\beta'} P_{\beta'} = a_\gamma P_\gamma + a_{\gamma'} P_{\gamma'}, \\ P_{\alpha'} &= a'_\beta P_\beta + a'_{\beta'} P_{\beta'} = a'_\gamma P_\gamma + a'_{\gamma'} P_{\gamma'}, \end{aligned} \tag{15.6–39}$$

the others being obtained by cyclic permutation.

Consider the quotients of two coefficients occurring in (15.6–39) with the same subscripts. Three of the four quotients

$$\frac{a_\beta}{a'_\beta}, \frac{a_{\beta'}}{a'_{\beta'}}, \frac{a_\gamma}{a'_\gamma}, \frac{a_{\gamma'}}{a'_{\gamma'}}$$

are determined by the remaining one, for the six functions (15.6–38) are each determined up to a multiplicative constant.

Let us consider the effect upon two solutions P_α, $P_{\alpha'}$ of analytic continuation along a simple closed path encircling the points $z = b$ and $z = c$ once. We may consider this path as a product of two loops, first about the point $z = b$ and then about the point $z = c$. As we pointed out in section 12.7.8 in the extended plane this loop is homotopic to a simple loop about $z = a$ percorsed in the opposite sense. Hence the analytic continuation of P_α and $P_{\alpha'}$ along the loops encircling $z = c$ and $z = a$ in the negative sense has the same effect as the analytic continuation along the loop about $z = b$ in the positive sense.

Performing the continuation along the loop about $z = b$ the branch P_α changes into

$$a_\beta e^{2\pi i\beta}P_\beta + a_{\beta'}e^{2\pi i\beta'}P_{\beta'}.$$

Performing the continuation about $z = c$ in the negative sense changes P_α into

$$a_\gamma e^{-2\pi i\gamma} + a_{\gamma'}e^{-2\pi i\gamma'}.$$

Since α is the exponent of P_α at $z = a$ an analytic continuation along a loop about $z = a$ in the negative sense yields the multiplier $e^{-2\pi i\alpha}$. Thus we obtain

$$\begin{aligned} a_\beta e^{2\pi i\beta}P_\beta + a_{\beta'}e^{2\pi i\beta'}P_{\beta'} &= a_\gamma e^{-2\pi i(\alpha+\gamma)}P_\gamma + a_\gamma e^{-2\pi i(\alpha+\gamma')}P_{\gamma'}, \\ a'_\beta e^{2\pi i\beta}P_\beta + a'_{\beta'}e^{2\pi i\beta'}P_{\beta'} &= a'_\gamma e^{-2\pi i(\alpha'+\gamma)}P_\gamma + a'_\gamma e^{-2\pi i(\alpha'+\gamma')}P_{\gamma'}. \end{aligned} \tag{15.6–40}$$

Eliminating $P_{\beta'}$, from the first equations of each pair (15.6–40) and (15.6–39), and also from the second equations of each of these pairs, we get

$$\begin{aligned} a_\beta \sin\pi(\beta'-\beta)P_\beta &= a_\gamma e^{-\pi i(\alpha+\beta+\gamma)}\sin\pi(\alpha+\beta'+\gamma)P_\gamma + \\ &\quad + a_{\gamma'}e^{-\pi i(\alpha+\beta+\gamma')}\sin\pi(\alpha+\beta'+\gamma')P_{\gamma'}, \\ a'_\beta \sin\pi(\beta'-\beta)P_\beta &= a'_\gamma e^{-\pi i(\alpha'+\beta+\gamma)}\sin\pi(\alpha'+\beta'+\gamma)P_\gamma + \\ &\quad + a'_{\gamma'}e^{-\pi i(\alpha'+\beta+\gamma')}\sin\pi(\alpha'+\beta'+\gamma')P_{\gamma'} \end{aligned} \tag{15.6–41}$$

Eliminating P_β from these relations we obtain a homogeneous linear relation between P_γ and $P_{\gamma'}$. Since these constitute a fundamental system the coefficients in the relation thus obtained vanish. Hence, from (15.6–41) and the corresponding expressions for $P_{\beta'}$ we get

$$\begin{aligned} \frac{a_\beta}{a'_\beta} &= \frac{a_\gamma}{a'_\gamma}\,\frac{e^{-\pi i\alpha}\sin\pi(\alpha+\beta'+\gamma)}{e^{-\pi i\alpha'}\sin\pi(\alpha'+\beta'+\gamma)} = \frac{a_{\gamma'}}{a'_{\gamma'}}\,\frac{e^{-\pi i\alpha}\sin\pi(\alpha+\beta'+\gamma')}{e^{-\pi i\alpha'}\sin\pi(\alpha'+\beta'+\gamma')}, \\ \frac{a_{\beta'}}{a'_{\beta'}} &= \frac{a_\gamma}{a'_\gamma}\,\frac{e^{-\pi i\alpha}\sin\pi(\alpha+\beta+\gamma)}{e^{-\pi i\alpha'}\sin\pi(\alpha'+\beta+\gamma)} = \frac{a_{\gamma'}}{a'_{\gamma'}}\,\frac{e^{-\pi i\alpha}\sin\pi(\alpha+\beta+\gamma')}{e^{-\pi i\alpha'}\sin\pi(\alpha'+\beta+\gamma')}. \end{aligned} \tag{15.6–42}$$

We find two times the value of the ratio

$$\frac{a_\gamma}{a'_\gamma}\bigg/\frac{a_{\gamma'}}{a'_{\gamma'}}.$$

The comparison of these values gives

$$\frac{\sin \pi(\alpha+\beta'+\gamma') \sin \pi(\alpha'+\beta'+\gamma)}{\sin \pi(\alpha'+\beta'+\gamma') \sin \pi(\alpha+\beta'+\gamma)} = \frac{\sin \pi(\alpha'+\beta+\gamma) \sin \pi(\alpha+\beta+\gamma')}{\sin \pi(\alpha+\beta+\gamma) \sin \pi(\alpha'+\beta+\gamma')},$$

and this equation is verified because of (15.6–4).

If we have a system of numbers a_β, $\alpha_{\beta'}$, a_γ, $a_{\gamma'}$, a'_β, $a'_{\beta'}$, a'_γ, $a'_{\gamma'}$, satisfying (15.6–42) there are six branches (15.6–38) presenting the desired behaviour at the singular points. Indeed, we may multiply the six branches by six constants, provided that the ratio of the quotients corresponding to the new coefficients have the same value as before. We can, therefore, consider a system of branches (15.6–38), being semi-regular at the singular points, where the coefficients are arbitrary, save for the mentioned restriction. If, however, the multipliers for the regular parts are given, that is to say, if the regular parts at the singular points take prescribed values, the coefficients are uniquely determined. Their evaluation requires more information about Riemann's equation than is needed in this section. The evaluation of the coefficients will be carried out in section 16.1.9 by an elegant method due to E. W. Barnes.

In order to find the monodromy group we focus our attention to P_α, $P_{\alpha'}$. An encircling about the point $z = a$ changes these functions into

$$e^{2\pi i\alpha}P_\alpha, \qquad e^{2\pi i\alpha'}P_{\alpha'} \tag{15.6–43}$$

respectively. As regards the effect of continuation along a simple loop about $z = b$ we express the functions P_α, $P_{\alpha'}$ in terms of P_β, $P_{\beta'}$, as in (15.6–39). After continuation along a loop encircling only the point $z = b$ they take the form

$$\begin{aligned} &a_\beta e^{2\pi i\beta}P_\beta + a_{\beta'} e^{2\pi i\beta'}P_{\beta'}, \\ &a'_\beta e^{2\pi i\beta}P_\beta + a'_{\beta'} e^{2\pi i\beta'}P_{\beta'}. \end{aligned} \tag{15.6–44}$$

If we replace P_β, $P_{\beta'}$ by their expressions in terms of P_α, $P_{\alpha'}$, by solving the equation (15.6–38) we have the desired solution. In order to facilitate the computation we write

$$\begin{aligned} P_\alpha &= a_\beta P_\beta + a_{\beta'} P_{\beta'}, \\ P_{\alpha'} &= \kappa_\beta a_\beta P_\beta + \kappa_{\beta'} a_{\beta'} P_{\beta'}, \end{aligned} \tag{15.6–45}$$

with

$$\kappa_\beta = \frac{a'_\beta}{a_\beta}, \qquad \kappa_{\beta'} = \frac{a'_{\beta'}}{a_{\beta'}}. \tag{15.6–46}$$

Solving the equations (15.6–45) for $a_\beta P_\beta$ and $a_{\beta'} P_{\beta'}$ we get

$$a_\beta P_\beta = \frac{\kappa_{\beta'} P_\alpha - P_{\alpha'}}{\kappa_{\beta'} - \kappa_\beta}, \qquad a_{\beta'} P_{\beta'} = \frac{\kappa_\beta P_\alpha - P_{\alpha'}}{\kappa_\beta - \kappa_{\beta'}}.$$

Inserting this into (15.6–44) we see that the analytic continuation along a loop about the point $z = b$ changes P_α, $P_{\alpha'}$ into

$$\frac{\kappa_{\beta'} e^{2\pi i\beta} - \kappa_\beta e^{2\pi i\beta'}}{\kappa_{\beta'} - \kappa_\beta} P_\alpha + \frac{e^{2\pi i\beta'} - e^{2\pi i\beta}}{\kappa_{\beta'} - \kappa_\beta} P_{\alpha'},$$
$$\frac{\kappa_\beta \kappa_{\beta'}(e^{2\pi i\beta} - e^{2\pi i\beta'})}{\kappa_{\beta'} - \kappa_\beta} P_\alpha + \frac{\kappa_{\beta'} e^{2\pi i\beta'} - \kappa_\beta e^{2\pi i\beta}}{\kappa_{\beta'} - \kappa_\beta} P_{\alpha'}. \tag{15.6–47}$$

The quotient $\kappa_{\beta'}/\kappa_\beta$ is equal to

$$\frac{\kappa_{\beta'}}{\kappa_\beta} = \frac{\sin \pi(\alpha+\beta'+\gamma) \sin \pi(\alpha'+\beta+\gamma)}{\sin \pi(\alpha'+\beta'+\gamma) \sin \pi(\alpha+\beta+\gamma)}. \tag{15.6–48}$$

With the aid of (4.6–13) and (15.6–4) this may be written as

$$\frac{\kappa_{\beta'}}{\kappa_\beta} = \frac{\Gamma(\alpha+\beta+\gamma)\Gamma(\alpha'+\beta'+\gamma')\Gamma(\alpha'+\beta'+\gamma)\Gamma(\alpha+\beta+\gamma')}{\Gamma(\alpha+\beta'+\gamma)\Gamma(\alpha'+\beta+\gamma')\Gamma(\alpha'+\beta+\gamma)\Gamma(\alpha+\beta'+\gamma')}. \tag{15.6–49}$$

In the transformation (15.6–47) a constant remains undetermined. A definite transformation is obtained if we replace P_α by $(\kappa_{\beta'} - \kappa_\beta)P_\alpha$. Then continuation about $z = a$ induces a transformation of the quotient $P_\alpha/P_{\alpha'}$ characterized by

$$\mathsf{T}_a = \begin{bmatrix} e^{2\pi i\alpha} & 0 \\ 0 & e^{2\pi i\alpha'} \end{bmatrix}. \tag{15.6–50}$$

Continuation about $z = b$ induces the transformation

$$\mathsf{T}_b = \begin{bmatrix} \dfrac{\kappa_\beta e^{2\pi i\beta'} - \kappa_{\beta'} e^{2\pi i\beta}}{\kappa_\beta - \kappa_{\beta'}} & e^{2\pi i\beta'} - e^{2\pi i\beta} \\ \dfrac{\kappa_\beta \kappa_{\beta'}(e^{2\pi i\beta} - e^{2\pi i\beta'})}{(\kappa_\beta - \kappa_{\beta'})^2} & \dfrac{\kappa_\beta e^{2\pi i\beta} - \kappa_{\beta'} e^{2\pi i\beta'}}{\kappa_\beta - \kappa_{\beta'}} \end{bmatrix}. \tag{15.6–51}$$

The substitutions T_a and T_b generate again the monodromy group of Riemann's differential equation. It should be noticed that in Jordan's method the transformation group for P_β/P_α is obtained. Since, however, P_β may be expressed linearly in terms of P_α and $P_{\alpha'}$ the groups obtained by the two methods are conjugate, i.e., identical as abstract groups.

15.6.6 – An alternative proof of Riemann's theorem

The knowledge of the monodromy group as obtained in the previous section enables us to give an independent proof of Riemann's theorem about contiguous functions. Again we need not know the explicit expressions for the solutions of Riemann's differential equation.

First we observe that the quotient $\kappa_\beta/\kappa_{\beta'}$, as given by (15.6–49) is the same for contiguous functions. In fact, in the numerator and the denominator the first gamma function is followed by its complementary function as is the third gamma function. As usual we call the gamma functions $\Gamma(\lambda)$ and $\Gamma(1-\lambda)$ complementary. If one of the exponents increases by one then the variable in the complementary function must be decreased by one. It follows from

$$\Gamma(1+\lambda)\Gamma(-\lambda) = \lambda\Gamma(\lambda)\Gamma(-\lambda) = -\Gamma(\lambda)\Gamma(1-\lambda)$$

that the product of two complementary functions changes sign. On the other hand each exponent occurs twice in the numerator as well as in the denominator and hence the quotient $\kappa_\beta/\kappa_{\beta'}$ remains unchanged if we pass from a function to one of its contiguous functions.

As a consequence in (15.6–50) and (15.6–51) the exponents appear only through the exponential multipliers, so that contiguous functions have the same exponential multipliers. Thus

Contiguous functions have the same monodromy group.

We consider three functions, characterized by the schemes

$$P_i = \begin{bmatrix} a & b & c \\ \alpha_i & \beta_i & \gamma_i \\ \alpha_i' & \beta_i' & \gamma_i' \end{bmatrix}, \qquad i = 1, 2, 3, \qquad (15.6\text{–}52)$$

whose corresponding exponents differ by an integer. The meaning of the scheme (15.6–52) is that every function P_i has singular points at a, b and c and at these points resp. the exponents α_i, α_i', β_i, β_i' and γ_i, γ_i'. We shall study these schemes more elaborately in section 16.1.1. The 3×6 branches

$$P_{i,\alpha_i}, P_{i,\alpha'_i}; P_{i,\beta_i}, P_{i,\beta'_i}; P_{i,\gamma_i}, P_{i,\gamma'_i}, \qquad i = 1, 2, 3, \qquad (15.6\text{–}53)$$

are defined as in the previous section. Without loss of generality we may suppose that the coefficients in (15.6–38) are the same for $i = 1, 2, 3$.

It is clear that the functions

$$Q_{ij} = \begin{vmatrix} P_{i,\alpha_i} & P_{j,\alpha_j} \\ P_{i,\alpha'_i} & P_{j,\alpha'_j} \end{vmatrix} = \begin{vmatrix} a_\beta & a_{\beta'} \\ a'_\beta & a'_{\beta'} \end{vmatrix} \begin{vmatrix} P_{i,\beta_i} & P_{j,\beta_j} \\ P_{i,\beta'_i} & P_{j,\beta'_j} \end{vmatrix} = \begin{vmatrix} a_\gamma & a_{\gamma'} \\ a'_\gamma & a'_{\gamma'} \end{vmatrix} \begin{vmatrix} P_{i,\gamma_i} & P_{j,\gamma_j} \\ P_{i,\gamma'_i} & P_{j,\gamma'_j} \end{vmatrix}$$

(15.6–54)

belong to the lower of the exponents $(\alpha_i+\alpha_j', \alpha_i'+\alpha_j)$ at $z = a$, $(\beta_i+\beta_j', \beta_i'+\beta_j)$ at $z = b$ and $(\gamma_i+\gamma_j', \gamma_i'+\gamma_j)$ at $z = c$. We shall denote them by α_{ij}, β_{ij} and γ_{ij} respectively.

We observe that the smallest of the two real numbers x and y may be expressed as

$$\min(x, y) = \tfrac{1}{2}(x+y-|x-y|).$$

Hence

$$\begin{aligned}\alpha_{ij} &= \tfrac{1}{2}(\alpha_i+\alpha_j+\alpha_i'+\alpha_j'-|\alpha_i-\alpha_j-(\alpha_i'-\alpha_j')|),\\ \beta_{ij} &= \tfrac{1}{2}(\beta_i+\beta_j+\beta_i'+\beta_j'-|\beta_i-\beta_j-(\beta_i'-\beta_j')|),\\ \gamma_{ij} &= \tfrac{1}{2}(\gamma_i+\gamma_j+\gamma_i'+\gamma_j'-|\gamma_i-\gamma_j-(\gamma_i'-\gamma_j')|).\end{aligned} \tag{15.6–55}$$

By hypothesis the numbers α_i and α_j differ by an integer. We may express this by

$$\alpha_i \equiv \alpha_j \pmod 1.$$

Hence

$$\begin{aligned}\alpha_i+\alpha_j &\equiv 2\alpha_i \equiv 2\alpha_j \pmod 1,\\ \alpha_i-\alpha_j &\equiv 0 \pmod 1,\end{aligned}$$

whence

$$\begin{aligned}\alpha_{ij} &\equiv \alpha_i+\alpha_i' \equiv \alpha_j+\alpha_j' \pmod 1,\\ \beta_{ij} &\equiv \beta_i+\beta_i' \equiv \beta_j+\beta_j' \pmod 1,\\ \gamma_{ij} &\equiv \gamma_i+\gamma_i' \equiv \gamma_j+\gamma_j' \pmod 1.\end{aligned} \tag{15.6–56}$$

In view of (15.6–4) we have

The sum of the exponents (15.6–55)

$$\alpha_{ij}+\beta_{ij}+\gamma_{ij} \tag{15.6–57}$$

is an integer.

In fact, this sum is $\equiv 0 \pmod 1$.

From (15.6–55) follows that this sum is also

$$\tfrac{1}{2}(2-|\alpha_i-\alpha_j-(\alpha_i'-\alpha_j')|-|\beta_i-\beta_j-(\beta_i'-\beta_j')|-|\gamma_i-\gamma_j-(\gamma_i'-\gamma_j')|)$$

and thus the sum turns out to be negative or zero.

From (15.6–56) we deduce

$$\alpha_{ij}-\alpha_{jk} \equiv \alpha_i+\alpha_i'-(\alpha_j+\alpha_j') \equiv 0 \pmod 1.$$

Hence

The exponents (15.6–55) *of a triad* $(\alpha_{ij}) = (\alpha_{12}, \alpha_{23}, \alpha_{31})$ *differ among themselves by an integer.*

Going back to (15.6–54) we readily see that the functions

$$(z-a)^{-\alpha_{ij}}(z-b)^{-\beta_{ij}}(z-c)^{-\gamma_{ij}}Q_{ij} \tag{15.6–58}$$

are regular at $z = a, b, c$ and have the exponent $\alpha_{ij}+\beta_{ij}+\gamma_{ij}$ at $z = \infty$. Since the Q_{ij} are regular at $z = \infty$, it follows that the functions (15.6–58) are polynomials of degree

$$N_{ij} = -(\alpha_{ij}+\beta_{ij}+\gamma_{ij}).$$

Up to now it is tacitly assumed that $z = \infty$ is not a singular point. Should this be the case, then we may apply a linear transformation to get the case considered above.

Expressed in terms of the functions (15.6–53) the functions (15.6–52) take the form

$$P_i = AP_{i,\alpha_i}+A'P_{i,\alpha'_i}, \quad i = 1, 2, 3, \tag{15.6–59}$$

with the same coefficients A, A'. They are connected by the relation

$$P_1 Q_{23}+P_2 Q_{31}+P_3 Q_{12} = 0, \tag{15.6–60}$$

or

$$\begin{aligned} &P_1(z-a)^{\alpha_{23}}(z-b)^{\beta_{23}}(z-c)^{\gamma_{23}}f_{23}+\\ &+P_2(z-a)^{\alpha_{31}}(z-b)^{\beta_{31}}(z-c)^{\gamma_{31}}f_{31}+\\ &+P_3(z-a)^{\alpha_{12}}(z-b)^{\beta_{12}}(z-c)^{\gamma_{12}}f_{12} = 0, \end{aligned}$$

where the f_{ij} are polynomials of degree N_{ij}.

If now (α, β, γ) are the lowest in the triads (α_{ij}), (β_{ij}) and (γ_{ij}) we have, in view of the last theorem stated above, that the numbers $\alpha_{ij}-\alpha$, $\beta_{ij}-\beta$, $\gamma_{ij}-\gamma$ are (non-negative) integers. We can remove a factor $(z-a)^\alpha$ $(z-b)^\beta$ $(z-c)^\gamma$ and find a relation

$$P_1 g_{23}+P_2 g_{31}+P_3 g_{12} = 0,$$

where the g_{ij} are polynomials of degree $N_{ij}+\alpha_{ij}-\alpha+\beta_{ij}-\beta+\gamma_{ij}-\gamma$. This concludes the proof of Riemann's theorem.

15.7 – Lamé's differential equation

15.7.1 – Lamé's equation

In certain problems of mathematical physics, namely in problems related to the theory of ellipsoidal harmonics, *Lamé's equation*

$$\boxed{w''-(h+n(n+1)\wp(z))w = 0} \tag{15.7–1}$$

plays an important part. Here $\wp(z)$ denotes the Weierstrass pe function studied in paragraph 5.3, h is an arbitrary constant and n is a positive integer.

This equation has regular singularities at the poles of $\wp(z)$, namely at the points $2m\omega+2m'\omega'$ (m and m' being integers) constituting a lattice in the z-plane. Because of the double periodicity of $\wp(z)$ it is sufficient to investigate the equation in a neighbourhood of $z = 0$.

With reference to (5.3–4) we see that the indicial equation (15.2–20) at $z = 0$ (and hence at every lattice point) is

$$\rho(\rho-1)-n(n+1) = 0,$$

with roots

$$\rho_0 = n+1, \qquad \rho_1 = -n.$$

The general Fuchsian theory (last theorem of section 15.2.2) ensures the existence of at least one regular solution

$$w_0(z) = z^{n+1}\varphi_0(z), \tag{15.7–2}$$

where $\varphi_0(z)$ is regular at $z = 0$.

The difference of the roots of the indicial equation is $2n+1$ and, this being an integer, the possibility of a second solution containing a logarithmic term has to be considered. Since (15.7–1) is invariant when we change the sign of z, one solution contains only odd and the other only even powers of z. Thus both solutions of a fundamental system are free from logarithms everywhere. This may be verified by computation. By the general method described in section 15.2.2 we may find a second solution by solving the equation (15.2–13), viz.,

$$w'+2\frac{w_0'(z)}{w_0(z)}w = 0, \tag{15.7–3}$$

having a solution $s'(z)$, with $s(z) = w_1(z)/w_0(z)$. The general solution of (15.7–3) is

$$w(z) = \frac{C}{w_0^2(z)},$$

C being a constant and, consequently,

$$w_1(z) = Cw_0(z)\int_{z_0}^{z}\frac{dt}{w_0^2(t)}, \tag{15.7–4}$$

z and z_0 being in a neighbourhood of $z = 0$. Since $1/w_0^2(z)$ is an even function, its residue at $z = 0$ is zero and, therefore, a logarithmic term cannot arise.

15.7.2 – The case $n = 1$.

If $n = 1$ the equation (15.7–1) becomes

$$w''-(h+2\wp(z))w = 0. \tag{15.7–5}$$

A solution may be obtained in terms of the Weierstrass sigma functions, introduced in section 4.12.2. Let a denote an arbitrary number, not being a period. It follows from (4.12–7) that the logarithmic derivative of the function

$$\varphi(z) = \frac{\sigma(z+a)}{\sigma(z)\sigma(a)}\exp(-z\zeta(a)) \tag{15.7–6}$$

is equal to

$$\frac{\varphi'(z)}{\varphi(z)} = \zeta(z+a)-\zeta(z)-\zeta(a), \tag{15.7–7}$$

where $\zeta(z)$ is the Weierstrass zeta function. Differentiating both members of (15.7–7) we get in view of (5.5–8)

$$\frac{\varphi''(z)}{\varphi(z)} - \left(\frac{\varphi'(z)}{\varphi(z)}\right)^2 = \wp(z) - \wp(z+a). \tag{15.7–8}$$

From (5.5–16) we infer that the expression on the right of (15.7–7) is

$$\frac{1}{2}\frac{\wp'(z)-\wp'(a)}{\wp(z)-\wp(a)}$$

and (5.5–3) may be read as

$$\wp(z+a)+\wp(z)+\wp(a) = \frac{1}{4}\left(\frac{\wp'(z)-\wp'(a)}{\wp(z)-\wp(a)}\right)^2.$$

It follows from (15.7–8) that

$$\frac{\varphi''(z)}{\varphi(z)} = \wp(a)+2\wp(z).$$

Hence $\varphi(z)$ is a solution of (15.7–5) if a satisfies the equation

$$\wp(a) = h. \tag{15.7–9}$$

Since also $\wp(-a) = h$, we obtain a second solution if we replace a by $-a$. Thus we have

A fundamental system of solutions of Lamé's equation (15.7–5) *is*

$$\boxed{\begin{aligned} w_0(z) &= \frac{\sigma(z+a)}{\sigma(z)\sigma(a)}\exp\left(-z\zeta(a)\right), \\ w_1(z) &= \frac{\sigma(z-a)}{\sigma(z)\sigma(a)}\exp\left(z\zeta(a)\right), \end{aligned}} \tag{15.7–10}$$

provided that a satisfies the equation (15.7–9) *and is not congruent to the half of a period.*

In fact, if this should be the case, then $w_0(a) \neq 0$ and $w_1(a) = 0$, i.e., the functions $w_0(z)$ and $w_1(z)$ would be linearly independent.

Let us now consider the case that a is congruent to ω_α, $\alpha = 1, 2, 3$, i.e., $h = e_\alpha$ (as follows from (5.6–12)). Without loss of generality we may assume that a is equal to one of the values ω_α. From (5.6–2), (5.6–3) and (5.6–11) follows that in this case the two solutions (15.7–10) coincide. In order to obtain a fundamental system in this case we consider $w_0(z)$ as a function of the variable a, i.e.,

$$f(a) = \frac{\sigma(z+a)}{\sigma(z)\sigma(a)}\exp\left(-z\zeta(a)\right). \tag{15.7–11}$$

Supposing, for the time being, $a \neq \omega_\alpha$, we have

$$\lim_{a \to w_\alpha} \frac{f(a)-f(\omega_\alpha)}{a-w_\alpha} = f'(w_\alpha)$$

and

$$\lim_{a \to w_\alpha} \frac{f(-a)-f(-\omega_\alpha)}{a-\omega_\alpha} = -f'(-\omega_\alpha).$$

Now

$$\frac{f(a)-f(-a)}{2(a-\omega_\alpha)}$$

is a solution of (15.7–5). Taking into account the fact that $f(\omega_\alpha)=f(-\omega_\alpha)$, it appears that

$$\lim_{a \to w_\alpha} \frac{f(a)-f(-a)}{2(a-\omega_\alpha)} = \tfrac{1}{2}(f'(\omega_\alpha)+f'(-\omega_\alpha))$$

is a solution of (15.7–5) if $h = e_\alpha$.

Differentiating (15.7–11) logarithmically with respect to a we get

$$\begin{aligned}\frac{f'(a)}{f(a)} &= \frac{\sigma'(z+a)}{\sigma(z+a)} - \frac{\sigma'(a)}{\sigma(a)} - z\zeta'(a)\\ &= \zeta(z+a)-\zeta(a)-z\zeta'(a).\end{aligned}$$

Making $a \to \omega_\alpha$ and taking into account (5.5–11) and (5.3–16) we find

$$\frac{f'(w_\alpha)}{f(w_\alpha)} = \zeta(z+\omega_\alpha)-\eta_\alpha+ze_\alpha,$$

or

$$f'(\omega_\alpha) = f(\omega_\alpha)\,(\zeta(z+\omega_\alpha)-\eta_\alpha+ze_\alpha).$$

Again

$$\begin{aligned}f'(-\omega_\alpha) &= f(\omega_\alpha)\,(\zeta(z-\omega_\alpha)+\eta_\alpha+ze_\alpha),\\ &= f(\omega_\alpha)(\zeta(z+\omega_\alpha)+ze_\alpha-\eta_\alpha),\end{aligned}$$

by (5.5–10). Thus we have (in view of (5.6–12))

If $h = e_\alpha$, $\alpha = 1, 2, 3$ *a fundamental system of solutions is*

$$\begin{aligned}w_0(z) &= \frac{\sigma_\alpha(z)}{\sigma(z)},\\ w_1(z) &= \frac{\sigma_\alpha(z)}{\sigma(z)}(\zeta(z+\omega_\alpha)+ze_\alpha).\end{aligned} \tag{15.7–12}$$

If h has not an exceptional value the functions of the fundamental system (15.7–10) are not doubly periodic, but consist of a doubly periodic function multiplied by an exponential factor. If, however, h has one of the exceptional values e_α the first solution of (15.7–12) is doubly periodic, but the second solution is not periodic.

It can be proved that if n is an arbitrary positive integer Lamé's equation is solved by

$$w(z) = \prod_{\nu=1}^{n} \frac{\sigma(z+a_\nu)}{\sigma(z)\sigma(a_\nu)} \exp\left(-z\zeta(a_\nu)\right),$$

where $a_1, \ldots, a_n$ are constants to be determined.

15.7.3 – Alternative forms of Lamé's equation

If we take $z_1 = \wp(z)$ as a new variable we easily find

$$\frac{d^2w}{dz_1^2}\wp'^2 + \frac{dw}{dz_1}\wp'' - (h+n(n+1)\wp) = 0.$$

From (5.4–1) we obtain by logarithmic differentiation

$$\frac{\wp''}{\wp'} = \frac{1}{2}\left(\frac{1}{\wp-e_1} + \frac{1}{\wp-e_2} + \frac{1}{\wp-e_3}\right)\wp'$$

and Lamé's equation appears in the algebraic form (omitting afterwards the subscript from z_1)

$$\boxed{w'' + \frac{1}{2}\left(\frac{1}{z-e_1} + \frac{1}{z-e_2} + \frac{1}{z-e_3}\right)w' - \frac{h+n(n+1)z}{4(z-e_1)(z-e_2)(z-e_3)}w = 0.}$$

(15.7–13)

This is a Fuchsian equation with singularities at e_1, e_2, e_3 and ∞. It is easily verified that the roots of the indicial equation are 0, $\frac{1}{2}$ at the finite singular points and $\frac{1}{2}(n+1)$, $-\frac{1}{2}n$ at $z = \infty$.

Another form of Lamé's equation is obtained by introducing the Jacobian elliptic functions. Let us write

$$z_1 = z\sqrt{e_1-e_3}$$

where the square root has been defined in (5.7–5). Then

$$\frac{d^2w}{dz^2} = \frac{d^2w}{dz_1^2}(e_1-e_3).$$

From (5.14–19) follows

$$\wp(z)-e_3 = \frac{e_1-e_3}{\mathrm{sn}^2\, z\sqrt{e_1-e_3}} = \frac{e_1-e_3}{\mathrm{sn}^2\, z_1}$$

and if we take $z_2 = z_1 + i\mathrm{K}'$ we get in view of (5.16–11)

$$\wp(z) - e_3 = (e_1 - e_3)k^2 \operatorname{sn}^2 z_2 .$$

Hence the differential equation (15.7–1) takes the form

$$\frac{d^2w}{dz_2^2} - \left(\frac{h}{e_1-e_3} + n(n+1)\left(\frac{e_3}{e_1-e_3} + k^2 \sin^2 z_2\right)\right) w = 0.$$

Omitting again the subscript from z_2 we finally have

$$\boxed{w'' - (A + n(n+1)k^2 \operatorname{sn}^2 z)w = 0,} \qquad (15.7\text{–}14)$$

the *Jacobian form* of Lamé's equation.

15.8 – Mathieu's equation

15.8.1 – Mathieu's equation as a limiting case of Lamé's equation

In the Jacobian form (15.7–14) of Lamé's equation we let

$$B = n(n+1)k^2$$

fixed as $k \to 0$ and $n \to \infty$. In the elliptic functions $2\mathrm{K} \to \pi$ and $2i\mathrm{K}' \to \infty$. for

$$\mathrm{K} = \int_0^{\frac{1}{2}\pi} \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}}, \qquad \mathrm{K}' = \int_0^{\frac{1}{2}\pi} \frac{d\theta}{\sqrt{1-k'^2\sin^2\theta}}, \qquad k'^2 = 1-k^2.$$

Hence the singularities, which are situated at the points $2m\mathrm{K} + (2m+1)i\mathrm{K}'$ (m and m' being integers) recede to infinity and we are left with the equation

$$w'' - (A + B\sin^2 z)w = 0. \qquad (15.8\text{–}1)$$

By introducing other constants this may be written as

$$\boxed{w'' + (\lambda - 2h^2 \cos 2z)w = 0.} \qquad (15.8\text{–}2)$$

This is *Mathieu's equation.*

The process which effects the coincidence of two or more singular points of a differential equation of the Fuchsian type is called *confluence*. Hence Mathieu's equation is a confluent equation of Lamé.

The theory of Mathieu's equation is rather involved and has given rise to an extensive litterature. In this paragraph we must confine ourselves to a sketchy treatment.

15.8.2 – Floquet's theory

The following analysis is aplicable to any linear differential equation with periodic coefficients which have only isolated singularities in the finite plane.

Let $w_0(z)$, $w_1(z)$ be a fundamental system of solutions at a regular point of any linear differential equation in which the coefficients have the period 2π. Since $w_0(z+2\pi)$ and $w_1(z+2\pi)$ are obviously solutions of the equation, they can be expressed in terms of the continuations of $w_0(z)$ and $w_1(z)$ by a matrix equation of the type

$$\begin{bmatrix} w_0(z+2\pi) \\ w_1(z+2\pi) \end{bmatrix} = \mathsf{A} \begin{bmatrix} w_0(z) \\ w_1(z) \end{bmatrix}. \tag{15.8–3}$$

Now we ask as in section 15.1.5 whether there is a *multiplicative solution*

$$v(z) = c_0 w_0(z) + c_1 w_1(z),$$

with

$$v(z+2\pi) = \sigma v(z). \tag{15.8–4}$$

Just as in section 15.1.5 we see that the number σ must satisfy the characteristic equation

$$\det(\mathsf{A} - \sigma \mathsf{E}) = 0. \tag{15.8–5}$$

Defining μ by the equation $\sigma = e^{2\pi\mu}$ and writing $\varphi(z)$ for $e^{-\mu z} v(z)$ we see that

$$\varphi(z+2\pi) = e^{-\mu(z+2\pi)} v(z+2\pi) = \varphi(z). \tag{15.8–6}$$

Thus we have proved *Floquet's theorem*

The differential equation (15.1–9) *where $p(z)$ and $q(z)$ are periodic with period 2π has a particular solution* $e^{\mu z}\varphi(z)$, *where $\varphi(z)$ is a periodic function with period 2π.*

In the particular case of Mathieu's equation a fundamental system of solutions is then

$$\begin{aligned} w_0(z) &= e^{\mu z}\varphi(z), \\ w_1(z) &= e^{-\mu z}\varphi(-z), \end{aligned} \tag{15.8–7}$$

since the equation is unaltered by writing $-z$ for z. Here μ is a definite function of λ and h^2.

15.8.3 – Mathieu's functions

The equation (15.8–2) has no finite singular points and, therefore, its solutions are valid for all finite values of z. If $w(z, \lambda)$ denotes a solution, so does $w(-z, \lambda)$ and it follows that

$$\tfrac{1}{2}(w(z, \lambda) + w(-z, \lambda)),\ \tfrac{1}{2}(w(z, \lambda) - w(-z, \lambda))$$

are again solutions. But the first is an even function and the second is an odd function. Thus it is sufficient to consider only even or odd solutions.

Two independent even solutions and, likewise, two independent odd solutions cannot exist.

Indeed, if the equation possessed two independent even solutions a solution satisfying the initial conditions $w(0) = 0$, $w'(0) = 1$ would not exist, which is in contradiction to the fact that the origin is an ordinary point. A similar argument shows that two independent odd solutions cannot exist.

A general theorem in the theory of linear differential equations of the second order in the real domain, the so called *oscillation theorem*, applied to Mathieu's equation states

If h is real there exists a non-decreasing sequence $\lambda_1, \lambda_2, \ldots,$ of values of the parameter λ tending to infinity, such that to each λ_m, $m = 1, 2, \ldots,$ corresponds a periodic solution $w(z, \lambda_m)$ with period 2π and satisfying the boundary conditions

$$w(-\pi, \lambda_m) = w(\pi, \lambda_m), \qquad w'(-\pi, \lambda_m) = w'(\pi, \lambda_m)$$

and every $w(z, \lambda_m)$ is characterized by the fact that in $-\pi < z < \pi$ it possesses m zeros or $m+1$ zeros, according as m is even or odd.

These values of λ are called the *characteristic values.*

The periodic solutions of Mathieu's equation with period 2π are called the *Mathieu functions of the first kind.*

Let us consider an even solution of (15.8–2) with period 2π. Suppose it is expanded in a cosine series (uniformly convergent in any bounded and closed set in the z-plane). If $W(z)$ is a solution so is $w(z+\pi)$. Since the equation cannot have two independent even solutions the functions $w(z+\pi)$ and $w(z)$ differ only in a multiplicative constant. It follows that the series are of the form

$$\sum_{\nu=0}^{\infty} c_\nu \cos(2\nu+1)z, \qquad \sum_{\nu=0}^{\infty} c_\nu \cos 2\nu z. \tag{15.8–8}$$

Similarly the odd solutions have the form

$$\sum_{\nu=0}^{\infty} c'_\nu \sin(2\nu+1)z, \qquad \sum_{\nu=0}^{\infty} c'_\nu \sin 2\nu z. \tag{15.8–9}$$

These solutions are said to be of the type C_1, C_0, S_1, S_0 respectively.

A degenerate form of Mathieu's equation is obtained if we take $h = 0$ viz.,

$$w'' + \lambda w = 0 \tag{15.8–10}$$

which admits of the succession of characteristic values $\lambda = m^2$, $m = 0$, 1, 2, . . ., and corresponding solutions are

$$\begin{array}{llll} 1, & \cos z, & \cos 2z, & \cos 3z, \ldots, \\ & \sin z, & \sin 2z, & \sin 3z, \ldots. \end{array}$$

The Mathieu functions which reduce to these functions as $h \to 0$ are denoted by

$$\begin{array}{llll} \mathrm{ce}_0(z), & \mathrm{ce}_1(z), & \mathrm{ce}_2(z), & \mathrm{ce}_3(z), \ldots, \\ & \mathrm{se}_1(z), & \mathrm{se}_2(z), & \mathrm{se}_3(z), \ldots \end{array}$$

and their expansions are

$$\begin{aligned} \mathrm{ce}_{2n+1}(z) &= \sum_{\nu=0}^{\infty} A_{n,\,2\nu+1} \cos(2\nu+1)z, & A_{n,\,2n+1} &= 1, \\ \mathrm{ce}_{2n}(z) &= \sum_{\nu=0}^{\infty} A_{n,\,2\nu} \cos 2\nu z, & A_{n,\,2n} &= 1, \\ \mathrm{se}_{2n+1}(z) &= \sum_{\nu=0}^{\infty} B_{n,\,2\nu+1} \sin(2\nu+1)z, & B_{n,\,2n+1} &= 1, \\ \mathrm{se}_{2n}(z) &= \sum_{\nu=0}^{\infty} B_{n,\,2\nu} \sin 2\nu z, & B_{n,\,2n} &= 1. \end{aligned} \tag{15.8–11}$$

15.8.4 – Recurrence relations

If a characteristic value of λ is known we may obtain the corresponding periodic solutions from recurrence relations connecting the coefficients in the expansions (15.8—8) or (15.9–9). Consider e.g., a solution of the type

$$C_1(z) = \sum_{\nu=0}^{\infty} c_\nu \cos(2\nu+1)z. \tag{15.8–12}$$

Differentiating two times

$$C_1''(z) = -\sum_{\nu=0}^{\infty} (2\nu+1)^2 c_\nu \cos(2\nu+1)z. \tag{15.8–13}$$

Inserting this into (15.8–2) we get

$$\sum_{\nu=0}^{\infty} (2\nu+1)^2 c_\nu \cos(2\nu+1)z - \lambda \sum_{\nu=0}^{\infty} c_\nu \cos(2\nu+1)z +$$
$$+h^2 \sum_{\nu=0}^{\infty} c_\nu \cos(2\nu+3)z + h^2 \sum_{\nu=0}^{\infty} c_\nu \cos(2\nu-1)z = 0$$

and we easily deduce the recurrence relations

$$\begin{gathered} (\lambda-1-h^2)c_0 - h^2 c_1 = 0, \\ ((2n+1)^2-\lambda)c_n + h^2(c_{n-1}+c_{n+1}) = 0, \qquad n = 1, 2, \ldots. \end{gathered} \tag{15.8–14}$$

Similarly we have for a solution of the type C_0

$$C_0(z) = \sum_{\nu=0}^{\infty} c_\nu \cos 2\nu z \tag{15.8–15}$$

the recurrence relations

$$\begin{aligned} \lambda c_0 - h^2 c_1 &= 0, \\ ((2n)^2 - \lambda)c_n + h^2(c_{n-1} + c_{n+1}) &= 0, \qquad n = 1, 2, \ldots. \end{aligned} \tag{15.8–16}$$

In the same way a solution of the type S_1 exists:

$$S_1(z) = \sum_{\nu=0}^{\infty} c'_\nu \sin (2\nu+1)z, \tag{15.8–17}$$

with the recurrence relations

$$\begin{aligned} (\lambda - 1 + h^2)c'_0 - h^2 c'_1 &= 0, \\ ((2n+1)^2 - \lambda)c'_n - h^2(c'_{n-1} + c'_{n+1}) &= 0, \qquad n = 1, 2, \ldots, \end{aligned} \tag{15.8–18}$$

and a solution of the type S_0:

$$S_0(z) = \sum_{\nu=0}^{\infty} c'_\nu \sin 2\nu z, \tag{15.8–19}$$

with the recurrence relations

$$\begin{aligned} (\lambda - 4)c'_1 - h^2 c'_2 &= 0, \\ ((2n)^2 - \lambda)c'_n + h^2(c'_{n-1} + c'_{n+1}) &= 0, \qquad n = 2, 3, \ldots. \end{aligned} \tag{15.8–20}$$

The recurrence relations may also serve to evaluate the characteristic values. We wish to illustrate this in the example of $\mathrm{ce}_0(z)$. According to (15.8–11) and (15.8–16) we now have

$$\begin{aligned} \lambda A_{00} - h^2 A_{02} &= 0, \\ -h^2 A_{00} + (\lambda - 4)A_{02} - h^2 A_{04} &= 0, \\ -h^2 A_{02} + (\lambda - 16)A_{04} - h^2 A_{06} &= 0. \end{aligned}$$

Now these equations must be consistent: the condition for their consistency is

$$D = \begin{vmatrix} \lambda & -h^2 & 0 & 0 & 0 & \ldots \\ -h^2 & \lambda-4 & -h^2 & 0 & 0 & \ldots \\ 0 & -h^2 & \lambda-16 & -h^2 & 0 & \ldots \\ 0 & 0 & -h^2 & \lambda-36 & -h^2 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{vmatrix} = 0.$$

As it stands this infinite determinant needs not to be convergent; it may, however, be made absolutely convergent by multiplying each row by an appropriate factor. By $D_1, D_2, \ldots$ we denote the determinants obtained by omitting the first row and column, the first two rows and columns, etc. Then

$$D = \lambda D_1 - h^4 D_2, \qquad D_1 = (\lambda-4)D_2 - h^4 D_3, \ldots.$$

It follows that

$$\lambda = h^4 \frac{D_2}{D_1}, \qquad \frac{D_1}{D_2} = (\lambda-4) - h^4 \frac{D_3}{D_2}, \ldots$$

and in this way we obtain the continuous fraction

$$\lambda = \frac{h^4 |}{\lambda-4} - \frac{h^4 |}{\lambda-16} - \frac{h^4 |}{\lambda-36} - \ldots \tag{15.8–21}$$

and it can be proved that the fraction the right is convergent. Given h^2 the number λ may be found by successive approximation.

15.8.5 – Ince's theorem

In the case of degenerence $h = 0$, to the same value $\lambda = m^2$ correspond two periodic solutions cos mz and sin mz, which constitute a fundamental system. Mathematicians were during a long time under the impression that also Mathieu's equation may admit of a fundamental system of two periodic solutions corresponding to appropriate values of λ and h. This view is not correct as has been proved by E. L. Ince in a very elegant way.

Suppose that $h \neq 0$ and let λ be such that Mathieu's equation has a solution of the type C_1. If $w_1(z)$ and $w_2(z)$ are independent solutions of (15.8–2) we have

$$w_1 w_2'' - w_2 w_1'' = 0,$$

or

$$w_1 w_2' - w_2 w_1' = c, \tag{15.8–22}$$

c being a constant. It follows that if w_1 is of type C_1 then w_2 is of type S_1 and not of type S_0 (provided w_2 is also periodic). From the first equations (15.8–14) and (15.8–18) we find by eliminating h

$$\begin{vmatrix} c_0 & c_1 \\ c_0' & c_1' \end{vmatrix} = 2c_0 c_0'.$$

Similarly the second equations give

$$c_n(c_{n-1}' + c_{n+1}') = c_n'(c_{n-1} + c_{n+1}),$$

or

$$\begin{vmatrix} c_n & c_{n+1} \\ c_n' & c_{n+1}' \end{vmatrix} = \begin{vmatrix} c_{n-1} & c_n \\ c_{n-1}' & c_n' \end{vmatrix},$$

whence for all values of n

$$\begin{vmatrix} c_n & c_{n+1} \\ c'_n & c'_{n+1} \end{vmatrix} = 2c_0 c'_0 .$$

If $c_0 = 0$ the remaining coefficients are zero and the solution is identically zero. Therefore $c_0 \neq 0$ and similarly $c'_0 \neq 0$. But in order that the series may be convergent it is necessary that

$$\lim_{n\to\infty} c_n = 0$$

which leads to a contradiction. Thus, except when $h = 0$, solutions of the type C_1 and S_1 cannot exist simultaneously. In the same way it may be proved that solutions of the type C_0 and S_0 do not exist simultaneously. Thus we have *Ince's theorem*

Except in the case that $h = 0$ *for every* h *in the equation* (15.8–2) *to every value of* λ *corresponds at most only one periodic solution.*

Assuming $h \neq 0$ we conclude that if one solution $w_1(z)$ is periodic the second solution $w_2(z)$ is definitely aperiodic. From (15.8–22) we get

$$w_2(z) = cw_1(z)\int_{z_0}^{z} \frac{d\zeta}{w_1^2(\zeta)} .$$

Now let

$$w_1(z) = C_1(z) = \sum_{\nu=0}^{\infty} c_\nu \cos(2\nu+1)z.$$

Then

$$w_1^2(z) = \sum_{\nu=0}^{\infty} a_\nu \cos 2\nu z$$

and since $w_1(z)$ is not zero at $z = 0$,

$$1/w_1^2(z) = \sum_{\nu=0}^{\infty} b_\nu \cos 2\nu z,$$

this series being convergent at least for sufficiently small values of $|z|$. Consequently

$$w_2(z) = c\Big(\sum_{\nu=0}^{\infty} c_\nu \cos(2\nu+1)z\Big)\Big(b_0 z + \sum_{\nu=1}^{\infty} d_\nu \sin 2\nu z\Big),$$

where, since $w_2(z)$ is known to be not periodic, b_0 is not zero. Therefore, with an appropriate choice of c

$$w_2(z) = zC_1(z) + S_1^*(z),$$

where $S_1^*(z)$ is a series of the type S_1. Thus $w_2(z)$ is not periodic, but quasi-periodic

$$w_2(z+2\pi) = w_2(z) + 2\pi w_1(z). \tag{15.8–23}$$

The nature of the second solution, when the first solution is of the type S_1, C_0, S_0 may be investigated in the same way.

The aperiodic functions associated with the periodic solutions in the above described manner are called *Mathieu's functions of the second kind.* They are denoted by

$$\begin{aligned}\mathrm{in}_m(z) &= \mathrm{ce}_m(z)\int_{z_0}^{z}\frac{d\zeta}{\mathrm{ce}_m^2(\zeta)},\\ \mathrm{jn}_m(z) &= \mathrm{se}_m(z)\int_{z_0}^{z}\frac{d\zeta}{\mathrm{se}_m^2(z)}, \qquad m = 0, 1, \ldots,\end{aligned} \tag{15.8–24}$$

where the value of z_0 is not very important.

15.8.6 – The integral equation of Whittaker

We conclude our account of Mathieu's equation with the proof of the statement that the Mathieu functions of the first kind satisfy certain integral equations. We confine ourselves to one example, viz., *that any Mathieu function, being either even or odd, satisfies an integral equation with symmetrical kernel*:

$$\varphi(z) = \kappa\int_{-\pi}^{\pi}\exp(2h\sin z\sin\theta)\varphi(\theta)d\theta. \tag{15.8–25}$$

It is advisable to write (15.8–2) as

$$w''+(\lambda-4h^2\cos^2 z)w = 0, \tag{15.8–26}$$

where λ is written rather than $\lambda+2h^2$. Now we consider the function

$$\Phi(z) = \int_{-\pi}^{\pi}\exp(2h\sin z\sin\theta)\varphi(\theta)d\theta,$$

where $\varphi(\theta)$ is regular between $-\pi$ and π, periodic with period 2π, and either even or odd. Then we have

$$\Phi''(z) - 4h^2\,\Phi(z)\cos^2 z = \int_{-\pi}^{\pi}\exp(2h\sin z\sin\theta)\times$$
$$\times(4h^2\cos^2 z\sin^2\theta-2h\sin z\sin\theta-4h^2\cos^2 z)\varphi(\theta)d\theta.$$

Since

$$\begin{aligned}\cos^2 z\sin^2\theta-\cos^2 z &= (1-\sin^2 z)(1-\cos^2\theta)-\cos^2 z\\ &= \sin^2 z\cos^2\theta-\cos^2\theta,\end{aligned}$$

the integral may be written as

$$\int_{-\pi}^{\pi} \exp(2h \sin z \sin\theta)(4h^2 \sin^2 z \cos^2\theta - 2h \sin z \sin\theta +$$
$$-4h^2 \cos^2\theta)\varphi(\theta)d\theta$$
$$= \int_{-\pi}^{\pi} \varphi(\theta) \frac{d}{d\theta}(2h \sin z \cos\theta \exp(2h \sin z \sin\theta)d\theta +$$
$$-\int_{-\pi}^{\pi} h^2 \varphi(\theta)\cos^2\theta d\theta = 2h\varphi(\theta) \sin z \cos\theta \exp(2h \sin z \sin\theta)\Big|_{-\pi}^{\pi} +$$
$$-\int_{-\pi}^{\pi} \varphi'(\theta)(2h \sin z \cos\theta \exp(2h \sin z \sin\theta)d\theta +$$
$$-\int_{-\pi}^{\pi} 4h^2 \varphi(\theta)\cos^2\theta d\theta.$$

Since $\varphi(\theta)$ is periodic the integrated part vanishes. Integrating again by parts we get

$$-\varphi'(\theta)\exp(2h \sin z \sin\theta)\Big|_{-\pi}^{\pi} +$$
$$+\int_{-\pi}^{\pi} \exp(2h \sin z \sin\theta)(\varphi'' - 4h^2 \varphi(\theta)\cos^2\theta)d\theta.$$

By hypothesis $\varphi(\theta)$ is either even or odd. In the first case $\varphi'(\pi) = \varphi'(-\pi) = 0$, in the second case $\varphi'(\pi) = \varphi'(-\pi)$.

Hence

$$\Phi''(z) + (\lambda - 4h^2 \cos^2 z)\Phi(z)$$
$$= \int_{-\pi}^{\pi} \exp(2h \sin z \sin\theta)(\varphi'' + (\lambda - 4h^2 \cos^2\theta)\varphi(\theta)d\theta.$$

Let now λ have a characteristic value and let φ denote an even or odd solution of Mathieu's equation. It follows that the integral on the right vanishes and that Φ is also a solution of Mathieu's equation, being even or odd according as φ is even or odd. Hence $\varphi = \kappa\Phi$ and this proves the assertion.

It is easily verified that Φ is not identically zero if φ does not vanish identically, for

$$\exp(2h \sin z \sin\theta) = \sum_{\nu=0}^{\infty} \frac{(2h)^\nu}{\nu!} \sin^\nu z \sin^\nu\theta$$

and since the series on the right is uniformly convergent if $-\pi \leqq \theta \leqq \pi$, we would have from $\Phi(z) = 0$ identically

$$0 = \sum_{\nu=0}^{\infty} \frac{(2h)^\nu}{\nu!} \sin^\nu z \int_{-\pi}^{\pi} \varphi(\theta)\sin^\nu\theta\, d\theta.$$

The series on the right is a power series in sin z and it follows that

$$\int_{-\pi}^{\pi} \varphi(\theta) \sin^n \theta \, d\theta = 0, \qquad n = 0, 1, 2, \ldots.$$

By the formulas mentioned at the end of section 3.13.5, written as

$$(-1)^n 2^{2n-1} \sin^{2n} \theta$$

$$= \cos 2n\theta - \binom{2n}{1} \cos (2n-2)\theta + \binom{2n}{2} \cos (2n-4)\theta + \ldots + (-1)^n \tfrac{1}{2} \binom{2n}{n}$$

and

$$(-1)^n 2^{2n} \sin^{2n+1} \theta$$

$$= \sin (2n+1)\theta - \binom{2n+1}{1} \sin (2n-1)\theta +$$

$$+ \binom{2n+1}{2} \sin (2n-3)\theta + \ldots + (-1)^n \binom{2n+1}{n} \sin \theta,$$

we find that

$$\int_{-\pi}^{\pi} \varphi(\theta) \cos 2n\theta \, d\theta = 0, \qquad \int_{-\pi}^{\pi} \varphi(\theta) \sin (2n+1)\theta \, d\theta = 0.$$

By hypothesis $\varphi(\theta)$ is either even or odd and we see that under the assumption that $\Phi(z) = 0$ identically all its Fourier coefficients vanish. By a well-known theorem in the theory of Fourier series this is absurd if $\varphi(\theta)$ does not vanish identically.

In quite the same way we may prove

Any even Mathieu function satisfies the integral equation

$$\varphi(z) = \kappa \int_{-\pi}^{\pi} \exp (2ih \cos z \cos \theta) \varphi(\theta) d\theta \qquad (15.8\text{–}27)$$

and any odd Mathieu function satisfies the integral equation

$$\varphi(z) = \kappa \int_{-\pi}^{\pi} \exp (2ih \sin z \sin \theta) \varphi(\theta) d\theta. \qquad (15.8\text{–}28)$$

These integral equations may be used to construct Mathieu functions.

Chapter 16

THE HYPERGEOMETRIC DIFFERENTIAL EQUATION

16.1 – The hypergeometric series

16.1.1 – Riemann's scheme

A differential equation of Riemann with singular points at $z = 0, 1, \infty$ and having an exponent equal to 0 at each of the points $z = 0$ and $z = 1$ is called a *hypergeometric differential equation.*

The restriction to the study to this type of differential equations, instead of the consideration of the general Riemann equation, does not imply a loss of generality. For we shall prove that by an appropriate transformation this latter equation can be brought into the hypergeometric form.

In order to study the effect of certain transformations Riemann introduced the scheme

$$P\begin{bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix} \tag{16.1–1}$$

for any solution of Riemann's equation. In this scheme the singular points are placed in the first row with the roots of the corresponding indicial equations beneath them; the independent variable is placed in the fourth column. It should be noticed that (16.1–1) does not stand for a single function, but for the class of all solutions of a differential equation. An equality

$$P\begin{bmatrix} a_1 & b_1 & c_1 & \\ \alpha_1 & \beta_1 & \gamma_1 & z_1 \\ \alpha_1' & \beta_1' & \gamma_1' & \end{bmatrix} = P\begin{bmatrix} a_2 & b_2 & c_2 & \\ \alpha_2 & \beta_2 & \gamma_2 & z_2 \\ \alpha_2' & \beta_2' & \gamma_2' & \end{bmatrix}, \qquad z_2 = f(z_1),$$

means that the class of solutions of the differential equation corresponding to the scheme on the left is also the class of solutions of the differential equation corresponding to the scheme on the right, if the change of variable has been effected as indicated. If a solution of one equation is provided by a solution of the other equation by multiplying it by a certain function, the same for all functions of the class represented by the scheme, we shall multiply the Riemannian scheme by this function.

Any solution of Riemann's equation can be obtained by combining linearly independent integrals of the type (15.6–38).

A linear transformation of the variable z transforms a solution (15.6–15) *into a similar solution of the transformed equation with the same exponents.*

Let

$$z \to \mathrm{A}z \tag{16.1–2}$$

denote this linear transformation; we may write symbolically

$$P\begin{bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix} = P\begin{bmatrix} \mathrm{A}a & \mathrm{A}b & \mathrm{A}c & \\ \alpha & \beta & \gamma & \mathrm{A}z \\ \alpha' & \beta' & \gamma' & \end{bmatrix}. \tag{16.1–3}$$

We recall that the meaning of this equality is the following: a solution of the original equation is also a solution in terms of Az of a Riemann equation with singularities at the transformed singular points, the exponents remaining unaltered.

The exponents of the integral in (15.6–15) remain unchanged if we add a number κ to α and α', a number λ to β and β' and subtract $\kappa+\lambda$ from γ and γ'. This leads to

$$\left(\frac{z-a}{z-c}\right)^{\kappa}\left(\frac{z-b}{z-c}\right)^{\lambda} P\begin{bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix} = P\begin{bmatrix} a & b & c & \\ \alpha+\kappa & \beta+\lambda & \gamma-\kappa-\lambda & z \\ \alpha'+\kappa & \beta'+\lambda & \gamma'-\kappa-\lambda & \end{bmatrix}, \tag{16.1–4}$$

By this transformation the fundamental identity (15.6–4) is not violated. It is clear that a permutation of the first three columns has no influence on the meaning of Riemann's scheme (16.1–1). It is also allowed to interchange the exponents of any column.

By means of the transformation

$$z \to \mathrm{A}z = \frac{(z-a)(b-c)}{(z-c)(b-a)} \tag{16.1–5}$$

we obtain, in accordance with (16.1–3),

$$P\begin{bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix} = P\begin{bmatrix} 0 & 1 & \infty & \\ \alpha & \beta & \gamma & \mathrm{A}z \\ \alpha' & \beta' & \gamma' & \end{bmatrix}.$$

Applying now (16.1–4) with $\kappa = -\alpha$, $\lambda = -\beta$, we find

$$P\begin{bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix} = (\mathrm{A}z)^{\alpha}(1-\mathrm{A}z)^{\beta} P\begin{bmatrix} 0 & 1 & \infty & \\ 0 & 0 & \alpha+\beta+\gamma & \mathrm{A}z \\ \alpha'-\alpha & \beta'-\beta & \alpha+\beta+\gamma' & \end{bmatrix} \tag{16.1–6}$$

and it appears that *any solution of the general Riemann equation can be obtained from a solution of the hypergeometric equation.*

In future we shall denote the solutions of a Riemann equation with singularities at $z = 0$, $z = 1$ and $z = \infty$ by the scheme

$$P\begin{bmatrix} \alpha & \beta & \gamma & \\ \alpha' & \beta' & \gamma' & z \end{bmatrix}. \tag{16.1–7}$$

16.1.2 – The hypergeometric differential equation

The hypergeometric differential equation may be obtained from Riemann's equation by making $a \to 0$, $b \to 1$ and $c \to \infty$. But we may also proceed in a straightforward manner to write down a differential equation with singularities at $z = 0, 1, \infty$ and having the appropriate exponents there. By the general theory of Fuchs the coefficient $p(z)$ in (15.1–9) has simple poles at $z = 0$, $z = 1$ and a simple zero at $z = \infty$ (see sections 15.2.2 and 15.5.1) Again $q(z)$ has double poles and a double zero in the corresponding points. Hence we may put

$$p(z) = \frac{A_1}{z} + \frac{A_2}{z-1},$$

$$q(z) = \frac{B_1}{z^2} + \frac{B_2}{(z-1)^2} + \frac{C_1}{z} + \frac{C_2}{z-1},$$

with $C_1 + C_2 = 0$, in accordance with (15.5–7). The indicial equations (15.5–16) and (15.5.17) at $z = 0$, 1 and ∞ are respectively

$$\begin{aligned} \rho(\rho-1)+A_1\rho+B_1 &= 0,\\ \rho(\rho-1)+A_2\rho+B_2 &= 0,\\ \rho(\rho-1)+(2-A_1-A_2)\rho+(B_1+B_2+C_2) &= 0. \end{aligned}$$

It is common practice to denote the exponents at $z = \infty$ by a, b and the exponents at $z = 0$ by 0, $1-c$. There is no danger of confusion with the notation in Riemann's equation, for the singular points are now fixed at $z = 0$, 1 and ∞. In terms of the original exponents we have

$$a = \alpha+\beta+\gamma, \qquad b = \alpha+\beta+\gamma', \qquad c = 1+\alpha-\alpha'. \tag{16.1–8}$$

Since the sum of all exponents is unity and one of the exponents at $z = 1$ is zero, the other is $c-a-b$. Inserting these values of the exponents into the indicial equations we easily find $B_1 = B_2 = 0$, $A_1 = c$, $A_2 = a+$ $+b-c+1$, $C_2 = -C_1 = ab$. Thus the hypergeometric differential equation may be written in the standard form

$$\boxed{z(1-z)w''-((a+b+1)z-c)w'-abw = 0.} \tag{16.1–9}$$

The corresponding Riemann scheme is

$$P\begin{bmatrix} 0 & 1 & a & \\ 1-c & c-a-b & b & z \end{bmatrix}. \tag{16.1–10}$$

It should be noticed that by (16.1–8) and (15.6–4) the differences of the exponents α, α'; β, β' and γ, γ' are determined by a, b and c, for we easily find

$$\alpha'-\alpha = 1-c, \qquad \beta'-\beta = c-a-b, \qquad \gamma'-\gamma = b-a. \tag{16.1–11}$$

Sometimes it is convenient to bring (16.1–9) into a simpler form by introducing the operator

$$\vartheta = z\frac{d}{dz} \tag{16.1–12}$$

which has the neat property that $\vartheta z^\kappa = \kappa z^\kappa$. Since

$$zw' = \vartheta w, \qquad z^2 w'' = \vartheta^2 w - \vartheta w,$$

we find (after multiplying the left member of (16.1–9) by z)

$$(1-z)(\vartheta^2 w - \vartheta w) - (a+b+1)z\vartheta w + c\vartheta w - abzw = 0,$$

or

$$\boxed{\vartheta(\vartheta+c-1)w = z(\vartheta+a)(\vartheta+b)w,} \tag{16.1–13}$$

the desired modified form of the hypergeometric differential equation.

16.1.3 The invariant of the hypergeometric differential equation

The invariant $R(z)$ of the equation (16.1–9) is obtained if we insert

$$p(z) = \frac{(a+b+1)z-c}{z(z-1)}, \qquad q(z) = \frac{ab}{z(z-1)}$$

into (15.1–29). Writing

$$R(z) = \frac{Az^2+Bz+C}{2z^2(z-1)^2} \tag{16.1–14}$$

we have

$$Az^2+Bz+C = 4abz(z-1) - ((a+b+1)z-c)^2 + \\ -2(a+b+1)z(z-1) + 2((a+b+1)z-c)(2z-1).$$

Comparing the coefficients of z^2 we find

$$A = 4ab - 2(a+b+1) + 4(a+b+1) - (a+b+1)^2 = 1-(a-b)^2.$$

Inserting $z = 0$ we get

$$C = 2c-c^2 = 1-(1-c)^2$$

and by taking $z = 1$ we find

$$A+B+C = 2(a+b-c+1)-(a+b-c+1)^2 = 1-(a+b-c)^2.$$

Taking into account (16.1–11) we see that A, B and C may be expressed in terms of the differences of the exponents of a Riemann equation with singularities at $z = 0, 1, \infty$.

Summing up we have

The invariant $R(z)$ *of the hypergeometric differential equation is the function* (16.1–14) *with*

$$\begin{aligned} A &= 1-(a-b)^2, \\ B &= (a-b)^2+(1-c)^2-(c-a-b)^2-1, \\ C &= 1-(1-c)^2. \end{aligned} \tag{16.1–15}$$

16.1.4 – The hypergeometric series

Since one of the exponents at $z = 0$ is equal to zero, the hypergeometric differential equation admits of a solution which is regular at $z = 0$. We insert the series

$$w(z) = \sum_{\nu=0}^{\infty} c_\nu z^\nu \tag{16.1–16}$$

into the equation (16.1–13). Since $\vartheta z^\kappa = \kappa z^\kappa$ we have

$$\vartheta(\vartheta+c-1)w(z) = \sum_{\nu=0}^{\infty} \nu(\nu+c-1)c_\nu z^\nu + \sum_{\nu=0}^{\infty} (\nu+c)(\nu+1)c_{\nu+1} z^{\nu+1}$$

and

$$z(\vartheta+a)(\vartheta+b)w(z) = \sum_{\nu=0}^{\infty} (\nu+a)(\nu+b)c_\nu z^{\nu+1}.$$

Equating coefficients of equal powers of z we find the recurrence relations

$$(n+c)(n+1)c_{n+1} = (n+a)(n+b)c_n, \qquad n = 0, 1, 2, \ldots. \tag{16.1–17}$$

If c is zero or a negative integer these relations can be solved only if a and b have special values. In all other cases we may take $c_0 = 1$ and the remaining coefficients are uniquely determined by

$$\frac{c_{n+1}}{c_n} = \frac{(a+n)(b+n)}{(c+n)(1+n)}, \tag{16.1–18}$$

whence

$$c_n = \frac{a(a+1)\ldots(a+n-1)b(b+1)\ldots(b+n-1)}{c(c+1)\ldots(c+n-1)} \frac{1}{n!}.$$

Thus we obtain the *hypergeometric series*

$$w(z) = 1+\frac{ab}{c}\frac{z}{1!}+\frac{a(a+1)b(b+1)}{c(c+1)}\frac{z^2}{2!}+\dots \qquad (16.1\text{–}19)$$

This series will be denoted by

$$F(a, b; c; z).$$

Notice that it is symmetric in a and b, as is the hypergeometric differential equation (16.1–9).

The series terminates after a finite number of terms if at least one of the numbers a, b is zero or a negative integer. In the remaining cases it is convergent if $|z| < 1$, for it follows from (16.1–18)

$$\lim_{n\to\infty}\frac{c_{n+1}}{c_n} = 1.$$

A function obtained by continuing analytically the series beyond its circumference of convergence is called a *hypergeometric function*.

It is convenient to introduce the symbol

$$(s)_n = \frac{\Gamma(s+n)}{\Gamma(s)} = \begin{cases} s(s+1)\dots(s+n-1), & \text{if} \quad n > 0, \\ 1, & \text{if} \quad n = 0. \end{cases} \qquad (16.1\text{–}20)$$

Clearly

$$(s)_n = (-1)^n n! \binom{-s}{n}, \qquad (16.1\text{–}21)$$

as follows from (2.16–19). Accordingly the hypergeometric series may be written as

$$\boxed{F(a, b; c; z) = \sum_{\nu=0}^{\infty}\frac{(a)_\nu(b)_\nu}{(c)_\nu}\frac{z^\nu}{\nu!} = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{\nu=0}^{\infty}\frac{\Gamma(a+\nu)\Gamma(b+\nu)}{\Gamma(c+\nu)}\frac{z^\nu}{\nu!},}$$

(16.1–22)

where c is different from $0, -1, -2, \dots$.

A direct consequence is the formula

$$\frac{d}{dz}F(a, b; c; z) = \frac{ab}{c}F(a+1, b+1; c+1; z). \qquad (16.1\text{–}23)$$

16.1.5 – Special examples

The hypergeometric series covers a great variety of well-known functions. Typical examples may be found by giving a, b and c particular values.

Since $(1)_n = n!$, we have

$$F(1, 1; 1; z) = \sum_{\nu=0}^{\infty} z^\nu = \frac{1}{1-z}, \tag{16.1–24}$$

the ordinary geometric series.

Taking account of (16.1–22) we have more generally

$$F(-s, 1; 1; z) = \sum_{\nu=0}^{\infty} (-1)^\nu \binom{s}{\nu} z^\nu = (1-z)^s, \tag{16.1–25}$$

the binomial series.

Observing that $(2)_n = (n+1)! = (n+1)n!$, we find

$$zF(1, 1; 2; z) = \sum_{\nu=0}^{\infty} \frac{z^{\nu+1}}{\nu+1} = \log\frac{1}{1-z}. \tag{16.1–26}$$

The relation

$$(s+1)_n = \frac{s+n}{s}(s)_n$$

yields for $s = \frac{1}{2}$ the equation

$$(\tfrac{3}{2})_n = (2n+1)(\tfrac{1}{2})_n$$

and so, by virtue of (2.16–23),

$$zF(\tfrac{1}{2}, 1; \tfrac{3}{2}; -z^2) = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{z^{2\nu+1}}{2\nu+1} = \arctan z. \tag{16.1–27}$$

Similarly, in view of (2.17–3),

$$zF(\tfrac{1}{2}, \tfrac{1}{2}; \tfrac{3}{2}; z^2) = \arcsin z. \tag{16.1–28}$$

It is easy to see that the series (14.5–11) for the function $\mathsf{K}(z)$ may be written as

$$\mathsf{K}(z) = \tfrac{1}{2}\pi \sum_{\nu=0}^{\infty} \frac{\binom{-\frac{1}{2}}{\nu}^2}{(\nu!)^2} z^\nu = \tfrac{1}{2}\pi \sum_{\nu=0}^{\infty} \frac{(\frac{1}{2})_\nu(\frac{1}{2})_\nu}{(1)_\nu} \frac{z^\nu}{\nu!},$$

whilst the series (14.5–16) for $\mathsf{E}(z)$ takes the form

$$\mathsf{E}(z) = \tfrac{1}{2}\pi \sum_{\nu=0}^{\infty} \frac{\binom{-\frac{1}{2}}{\nu}\binom{\frac{1}{2}}{\nu}}{(\nu!)^2} z^\nu = \tfrac{1}{2}\pi \sum_{\nu=0}^{\infty} \frac{(\frac{1}{2})_\nu(-\frac{1}{2})_\nu}{(1)_\nu} \frac{z^\nu}{\nu!}.$$

Thus we have

$$\boxed{\begin{aligned}\mathsf{K}(z) &= \tfrac{1}{2}\pi F(\tfrac{1}{2}, \tfrac{1}{2}; 1; z),\\ \mathsf{E}(z) &= \tfrac{1}{2}\pi F(\tfrac{1}{2}, -\tfrac{1}{2}; 1; z).\end{aligned}} \tag{16.1–29}$$

Finally we consider the series (15.4–37) for the function $F_{-(\kappa+1)}(z)$. The coefficient of z^{-2n} may be written as

$$\frac{(\kappa+1)(\kappa+3)\dots(\kappa+2n-1)(\kappa+2)(\kappa+4)\dots(\kappa+2n)}{2^n n! 2^n(\kappa+\frac{3}{2})(\kappa+\frac{3}{2}+1)\dots(\kappa+\frac{3}{2}+n-1)}$$

$$= \frac{(\frac{1}{2}\kappa+\frac{1}{2})_n(\frac{1}{2}\kappa+1)_n}{(\kappa+\frac{3}{2})_n}\,\frac{1}{n!}.$$

Hence we may bring Legendre's function of the second kind into the hypergeometric form

$$\boxed{Q_\kappa(z) = \frac{\Gamma(\kappa+1)\sqrt{\pi}}{2^{\kappa+1}\Gamma(\kappa+\frac{3}{2})}\,\frac{1}{z^{\kappa+1}}\,F(\tfrac{1}{2}\kappa+\tfrac{1}{2}, \tfrac{1}{2}\kappa+1; \kappa+\tfrac{3}{2}, z^{-2}).} \tag{16.1–30}$$

An expression for Legendre's function of the first kind in terms of hypergeometric series may be obtained from (15.4–42).

A simpler expression for $P_\kappa(z)$ can be obtained by a simple transformation. It is readily seen that the solutions of Legendre's equation (15.4–27) are characterized by the scheme

$$P\begin{bmatrix} -1 & \infty & 1 & \\ 0 & -\kappa & 0 & z \\ 0 & \kappa+1 & 0 & \end{bmatrix}.$$

By means of the substitution $z \to \frac{1}{2}(1-z)$ this changes into

$$P\begin{bmatrix} 0 & 1 & \infty & \\ 0 & 0 & -\kappa & \frac{1}{2}(1-z) \\ 0 & 0 & \kappa+1 & \end{bmatrix}.$$

Since $P_\kappa(z)$ is regular at $z = 1$ and takes the value 1 there, we conclude that

$$\boxed{P_\kappa(z) = F(-\kappa, \kappa+1; 1; \tfrac{1}{2}(1-z)).} \tag{16.1–31}$$

This is *Murphy's expression* for Legendre's function of the first kind.

16.1.6 – Kummer's twenty four series

By linear transformation of the variable z we may carry the singular points of Riemann's equation (15.6–5) into the points $z = 0, 1, \infty$ respectively. An equation with these singular points is characterized by the scheme

$$P\begin{bmatrix} 0 & 1 & \infty & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix} = P\begin{bmatrix} \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{bmatrix}. \tag{16.1–32}$$

This can be reduced to the hypergeometric scheme by applying (16.1–4) and this reduction can be performed in many ways. For we may take for κ one of the values α, α', for λ one of the values β, β', giving four possibilities. Moreover, we can permute the singular points 0, 1, ∞ by means of the transformations of the group of the anharmonic ratio

$$z \to z, \frac{1}{z}, \; 1-z, \frac{1}{1-z}, \frac{z}{z-1}, \frac{z-1}{z}. \tag{16.1–33}$$

Thus we obtain 24 transforms of the original equation and, therefore, 24 possibilities to represent a solution of Riemann's general equation as solutions of certain hypergeometric equations. In particular we may start with a hypergeometric equation and so we have *Kummer's theorem*:

Any solution of the hypergeometric equation (16.1–9) *can be represented by solutions of twenty four (formally) different hypergeometric equations.*

This theorem may also be stated as follows

There are twenty four hypergeometric series which provide a solution of a given hypergeometric differential equation.

It is not difficult to obtain these series explicitly.

The hypergeometric series (16.1–23) is a solution corresponding to the exponent 0 at $z = 0$. We shall denote this solution by $w_{01}(z)$. Since

$$P\begin{bmatrix} 0 & 0 & a & \\ 1-c & c-a-b & b & z \end{bmatrix} = P\begin{bmatrix} 0 & c-a-b & b & \\ 1-c & 0 & a & z \end{bmatrix}$$
$$= (z-1)^{c-a-b} P\begin{bmatrix} 0 & 0 & c-a & \\ 1-c & a+b-c & c-b & z \end{bmatrix},$$

the solution regular at $z = 0$ may also be written as

$$w_{02}(z) = (1-z)^{c-a-b} F(c-a, c-b; c; z), \tag{16.1–34}$$

where we have taken $1-z$ rather than $z-1$ in order to obtain a function which takes the value 1 at $z = 0$.

The transformation $z \to z/(z-1)$ interchanges the points 1, ∞ and leaves $z = 0$ invariant. We have now

$$P\begin{bmatrix} 0 & 0 & a & \\ 1-c & c-a-b & b & z \end{bmatrix} = P\begin{bmatrix} 0 & a & 0 & z \\ 1-c & b & c-a-b & \overline{z-1} \end{bmatrix}$$
$$= \left(\frac{z}{z-1} - 1\right)^a P\begin{bmatrix} 0 & 0 & a & z \\ 1-c & b-a & c-b & \overline{z-1} \end{bmatrix}.$$

It follows that

$$w_{03}(z) = (1-z)^{-a} F\left(a, c-b; c; \frac{z}{z-1}\right). \tag{16.1–35}$$

Since there is symmetry between a and b, we also have a solution

$$w_{04}(z) = (1-z)^{-b}F\left(b, c-a; c; \frac{z}{z-1}\right). \qquad (16.1\text{–}36)$$

A solution corresponding to the exponent $1-c$ at $z = 0$ may be found by observing that

$$P\begin{bmatrix} 0 & 0 & a & z \\ 1-c & c-a-b & b & \end{bmatrix} = P\begin{bmatrix} 1-c & 0 & a & z \\ 0 & c-a-b & b & \end{bmatrix}$$

$$= z^{1-c}P\begin{bmatrix} 0 & 0 & a-c+1 & z \\ c-1 & c-a-b & b-c+1 & \end{bmatrix}.$$

The desired solution may be taken as

$$w_{01'}(z) = z^{1-c}F(a-c+1, b-c+1; 2-c; z). \qquad (16.1\text{–}37)$$

Proceeding as before we may get from this solution three other solutions $w_{02'}(z)$, $w_{03'}(z)$, $w_{04'}(z)$.

In order to obtain a solution corresponding to the exponent 0 at $z = 1$ we apply the transformation $z \to 1-z$ which interchanges the points 0, 1 but leaves $z = \infty$ invariant. From

$$P\begin{bmatrix} 0 & 0 & a & z \\ 1-c & c-a-b & b & \end{bmatrix} = P\begin{bmatrix} 0 & 0 & a & 1-z \\ c-a-b & 1-c & b & \end{bmatrix}$$

we deduce

$$w_{11}(z) = F(a, b; a+b-c+1; 1-z) \qquad (16.1\text{–}38)$$

and a solution corresponding to the exponent $c-a-b$ at $z = 1$ is represented by

$$w_{11'}(z) = (1-z)^{c-a-b}F(c-a, c-b; c-a-b+1; 1-z). \qquad (16.1\text{–}39)$$

Either of these solutions gives rise to three other expressions.

Finally we perform the transformation $z \to 1/z$, interchanging the points 0 and ∞, leaving $z = 1$ invariant. Now we have

$$P\begin{bmatrix} 0 & 0 & a & z \\ 1-c & c-a-b & b & \end{bmatrix} = P\begin{bmatrix} a & 0 & 0 & z^{-1} \\ b & c-a-b & 1-c & \end{bmatrix}$$

$$= z^{-a}P\begin{bmatrix} 0 & 0 & a & z^{-1} \\ b-a & c-a-b & a-c+1 & \end{bmatrix}.$$

Hence a solution corresponding to the exponent a at $z = \infty$ is

$$w_{\infty 1}(z) = z^{-a}F(a, a-c+1; a-b+1; z^{-1}) \qquad (16.1\text{–}40)$$

and by reasons of symmetry a solution corresponding to the exponent b

at $z = \infty$ is

$$w_{\infty 1'}(z) = z^{-b}F(b, b-c+1; b-a+1; z^{-1}). \qquad (16.1\text{–}41)$$

For the sake of completeness we shall list all series obtainable by the method described above.

$$\begin{aligned}
w_{01}(z) &= F(a, b; c; z),\\
w_{02}(z) &= (1-z)^{c-a-b}F(c-a, c-b; c; z),\\
w_{03}(z) &= (1-z)^{-a}F(a, c-b; c; z/(z-1)),\\
w_{04}(z) &= (1-z)^{-b}F(b, e-a; c; z/(z-1));
\end{aligned}$$

$$\begin{aligned}
w_{01'}(z) &= z^{1-c}F(a-c+1, b-c+1; 2-c; z),\\
w_{02'}(z) &= z^{1-c}(1-z)^{c-a-b}F(1-a, 1-b; 2-c; z),\\
w_{03'}(z) &= z^{1-c}(1-z)^{c-a-1}F(a-c+1, 1-b; 2-c; z/(z-1)),\\
w_{04'}(z) &= z^{1-c}(1-z)^{c-b-1}F(b-c+1, 1-a; 2-c; z/(z-1));
\end{aligned}$$

$$\begin{aligned}
w_{11}(z) &= F(a, b; a+b-c+1; 1-z),\\
w_{12}(z) &= z^{1-c}F(a-c+1, b-c+1; a+b-c+1; 1-z),\\
w_{13}(z) &= z^{-a}F(a, a-c+1; a+b-c+1; (z-1)/z),\\
w_{14}(z) &= z^{-b}F(b, b-c+1; a+b-c+1; (z-1)/z);
\end{aligned}$$

$$\begin{aligned}
w_{11'}(z) &= (1-z)^{c-a-b}F(c-a, c-b; c-a-b+1; 1-z),\\
w_{12'}(z) &= z^{1-c}(1-z)^{c-a-b}F(1-a, 1-b; c-a-b+1; 1-z),\\
w_{13'}(z) &= z^{a-c}(1-z)^{c-a-b}F(1-a, c-a; c-a-b+1; (z-1)/z),\\
w_{14'}(z) &= z^{b-c}(1-z)^{c-a-b}F(1-b, c-b; c-a-b+1; (z-1)/z);
\end{aligned}$$

$$\begin{aligned}
w_{\infty 1}(z) &= z^{-a}F(a, a-c+1; a-b+1; z^{-1}),\\
w_{\infty 2}(z) &= z^{-a}(1-1/z)^{c-a-b}F(1-b, c-b; a-b+1; z^{-1}),\\
w_{\infty 3}(z) &= z^{-a}(1-1/z)^{-a}F(a, c-b; a-b+1; 1/(1-z)),\\
w_{\infty 4}(z) &= z^{-a}(1-1/z)^{c-a-1}F(1-b, a-c+1; a-b+1; 1/(1-z));
\end{aligned}$$

$$\begin{aligned}
w_{\infty 1'}(z) &= z^{-b}F(b, b-c+1; b-a+1; z^{-1}),\\
w_{\infty 2'}(z) &= z^{-b}(1-1/z)^{c-a-b}F(1-a, c-a; b-a+1; z^{-1}),\\
w_{\infty 3'}(z) &= z^{-b}(1-1/z)^{-b}F(b, c-a; b-a+1; 1/(1-z)),\\
w_{\infty 4'}(z) &= z^{-b}(1-1/z)^{c-b-1}F(1-a, b-c+1; b-a+1; 1/(1-z)).
\end{aligned}$$

16.1.7 – The hypergeometric integrals

Analytic continuation of the hypergeometric series beyond the circle of convergence may be effected by representing it by a definite integral. Starting from (16.1–22) we may write (in view of (4.7–37) and (4.7–38))

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)} F(a, b; c; z) = \sum_{\nu=0}^{\infty} \frac{(b)_\nu}{\nu!} \frac{\Gamma(a+\nu)\Gamma(c-a)}{\Gamma(c+\nu)} z^\nu$$

$$= \sum_{\nu=0}^{\infty} \binom{-b}{\nu} (-1)^\nu z^\nu \int_0^1 u^{a+\nu-1}(1-u)^{c-a-1}\,du$$

$$= \int_0^1 u^{a-1}(1-u)^{c-a-1} \sum_{\nu=0}^{\infty} \binom{-b}{\nu} (-1)^\nu (uz)^\nu\,du,$$

valid under the assumptions Re $a > 0$, Re $(c-a) > 0$, $|z| < 1$. The series within the sign of integration is a binomial series. We finally have

$$\begin{aligned} \frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)} F(a, b; c; z) &= \int_0^1 u^{a-1}(1-u)^{c-a-1}(1-zu)^{-b}\,du, \\ \frac{\Gamma(b)\Gamma(c-b)}{\Gamma(c)} F(a, b; c; z) &= \int_0^1 u^{b-1}(1-u)^{c-b-1}(1-zu)^{-a}\,du. \end{aligned} \tag{16.1–42}$$

The second equation is a consequence of the fact that the hypergeometric series is symmetric with respect to a and b; it is valid under the assumptions Re $b > 0$, Re $(c-b) > 0$, $|z| < 1$.

Proceeding as in section 2.9.1 we may show that the above integrals are regular at every point z of the region obtained from the z-plane by slitting it along the real axis from $+1$ to $+\infty$. The integrals (16.1–42) were already known to Euler, and are, therefore, referred to as the *Eulerian integrals*. They reduce to the integrals for the beta function if b or a is zero.

Since every solution of the hypergeometric differential equation listed at the end of the previous sections corresponds to two Eulerian integrals, the number of these integrals representing a solution of the given hypergeometric differential equation is fourty eight.

For many applications it is convenient to introduce the variable $t = 1/u$. The second integral of (16.1–42) takes the form

$$\int_1^\infty t^{a-c}(t-1)^{c-b-1}(t-z)^{-a}\,dt. \tag{16.1–43}$$

It is an integral of the type occurring in (15.6–15). As we pointed out in section 16.1.2 we may obtain the hypergeometric differential equation from Riemann's equation by making the singular points a, b and c tend to 0, 1 and ∞ respectively. If in (15.6–15) we divide by $c^{\gamma+\mu_c-1} = c^{\alpha+\beta+\gamma+\gamma'-1}$ we obtain, after performing the limiting process, the integral

$$z^\alpha(z-1)^\beta \int_C t^{\mu_a-1}(t-1)^{\mu_b-1}(t-z)^{\mu+1}\,dt.$$

Inserting the appropriate exponents we find (16.1–43) in the more general form

$$\int_C t^{a-c}(t-1)^{c-b-1}(t-z)^{-a}\,dt, \tag{16.1–44}$$

where the path C is such that $\int dV = 0$, $V(t)$ being the function

$$V(t) = t^{a-c+1}(t-1)^{c-b}(t-z)^{-a-1}. \tag{16.1–45}$$

It is clear that $V(t)$ vanishes at the points 1 and ∞, provided that $\operatorname{Re}(c-b) > 0$, $\operatorname{Re} b > 0$ and z not on the real line $1 \leqq x \leqq \infty$. This proves again that (16.1–43) is a solution of the hypergeometric differential equation.

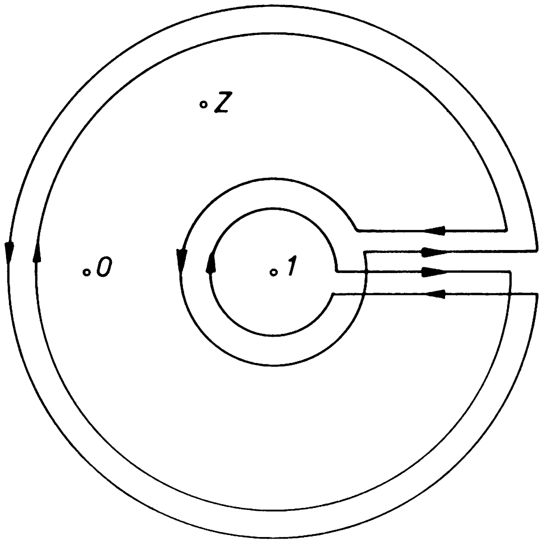


Fig. 16.1–1. The Jordan-Pochhammer loop $(\infty_+, 1_+, \infty_-, 1_-)$

In more unfavourable instances we take for C a double loop of the Jordan-Pochhammer type (fig. 16.1–1) which does not encircle the point z. From (15.6–22) we get

$$\boxed{\begin{aligned}&\frac{\Gamma(b)\Gamma(c-b)}{\Gamma(c)}(1-e^{2\pi ib})(1-e^{2\pi i(c-b)})F(a, b; c; z)\\ &= \int^{(\infty+,\,1+,\,\infty-,\,1-)} t^{a-c}(t-1)^{c-b-1}(t-z)^{-a}\,dt,\end{aligned}} \tag{16.1–46}$$

For applications in the next section we need the following result. If we make the assumptions $\operatorname{Re} a > 0$, $\operatorname{Re}(c-a-b) > 0$ then the limit of the first integral on the right of (16.1–42) as $z \to 1$ is

$$\int_0^1 u^{a-1}(1-u)^{c-a-b-1}\,du = \frac{\Gamma(a)\Gamma(c-a-b)}{\Gamma(c-b)}.$$

It follows that

$$\boxed{\lim_{z\to 1} F(a, b; c; z) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}.} \qquad (16.1\text{–}47)$$

The restrictions on the parameters a, b and c can be relaxed as we shall see in section 16.3.5.

16.1.8 – Analytic continuation of the hypergeometric series

We proceed to solve the problem of expressing an analytic continuation of the hypergeometric series in terms of a fundamental system about the singular point $z = 1$ or about $z = \infty$.

Assuming that $c-a-b$ is not an integer we see from Kummer's table (section 16.1.6) that the functions

$$w_{11}(z) = F(a, b; a+b-c+1; 1-z) \qquad (16.1\text{–}48)$$

and

$$w_{11'}(z) = (1-z)^{c-a-b}F(c-a, c-b; c-a-b+1; 1-z) \qquad (16.1\text{–}49)$$

constitute a fundamental system of solutions of the differential equation (16.1–9), valid in the open disc $|z-1| < 1$. We make the second function definite by giving $(1-z)^{c-a-b}$ its principal value.

In the intersection of the discs $|z| < 1$, $|1-z| < 1$ we must have the identity

$$w_{10}(z) = Aw_{11}(z)+Bw_{11'}(z), \qquad (16.1\text{–}50)$$

where A and B are constants. Supposing that $\operatorname{Re} a > 0$, $\operatorname{Re}(a+b) < \operatorname{Re} c < 1$ we may apply (16.1–47).

If we make z tend to 1 we find

$$A = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}. \qquad (16.1\text{–}51)$$

If we make z tend to zero we find from (16.1–47) by appropriately changing the parameters

$$1 = A\Gamma(1-c)\frac{\Gamma(1+a+b-c)}{\Gamma(1+b-c)\Gamma(1+a-c)} + B\Gamma(1-c)\frac{\Gamma(1+c-a-b)}{\Gamma(1-b)\Gamma(1-a)}.$$

It is apparent that the coefficient of $B\Gamma(1-c)$ is obtained from that of

$A\Gamma(1-c)$ by replacing $c-a$ and $c-b$ by a and b respectively. Thus we conclude that

$$B = \frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}. \tag{16.1–52}$$

It is not difficult to verify this result using (4.6–13) and effecting some elementary manipulations with circular function. This we may state

The function

$$\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} F(a, b; a+b-c+1; 1-z)+ \\ + \frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)} (1-z)^{c-a-b} F(c-a, c-b; c-a-b+1; 1-z) \tag{16.1–53}$$

is an analytic continuation of the hypergeometric series $F(a, b, c; z)$ into the region $|z-1| < 1$.

A fundamental system of solutions valid in the region $|z| > 1$ is given by the functions

$$w_{\infty 1}(z) = z^{-a}F(a, a-c+1; a-b+1; z^{-1}) \tag{16.1–54}$$

and

$$w_{\infty 1'}(z) = z^{-b}F(b, b-c+1; b-a+1; z^{-1}). \tag{16.1–55}$$

The hypergeometric series occurring on the right may be expressed as Eulerian integrals which are single valued in the region obtained from the *z-plane* by cutting it along the segment $0 \leqq x \leqq 1$. On the other hand a single valued continuation of $w_{01}(z)$ is represented by an Eulerian integral in the region obtained by omitting the line $1 \leqq x \leqq \infty$ from the z-plane. In order to avoid ambiguities we shall slit the z-plane from 0 to ∞ along the positive real axis and agree that $0 < \arg z < 2\pi$, $z^{-a} = \exp(-a \log|z| + -ai \arg z)$, $z^{-b} = \exp(-b \log|z| - bi \arg z)$.

Suppose now that

$$Aw_{\infty 1}(z) + Bw_{\infty 1'}(z)$$

is an analytic continuation of $w_{01}(z)$ into the slit region $|z| > 1$. Agreeing that $w_{01}(1)$ stands for $\lim_{z\to 1} w_{01}(z)$, etc., we have, evidently,

$$w_{01}(1) = Aw_{\infty 1}(1) + Bw_{\infty 1}(1),$$

as z tends to 1 from above, and

$$w_{01}(1) = Ae^{-2\pi ia}w_{\infty 1}(1) + Be^{-2\pi ib}w_{\infty 1'}(1),$$

as z tends to 1 from below. Assuming that $b-a$ is not an integer, it follows that

$$A = \frac{e^{2\pi ib}-1}{e^{2\pi i(b-a)}-1} \frac{w_{01}(1)}{w_{\infty 1}(1)} = e^{\pi ia} \frac{\sin \pi b}{\sin \pi(b-a)} \frac{w_{01}(1)}{w_{\infty 1}(1)}.$$

Applying (16.1–47) to $w_{\infty 1}(z)$ we get

$$w_{\infty 1}(1) = \frac{\Gamma(a-b+1)\Gamma(c-a-b)}{\Gamma(1-b)\Gamma(c-b)}.$$

Hence, by virtue of (4.6–13),

$$A = e^{\pi i a}\frac{\Gamma(c)\Gamma(1-b)}{\Gamma(c-a)\Gamma(1-(b-a))}\,\frac{\sin \pi b}{\sin \pi(b-a)} = e^{\pi i a}\frac{\Gamma(c)\Gamma(b-a)}{\Gamma(c-a)\Gamma(b)}.$$

Interchanging a and b yields

$$B = e^{\pi i b}\frac{\Gamma(c)\Gamma(a-b)}{\Gamma(c-b)\Gamma(a)}.$$

If we agree that $-z = ze^{-\pi i}$, then $|\arg(-z)| < \pi$ and

$$(-z)^{-a} = z^{-a}e^{\pi i a}, \qquad (-z)^{-b} = z^{-b}e^{\pi i b}.$$

Thus we may state

The function

$$\begin{aligned}\frac{\Gamma(c)\Gamma(b-a)}{\Gamma(b)\Gamma(c-a)}&(-z)^{-a}F(a, a-c+1; a-b+1; z^{-1})+\\ &+\frac{\Gamma(c)\Gamma(a-b)}{\Gamma(a)\Gamma(c-b)}(-z)^{-b}F(b, b-c+1; b-a+1; z^{-1})\end{aligned} \qquad (16.1\text{–}56)$$

is an analytic continuation of the hypergeometric series $F(a, b; c; z)$ into the region $|z| > 1$, $|\arg(-z)| < \pi$.

The restrictions on the parameters are quite unnecessary for the truth of this result; they arise on account of the particular method of proof adopted. The restrictions can be removed by an appeal to the results of section 16.3.1 An alternative proof valid for all values of the parameters for which the expressions involved have a meaning will be given in section 16.5.3 by a wholly different method.

16.1.9 – Evaluation of Riemann's coefficients

It is clear that Riemann's coefficient occurring in (15.6–39) are uniquely determined if there is no doubt about the multiplier of the functions (15.6–38). In other words: if the values of the regular parts $(z-a)^{-\alpha}P_\alpha$ at $z = a$ etc. are determined without ambiguity. This is the case if we define in accordance with (16.1–6) and (16.1–8)

$$\begin{aligned}P_\alpha(z) &= (\mathrm{A}z)^\alpha(1-\mathrm{A}z)^\beta F(\alpha+\beta+\gamma, \alpha+\beta+\gamma'; 1+\alpha-\alpha'; \mathrm{A}z),\\ P_\beta(z) &= (\mathrm{B}z)^\beta(1-\mathrm{B}z)^\gamma F(\alpha+\beta+\gamma, \alpha'+\beta+\gamma; 1+\beta-\beta'; \mathrm{B}z),\\ P_\gamma(z) &= (\mathrm{C}z)^\gamma(1-\mathrm{C}z)^\alpha F(\alpha+\beta+\gamma, \alpha+\beta'+\gamma; 1+\gamma-\gamma'; \mathrm{C}z),\end{aligned} \qquad (16.1\text{–}57)$$

with

$$\mathsf{A}z = \frac{(z-a)(b-c)}{(z-c)(b-a)},$$
$$\mathsf{B}z = \frac{(z-b)(c-a)}{(z-a)(c-b)}, \qquad (16.1\text{–}58)$$
$$\mathsf{C}z = \frac{(z-c)(a-b)}{(z-b)(a-c)}.$$

The functions $P_{\alpha'}(z)$, $P_{\beta'}(z)$ and $P_{\gamma'}(z)$ arise from those listed in (16.1–57) by interchanging α and α', etc.

Between the functions (16.1–58) exists the relations

$$\mathsf{A}z\,\mathsf{B}z\,\mathsf{C}z = -1, \qquad (16.1\text{–}59)$$

$$1-\mathsf{A}z = \frac{1}{\mathsf{C}z}, \qquad 1-\mathsf{B}z = \frac{1}{\mathsf{A}z}, \qquad 1-\mathsf{C}z = \frac{1}{\mathsf{B}z}. \qquad (16.1\text{–}60)$$

Each of the functions (15.1–57) can be expressed in four different ways. In the first place we observe that we may interchange β and γ in the first equation (16.1-57), yielding an expression whose regular part at $z = a$ takes the same value there as that of P_α. Next we interchange the singular points b and c and at the same time the exponents β and γ. Then $\mathsf{A}z$ is replaced by $1/\mathsf{B}z$ and it follows that

$$P_\alpha(z) = \left(\frac{-1}{\mathsf{B}z}\right)^\alpha \left(1-\frac{1}{\mathsf{B}z}\right)^\gamma F\left(\alpha+\beta+\gamma, \alpha+\beta'+\gamma; 1+\alpha-\alpha'; \frac{1}{\mathsf{B}z}\right),$$

for the regular part has the right value. Indeed, the contribution to the regular part by $\mathsf{A}z$ at $z = a$ is $(b-c)/(a-c)(b-a)$ and that by $1/\mathsf{B}z$ is $(c-b)/(a-b)(c-a)$. Using (16.1–59) and (16.1–60) we may represent $P_\alpha(z)$ in the four ways

$$P_\alpha(z) = (\mathsf{A}z)^\alpha(\mathsf{C}z)^{-\beta}F(\alpha+\beta+\gamma, \alpha+\beta+\gamma'; 1+\alpha-\alpha'; \mathsf{A}z),$$
$$P_\alpha(z) = (\mathsf{A}z)^\alpha(\mathsf{C}z)^{-\beta'}F(\alpha+\beta'+\gamma, \alpha+\beta'+\gamma'; 1+\alpha-\alpha'; \mathsf{A}z),$$
$$P_\alpha(z) = (-\mathsf{B}z)^{-\alpha}(\mathsf{C}z)^{\gamma}F\left(\alpha+\beta+\gamma, \alpha+\beta'+\gamma; 1+\alpha-\alpha'; \frac{1}{\mathsf{B}z}\right), \qquad (16.1\text{–}61)$$
$$P_\alpha(z) = (-\mathsf{B}z)^{-\alpha}(\mathsf{C}z)^{\gamma'}F\left(\alpha+\beta+\gamma', \alpha+\beta'+\gamma'; 1+\alpha-\alpha'; \frac{1}{\mathsf{B}z}\right).$$

By cyclic permutations we obtain a table of 24 solutions of Riemann's equation analogous to Kummer's table in section 16.1.6.

There still remains the task to give an accurate definition of the many-valued functions which occur in the expressions for P_α, etc. The transformation $w = \mathsf{A}z$ carries the points a, b and c into the points 0, 1, ∞, respectively, in the w-plane. Let us suppose that the singular points a, b and c in the z-plane are so placed that, when we go round the circumference through a, b and c in the counterclockwise sense we pass from a to b to c. Then the interior of the circumference is mapped onto the upper half of the w-plane. If we delete from the w-plane the negative real axis (including the origin) we can give to arg w a uniquely defined value between $-\pi$ and π. The corresponding point z varies throughout the extended z-plane cut along the smallest closed arc from c to a. Further arg $\mathsf{A}z$ tends to π if z tends to an interior point of this arc from the inside of the circumference, and to $-\pi$ if it tends to such a point from the outside. On the supplementary arc we have arg $\mathsf{A}z = 0$. In a similar way we may define arg $\mathsf{B}z$ by cutting the z-plane along an arc from a to b and arg $\mathsf{C}z$ by cutting the z-plane along an arc from b to c. In accordance with this we define

$$-\mathsf{A}z = e^{\mp\pi i}\mathsf{A}z, \tag{16.1–62}$$

etc. the upper or lower sign being taken as z lies inside or outside this circumference. In the case that a, b and c are collinear we consider the part to the left of the line through these points as the interior, the line being percorsed in the direction from a to b to c.

Following E. W. Barnes we may evaluate Riemann's coefficients by the aid of the second theorem of section 16.1.8. We apply it to the first expression of $P_\alpha(z)$ listed in (16.1–61). By inserting the values of the parameters we may simplify the result by using the relation (15.6–4). Taking into account (16.1–62) and (16.1–59) we conclude that inside the circle the function $P_\alpha(z)$ is represented by the single-valued function (16.1–57) if $|\mathsf{A}z| < 1$ and by

$$\begin{aligned}&\frac{\Gamma(1+\alpha-\alpha')\Gamma(\gamma'-\gamma)}{\Gamma(\alpha+\beta+\gamma')\Gamma(\alpha+\beta'+\gamma')}\, e^{\pi i\alpha}(-\mathsf{A}z)^{-\gamma}(\mathsf{B}z)^{\beta}\\ &\qquad\times F\left(\alpha+\beta+\gamma, \alpha'+\beta+\gamma; 1+\gamma-\gamma'; \frac{1}{\mathsf{A}z}\right)+\\ &+\frac{\Gamma(1+\alpha-\alpha')\Gamma(\gamma-\gamma')}{\Gamma(\alpha+\beta+\gamma)\Gamma(\alpha+\beta'+\gamma)}\, e^{\pi i\alpha}(-\mathsf{A}z)^{-\gamma'}(\mathsf{B}z)^{\beta}\\ &\qquad\times F\left(\alpha+\beta+\gamma', \alpha'+\beta+\gamma'; 1+\gamma'-\gamma; \frac{1}{\mathsf{A}z}\right),\end{aligned} \tag{16.1–63}$$

if $|\mathsf{A}z| > 1$.

If z is outside the circle we must replace $e^{\pi i\alpha}$ by $e^{-\pi i\alpha}$. It is readily seen that the functions occurring on the right of (16.1–63) arise from the last two functions listed in (16.1–61) by performing two times a cyclic permutation. It follows

The coefficients a_γ, $a_{\gamma'}$, *in the relation*

$$P_\alpha = a_\gamma P_\gamma + a_{\gamma'} P_{\gamma'} \tag{16.1–64}$$

have the values

$$\begin{aligned} a_\gamma &= e^{\pi i\alpha} \frac{\Gamma(1+\alpha-\alpha')\Gamma(\gamma'-\gamma)}{\Gamma(\alpha+\beta+\gamma)\Gamma(\alpha+\beta'+\gamma')}, \\ a_{\gamma'} &= e^{\pi i\alpha} \frac{\Gamma(1+\alpha-\alpha')\Gamma(\gamma-\gamma')}{\Gamma(\alpha+\beta+\gamma)\Gamma(\alpha+\beta'+\gamma)}, \end{aligned} \tag{16.1–65}$$

provided that the functions are evaluated in the interior of the circumference through the singular points.

In quite the same way we can handle the third expression for $P_\alpha(z)$ listed in (16.1–61). It follows that if $|\mathrm{B}z| < 1$ the function $P_\alpha(z)$ is expressed by

$$\begin{aligned} &\frac{\Gamma(1+\alpha-\alpha')\Gamma(\beta'-\beta)}{\Gamma(\alpha+\beta'+\gamma)\Gamma(\alpha+\beta'+\gamma')} e^{-\pi i\beta}(\mathrm{B}z)^{\beta}(\mathrm{A}z)^{-\gamma} \\ &\qquad\qquad \times F(\alpha+\beta+\gamma, \alpha'+\beta+\gamma; 1+\beta-\beta'; \mathrm{B}z)+ \\ &\frac{\Gamma(1+\alpha-\alpha')\Gamma(\beta-\beta')}{\Gamma(\alpha+\beta+\gamma)\Gamma(\alpha+\beta+\gamma')} e^{-\pi i\beta'}(\mathrm{B}z)^{\beta'}(\mathrm{A}z)^{-\gamma} \\ &\qquad\qquad \times F(\alpha+\beta'+\gamma, \alpha'+\beta'+\gamma'; 1+\beta'-\beta; \mathrm{B}z) \end{aligned} \tag{16.1–66}$$

provided that z is in the interior of the circumference through the singular points. If z is outside we have to change the sign of the exponents of the exponential factors. On the right appear the functions P_β, $P_{\beta'}$, whence

The coefficients a_β, $a_{\beta'}$ *in the relation*

$$P_\alpha = a_\beta P_\beta + a_{\beta'} P_{\beta'} \tag{16.1–67}$$

have the values

$$\begin{aligned} a_\beta &= e^{-\pi i\beta} \frac{\Gamma(1+\alpha-\alpha')\Gamma(\beta'-\beta)}{\Gamma(\alpha+\beta'+\gamma)\Gamma(\alpha+\beta'+\gamma')}, \\ a_{\beta'} &= e^{-\pi i\beta'} \frac{\Gamma(1+\alpha-\alpha')\Gamma(\beta-\beta')}{\Gamma(\alpha+\beta+\gamma)\Gamma(\alpha+\beta+\gamma')}, \end{aligned} \tag{16.1–68}$$

provided that z is inside the circumference through the singular points.

The coefficients a'_γ, $a'_{\gamma'}$, a'_β, $a'_{\beta'}$ in corresponding relations are obtained by interchanging α and α'.

The values of a_β, $a_{\beta'}$ are not quite symmetric with those obtained for a_γ, $a_{\gamma'}$. This is explained by the fact that P_β, $P_{\beta'}$ are single valued in the z-plane cut along the arc from a to b, P_γ, $P_{\gamma'}$ are single valued in the plane cut along the arc from b to c.

It is now an easy matter to verify the relations (15.6–42). We find with the aid of (4.6–13)

$$\frac{a_\beta}{a'_\beta}\Big/\frac{a_\gamma}{a'_\gamma} = \frac{a_\beta}{a_\gamma}\frac{a'_\gamma}{a'_\beta}$$

$$= \frac{e^{-\pi i\alpha}\Gamma(\alpha+\beta+\gamma')\Gamma(\alpha+\beta'+\gamma')}{e^{-\pi i\alpha'}\Gamma(\alpha+\beta'+\gamma)\Gamma(\alpha'+\beta+\gamma')} = \frac{e^{-\pi i\alpha}\sin\pi(\alpha+\beta'+\gamma)}{e^{-\pi i\alpha'}\sin\pi(\alpha'+\beta'+\gamma)},$$

in accordance with the first equation of (15.6–42).

16.1.10 – The monodromy group of the hypergeometric differential equation

Jordan's method for obtaining the monodromy group of Riemann's equation may be specialized to the case of the hypergeometric differential equation. Inserting the appropriate exponents we obtain the first functions of (15.6–31) in the form

$$\begin{aligned} w_0 &= \int^{(0+,z+,0-,z-)} t^{a-c}(t-1)^{c-b-1}(t-z)^{-a}\,dt,\\ w_1 &= \int^{(1+,z+,1-,z-)} t^{a-c}(t-1)^{c-b-1}(t-z)^{-a}\,dt. \end{aligned} \tag{16.1–69}$$

The analytic continuation along a loop surrounding the point $z = 0$ once induces a linear transformation of the quotient w_1/w_0 characterized by the matrix

$$S_0 = \begin{bmatrix} 1 & e^{2\pi i(c-a-b)} - e^{2\pi i a} \\ 0 & e^{-2\pi i c} \end{bmatrix}. \tag{16.1–70}$$

The analytic continuation along a closed path surrounding $z = 1$ once induces a linear transformation characterized by the matrix

$$S_1 = \begin{bmatrix} e^{2\pi i(c-a-b)} & 0 \\ e^{-2\pi i c} - e^{-2\pi i a} & 1 \end{bmatrix}. \tag{16.1–71}$$

The transformations S_0 and S_1 generate the monodromy group of the hypergeometric equation.

By taking $c = 1$, $a = \frac{1}{2}$, $b = \frac{1}{2}$ we get

$$S_0 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \qquad S_1 = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}, \tag{16.1–72}$$

the generators of the congruence group mod 2, (section 14.2.6). In fact, in this case the integrals (16.1–69) reduce to Legendre's complete elliptic integrals, for it follows from (15.6–22) that

$$\begin{aligned} w_0 &= -4\int_0^z t^{-\frac{1}{2}}(t-1)^{-\frac{1}{2}}(t-z)^{-\frac{1}{2}}\,dt, \\ w_1 &= -4\int_1^z t^{-\frac{1}{2}}(t-1)^{-\frac{1}{2}}(t-z)^{-\frac{1}{2}}\,dt. \end{aligned} \tag{16.1–73}$$

In the first integral we perform the substitution $t = u^2 z$; we get

$$w_0 = -8\int_0^1 \frac{du}{\sqrt{(1-u^2)(1-zu^2)}} = -8\mathrm{K}(z).$$

Substituting $t = 1-u^2(1-z)$ in the second integral yields

$$w_1 = -8i\int_0^1 \frac{du}{\sqrt{(1-u^2)(1-(1-z)u^2)}} = -8i\mathrm{K}'(z).$$

It follows that

$$w_1/w_0 = i\mathrm{K}'(z)/\mathrm{K}(z).$$

16.1.11 - Uniformization of the hypergeometric function

The analytic continuation of the hypergeometric series $F(a, b; c; z)$ defines an analytic function of the variable z, the hypergeometric function, also denoted by $F(a, b; c; z)$. The singular points are at $z = 0$, $z = 1$, $z = \infty$. By effecting the continuation without crossing the cut from 1 to ∞ along the positive real axis we obtain the *principal branch* of the function. It is holomorphic throughout the region obtained from the z-plane by deleting the positive real axis from 1 to ∞. In general the point $z=0$ is singular for all other branches of the analytic function.

The possibility to uniformize this function by expressing z as a modular function has been announced by H. Poincaré. As W. Wirtinger has shown this may be done explicitly by means of the function $\lambda(\tau)$ introduced in section 14.3.3.

The principal branch may be represented by Euler's integral (16.1–42)

$$B(a, c-a)F(a, b; c; z) = \int_0^1 t^{a-1}(1-t)^{c-a-1}(1-zt)^{-b}\,dt, \tag{16.1–74}$$

valid under the condition $\operatorname{Re} a > 0$, $\operatorname{Re}(c-a) > 0$. If these conditions are not satisfied we take the integral along a Jordan-Pochhammer contour.

As stated in section 14.5.9 the equation

$$\lambda(\tau) = z \tag{16.1–75}$$

has always a solution in the upper half of the τ-plane if $z \neq 0$ and $\neq 1$.

We may even take τ in the fundamental region plotted in fig. 14.3–6. Now we consider z as the square of the modulus of Jacobian elliptic functions. If we replace t by the function

$$t = \mathrm{sn}^2\, \mathsf{K}u, \tag{16.1–76}$$

where in accordance with (5.15–9) the quarter period K is defined by

$$\mathsf{K} = \tfrac{1}{2}\pi\vartheta_3^2(0|\tau), \tag{16.1–77}$$

we readily obtain

$$B(a, c-a)F(a, b; c; \lambda(\tau)) = 2\mathsf{K}\int_0^1 \mathrm{sn}^{2a-1}\, \mathsf{K}u\, \mathrm{cn}^{2(c-a)-1}\, \mathsf{K}u\, \mathrm{dn}^{1-2b}\, \mathsf{K}u\, du. \tag{16.1–78}$$

In order to exhibit the dependence on the parameter τ explicitly we introduce the theta functions. From (5.15–10) and (5.15–8) we obtain

$$\begin{aligned} \mathrm{sn}\, \mathsf{K}u &= \frac{\vartheta_3(0|\tau)\vartheta_1(\frac{1}{2}u|\tau)}{\vartheta_2(0|\tau)\vartheta_4(\frac{1}{2}u|\tau)}, \\ \mathrm{cn}\, \mathsf{K}u &= \frac{\vartheta_4(0|\tau)\vartheta_2(\frac{1}{2}u|\tau)}{\vartheta_2(0|\tau)\vartheta_4(\frac{1}{2}u|\tau)}, \\ \mathrm{dn}\, \mathsf{K}u &= \frac{\vartheta_4(0|\tau)\vartheta_3(\frac{1}{2}u|\tau)}{\vartheta_3(0|\tau)\vartheta_4(\frac{1}{2}u|\tau)}. \end{aligned} \tag{16.1–79}$$

Substituting (16.1–77) and (16.1–79) into (16.1–78) yields

$$B(a, c-a)F(a, b; c; \lambda(\tau)) = \pi\vartheta_2^{1-2a}(0|\tau)\vartheta_3^{1+2a}(0|\tau)\vartheta_4^{1-2a}(0|\tau)\times$$
$$\times \int_0^1 \vartheta_1^{2a-1}(\tfrac{1}{2}u|\tau)\left(\frac{\vartheta_2(\frac{1}{2}u|\tau)}{\vartheta_2(0|\tau)}\right)^{2(c-a)-1}\left(\frac{\vartheta_3(\frac{1}{2}u|\tau)}{\vartheta_3(0|\tau)}\right)^{1-2b}\left(\frac{\vartheta_4(\frac{1}{2}u|\tau)}{\vartheta_4(0|\tau)}\right)^{1-2(c-b)} du.$$

We may bring this result in a more elegant form by the aid of (5.9–14) and we obtain *Wirtinger's theorem* which states

If we express the variable z by the modular function

$$z = \lambda(\tau) = \frac{\vartheta_3^2(0|\tau)}{\vartheta_2^2(0|\tau)}, \tag{16.1–80}$$

the analytic function $F(a, b; c; z)$ becomes a single-valued function of the variable τ throughout the upper half of the τ-plane expressed by

$$\frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)}\pi^{2a}\vartheta_3^{4a}(0|\tau)\int_0^1 \Theta(u|\tau)du, \tag{16.1–81}$$

with

$$\Theta(u|\tau) = \left(\frac{\vartheta_1(\frac{1}{2}u|\tau)}{\vartheta_1'(0|\tau)}\right)^{2a-1} \left(\frac{\vartheta_2(\frac{1}{2}u|\tau)}{\vartheta_2(0|\tau)}\right)^{2(c-a)-1} \left(\frac{\vartheta_3(\frac{1}{2}u|\tau)}{\vartheta_3(0|\tau)}\right)^{1-2b} \left(\frac{\vartheta_4(\frac{1}{2}u|\tau)}{\vartheta_4(0|\tau)}\right)^{1-2(c-b)}. \tag{16.1–82}$$

In order to prove that (16.1–81) is actually a single-valued function we express the theta functions as infinite products in terms of the variable

$$q = e^{\pi i\tau}, \qquad \text{Im}\,\tau > 0.$$

These products are listed in section 5.9.3 and their factors are of the form

$$1 \pm q^n, \qquad 1 \pm q^n e^{\pm\pi iu},$$

where n is a positive integer. The region of convergence is the interior of the unit disc in the q-plane. It is clear that the zeros of these factors are on the circumference $|q| = 1$. Hence the branch points of the various factors occurring in (16.1–82) are on the real τ-axis. We may select the initial values of all functions under consideration in such a way that (16.1–81) represents the principal branch of the hypergeometric function, if τ is in the fundamental domain mentioned above. Reflecting this domain in either of each sides and repeating this process undefinitely we obtain the other branches of the hypergeometric function and an insight in the structure of its Riemann surface.

16.2 – Hypergeometric polynomials

16.2.1 – The Jacobian polynomials

The hypergeometric series (16.1–23) terminates after a finite number of terms if at least one of the numbers a or b is zero or a negative integer. It is called a *hypergeometric polynomial.* These polynomials possess remarkable properties and have been the subject matter of an extensive research.

A well-known example is Murphy's expression (16.1–31) for Legendre's polynomial of degree n:

$$P_n(z) = F(-n, n+1; 1; \tfrac{1}{2}(1-z)). \tag{16.2–1}$$

The general case

$$F((-n, b; c; \tfrac{1}{2}(1-z)), \tag{16.2–2}$$

where n is an integer ≥ 0, takes a more symmetrical form with respect to $z = 1$ and $z = -1$ if we apply the transformation (16.1–35) to (16.2–2). We get

$$\frac{1}{2^n}(z+1)^n F\left(-n, c-b; c; \frac{z-1}{z+1}\right) = \frac{1}{2^n}\sum_{\nu=0}^{\infty} A_\nu (z-1)^\nu (z+1)^{n-\nu}, \tag{16.2–3}$$

where the coefficients A_k, $k = 0, 1, \ldots$, are determined as follows. By virtue of (16.1–22) we have

$$A_k = \frac{(-n)_k(c-b)_k}{(c)_k k!} = \frac{n!}{(n-k)!} \frac{\binom{b-c}{k}}{(c)_k}, \qquad k = 0, 1, \ldots.$$

The part of the denominator involving c may be eliminated if we multiply by

$$\binom{n+c-1}{n} = \frac{c(c+1)\ldots(c+n-1)}{n!},$$

for

$$\frac{n!\binom{n+c-1}{n}}{(c)_k} = \frac{c(c+1)\ldots(c+n-1)}{c(c+1)\ldots(c+k-1)} = (c+k)\ldots(c+n-1)$$
$$= (n-k)!\binom{c+n-1}{n-k},$$

whence

$$A_k = \binom{c+n-1}{n-k}\binom{b-c}{k}.$$

We can give this coefficient a more satisfactory appearance if we introduce the numbers α and β by putting

$$c-1 = \alpha, \qquad b-c = n+\beta. \tag{16.2–4}$$

Thus we find a polynomial $P_n^{(\alpha,\beta)}(z)$ defined by

$$\boxed{P_n^{(\alpha,\beta)}(z) = \frac{1}{2^n}\sum_{\nu=0}^{n}\binom{n+\alpha}{n-\nu}\binom{n+\beta}{\nu}(z-1)^\nu(z+1)^{n-\nu}.} \tag{16.2–5}$$

These polynomials are named after Jacobi. The first polynomials are

$$P_0^{(\alpha,\beta)}(z) = 1,\ P_1^{(\alpha,\beta)}(z) = \tfrac{1}{2}((\alpha+\beta+2)z+(\alpha-\beta)),$$
$$P_2^{(\alpha,\beta)}(z) = \tfrac{1}{8}((\alpha+\beta)^2+7(\alpha+\beta)+12)z^2+2(\alpha-\beta)(\alpha+\beta+3)z+$$
$$+(\alpha-\beta)^2-(\alpha+\beta)-4)).$$

For $\alpha = \beta = 0$ the Jacobian polynomials reduce to the Legendre polynomials.

Substituting $z = 1$ into (16.2–5) reveals that only the first term on the right survives and we get

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}. \tag{16.2–6}$$

If in (16.2–5) we replace z by $-z$ and ν by $n-\nu$ we find

$$P^{(\alpha,\beta)}(-z) = \frac{1}{2^n}\sum_{\nu=0}^{n}\binom{n+\alpha}{\nu}\binom{n+\beta}{n-\nu}(-z-1)^{n-\nu}(-z+1)^{\nu}$$

whence

$$\boxed{P_n^{(\alpha,\beta)}(-z) = (-1)^n P_n^{(\beta,\alpha)}(z).} \tag{16.2–7}$$

For many purposes it is necessary to know the coefficient k_n of z^n in (16.2–5). This may be found by the following consideration. On multiplying the expansions

$$(1+z)^{\lambda} = \sum_{\nu=0}^{\infty}\binom{\lambda}{\nu}z^{\nu}, \qquad (1+z)^{\mu} = \sum_{\nu=0}^{\infty}\binom{\mu}{\nu}z^{\nu},$$

we readily find that the coefficient of z^n in the expansion of

$$(1+z)^{\lambda}(1+z)^{\mu} = (1+z)^{\lambda+\mu}$$

is

$$\sum_{\nu=0}^{n}\binom{\lambda}{n-\nu}\binom{\mu}{\nu} = \binom{\lambda+\mu}{n}.$$

Since

$$\binom{n+\xi}{n} = \frac{(n+\xi)\dots(1+\xi)}{n!} = \frac{\Gamma(n+\xi+1)}{n!\,\Gamma(\xi+1)},$$

we may infer that the coefficient of z^n in $P_n^{(\alpha,\beta)}(z)$ is

$$k_n = \frac{1}{2^n n!}\,\frac{\Gamma(2n+\alpha+\beta+1)}{\Gamma(n+\alpha+\beta+1)}. \tag{16.2–8}$$

The Jacobian polynomials have real coefficients if α and β are real. The coefficient k_n is certainly different from zero if $\alpha > -1$, $\beta > -1$. Under these restrictions the degree of the polynomials $P_n^{(\alpha,\beta)}(z)$ is precisely equal to n.

An alternative expression for the Jacobian polynomials is the modified Murphy expression

$$P_n^{(\alpha,\beta)}(z) = \binom{n+\alpha}{n} F(-n, n+\alpha+\beta+1; \alpha+1; \tfrac{1}{2}(1-z)) \tag{16.2–9}$$

the corresponding Riemann scheme being

$$P\begin{bmatrix} 0 & 0 & -n & \\ \alpha & -\beta & n+\alpha+\beta+1 & \tfrac{1}{2}(1-z) \end{bmatrix}. \tag{16.2–10}$$

In the hypergeometric differential equation (16.1–9) we introduce the variable $z' = 1-2z$. Omitting afterwards the primes we get

$$(1-z^2)w'' + (a+b+1-(a+b+2c+1)z)w' - abw = 0.$$

It follows that *the Jacobian polynomials are solutions of the differential equation*

$$\boxed{(1-z^2)w''+(\beta-\alpha-(2+\alpha+\beta)z)w'+n(n+\alpha+\beta+1)w = 0.} \qquad (16.2\text{–}11)$$

16.2.2 – The Rodrigues formula

A remarkable feature of the Jacobian polynomials is the fact that they can be defined by a formula which is a generalization of the Rodrigues formula (3.13–27).

Let k be one of the numbers $0, 1, \ldots, n$. We have

$$\begin{aligned}\frac{d^k}{dz^k}(z+1)^{n+\beta} &= (n+\beta)(n+\beta-1)\ldots(n+\beta-k+1)(z+1)^{n+\beta-k}\\ &= k!\binom{n+\beta}{k}(z+1)^{n+\beta-k}.\end{aligned}$$

Similarly

$$\frac{d^{n-k}}{dz^{n-k}}(z-1)^{n+\alpha} = (n-k)!\binom{n+\alpha}{n-k}(z-1)^{\alpha+k}.$$

It follows that (16.2–5) can be put in the form

$$P_n^{(\alpha,\beta)}(z) = \frac{(z-1)^{-\alpha}(z+1)^{-\beta}}{2^n n!}\sum_{\nu=0}^{n}\binom{n}{\nu}\frac{d^{n-\nu}}{dz^{n-\nu}}(z-1)^{n+\alpha}\frac{d^\nu}{dz^\nu}(z+1)^{n+\beta}.$$

By virtue of Leibniz's rule for the derivative of a product we have

$$\boxed{P_n^{(\alpha,\beta)}(z) = \frac{1}{2^n n!\,p(z)}\frac{d^n}{dz^n}(p(z)(z^2-1)^n),} \qquad (16.2\text{–}12)$$

where the weight function $p(z)$ is given by

$$\boxed{p(z) = (1-z)^\alpha(1+z)^\beta.} \qquad (16.2\text{–}13)$$

In the case of Legendre's polynomials $p(z) = 1$ identically.

16.2.3 – Orthogonality

For the considerations in this section the assumptions $\alpha > -1, \beta > -1$ are essential. Proceeding as in section 3.14.2 we may derive relations of orthogonality for Jacobian polynomials. We start with the function

$$u(x) = p(x)(x^2-1)^n, \qquad (16.2\text{–}14)$$

where $p(z)$ is the function (16.2–13) and n a non-negative integer. Let m

denote one of the numbers 0, 1, . . ., n. By integrating by parts successively we find

$$\int_{-1}^{1} x^m u^{(n)}(x)dx = (-1)^m m! \int_{-1}^{1} u^{(n-m)}(x)dx.$$

As long as $m < n$ the function $u^{(n-m-1)}$ vanishes at $x = \pm 1$, for it has still a factor x^2-1. Hence

$$\int_{-1}^{1} u^{(n-m)}(x)dx = u^{(n-m-1)}(x)\Big|_{-1}^{+1} = 0.$$

Since according to (16.2–12)

$$u^{(n)}(x) = 2^n n! p(x) P_n^{(\alpha,\beta)}(x),$$

we conclude that

$$\int_{-1}^{1} p(x) x^m P_n^{(\alpha,\beta)}(x)dx = 0, \qquad m < n. \qquad (16.2\text{–}15)$$

There remains the task of the evaluation of this integral in the case that $m = n$. As above we find

$$\int_{-1}^{1} x^n u^{(n)}(x)dx = (-1)^n n! \int_{-1}^{1} u(x)dx$$

$$= (-1)^n n! \int_{-1}^{1} (1-x)^\alpha (1+x)^\beta (x^2-1)^n dx = n! \int_{-1}^{1} (1-x)^{\alpha+n}(1+x)^{\beta+n} dx.$$

If we put $1-x = 2t$ we find from (4.7–38) and (4.7–37)

$$\int_{-1}^{1} (1-x)^{n+\alpha}(1+x)^{n+\beta} dx = 2^{2n+\alpha+\beta+1} \int_{0}^{1} t^{n+\alpha}(1-t)^{n+\beta} dt$$

$$= 2^{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(2n+\alpha+\beta+2)}$$

and it follows that

$$\int_{-1}^{1} p(x) x^n P_n^{(\alpha,\beta)}(x)dx = 2^{n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(2n+\alpha+\beta+2)}. \qquad (16.2\text{–}16)$$

It is now easy to see that (16.2–15) and (16.2–16) combine into *the relation of orthogonality* (taking into account (16.2–8))

$$\int_{-1}^{1} p(x) P_m^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(x)dx = \begin{cases} 0, & \text{if } m < n, \\ \dfrac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \dfrac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)n!}, & \text{if } m = n. \end{cases} \qquad (16.2\text{–}17)$$

If $\alpha = \beta = 0$ this reduces to the corresponding relations for the Legendre functions.

16.2.4 – Recurrent relations

Proceeding as in section 3.14.3 we may derive a recurrence formula for the Jacobian polynomials (with the same parameters α and β) which reduces to (3.14–11) if $\alpha = \beta = 0$. Since under the assumption $\alpha > -1$, $\beta > -1$ the polynomial $P_n^{(\alpha,\beta)}(z)$ is a polynomial of exactly the degree n the first theorem of section 3.14.3 remains valid for Jacobian polynomials.

We start again with (3.14–9), where now $P_k(z)$ stands for $P_k^{(\alpha,\beta)}(z)$. Multiplying both members by $p(z)P_m^{(\alpha,\beta)}(z)$, $m = 0, \ldots, n-2$ and integrating between -1 and $+1$ we find that only the coefficients c_{n-1}, c_n, c_{n+1} survive. (In the case of Legendre polynomials also $c_n = 0$, because $zP_n^2(z)$ is an odd function). Thus we have

$$zP_n^{(\alpha,\beta)}(z) = c_{n-1}P_{n-1}^{(\alpha,\beta)}(z) + c_nP_n^{(\alpha,\beta)}(z) + c_{n+1}P_{n+1}^{(\alpha,\beta)}(z). \qquad (16.2\text{–}18)$$

Equating coefficients of z^{n+1} in (16.2–18) yields

$$k_n = c_{n+1}k_{n+1},$$

where k_n is the coefficient of z^n in $P_n^{(\alpha,\beta)}(z)$. It follows that $c_{n+1} \neq 0$ and we may write

$$\boxed{P_{n+1}^{(\alpha,\beta)}(z) = (A_nz + B_n)P_n^{(\alpha,\beta)}(z) + C_nP_{n-1}^{(\alpha,\beta)}(z),} \quad n > 0 \qquad (16.2\text{–}19)$$

with

$$A_n = \frac{k_{n+1}}{k_n}. \qquad (16.2\text{–}20)$$

Let h_n denote the coefficient of z^{n-1} in $P_n^{(\alpha,\beta)}(z)$. By equating coefficients of z^n in (16.2–19) we find

$$h_{n+1} = A_nh_n + B_nk_n$$

whence

$$B_n = \frac{h_{n+1}}{k_n} - A_n\frac{h_n}{k_n} = A_n\left(\frac{h_{n+1}}{k_{n+1}} - \frac{h_n}{k_n}\right). \qquad (16.2\text{–}21)$$

By g_n we shall denote the value of the integral (16.2–17) for $m = n$. Multiplying both members of (16.2–19) by $p(z)P_{n-1}^{(\alpha,\beta)}(z)$ and integrating from -1 to $+1$ yields by virtue of (16.2–17)

$$\begin{aligned} C_ng_{n-1} &= A_n\int_{-1}^{1} p(x)xP_{n-1}^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(x)dx \\ &= A_nk_{n-1}\int_{-1}^{1} p(x)x^nP_n^{(\alpha,\beta)}(x)dx \\ &= A_n\frac{k_{n-1}}{k_n}\int_{-1}^{1} p(x)(P_n^{(\alpha,\beta)}(x))^2\,dx = \frac{A_n}{A_{n-1}}\,g_n, \end{aligned}$$

whence

$$C_n = \frac{A_n}{A_{n-1}} \frac{g_n}{g_{n-1}}. \qquad (16.2\text{–}22)$$

The values of the coefficients k_n and g_n may be taken from (16.2–8) and (16.2–17) respectively. The evaluation of h_n requires another relation between binomial coefficients. Differentiating

$$(1+z)^\lambda = \sum_{\nu=0}^{\infty} \binom{\lambda}{\nu} z^\nu$$

yields

$$\lambda z(1+z)^{\lambda-1} = \sum_{\nu=0}^{\infty} \nu \binom{\lambda}{\nu} z^\nu.$$

The coefficient of z^n in

$$\lambda z(1+z)^{\lambda-1}(1+z)^\mu = \lambda z(1+z)^{\lambda+\mu-1}$$

is

$$\sum_{\nu=0}^{n} \nu \binom{\lambda}{\nu}\binom{\mu}{n-\nu} = \lambda \binom{\lambda+\mu-1}{n-1}, \qquad (16.2\text{–}23)$$

the desired relation.

Since

$$\begin{aligned}(z-1)^m(z+1)^{n-m} &= (z^m - mz^{m-1} + \ldots)(z^{n-m} + (n-m)z^{n-m-1} + \ldots) \\ &= z^n + (n-2m)z^{n-1} + \ldots,\end{aligned}$$

it is clear from (16.2–5) that h_n is equal to

$$\begin{aligned}\frac{1}{2^n}\sum_{\nu=0}^{n}\binom{n+\alpha}{n-\nu}\binom{n+\beta}{\nu}(n-2\nu) &= \frac{1}{2^n}\sum_{\nu=0}^{n}(n-\nu)\binom{n+\alpha}{n-\nu}\binom{n+\beta}{\nu} + \\ &\quad - \frac{1}{2^n}\sum_{\nu=0}^{\infty}\nu\binom{n+\alpha}{n-\nu}\binom{n+\beta}{\nu} \\ &= \frac{1}{2^n}(n+\alpha)\binom{2n+\alpha+\beta-1}{n-1} - \frac{1}{2^n}(n+\beta)\binom{2n+\alpha+\beta-1}{n-1} \\ &= \frac{1}{2^n}(\alpha-\beta)\binom{2n+\alpha+\beta-1}{n-1}.\end{aligned}$$

Summing up we have

$$\begin{aligned}k_n &= \frac{1}{2^n}\binom{2n+\alpha+\beta}{n} = \frac{1}{2^n n!}\frac{\Gamma(2n+\alpha+\beta+1)}{\Gamma(n+\alpha+\beta+1)}, \\ h_n &= \frac{\alpha-\beta}{2^n}\binom{2n+\alpha+\beta-1}{n-1} = \frac{\alpha-\beta}{2^n(n-1)!}\frac{\Gamma(2n+\alpha+\beta)}{\Gamma(n+\alpha+\beta+1)}, \qquad (16.2\text{–}24) \\ g_n &= \frac{2^{\alpha+\beta+1}}{(2n+\alpha+\beta+1)n!}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)}.\end{aligned}$$

By easy computation we may evaluate the constants occurring in (16.2–19). They are

$$\boxed{\begin{aligned} A_n &= \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{2(n+1)(n+\alpha+\beta+1)}, \\ B_n &= \frac{(\alpha^2-\beta^2)(2n+\alpha+\beta+1)}{2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}, \\ C_n &= \frac{(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)}{(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}. \end{aligned}} \qquad (16.2\text{–}25)$$

In order to find a generalization of (3.14–16) for the Jacobian polynomials with indices α and β we observe that the coefficient of z^{n+1} in

$$(z^2-1)\frac{d}{dz}P_n^{(\alpha,\beta)}(z)-nP_n^{(\alpha,\beta)}(z)$$

is $nk_n-nk_n = 0$. Hence we have a relation

$$(x^2-1)\frac{d}{dx}P_n^{(\alpha,\beta)}(x)-nP_n^{(\alpha,\beta)}(x) = \sum_{\nu=0}^{n} c_\nu P_\nu^{(\alpha,\beta)}(x).$$

Next we multiply both members of this equation by $p(x)P_m^{(\alpha,\beta)}(x)$, $m = 0, 1, \ldots, n-2$. Integrating from -1 to $+1$ yields for the expression on the left thus obtained

$$\int_{-1}^{1} p(x)(x^2-1)\frac{d}{dx}P_n^{(\alpha,\beta)}(x)P_m^{(\alpha,\beta)}(x)dx$$
$$= \int_{-1}^{1} p(x)(x^2-1)\left(\frac{d}{dx}P_n^{(\alpha,\beta)}(x)\right)\tilde{P}_m'(x)dx$$

where $\tilde{P}_m(x)$ is a polynomial whose derivative is $P_m^{(\alpha,\beta)}(x)$. Integrating by parts we get

$$p(x)(x^2-1)\frac{d}{dx}P_n^{(\alpha,\beta)}(x)\tilde{P}_m(x)\Big|_{-1}^{1}$$
$$-\int_{-1}^{1}\frac{d}{dx}\left(p(x)(x^2-1)\frac{d}{dx}P_n^{(\alpha,\beta)}(x)\right)\tilde{P}_m(x)dx.$$

The first term vanishes. The second term may be evaluated if we bring the differential equation (16.2–11) into another form. Since

$$\beta-\alpha-(2+\alpha+\beta)z = (1+\beta)(1-z)-(1+\alpha)(1-z),$$

we may bring this equation into the form

$$((1-z)^{\alpha+1}(1+z)^{\beta+1}w')'+n(n+\alpha+\beta+1)(1-z)^{\alpha}(1+z)^{\beta}w = 0,$$

that is to say

$$\boxed{(p(z)(1-z^2)w')'+\lambda_n p(z)w=0,} \tag{16.2.-26}$$

with

$$\lambda_n = n(n+\alpha+\beta+1). \tag{16.2-27}$$

Hence the above integral is equal to

$$-\lambda_n \int_{-1}^{1} p(x)P_n^{(\alpha,\beta)}(x)\tilde{P}_m(x)dx = 0$$

since the degree of $\hat{P}_m(x)$ does not exceed $n-1$.

As a consequence we have

$$(z^2-1)\frac{d}{dz}P_n^{(\alpha,\beta)}(z)-nzP_n^{(\alpha,\beta)}(z) = a_nP_n^{(\alpha,\beta)}(z)+b_nP_{n-1}^{(\alpha,\beta)}(z). \tag{16.2-28}$$

Equating coefficients of z^n on both sides yields

$$(n-1)h_n - nh_n = a_nk_n,$$

whence

$$a_n = -\frac{h_n}{k_n}.$$

On taking $z = 1$ and observing that by virtue of (16.2-6)

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$$

we get

$$-n\binom{n+\alpha}{n} = a_n\binom{n+\alpha}{n} + b_n\binom{n+\alpha-1}{n-1}$$

and so

$$b_n = -(a_n+n)\frac{n+\alpha}{n}. \tag{16.2-29}$$

Inserting the known values for k_n and h_n we finally have in (16.2-28)

$$\boxed{\begin{aligned} a_n &= (\beta-\alpha)\frac{n}{2n+\alpha+\beta}, \\ b_n &= -2\frac{(n+\alpha)(n+\beta)}{2n+\alpha+\beta}. \end{aligned}} \tag{16.2-30}$$

16.2.5 – Gegenbauer polynomials

A natural generalization of the definitions of Legendre polynomials by means of the expansion (3.13–29) of a generating function is the definition of functions $C_n^\lambda(z)$, $n = 0, 1, \ldots$, by means of

$$\boxed{(1-2zw+w^2)^{-\lambda} = \sum_{v=0}^{\infty} C_v^\lambda(z) w^v.} \quad \lambda \neq 0. \tag{16.2–31}$$

Expanding the function on the left by the binomial theorem we get

$$\sum_{v=0}^{\infty} (-1)^v \binom{-\lambda}{v} w^v(2z-w)^v = \sum_{v=0}^{\infty} \frac{(\lambda)_v}{v!} w^v(2z-w)^v.$$

The coefficient of w^n is

$$\frac{(\lambda)_n}{n!}(2z)^n - \frac{(\lambda)_{n-1}}{(n-1)!}\binom{n-1}{1}(2z)^{n-2} + \ldots = \sum_{v=0}^{[\frac{1}{2}n]} \frac{(\lambda)_{n-v}}{(n-2v)!}(-1)^v(2z)^{n-2v}.$$

Thus

$$\boxed{C_n^\lambda(z) = \sum_{v=0}^{[\frac{1}{2}n]} \frac{(-1)^v(\lambda)_{n-v}}{v!(n-2v)!}(2z)^{n-2v}} \tag{16.2–32}$$

is a polynomial. The coefficient of the leading term is $2^n(\lambda)_n/n!$ and it appears that the degree of this polynomial is precisely equal to n if λ is neither zero, nor a negative integer. The above polynomials are called *Gegenbauer polynomials.* They reduce to Legendre polynomials if $\lambda = \frac{1}{2}$. The first polynomials are

$$C_0^\lambda(z) = 1, \qquad C_1^\lambda(z) = 2\lambda z, \qquad C_2^\lambda(z) = 2\lambda(\lambda+1)z^2 - \lambda,$$
$$C_3^\lambda(z) = \tfrac{4}{3}\lambda(\lambda+1)(\lambda+2)z^3 - 2\lambda(\lambda+1)z,$$
$$C_4^\lambda(z) = \tfrac{2}{3}\lambda(\lambda+1)(\lambda+2)(\lambda+3)z^4 - 2\lambda(\lambda+1)(\lambda+2)z^2 + \tfrac{1}{2}\lambda(\lambda+1),$$

etc. We wish to derive some properties of the Gegenbauer polynomials by using the generating function. In (16.2–31) we replace z by $-z$ and w by $-w$ and we get

$$\boxed{C_n^\lambda(-z) = (-1)^n C_n^\lambda(z).} \tag{16.2–33}$$

This means that $C_n^\lambda(z)$ contains only even or odd powers of z according as n is even or odd.

By taking $z = 1$ we deduce from

$$(1-w)^{-2\lambda} = \sum_{v=0}^{\infty} (-1)^v \binom{-2\lambda}{v} w^v = \sum_{v=0}^{\infty} \frac{(2\lambda)_v}{v!} w^v$$

that

$$\boxed{C_n^\lambda(1) = \frac{(2\lambda)_n}{n!},} \tag{16.2–34}$$

A recurrent relation like (3.14–3) may be obtained by differentiating both members of (16.2–31) with respect to w. We find

$$(1-2zw+w^2)\sum_{\nu=1}^{\infty} \nu C_\nu^\lambda(z)w^{\nu-1} = 2\lambda(z-w)\sum_{\nu=0}^{\infty} C_\nu^\lambda(z)w^\nu. \tag{16.2–35}$$

Comparing terms with equal powers of w we find

$$(n+1)C_{n+1}^\lambda(z)-(2n+2\lambda)zC_n^\lambda(z)+(n+2\lambda-1)C_{n-1}^\lambda(z) = 0, \tag{16.2–36}$$

valid for $n \geqq 0$, if we agree that $C_{-1}^\lambda(z)$ stands for the constant 0.

Differentiating (16.2–31) with respect to z yields

$$2\lambda(1-2zw+w^2)^{-(\lambda+1)} = \sum_{\nu=1}^{\infty} \frac{d}{dz} C_\nu^\lambda(z)w^{\nu-1}, \tag{16.2–37}$$

or

$$\sum_{\nu=0}^{\infty} 2\lambda C_\nu^{\lambda+1}(z)w^\nu = \sum_{\nu=0}^{\infty} \frac{d}{dz} C_{\nu+1}^\lambda(z)w^\nu. \tag{16.2–38}$$

It follows that

$$\boxed{\frac{d}{dz} C_n^\lambda(z) = 2\lambda C_{n-1}^{\lambda+1}(z).} \tag{16.2–39}$$

As usual we denote the leading coefficient in the polynomial $C_n^\lambda(z)$ by k_n. Differentiating this polynomial n times we have by virtue of (16.2–39)

$$n!k_n = 2^n(\lambda)_n C_0^{\lambda+n}(z) = 2^n(\lambda)_n,$$

whence

$$k_n = \frac{2^n(\lambda)_n}{n!} \tag{16.2–40}$$

in accordance with the remark following the formula (16.2–32).

We may bring (16.2–35) into the form

$$2\lambda(z-w)(1-2zw+w^2)^{-(\lambda+1)} = \sum_{\nu=1}^{\infty} \nu C_\nu^\lambda(z)w^{\nu-1}. \tag{16.2–41}$$

Combining this with (16.2–37) we get

$$(z-w)\sum_{\nu=1}^{\infty} \frac{d}{dz} C_\nu^\lambda(z)w^{\nu-1} = \sum_{\nu=1}^{\infty} \nu C_\nu^\lambda(z)w^{\nu-1},$$

whence, equating equal powers of w,

$$\boxed{z\frac{d}{dz}C_n^\lambda(z)-\frac{d}{dz}C_{n-1}^\lambda(z) = nC_n^\lambda(z).} \tag{16.2–42}$$

This equation corresponds to (3.14–14).

Multiplying both members of (16.2–37) by $1-w^2$ and those of (16.2–41) by $-2w$ we find, after adding corresponding members of the equations thus obtained,

$$2\lambda(1-2zw+w^2)^{-\lambda} = \sum_{\nu=1}^{\infty}\frac{d}{dz}C_\nu^\lambda(z)w^{\nu-1}(1-w^2)-\sum_{\nu=1}^{\infty}2\nu C_\nu^\lambda(z)w^\nu$$

or

$$\sum_{\nu=0}^{\infty}(2\lambda+2\nu)C_\nu^\lambda(z)w^\nu = \sum_{\nu=0}^{\infty}\frac{d}{dz}C_{\nu+1}^\lambda(z)w^\nu-\sum_{\nu=0}^{\infty}\frac{d}{dz}C_{\nu-1}^\lambda(z)w^\nu$$

where $C_{-1}^\lambda(z)$ stands for a polynomial which is identically zero. It follows that

$$\boxed{\frac{d}{dz}C_{n+1}^\lambda(z)-\frac{d}{dz}C_{n-1}^\lambda(z) = (2n+2\lambda)C_n^\lambda(z).} \tag{16.2–43}$$

This equation corresponds to (3.14–13).

Eliminating $dC_{n-1}^\lambda(z)/dz$ from (16.2–42) and (16.2–43) yields

$$\boxed{\frac{d}{dz}C_{n+1}^\lambda(z)-z\frac{d}{dz}C_n^\lambda(z) = (n+2\lambda)C_n^\lambda(z).} \tag{16.2–44}$$

This equation corresponds to (3.14–15).

The method for obtaining (3.14–16) is also applicable to Gegenbauer polynomials and we obtain

$$(z^2-1)\frac{d}{dz}C_n^\lambda(z) = nzC_n^\lambda(z)-(n-1+2\lambda)C_{n-1}^\lambda(z). \tag{16.2–45}$$

Differentiating both members of this equation yields

$$(z^2-1)\frac{d^2}{dz^2}C_n^\lambda(z)+2z\frac{d}{dz}C_n^\lambda(z)$$
$$= nC_n^\lambda(z)+nz\frac{d}{dz}C_n^\lambda(z)-(n-1+2\lambda)\frac{d}{dz}C_{n-1}^\lambda(z).$$

Taking (16.2–42) into account we deduce, by eliminating $dC_{n-1}^\lambda(z)/dz$.

The Gegenbauer polynomials $C_n^\lambda(z)$ are solutions of the differential equation

$$\boxed{(1-z^2)w''-(2\lambda+1)zw'+n(n+2\lambda)w = 0.} \tag{16.2–46}$$

16.2.6 – Ultraspherical polynomials and Gegenbauer polynomials

The differential equation (16.2–11) reduces to (16.2–46) if we take $\alpha = \beta = \lambda - \frac{1}{2}$. Since by appropriately changing the independent variable this equation can be transformed into a hypergeometric differential equation we have in view of (16.2–34) a formula like (16.2–9), viz.,

$$C_n^\lambda(z) = \frac{(2\lambda)_n}{n!} F(-n, n+\lambda; \lambda+\tfrac{1}{2}; \tfrac{1}{2}(1-z)). \tag{16.2–47}$$

The multiplier of the hypergeometric series in (16.2–9) is now

$$\binom{n+\lambda-\frac{1}{2}}{n} = \frac{(\lambda+\frac{1}{2})_n}{n!}.$$

This leads to the relation

$$\boxed{C_n^\lambda(z) = \frac{(2\lambda)_n}{(\lambda+\frac{1}{2})_n} P_n^{(\lambda-\frac{1}{2}, \lambda-\frac{1}{2})}(z).} \tag{16.2–48}$$

The Jacobi polynomials with $\alpha = \beta$ are called *ultraspherical polynomials.* Hence Gegenbauer polynomials and ultraspherical polynomials are essentially the same mathematical entities. Results obtained for the Gegenbauer polynomials may be translated into results on the ultraspherical polynomials, or conversely. Certain properties of the ultraspherical polynomials are due solely to their being Jacobian polynomials, others are a consequence of the equality of the parameters α and β.

We can now immediately write down the *relations of orthogonality for the Gegenbauer polynomials.* Taking into account (4.6–26) we have

$$\boxed{\begin{aligned}&\int_{-1}^{1} (1-x^2)^{\lambda-\frac{1}{2}} C_m^\lambda(x) C_n^\lambda(x)dx \\ &\qquad = \begin{cases} 0, & \text{if} \quad m \neq n, \\ 2^{1-2\lambda}\pi \dfrac{\Gamma(n+2\lambda)}{(n+\lambda)n!\Gamma^2(\lambda)}, & \text{if} \quad m = n. \end{cases}\end{aligned}} \tag{16.2–49}$$

It is readily seen that this formula implies (3.14–7) if we take $\lambda = \frac{1}{2}$.

16.2.7 – Čebishev polynomials

The expansion (16.2–31) breaks down if $\lambda = 0$. We may, however, consider

$$\frac{(1-2zw+w^2)^{-\lambda}-1}{\lambda}$$

and make λ tend to zero. Then we find

$$\boxed{-\log(1-2zw+w^2) = \sum_{\nu=0}^{\infty} C_\nu^0(z)w^\nu,} \qquad (16.2\text{–}50)$$

with

$$C_n^0(z) = \lim_{\lambda\to 0} \frac{C_n^\lambda(z)}{\lambda}, \qquad n > 0; \qquad C_0^0(z) = 0. \qquad (16.2\text{–}51)$$

Hence, in accordance with (16.2–32),

$$\boxed{C_n^0(z) = \sum_{\nu=0}^{[\frac{1}{2}n]} \frac{(-1)^\nu(n-\nu-1)!}{\nu!(n-2\nu)!}(2z)^{n-2\nu},} \qquad n > 0. \qquad (16.2\text{–}52)$$

The first polynomials are

$$C_1^0(z) = 2z, \qquad C_2^0(z) = 2z^2-1, \qquad C_3^0(z) = \tfrac{8}{3}z^3-2z,$$
$$C_4^0(z) = 4z^4-4z^2+\tfrac{1}{2}, \ldots.$$

The recurrent relation (16.2–36) is still valid if $\lambda = 0$, but only if $n \geqq 2$. In addition we have

$$C_2^0(z) = zC_1^0(z)-1. \qquad (16.2\text{–}53)$$

The expression in terms of ultraspherical polynomials takes the form

$$C_n^0(z) = \frac{2^{2n+1}(n!)^2}{n(2n)!} P_n^{(-\frac{1}{2},-\frac{1}{2})}(z). \qquad (16.2\text{–}54)$$

New light is shed on these polynomials if in (16.2–50) we take $w = z > 0$, $z = \cos\varphi$. This is done automatically if in the logarithmic series

$$-\log(1-z) = \sum_{\nu=1}^{\infty} \frac{z^\nu}{\nu}$$

we substitute $z = r\,e^{i\varphi}$ and take the real part. Then we find

$$-\log(1-2r\cos\varphi+r^2) = 2\sum_{\nu=1}^{\infty} r^\nu \frac{\cos\nu\varphi}{\nu},$$

whence

$$C_n^0(\cos\varphi) = \frac{2}{n}\cos n\varphi, \qquad n > 0. \qquad (16.2\text{–}55)$$

It is common practice to write

$$\boxed{\cos n\varphi = T_n(\cos\varphi).} \qquad (16.2\text{–}56)$$

The polynomials $T_n(z)$ are called *Čebishev polynomials.* They are essentially Gegenbauer polynomials, for we have

$$T_n(z) = \frac{n}{2} C_n^0(z), \qquad n > 0. \tag{16.2–57}$$

The first polynomials are

$$T_0(z) = 1, \quad T_1(z) = z, \quad T_2(z) = 2z^2-1, \quad T_3(z) = 4z^3-3z,$$
$$T_4(z) = 8z^4-8z^2+1.$$

These polynomials can be obtained in a straight-forward manner by expanding $\cos n\varphi$ as a polynomial in $\cos \varphi$ by elementary computation. This may be done in a most easy way, starting with De Moivre's formula

$$(\cos \varphi + i \sin \varphi)^n = \cos n\varphi + i \sin n\varphi \tag{16.2–58}$$

and equating the real parts. We find

$$\cos n\varphi = \sum_{\nu=0}^{[\frac{1}{2}n]} (-1)^\nu \binom{n}{2\nu} \cos^{n-2\nu} \varphi \sin^{2\nu} \varphi,$$

whence

$$T_n(z) = \sum_{\nu=0}^{[\frac{1}{2}n]} (-1)^\nu \binom{n}{2\nu} z^{n-2\nu}(1-z^2)^\nu. \tag{16.2–59}$$

The elementary relations of orthogonality for the circular functions

$$\int_0^\pi \cos m\varphi \cos n\varphi \, d\varphi = \begin{cases} 0, & \text{if } \; m \neq n, \\ \frac{1}{2}\pi, & \text{if } \; m = n > 0, \\ \pi, & \text{if } \; m = n = 0, \end{cases} \tag{16.2–60}$$

lead to the *relations of orthogonality for the Čebishev polynomials*

$$\boxed{\int_{-1}^1 (1-x^2)^{-\frac{1}{2}} T_m(x) T_n(x) dx = \begin{cases} 0, & \text{if } \; m \neq n, \\ \frac{1}{2}\pi, & \text{if } \; m = n > 0, \\ \pi, & \text{if } \; m = n = 0. \end{cases}} \tag{16.2–61}$$

We may obtain them by a limiting process from (16.2–49) by multiplying first by r^2/λ^2 and making λ tend to zero, provided that $n > 0$. The case $n = 0$ is trivial.

From (16.2–58) we also find (replacing first n by $n+1$)

$$\sin (n+1)\varphi = \sum_{\nu=0}^{[\frac{1}{2}n]} (-1)^\nu \binom{n+1}{2\nu+1} \cos^{n-2\nu} \varphi \sin^{2\nu+1} \varphi$$

and it follows that

$$\boxed{\frac{\sin (n+1)\varphi}{\sin \varphi} = U_n(\cos \varphi)} \tag{16.2–62}$$

is a polynomial in $\cos\varphi$, whence

$$U_n(z) = \sum_{\nu=0}^{[\frac{1}{2}n]} (-1)^\nu \binom{n+1}{2\nu+1} z^{n-2\nu}(1-z^2)^\nu. \tag{16.2–63}$$

The polynomials $U_n(z)$ are called *Čebishev polynomials of the second kind.* They are Gegenbauer polynomials as follows from the expansion

$$\frac{1}{1-z} = \sum_{\nu=0}^{\infty} z^\nu$$

if we replace z by $re^{i\varphi}$. Equating imaginary parts yields

$$\frac{1}{1-2r\cos\varphi+r^2} = \sum_{\nu=0}^{\infty} r^\nu \frac{\sin(\nu+1)\varphi}{\sin\varphi},$$

whence

$$C_n^1(\cos\varphi) = U_n(\cos\varphi). \tag{16.2–64}$$

The first polynomials of the second kind are

$$U_0(z) = 1,\quad U_1(z) = 2z,\quad U_2(z) = 4z^2-1,\quad U_3(z) = 8z^3-4z,$$
$$U_4(z) = 16z^4-12z^2+1, \ldots .$$

The *relations of orthogonality for these functions* are

$$\boxed{\int_{-1}^{1} (1-x^2)^{\frac{1}{2}} U_m(x) U_n(x)\,dx = \begin{cases} 0, & \text{if } m \neq n, \\ \frac{1}{2}\pi, & \text{if } m = n. \end{cases}} \tag{16.2–65}$$

They are consequences of

$$\int_0^\pi \sin(m+1)\varphi \sin(n+1)\varphi\, d\varphi = \begin{cases} 0, & \text{if } m \neq n, \\ \frac{1}{2}\pi, & \text{if } m = n, \end{cases} \tag{16.2–66}$$

and can be obtained from (16.2–49) by taking $\lambda = 1$.

Finally we wish to list some relations between the polynomials of the first and the second kind, viz.,

$$\frac{d}{dz} T_{n+1}(z) = (n+1)U_n(z),$$
$$T_n(z) = U_n(z) - zU_{n-1}(z),$$
$$(1-z^2)U_{n-1}(z) = zT_n(z) - T_{n+1}(z).$$

They are direct consequences of elementary relations between circular functions.

16.3 – The hypergeometric series as functions of the parameters

16.3.1 – Regularity

The series

$$F(a, b; c; z) = \sum_{\nu=0}^{\infty} \frac{(a)_\nu (b)_\nu}{(c)_\nu} \frac{z^\nu}{\nu!} \tag{16.3–1}$$

is absolutely and uniformly convergent for $|z| \leqq R < 1$, independent of the choice of the parameters a, b, c, as long as c is neither zero nor a negative integer. In order to avoid the difficulty created by the possibility of zero denominators on the right of (16.3–1) we prefer to investigate the series

$$\frac{F(a, b; c; z)}{\Gamma(c)} = \sum_{\nu=0}^{\infty} \frac{(a)_\nu (b)_\nu}{\Gamma(c+\nu)} \frac{z^\nu}{\nu!}. \tag{16.3–2}$$

Recalling (4.6–33), that is the formula

$$\lim_{n\to\infty} \frac{\Gamma(s+n)}{\Gamma(n)n^s} = 1, \tag{16.3–3}$$

we find that

$$\lim_{n\to\infty} \left| \frac{(a)_n (b)_n}{\Gamma(c+n)} \frac{z^{\frac{1}{2}n}}{n!} \right|$$

$$= \frac{1}{\Gamma(a)\Gamma(b)} \lim_{n\to\infty} \left| \frac{\Gamma(a+n)}{\Gamma(n)n^a} \right| \left| \frac{\Gamma(b+n)}{\Gamma(n)n^b} \right| \left| \frac{\Gamma(n)n^c}{\Gamma(c+n)} \right| \frac{|z|^{\frac{1}{2}n}}{|n^{1+c-a-b}|},$$

$$= \frac{1}{\Gamma(a)\Gamma(b)} \lim_{n\to\infty} \frac{|z|^{\frac{1}{2}n}}{|n^{1+c-a-b}|} = 0,$$

uniformly in $|z|$ as long as $|z| \leqq R < 1$. Now $(a)_n$, $(b)_n$ and, if n is sufficiently large, also $\Gamma(c+n)$ are holomorphic functions of their variables a, b and c respectively. Hence if a, b and c vary throughout a bounded and closed set the moduli of $(a)_n$, $(b)_n$ and $1/\Gamma(c+n)$ are bounded for fixed n. Hence there is a constant K, independent of a, b, c and R such that

$$\left| \frac{(a)_n (b)_n}{\Gamma(c+n)} \right| \frac{|z|^n}{n!} < K|z|^{\frac{1}{2}n}.$$

The series

$$\sum_{\nu=0}^{\infty} K|z|^{\frac{1}{2}\nu}$$

is convergent for $|z| < 1$. Hence, if the range of any variable a, b or c is a closed and bounded set, the series (16.3–2) is uniformly convergent. It follows (section 2.20.3)

If $|z| < 1$ *the function* $F(a, b; c; z)$ *is holomorphic in* a, b *and* c *separately for all finite* a, b, c *except for simple poles at* $c = 0, -1, -2, \ldots$.

The residues at these poles may be found by evaluating the limit of

$$\begin{aligned}(c+n)F(a, b; c; z) &= (c+n)\sum_{\nu=0}^{n} \frac{(a)_\nu(b)_\nu}{(c)_\nu}\frac{z^\nu}{\nu!} \\ &+ \frac{(a)_{n+1}(b)_{n+1}}{(c)_n}\frac{z^{n+1}}{(n+1)!}\sum_{\nu=0}^{\infty}\frac{(a+n+1)_\nu(b+n+1)_\nu}{(c+n+1)_\nu}\frac{z^\nu}{(n+2)_\nu},\end{aligned} \tag{16.3–4}$$

as $c \to -n$. The last series is dominated by the series

$$F(|a|+n+1, |b|+n+1; -|c|+n+1; z)$$

and, as a consequence, we may perform the limiting process term-by-term. Thus we find

If we consider the series $F(a, b; c; z)$ *as a function of the third parameter* c, *it presents simple poles at* $c = -n$, $n = 0, 1, \ldots,$ *with residues*

$$\operatorname*{Res}_{c=-n} F(a, b; c; z) = A_n z^{n+1} F(a+n+1, b+n+1; 2+n; z) \tag{16.3–5}$$

with

$$A_n = (-1)^n \frac{(a)_{n+1}(b)_{n+1}}{n!(n+1)!}. \tag{16.3–6}$$

The first theorem of this section is very important. For if we have proved a certain relation between hypergeometric series under restrictive conditions for the parameters, the result remains true if these conditions are relaxed, provided the various expressions involved have a meaning in the general case.

16.3.2 – Gauss's relations between contiguous series

The hypergeometric series considered as functions of the parameters a, b and c satisfy certain functional relations discovered by Gauss. They are relations between the six series

$$F(a\pm 1, b; c; z), F(a, b\pm 1; c; z), F(a, b; c\pm 1; z) \tag{16.3–7}$$

and the series $F(a, b; c; z)$. The series (16.3–7), briefly denoted by

$$F_{a+}, F_{a-}, F_{b+}, F_{b-}, F_{c+}, F_{c-}$$

are said to be *contiguous* to F. They are also contiguous in the sense of Riemann (section 15.6.3), but the converse is not true; there are more contiguous functions in the sense of Riemann. Gauss's theorem is a particular instance of Riemann's general theorem about contiguous functions.

It is readily seen that to pass from F to F_{a+} we must multiply the coefficient of z^n by $a+n$ and remove the factor a. This is effected by the operator $\vartheta+a$, where ϑ has the same meaning as in section 16.1.2. To pass from F to F_{c-}, we must multiply the coefficient of z^n by $n+c-1$ and remove the factor $c-1$. This is effected by the operator $\vartheta+c-1$. Thus

$$aF_{a+} = (\vartheta+a)F, \tag{16.3–8}$$

$$bF_{b+} = (\vartheta+b)F, \tag{16.3–9}$$

$$(c-1)F_{c-} = (\vartheta+c-1)F. \tag{16.3–10}$$

Next we make use of equation (16.1–13) for F. The corresponding equation for F_{a-} is

$$(\vartheta(\vartheta+c-1)-z(\vartheta+a-1)(\vartheta+b))F_{a-} = 0,$$

or

$$(\vartheta+c-a-z(\vartheta+b))(\vartheta+a-1)F_{a-} = (c-a)(a-1)F_{a-}.$$

With the aid of (16.3–8) with $a-1$ written for a we get

$$(c-a)F_{a-} = (1-z)\vartheta F+(c-a-bz)F. \tag{16.3–11}$$

The proof breaks down if $a = 1$. But we may assume $a \neq 1$ and then make a tending to 1, as follows from the first theorem of the previous section.

From (16.3–9) we obtain a similar formula with a and b interchanged. The equation for F_{c+} corresponding to (16.1–13) is

$$(\vartheta(\vartheta+c)-z(\vartheta+a)(\vartheta+b))F_{c+} = 0,$$

or

$$(\vartheta-z(\vartheta+a+b-c))(\vartheta+c)F_{c+} = (c-a)(c-b)zF_{c+}.$$

With the aid of (16.3–10) we get

$$(c-a)(c-b)F_{c+} = c((1-z)F'+(c-a-b)F). \tag{16.3–12}$$

Summing up we have

$$\begin{aligned} zF' &= a(F_{a+}-F), \\ zF' &= b(F_{b+}-F), \\ zF' &= (c-1)(F_{c-}-F), \\ z(1-z)F' &= (c-a)F_{a-}+(a-c+bz)F, \\ z(1-z)F' &= (c-b)F_{b-}+(b-c+az)F, \\ c(1-z)F' &= (c-a)(c-b)F_{c+}+c(a+b-c)F. \end{aligned} \tag{16.3–13}$$

Thus we have seen that each of the functions (16.3–7) can be expressed in terms of F and the derivative of F. By eliminating this derivative F' we get fifteen relations between F and any two of its contiguous functions.

Particular examples of the relations (16.3–13) are Legendre's relations for the complete elliptic integrals. In fact $\mathsf{K}(z)$ and $\mathsf{E}(z)$ are contiguous, as is seen from (16.1–29). The first of (16.3–13) yields (14.5–24), the fourth yields (14.5–28).

16.3.3 – Mixed relations for Jacobian polynomials

In the recursive relations for Jacobian polynomials derived in section 16.2.4 only polynomials with the same parameters occur. The Gaussian relations between contiguous functions enable us to derive relations between Jacobian polynomials which present a shift about unity in the parameters. We shall give some striking examples.

First we find by eliminating F' from the first two relations of (16.3–13) the Gaussian relation between contiguous functions

$$(b-a)F = bF_{b+}-aF_{a+}. \qquad (16.3\text{–}14)$$

Replacing the variable z by $\frac{1}{2}(1-z)$, the parameters a, b and c by $-n$, $n+\alpha+\beta+1$ and $\alpha+1$ respectively, we find, taking into account (16.2–9),

$$(2n+\alpha+\beta+1)P_n^{(\alpha,\beta)}(z) = (n+\alpha+\beta+1)P_n^{(\alpha,\beta+1)}(z)+(n+\alpha)P_{n-1}^{(\alpha,\beta+1)}(z),$$

which becomes, after a shift from β to $\beta-1$,

$$(2n+\alpha+\beta)P_n^{(\alpha,\beta-1)}(z) = (n+\alpha+\beta)P_n^{(\alpha,\beta)}(z)+(n+\alpha)P_{n-1}^{(\alpha,\beta)}(z). \qquad (16.3\text{–}15)$$

This remains true if we replace z by $-z$. On applying (16.2–7) and interchanging afterwards α and β, we also have

$$(2n+\alpha+\beta)P_n^{(\alpha-1,\beta)}(z) = (n+\alpha+\beta)P_n^{(\alpha,\beta)}(z)-(n+\beta)P_{n-1}^{(\alpha,\beta)}(z). \qquad (16.3\text{–}16)$$

Subtracting corresponding members of (16.3–15) and (16.3–16) yields

$$\boxed{P_n^{(\alpha,\beta-1)}(z)-P_n^{(\alpha-1,\beta)}(z) = P_{n-1}^{(\alpha,\beta)}(z).} \qquad (16.3\text{–}17)$$

From the fourth and the fifth equation of (16.3–13) we find the Gaussian relation between contiguous functions

$$(c-a)F_{a-}-(c-b)F_{b-} = (b-a)(1-z)F. \qquad (16.3\text{–}18)$$

Proceeding as above we now find

$$\tfrac{1}{2}(1+z)(2n+\alpha+\beta+1)P_n^{(\alpha,\beta)}(z) = (n+1)F_{n+1}^{(\alpha,\beta-1)}(z)+(n+\beta)P_n^{(\alpha,\beta-1)}(z)$$

or, after a shift from β to $\beta+1$,

$$\tfrac{1}{2}(1+z)(2n+\alpha+\beta+2)P_n^{(\alpha,\beta+1)}(z) = (n+1)F_{n+1}^{(\alpha,\beta)}(z)+(n+\beta+1)P_n^{(\alpha,\beta)}(z). \qquad (16.3\text{–}19)$$

Replacing z by $-z$ and interchanging α and β we have, by virtue of (16.2–7),

$$\tfrac{1}{2}(1-z)(2n+\alpha+\beta+2)P_n^{(\alpha+1,\beta)}(z) = -(n+1)P_{n+1}^{(\alpha,\beta)}(z)+(n+\alpha+1)P_n^{(\alpha,\beta)}(z). \tag{16.3–20}$$

Adding corresponding members of (16.3–19) and (16.3–20) we get

$$\boxed{(1-z)P_n^{(\alpha+1,\beta)}(z)+(1+z)P_n^{(\alpha,\beta+1)}(z) = 2P_n^{(\alpha,\beta)}(z).} \tag{16.3–21}$$

The fact that in this way many relations for Jacobian polynomials may be obtained needs no further comment.

16.3.4 - The behaviour of the hypergeometric series on the circumference of convergence

The convergence of the hypergeometric series on the circumference of convergence depends on the values of the parameters a, b and c. We may assume that neither a nor b is zero or a negative integer, for then the series is a polynomial and there is no problem of convergence.

For the solution of our problem it is more convenient to consider the series

$$\frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} F(a, b; c; z) = \sum_{\nu=0}^{\infty} \frac{\Gamma(a+\nu)\Gamma(b+\nu)}{\Gamma(c+\nu)} \frac{z^\nu}{\nu!}. \tag{16.3–22}$$

Denoting the coefficient of z^n in the series on the right by c_n we have

$$c_n n^{1+c-a-b} = \frac{\Gamma(a+n)}{\Gamma(n)n^a} \frac{\Gamma(b+n)}{\Gamma(n)n^b} \frac{\Gamma(n)n^c}{\Gamma(c+n)} \tag{16.3–23}$$

and it follows from (16.3–3) that the expression on the right tends to 1 as $n \to \infty$. As a consequence the series $\Sigma|c_\nu|$ behaves like

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu^{1+\mathrm{Re}\,(c-a-b)}}.$$

Hence

The series (16.3–22) *is absolutely convergent on* $|z| = 1$ *if*

$$\mathrm{Re}\,(c-a-b) > 0 \tag{16.3–24}$$

and absolutely divergent if

$$\mathrm{Re}\,(c-a-b) \leqq 0. \tag{16.3–25}$$

If

$$\mathrm{Re}\,(c-a-b) \leqq -1 \tag{16.3–26}$$

the coefficients c_n do not tend to zero as $n \to \infty$ and the series turns out to be divergent at every point of the circumference $|z| = 1$.

Let us now consider the case

$$-1 < \mathrm{Re}\,(c-a-b) \leqq 0. \tag{16.3–27}$$

It is clear that the series F_{c+} and F_{a-} satisfy the conditions needed for convergence. By eliminating F' from the fourth and the sixth relation (16.3–13) we get

$$c(1-z)F = z(b-c)F_{c+} + cF_{a-}. \tag{16.3–28}$$

It follows that under the conditions (16.3–27) the series (16.3–22) is convergent on $|z| = 1$, except possibly at $z = 1$. However, the convergence is not absolute.

There remains the case $z = 1$. This is not quite so easy and we need more information than is contained in (16.3–3). For this purpose we use Stirlings's theorem for the gamma function. The formula (4.9–3) may be written in the form

$$\log \Gamma(z) = (z-\tfrac{1}{2}) \log z - z + \log \sqrt{2\pi} + \mu(z)$$

and it follows that

$$\log \frac{\Gamma(z+a)}{\Gamma(z)z^a} = (z+a-\tfrac{1}{2}) \log \left(1+\frac{a}{z}\right) - a + \mu(z+a) - \mu(z).$$

Now

$$(z+a-\tfrac{1}{2}) \log \left(1+\frac{a}{z}\right) = a + O(|z|^{-1})$$

and (4.9–10) states that

$$\mu(z) = O(|z|^{-1})$$

and is $\mu(z+a)$, provided that $z \to \infty$ receding indefinitely from the negative real axis.

Thus

$$\frac{\Gamma(z+a)}{\Gamma(z)z^a} = 1 + O(|z|^{-1}). \tag{16.3–29}$$

In particular

$$\frac{\Gamma(a+n)}{\Gamma(n)n^a} = 1 + O(n^{-1}). \tag{16.3–30}$$

The convergence of the series

$$\sum_{\nu=0}^{\infty} c_\nu = \sum_{\nu=0}^{\infty} \frac{\Gamma(a+\nu)\Gamma(b+\nu)}{\Gamma(c+\nu)\nu!} \tag{16.3–31}$$

may be examined by comparing it with the series

$$\sum_{\nu=0}^{\infty}(-1)^{\nu}\binom{s}{\nu} = \sum_{\nu=0}^{\infty}\frac{\Gamma(-s+\nu)}{\Gamma(-s)\nu!} \tag{16.3–32}$$

with $s = c-a-b$. For the time being we suppose that $s \neq 0$. By multiplying the series

$$(1-z)^{s} = \sum_{\nu=0}^{\infty}(-1)^{\nu}\binom{s}{\nu} z^{\nu} \tag{16.3–33}$$

by

$$(1-z)^{-1} = \sum_{\nu=0}^{\infty} z^{\nu}$$

we easily find that the sum of the first $n+1$ coefficients in (16.3–33) is the coefficient of z^n of the expansion of $(1-z)^{s-1}$, that is to say, the sum of the first $n+1$ terms of (16.3–32) is equal to

$$(-1)^{n}\binom{s-1}{n} = \frac{\Gamma(-s+n+1)}{\Gamma(-s+1)n!}.$$

From (16.3–3) follows that

$$\frac{\Gamma(-s+n+1)}{\Gamma(n+1)(n+1)^{-s}} \to 1$$

as $n \to \infty$, and since $0 \leq \operatorname{Re}(-s)$, $s \neq 0$, by assumption, the sum of the first $n+1$ terms of (16.3–32) does not tend to a finite limiting value. Hence this series is divergent.

Next we observe

$$\frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)\Gamma(a+b-c+n)} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(c)\Gamma(a+b-c)}(1+O(n^{-1}))$$

It follows that the series (16.3–31) behaves like the series

$$\sum_{\nu=0}^{\infty}(-1)^{\nu}\binom{s}{\nu}(1+O(\nu^{-1})).$$

It is clear that the series of error terms, being the series

$$\sum_{\nu=0}^{\infty}(-1)^{\nu}\binom{s}{\nu}O(\nu^{-1}),$$

is convergent and from the above considerations we may conclude that the series (16.3–31) is divergent. If $a+b-c = 0$ we compare the given series with the divergent series

$$\sum_{\nu=1}^{\infty}\nu^{-1}.$$

As above we find that the series (16.2–31) is also divergent in this case. Summing up we have

On the circumference of convergence $|z| = 1$ *the hypergeometric series is absolutely convergent, if* $\operatorname{Re}(c-a-b) > 0$;
conditionally convergent, if $-1 < \operatorname{Re}(c-a-b) \leqq 0$, *and* $z \neq 1$, *and divergent at* $z = 1$;
divergent at every point of $|z| = 1$ *if* $\operatorname{Re}(c-a-b) < -1$.

This theorem implies Abel's theorem about the behaviour of the binomial series on the circumference of convergence. In view of (16.1–25) we may state

On the circumference of convergence the binomial series (2.16–20) *is absolutely convergent, if* $\operatorname{Re} s > 0$;
conditionally convergent, if $-1 < \operatorname{Re} s \leqq 0, z \neq -1$, *and divergent at* $z = -1$;
divergent at each point of the circumference if $\operatorname{Re} s < -1$.

16.3.5 – Gauss's method of evaluating $F(a, b; c; 1)$

If $\operatorname{Re}(c-a-b) > 0$ the series $F(a, b; c; 1)$ is convergent. If, moreover, $\operatorname{Re} a > 0$ the relation (16.1–47) is valid. Hence by making z tend to 1 on the real axis from the left we find by virtue of Abel's theorem of section 1.8.1

$$\boxed{F(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}.} \tag{16.3–34}$$

Since both sides of this equation are holomorphic functions of a the restriction $\operatorname{Re} a > 0$ may be dropped.

An interesting straight-forward method of evaluating $F(a, b; c; 1)$ is due to Gauss. First we observe that the series F and F_{c+} occurring in the last relation listed in (16.3–13) are convergent at $z = 1$, as is $(1-z)F'$. If c_n has the same meaning as in section (16.3.4) we have

$$\frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} F(a, b; c; z) = \sum_{\nu=1}^{\infty} c_\nu z^\nu$$

and from (16.3–3) follows

$$\lim_{n\to\infty} nc_n n^{c-a-b} = 1.$$

Hence $nc_n \to 0$ as $n \to \infty$. But

$$(1-z)F' = \frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} \sum_{\nu=1}^{\infty} (\nu c_\nu - (\nu-1)c_{\nu-1})z^{\nu-1}.$$

On the right we take $z = 1$ to get

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} \sum_{\nu=1}^{\infty} (\nu c_\nu - (\nu-1)c_{\nu-1}) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} \lim_{n\to\infty} nc_n = 0,$$

whence

$$\lim (1-z)F' = 0,$$

if z tends to 1 along the real axis from the left. Thus we see that the last equation of (16.3–13) reduces to

$$c(c-a-b)F(a, b; c; 1) = (c-a)(c-b)F(a, b; c+1; 1).$$

Repeating this process we find for every positive integer n

$$F(a, b; c; 1) = \frac{(c-a)_n(c-b)_n}{(c)_n(c-a-b)_n} F(a, b; c+n; 1).$$

If a, b and c are fixed numbers the series $F(a, b; c+n; 1)$ is uniformly convergent with respect to n. Since each term, except the first, tends to zero, we have

$$\lim_{n\to\infty} F(a, b; c+n; 1) = 1.$$

On the other hand

$$\begin{aligned}\frac{(c-a)_n(c-b)_n}{(c)_n(c-a-b)_n} &= \prod_{\nu=0}^{n-1} \frac{(c-a+\nu)(c-b+\nu)}{(c+\nu)(c-a-b+\nu)} \\ &= \prod_{\nu=0}^{n-1} \frac{\nu^2+(2c-a-b)\nu+(c-a)(c-b)}{\nu^2+(2c-a-b)\nu+c(c-a-b)} \\ &= \prod_{\nu=0}^{n-1} \left(1+\frac{ab}{\nu^2+(2c-a-b)\nu+c(c-a-b)}\right).\end{aligned}$$

By making n tend to infinity we obtain a convergent infinite product (section 4.1.1). From (4.6–8) we now conclude that (16.3–34) is true. It should be noticed that in this reasoning only the assumption $\operatorname{Re}(c-a-b) > 0$ has been used. This assumption cannot be relaxed.

16.4 – The fundamental system in the case that the third parameter is an integer

16.4.1 – General considerations

If c is not an integer the series

$$\begin{aligned} w_{01}(z) &= F(a, b; c; z) \\ w_{01'}(z) &= z^{1-c}F(a-c+1, b-c+1; 2-c; z) \end{aligned} \qquad (16.4\text{–}1)$$

constitute a fundamental system of solutions of the hypergeometric differential equation (16.1–9), valid in the disc $|z| < 1$ and corresponding to the exponents 0 and $1-c$ respectively. If, however, c is an integer the series (16.4–1) do not represent a fundamental system. For if $c \leqq 0$ the first series has no meaning, if $c \geqq 2$ the second fails to make sense and if $c = 1$ the two series coincide. In these exceptional cases we are up to now in possession of only one solution and there arises the problem of finding a second independent solution.

Let us consider any linear combination $w(z; c)$ of the functions (16.4–1), assuming first, of course, that c is not an integer. If primes indicate differentiation with respect to z, we have

$$z(1-z)w''(z; c)-((a+b+1)z-c)w'(z; c)-abw(z; c) = 0. \qquad (16.4\text{–}2)$$

As we know the function $w(z; c)$, considered as a function of the variable c, is regular at all points different from 0, -1, $-2, \ldots$ etc. (section 16.3.1). Differentiating we get the left hand member of (16.4–2) with respect to c

$$z(1-z)\frac{\partial w''}{\partial c} - ((a+b+1)z-c)\frac{\partial w'}{\partial c} - ab\frac{\partial w}{\partial c} = -w'(z, c).$$

If we are able to select $w(z; c)$ in such a way that

$$\lim_{c\to m} w(z; c) = 0, \qquad (16.4\text{–}3)$$

m being an integer, the passing to the limit being uniformly with respect to z if $|z| \leqq R < 1$, then also

$$\lim_{c\to m} w'(z; c) = 0, \qquad (16.4\text{–}4)$$

uniformly in $|z| \leqq R < 1$. Since differentiation with respect to z and with respect to c is interchangeable, we conclude that

$$\frac{\partial}{\partial c} w(z; c)|_{c=m} \qquad (16.4\text{–}5)$$

is again a solution of the given hypergeometric differential equation. In the next sections we consider the cases $m = 1$ and $m \neq 1$ separately.

16.4.2 – The case that the third parameter is unity

If c is not an integer then

$$w(z; c) = -z^{1-c}F(a-c+1, b-c+1; 2-c; z)+F(a, b; c; z)$$

is a solution of (16.1–9) which satisfies (16.4–3).

Now

$$\frac{\partial}{\partial c} w(z;c) = z^{1-c}F(a-c+1, b-c+1; 2-c; z)\log z + \\ -z^{1-c}\frac{\partial}{\partial c} F(a-c+1, b-c+1; 2-c; z) + \frac{\partial}{\partial c} F(a, b; c; z)$$

and by making $c \to 1$ we have

$$\frac{\partial}{\partial c} w(z;c)\bigg|_{c=1} = F(a, b; 1; z)\log z + \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial b} + 2\frac{\partial}{\partial c}\right) F(a, b; c; z)\bigg|_{c=1}.$$

This solution has a logarithmic singularity at $z = 0$ and is, therefore, independent of $F(a, b; 1; z)$.

Thus

If $c = 1$ then a fundamental system of solutions of the hypergeometric equation at $z = 0$ is given by the functions

$$\begin{gathered} F(a, b; 1; z) \\ F(a, b; 1; z)\log z + F^*(a, b; 1; z) \end{gathered} \tag{16.4–6}$$

with

$$F^*(a, b; 1; z) = \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial b} + 2\frac{\partial}{\partial c}\right) F(a, b; c; z)\bigg|_{c=1}. \tag{16.4–7}$$

Because of the uniform convergence it is legitimate to differentiate the series $F(a, b; c; z)$ term by term, where a, b and c are variables which take after the process of differentiation their given values. We notice that

$$\frac{d}{dz}(z)_n = \frac{d}{dz}\frac{\Gamma(z+n)}{\Gamma(z)} = (\Psi(z+n) - \Psi(z))(z)_n,$$

where $\psi(z)$ is the Gaussian function (4.8–1). A straightforward computation yields (neglecting a term involving $F(a, b; 1; z)$).

$$\sum_{\nu=0}^{\infty} \frac{(a)_\nu (b)_\nu}{(1)_\nu}\frac{z^\nu}{\nu!}(\log z + \Psi(a+\nu) + \Psi(b+\nu) - 2\Psi(1+\nu)). \tag{16.4–8}$$

If a or b is zero or a negative integer the series on the right terminates after a finite number of terms.

16.4.3 – The case that the third parameter is an integer less than unity

Now we focus our attention on the case that $c = -m$, $m = 0, 1, 2, \ldots$. In view of (16.3–5) it is natural to consider the linear combination

$$w(z; c) = -A_m z^{1+m} F(a-c+1, b-c+1; 2-c; z) + (c+m)F(a, b; c; z), \tag{16.4–9}$$

where c is a variable and A_m the constant (16.3–6) (with $n = m$). It is clear that (16.4–3) is satisfied.

Since

$$\frac{\partial}{\partial c} w(z; c)\bigg|_{c=-m} = A_m z^{1+m} F(a+m+1, b+m+1; 2+m; z) \log z + \\ -A_m z^{1+m} \frac{\partial}{\partial c} F(a-c+1, b-c+1; 2-c; z)\bigg|_{c=-m} \\ + \frac{\partial}{\partial c}(c+m)F(a, b; c; z)\bigg|_{c=-m}$$

we may state

If the third parameter in the hypergeometric differential equation (16.1–9) *is an integer* $-m$ *less than unity then a fundamental system of solutions consists of the functions*

$$z^{1+m}F(a+m+1, b+m+1; 2+m; z), \\ A_m z^{1+m} F(a+m+1, b+m+1; 2+m; z) \log z + \\ + F^*(a+m+1, b+m+1; 2+m; z) \tag{16.4–10}$$

with

$$A_m = (-1)^m \frac{(a)_{m+1}(b)_{m+1}}{m!(m+1)!} \tag{16.4–11}$$

and

$$F^*(a+m+1, b+m+1; 2+m; z) \\ = -A_m z^{1+m} \frac{\partial}{\partial c} F(a-c+1, b-c+1; 2-c; z)\bigg|_{c=-m} \\ + \frac{\partial}{\partial c}(c+m)F(a, b; c; z)\bigg|_{c=-m}. \tag{16.4–12}$$

It is clear that the second solution will be free from the logarithm if $A_m = 0$, i.e., if a or b takes one of the values $0, -1, \ldots, -m$. Otherwise stated

There is no solution with logarithmic singularity at $z = 0$ *if and only if*

$$\prod_{\nu=0}^{m} (a+\nu)(b+\nu) = 0. \tag{16.4–13}$$

A fundamental system is now given by (16.4–10) with $c = -m$. In this case the first series makes sense for we can solve the recurrent relations (16.1–17).

We proceed to evaluate (16.4–10), supposing that $A_m \neq 0$. Observing that

$$(a+m+1)_n = \frac{(a)_{m+1+n}}{(a)_{m+1}} \tag{16.4–14}$$

which is equal to $(m+1+n)!/(m+1)!$ if $a = 1$, we may write the logarithmic term

$$A_m \sum_{\nu=0}^{\infty} \frac{(a+m+1)_\nu (b+m+1)_\nu}{(1+m+1)_\nu} \frac{z^{m+\nu+1}}{\nu!} \log z$$

as

$$\frac{(-1)^m}{m!} \sum_{\nu=m+1}^{\infty} \frac{(a)_\nu (b)_\nu}{(\nu-m-1)!} \frac{z^\nu}{\nu!} \log z.$$

The second term can be found from

$$-A_m z^{1+m} \frac{\partial}{\partial c} F(a-c+1, b-c+1; 2-c; z)$$

$$= A_m z^{1+m} \sum_{\nu=0}^{\infty} \frac{(a-c+1)_\nu (b-c+1)_\nu}{(2-c)_\nu} \frac{z^\nu}{\nu!}$$

$$\times (\Psi(a-c+1-\nu)+\Psi(b-c+1-\nu)-\Psi(z-c+\nu)+*),$$

by making $c \to -m$. The asterisk denotes terms which are independent of ν and in the limit their contribution is a constant times the first series (16.4–10). Since this is a solution we may neglect it. The other terms yield the series

$$A_m \sum_{\nu=0}^{\infty} \frac{(a+m+1)_\nu (b+m+1)_\nu}{(1+m+1)_\nu} \frac{z^{m+\nu+1}}{\nu!} \times$$

$$\times (\Psi(a+m+1+\nu)+\Psi(b+m+1+\nu)-\Psi(1+m+1=\nu))$$

and with the aid of (16.3–14) this may be written as

$$\frac{(-1)^m}{m!} \sum_{\nu=m+1}^{\infty} \frac{(a)_\nu (b)_\nu}{(\nu-m-1)!} \frac{z^\nu}{\nu!} (\Psi(a+\nu)+\Psi(b+\nu)-\Psi(1+\nu)).$$

Taking (16.3–4) into account, we see that the last term in (16.4–12) becomes

$$\sum_{\nu=0}^{m} \frac{(a)_\nu (b)_\nu}{(-m)_\nu} \frac{z^\nu}{\nu!}$$

$$+ \frac{\partial}{\partial c} \frac{(a)_{m+1}(b)_{m+1}}{(c)_m} \bigg|_{c=-m} \frac{z^{1+m}}{(1+m)!} F(a+m+1, b+m+1; 2+m; z)+$$

$$-A_m \sum_{\nu=0}^{\infty} \frac{(a+m+1)_\nu (b+m+1)_\nu}{(1+m+1)_\nu} \frac{z^{m+\nu+1}}{\nu!} (\Psi(1+\nu)-\Psi(1)).$$

Neglecting a constant multiple of the first series (16.4–10) and observing that

$$(-m)_n = (-1)^n \frac{m!}{(m-n)!}, \qquad n \leq m,$$

we get for the remaining term the expression

$$\sum_{\nu=0}^{m} (a)_\nu (b)_\nu \frac{(m-\nu)!}{m!} (-1)^\nu \frac{z^\nu}{\nu!} - \frac{(-1)^m}{m!} \sum_{\nu=m+1}^{\infty} \frac{(a)_\nu (b)_\nu}{(\nu-m+1)!} \frac{z^\nu}{\nu!} \Psi(-m+\nu).$$

Multiplying by $(-1)^m m!$ we may infer that *a second solution can be represented by*

$$\sum_{\nu=0}^{m} (-1)^{\nu+m} (a)_\nu (b)_\nu (m-\nu)! \frac{z^\nu}{\nu!} +$$

$$+ \sum_{\nu=m+1}^{\infty} \frac{(a)_\nu (b)_\nu}{(\nu-m-1)!} \frac{z^\nu}{\nu!} (\log z + \Psi(a+\nu) + \Psi(b+\nu) +$$

$$- \Psi(-m+\nu) - \Psi(1+\nu)). \tag{16.4–15}$$

If a or b is zero or a negative integer the last series terminates after a finite number of terms.

46.4.4 The case that the third parameter is an integer exceeding unity

The case for which the third parameter is an integer $m > 1$ may be reduced to the previous case by the following remark; if $w(z)$ is a solution of a hypergeometric differential equation, then the function $w_1(z)$ defined by

$$w(z) = z^{1-c} w_1(z) \tag{16.4–16}$$

satisfies a hypergeometric differential equation with parameters $a_1 = a-c+1$, $b_1 = b-c+1$, $c_1 = 2-c$, whereas c_1 takes the integral value $-m_1 = 2-m$. Applying to $F(a_1, b_1; c_1; z)$ the results of the previous sections we first have

If the third parameter in the hypergeometric differential eluation (16.1–9) *is an integer in exceeding unity a fundamental system of solutions at* $z = 0$ *consists of the functions*

$$\begin{gathered} F(a, b; m; z) \\ B_m F(a, b; m; z) \log z + F^*(a, b; m; z) \end{gathered} \tag{16.4–17}$$

with

$$B_m = (-1)^m \frac{(a-m+1)_{m-1}(b-m+1)_{m-1}}{(m-2)!(m-1)!} \tag{16.4–18}$$

and

$$F^*(a, b; m; z) = B_m \frac{\partial}{\partial c} F(a, b; c; z)\Big|_{c=m} +$$

$$-z^{1-m} \frac{\partial}{\partial c}(c-m)F(a-c+1, b-c+1; 2-c; z)\Big|_{c=m}. \qquad (16.4\text{–}19)$$

In addition we have

There is no solution with logarithmic singularity at $z = 0$ *and only if*

$$\prod_{\nu=1}^{m-1}(a-\nu)(b-\nu) = 0. \qquad (16.4\text{–}20)$$

In the case that $B_m \neq 0$ we may obtain from (16.4–15) a second solution in the form

$$\sum_{\nu=0}^{m-2}(-1)^{\nu+m}(a-m+1)_\nu(b-m+1)_\nu(m-2-\nu)!\frac{z^{\nu-m+1}}{\nu!} +$$

$$+C_m\sum_{\nu=0}^{\infty}\frac{(a)_\nu(b)_\nu}{(\nu+m-1)!}\frac{z^\nu}{\nu!}(\log z+\Psi(a+\nu)+\Psi(b+\nu)-\Psi(m+\nu)-\Psi(1+\nu)). \qquad (16.4\text{–}21)$$

where $C_m = (a-m+1)_{m-1}(b-m+1)_{m-1}$

It is interesting to notice that the expansion still makes sense for $m = 1$, for it reduces to (16.4–8).

Finally we make the remark that exceptional cases at the singularity $z = 1$ occur if $c-a-b$ is an integer and that there is an exceptional case at $z = \infty$ if $a-b$ is an integer. By the well-known transformation these cases can be reduced to those considered above.

16.5 – Barnes's contour integrals

16.5.1 – Introduction

A very powerful tool which has found many applications in the theory of the hypergeometric functions is provided by a certain contour integral introduced by S. Pincherle and R. Mellin. It is usually referred to as Barnes's integral, because E. W. Barnes has shown the great value of this integral for the study of the hypergeometric functions.

The idea is simple. In section 3.7.2 we obtained by means of a certain standard function $\varphi(s)$ a sum formula for a series associated with a certain function $f(s)$. The general result is stated in equation (3.7–10). A useful standard function is the function $\Gamma(-s)$. It has simple poles at

$s = n$, $n = 0, 1, 2, \ldots$ and the residues are

$$\operatorname*{Res}_{s=n} \Gamma(-s) = \frac{(-1)^{n+1}}{n!}, \tag{16.5–1}$$

as follows from (4.6–12). Assuming that $f(s)$ is holomorphic in a region containing the closed contour C, we may conclude that

$$\sum_{\nu} (-1)^{\nu+1} \frac{f(\nu)}{\nu!} = \frac{1}{2\pi i} \int_C f(s)\Gamma(-s)ds,$$

the terms on the left corresponding to the poles of $\Gamma(-s)$ within the contour. It is understood that the winding-number of C with respect to these poles is one. If we take

$$f(s) = \frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)}(-z)^s,$$

where

$$(-z)^s = \exp(s \log(-z)),$$
$$\log(-z) = \log|z| + i \arg(-z), \ |\arg(-z)| < \pi, \tag{16.5–2}$$

then a sum of the type

$$-\sum_{\nu} \frac{\Gamma(a+\nu)\Gamma(b+\nu)}{\Gamma(c+\nu)} \frac{z^\nu}{\nu!} \tag{16.5–3}$$

appears. Apart from a multiplicative constant these terms are those of a hypergeometric series. If we wish to produce all terms of the hypergeometric series the contour C cannot be a closed path in the finite plane, but must extend to infinity. Following Barnes we take for C a path which starts at $-i\infty$ and goes along the imaginary axis to $+i\infty$, curving to put the poles of $\Gamma(a+s)$ and $\Gamma(b+s)$ to the left of the path and to put the poles of $\Gamma(-s)$ to the right of the path, (fig. 16.5–1). This is possible, provided that neither a nor b is a negative integer or zero. A contour of this type will be called a *Barnes contour.*

The main theorem may be stated as

The function

$$F(z) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)} \Gamma(-s)(-z)^s ds \tag{16.5–4}$$

where the integral is taken along a Barnes contour is holomorphic in the region $|\arg(-z)| < \pi$ (that is the complex plane slit along the positive

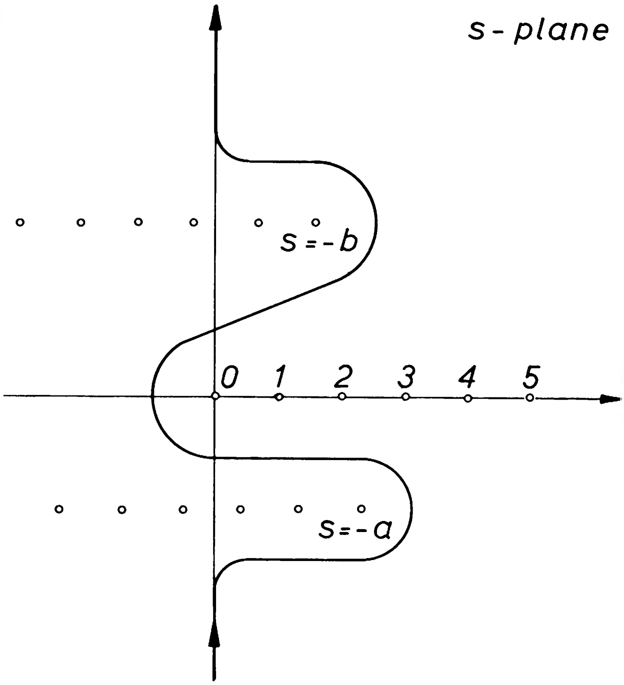


Fig. 16.5–1. A Barnes contour

real axis) *and, if c is not an integer* $\leqq 0$, *represents the series*

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} F(a, b; c; z) \tag{16.5–5}$$

when $|z| < 1$.

This theorem implies the statement: The function $F(a, b; c; z)$, defined by the power series

$$\frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{\nu=0}^{\infty} \frac{\Gamma(a+\nu)\Gamma(b+\nu)}{\Gamma(c+\nu)} \frac{z^\nu}{\nu!},$$

when $|z| < 1$, can be continued by analytically as a single valued function throughout the whole z-plane cut along the real axis from 1 to ∞ by means of (16.5–4).

16.5.2 – Proof of the regularity

First we shall prove the following more general theorem

The integral

$$\boxed{F(z) = \frac{1}{2\pi i} \int_{\alpha-i\infty}^{\alpha+i\infty} \frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)} \Gamma(-s)(-z)^s ds} \tag{16.5–6}$$

represents a holomorphic function in the region $|\arg(-z)| < \pi$, *provided that the path of integration which is a line parallel to the imaginary axis does not contain any pole of the integrand.*

By virtue of (4.6–13) we may write the integrand as

$$-\Phi(s)\frac{\pi}{\sin \pi s}(-z)^s \tag{16.5–7}$$

with

$$\Phi(s) = \frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)\Gamma(1+s)}. \tag{16.5–8}$$

Our main task will be the investigation of the behaviour of the integrand for large values of $|s|$.

Let δ denote a positive number $< \pi$. For every z in the set $|z| > 0$, $|\arg(-z)| \leqq \pi - \delta$ and for every $s = p+qi$ we have

$$\begin{aligned}(-z)^s &= \left|\exp\left((p+iq)(\log|z|+i\arg(-z)\right)\right| \\ &= \exp\left(p\log|z| - q\arg(-z)\right) \leqq \exp\left(p\log|z| + (\pi-\delta)|q|\right) \\ &= |z|^p e^{(\pi-\delta)|q|}.\end{aligned}$$

From (1.10–17) we deduce

$$\begin{aligned}\frac{\pi}{|\sin \pi s|} &= \frac{\pi}{\sqrt{\sin^2 \pi p + \sinh^2 \pi q}} \leqq \frac{\pi}{|\sinh \pi q|} \\ &= \frac{\pi}{\sinh \pi|q|} = \frac{2\pi}{1-e^{-2\pi|q|}} e^{-\pi|q|}.\end{aligned}$$

Hence, to a given number $q_0 > 0$ corresponds a number K such that

$$\frac{\pi}{|\sin \pi s|} < Ke^{-\pi|q|},$$

provided that $|q| > q_0$. Combining this with the above result we see that for $|q| > q_0$

$$\left|\frac{\pi}{\sin \pi s}(-z)^s\right| < K|z|^p e^{-|q|\delta}. \tag{16.5–9}$$

Next we turn our attention to $\Phi(s)$. Writing this function as

$$\frac{\Gamma(a+s)}{\Gamma(s)}\,\frac{\Gamma(b+s)}{\Gamma(s)}\,\frac{\Gamma(s)}{\Gamma(c+s)}\,\frac{\Gamma(s)}{\Gamma(1+s)},$$

we deduce from (16.3–29) that

$$\Phi(s) = s^{\kappa}(1+o(1)), \tag{16.5–10}$$

where κ stands for $a+b-c-1$, uniformly in $|\arg s| \leqq \pi-\delta$. Now

$$|s^{\kappa}| \leqq |\exp(\kappa \log s)| = \exp((\operatorname{Re}\kappa)\log|s|-(\operatorname{Im}\kappa)\arg s).$$

For $|s| \geqq 1$, $|\arg s| \leqq \pi-\delta$ it follows that

$$|s^{\kappa}| \leqq \exp(|\operatorname{Re}\kappa|\log|s|+(\pi-\delta)|\operatorname{Im}\kappa|) < K_1|s|^{|\operatorname{Re}\kappa|}, \qquad (16.5\text{–}11)$$

where K_1 is a constant independent of s. It follows from (16.5–10) that there is a constant $R \geqq 1$ such that

$$|\Phi(s)| < 2|s^{\kappa}|$$

provided that $|s| \geqq R$, $|\arg s| \leqq \pi-\delta$. Combining this with (16.5–11) we find

$$|\Phi(s)| < 2K_1|s|^{|\operatorname{Re}\kappa|} < K_2(1+|s|)^{|\operatorname{Re}\kappa|}, \qquad (16.5\text{–}12)$$

where K_2 is again a constant independent of s, provided that $|s| \geqq R$, $|\arg s| \leqq \pi-\delta$.

Now we are sufficiently prepared to prove the assertion stated above. Let w_1, w_2 denote two numbers such that either $w_2 > w_1 > q_0$ or $w_1 < w_2 < -q_0$, where q_0 is a certain positive constant. We may take q_0 in such a way that for $s = \alpha+iq$, $|q| > q_0$ the estimate (16.5–12) is valid. By virtue of (16.5–9) there is a constant, again denoted by K, such that for $|z| > 0$, $|\arg(-z)| \leqq \pi-\delta$ we have

$$\left|\frac{1}{2\pi i}\int_{\alpha+iw_1}^{\alpha+iw_2}\frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)}\Gamma(-s)(-z)^s\,ds\right|$$
$$< K\int_{w_1}^{w_2}(1+(|s|))^{|\operatorname{Re}\kappa|}|z|^{\alpha}e^{-|q|\delta}\,dq.$$

It is clear that the integral

$$\int_{-\infty}^{\infty}(1+|s|)^{|\operatorname{Re}\kappa|}e^{-|q|\delta}\,dq$$

is convergent. It follows that, if z is restricted to a closed and bounded set, to a given $\varepsilon > 0$ corresponds a number $q_0 > 0$ such that

$$\left|\frac{1}{2\pi i}\int_{\alpha+iw_1}^{\alpha+iw_2}\frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)}\Gamma(-s)(-z)^s\,ds\right| < \varepsilon,$$

that is to say, the integral (16.5–6) is uniformly convergent.

Reasoning as in section 2.20.5 we may conclude that the assertion stated at the beginning of this section is true.

16.5.3 – Evaluation of Barnes's integral

The evaluation of the integral (16.5–4) is made possible by the following assertion

The difference

$$\frac{1}{2\pi i}\left(\int_{\beta-i\infty}^{\beta+i\infty}-\int_{\alpha-i\infty}^{\alpha+i\infty}\right), \qquad \alpha<\beta,$$

of two integrals of the type (16.5–6) *is equal to the sum of the residues of the integrand between the vertical lines* Re $s=\alpha$, Re $s=\beta$, *provided, of course, that none of these lines pass through a pole of the integrand.*

This assertion is a direct consequence of the fact that

$$\lim_{w\to\infty}\int_{\alpha+iw}^{\beta+iw}\frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)}\Gamma(-s)(-z)^s ds = 0. \qquad (16.5\text{–}13)$$

In fact, if $|w|$ is sufficiently large, the rectangle with vertices $\alpha-iw$, $\beta-iw$, $\beta+iw$, $\alpha+iw$ encloses all the poles of the integrand between the two lines under consideration.

For the proof of (16.5–13) we employ the same technique as in the previous section. If $|w|$ is sufficiently large and $|\arg(-z)|\leqq\pi-\delta$, $|z|>0$ we know that there is a constant K, independent of z and s, such that

$$\left|\int_{\alpha+iw}^{\beta+iw}\frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)}\Gamma(-s)(-z)^s ds\right| < K\int_{\alpha}^{\beta}(1+|s|)^{|\mathrm{Re}\,\kappa|}|z|^p e^{-|w|\delta}dp.$$

If $\gamma=\max(|\alpha|,|\beta|)$ then on the path of integration $|s|\leqq\gamma+|w|$. Further $|z|^p\leqq R_0^\gamma$, if z is taken from a closed and bounded set, R_0 being an appropriate constant. By Darboux's theorem of section 2.4.3 the integral on the right is dominated by

$$(1+\gamma+|w|)^{|\mathrm{Re}\,\kappa|}R_0^\gamma e^{-|w|\delta}(\beta-\alpha)$$

and, therefore, tends to zero as $|w|\to\infty$. This concludes the proof of (16.5–13).

Next we take $\alpha=n$ in (16.5–6). If n is a sufficiently large integer the poles of $\Gamma(a+s)\Gamma(b+s)$ are on the left of the path of integration. By a small indentation we can achieve that no pole of $\Gamma(-s)$ is on it. We contend that the integral tends to zero as $n\to\infty$, provided that $|z|<1$. This follows from

$$\left|\int_{n-i\infty}^{n+i\infty}\frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)}\Gamma(-s)(-z)^s ds\right| < K\int_{-\infty}^{\infty}(1+|s|)^{|\mathrm{Re}\,\kappa|}|z|^n e^{-|q|\delta}dq$$

$$< K|z|^n\int_{-\infty}^{\infty}(1+n+|q|)^{|\mathrm{Re}\,\kappa|}e^{-|q|\delta}dq.$$

The last integral is convergent.

A direct consequence is that the integral (16.5–4) taken along a Barnes contour is equal to the negative sum of the residues at the poles of the integrand to the right of the contour and this is precisely the last assertion of section 16.6.1.

A very remarkable result presents itself if we evaluate Barnes' integral for $|z| > 1$. Let $\alpha = -n$, where n is again a positive integer. If a or b is a positive integer the path of integration is indented, so that the pole, which should otherwise lie on it, lies to its left. We now have

$$\left|\int_{-n-i\infty}^{-n+i\infty} \frac{\Gamma(a+s)\Gamma(b+s)}{\Gamma(c+s)} \Gamma(-s)(-z)^s ds\right| = \left|\int_{-n-i\infty}^{-n+i\infty} \Phi(s) \frac{\pi}{\sin \pi s} (-z)^s ds\right|$$

$$= \left|\int_{-i\infty}^{i\infty} \Phi(s-n) \frac{\pi}{\sin \pi s} (-z)^{s-n} ds\right|$$

$$< K|z|^{-n}\int_{-\infty}^{\infty} (1+|s-n|)^{|\mathrm{Re}\,\kappa|}e^{-|q|\delta}dq$$

and this tends to zero as $n \to \infty$ under the assumption $|z| > 1$.

It is now clear that the Barnes integral (16.5–4) for $|z| > 1$ is equal to the sum of the residues of the integrand at the poles of $\Gamma(a+s)\Gamma(b+s)$. These poles are $-a-n$ and $-b-n$, $n = 0, 1, \ldots$ and are simple if we assume that $a-b$ is not an integer. Under this assumption the integral is equal to

$$\sum_{\nu=0}^{\infty} \frac{\Gamma(b-a-\nu)\Gamma(a+\nu)}{\Gamma(c-a-\nu)} \frac{(-1)^\nu}{\nu!}(-z)^{-a-\nu} + \\ + \sum_{\nu=0}^{\infty} \frac{\Gamma(a-b-\nu)\Gamma(b+\nu)}{\Gamma(c-b-\nu)} \frac{(-1)^\nu}{\nu!}(-z)^{-b-\nu}.$$

Now, in view of (4.6–13)

$$\Gamma(b-a-n)\Gamma(a-b+1+n) = \frac{\pi(-1)^n}{\sin \pi(b-a)} = (-1)^n\Gamma(b-a)\Gamma(a-b+1)$$

and a similar equation holds if we interchange a and b.

Hence the above expression takes the form

$$\frac{\Gamma(b-a)}{\Gamma(c-a)}(-z)^{-a}\frac{\Gamma(a-b+1)}{\Gamma(a-c+1)}\sum_{\nu=0}^{\infty}\frac{\Gamma(a+\nu)\Gamma(a-c+1+\nu)}{\Gamma(a-b+1+\nu)}\frac{z^{-\nu}}{\nu!} + \\ + \frac{\Gamma(a-b)}{\Gamma(c-b)}(-z)^{-b}\frac{\Gamma(b-a+1)}{\Gamma(b-c+1)}\sum_{\nu=0}^{\infty}\frac{\Gamma(b+\nu)\Gamma(b-c+1+\nu)}{\Gamma(b-a+1+\nu)}\frac{z^{-\nu}}{\nu!}$$

$$= \frac{\Gamma(a)\Gamma(b-a)}{\Gamma(c-a)}(-z)^{-a}F(a, a-c+1; a-b+1; z^{-1}) + \\ + \frac{\Gamma(b)\Gamma(a-b)}{\Gamma(c-b)}(-z)^{-b}F(b, b-c+1; b-a+1; z^{-1}), \ |\arg(-z)| < \pi.$$

This is, however, the result stated in (16.1–56). In this proof the parameters are only restricted to the condition that the various expressions have a meaning.

16.5.4 – The complementary representation of the hypergeometric series

An alternative representation of $F(a, b; c; z)$ is provided by the integral

$$\boxed{\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma(a+s)\Gamma(b+s)\Gamma(c-a-b-s)\Gamma(-s)(1-z)^s ds} \qquad (16.5\text{–}14)$$

which we shall call a complementary representation, because in the integrand occurs the variable $1-z$. The path of integration is again a Barnes contour, curved in such a way that the poles of $\Gamma(a+s)\Gamma(b+s)$ are to the left and of $\Gamma(c-a-b-s)\Gamma(-s)$ to the right. This is always possible if neither $c-a$, nor $c-b$ is an integer.

Proceeding as in the sections 16.5.2 and 16.5.3 it can be proved that the integral (16.5–14) represents a holomorphic function in the region $|\arg(1-z)| < \pi, z \neq 1$, and is equal to the negative sum of the residues of the integrand at the poles to the right of the path of integration, provided that $|1-z| < 1$. These poles are at $s = n$ and $s = c-a-b+n$, $n = 0, 1, 2, \ldots$. Supposing that $c-a-b$ is not an integer the value of the integral turns out to be

$$\sum_{\nu=0}^{\infty} \Gamma(a+\nu)\Gamma(b+\nu)\Gamma(c-a-b-\nu)\frac{(-1)^\nu}{\nu!}(1-z)^\nu +$$
$$+\sum_{\nu=0}^{\infty} \Gamma(c-a+\nu)\Gamma(c-b+\nu)\Gamma(a+b-c-\nu)\frac{(-1)^\nu}{\nu!}(1-z)^{c-a-b+\nu}.$$

Since

$$\Gamma(c-a-b-n)\Gamma(a+b-c+1+n) = (-1)^n\Gamma(c-a-b)\Gamma(a+b-c+1)$$

and

$$\Gamma(a+b-c-n)\Gamma(c-a-b+1+n) = (-1)^n\Gamma(a+b-c)\Gamma(c-a-b-1)$$

the above expression may be put in the form

$$\Gamma(c-a-b)\Gamma(a+b-c+1)\sum_{\nu=0}^{\infty}\frac{\Gamma(a+\nu)\Gamma(b+\nu)}{\Gamma(a+b-c+1+\nu)}\frac{(-1-z)^\nu}{\nu!} +$$
$$+\Gamma(a+b-c)\Gamma(c-a-b+1)(1-z)^{c-a-b}\sum_{\nu=0}^{\infty}\frac{\Gamma(c-a+\nu)\Gamma(c-b+\nu)}{\Gamma(c-a-b+1+\nu)}\frac{(1-z)^\nu}{\nu!}$$

$$= \frac{\Gamma(c-a-b)}{\Gamma(a)\Gamma(b)} F(a, b; a+b-c+1; 1-z)+$$

$$+ \frac{\Gamma(a+b-c)}{\Gamma(c-a)\Gamma(c-b)} (1-z)^{c-a-b} F(c-a, c-b; c-a-b+1; 1-z).$$

Comparing this with (16.1–53) we may infer that

The integral (16.5–14) *is an analytic continuation of the series*

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} \Gamma(c-a)\Gamma(c-b)F(a, b; c; z) \tag{16.5–15}$$

into the region $|\arg(1-z)| < \pi$, $z \neq 1$.

The equation (16.5–14) has been obtained under restrictive assumptions about the parameters. In view of the results of section 16.3.1 it is valid for all values of the parameters for which it makes sense.

16.5.5 – Legendre's complete elliptic integrals

If we make a and b tend to $\frac{1}{2}$ and c tend to 1 in the integrals (16.5–4) and (16.5–14) we find from (16.1–29) two representations of Legendre's elliptic integral of the first kind as a Barnes integral, viz.,

$$\boxed{2\mathsf{K}(z) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \frac{\Gamma^2(\frac{1}{2}+s)}{\Gamma(1+s)} \Gamma(-s)(-z)^s ds,} \qquad |\arg(-z)| < \pi, \tag{16.5–16}$$

and

$$\boxed{2\pi\mathsf{K}(z) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma^2(\tfrac{1}{2}+s)\Gamma^2(-s)(1-z)^s ds,} \qquad |\arg(1-z)| < \pi. \tag{16.5–17}$$

Replacing z by $1-z$ in (16.5–17) we obtain from (14.5–12) a useful representation of Legendre's elliptic integral of the second kind

$$\boxed{2\pi\mathsf{K}'(z) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma^2(\tfrac{1}{2}+s)\Gamma^2(-s)z^s ds,} \qquad |\arg z| < \pi. \tag{16.5–18}$$

Assuming $|z| < 1$ the integral on the right of (16.5–16) is equal to the negative sum of the residues of the integrand at the points $s = n$, $n = 0$, 1, . . . and yields, therefore, the power series expansion

$$2\mathsf{K}(z) = \sum_{\nu=0}^{\infty} \frac{\Gamma^2(\frac{1}{2}+\nu)}{\Gamma^2(1+\nu)} z^\nu, \tag{16.5–19}$$

in accordance with (14.5–11).

In a similar way we may find an expansion of $\mathsf{K}'(z)$ starting with (16.5–18) and assuming again $|z| < 1$. By virtue of (3.3–4) the residue of the integrand at $s = n$ is

$$\frac{d}{ds}(s-n)^2\Gamma^2(-s)\Gamma^2(s+\tfrac{1}{2})z^s\Big|_{s=n}$$

$$= \Gamma^2(n+\tfrac{1}{2})z^n\frac{d}{ds}(s-n)^2\Gamma^2(-s)\Big|_{s=n} + \frac{2}{(n!)^2}\Gamma'(n+\tfrac{1}{2})\Gamma(n+\tfrac{1}{2})z^n +$$

$$+\frac{2}{(n!)^2}\Gamma^2(n+\tfrac{1}{2})z^n\log z,$$

because in view of (16.5–1)

$$\lim_{s\to n}(s-n)^2\Gamma^2(-s) = \frac{1}{(n!)^2}.$$

In order to evaluate $d/ds\,(s-n)^2\Gamma^2(-s)|_{s=n}$ we observe that

$$\Gamma(-s) = \frac{\Gamma(-s+n+1)}{(-s)(-s+1)\dots(-s+n)} = (-1)^{n+1}\frac{\Gamma(-s+n+1)}{s(s-1)\dots(s-n)},$$

whence

$$\frac{d}{ds}(s-n)^2\Gamma^2(-s)\Big|_{s=n} = -2\left(\frac{\Gamma'(1)}{\Gamma(1)}+\frac{1}{1}+\dots+\frac{1}{n}\right)\frac{\Gamma^2(1)}{(n!)^2}. \qquad (16.5\text{–}20)$$

Introducing the function (4.8–1) we may write the residue of the integrand at $s = n$ as

$$\frac{\Gamma^2(n+\tfrac{1}{2})}{\Gamma^2(n+1)}z^n\left(\log z+2\Psi(n+\tfrac{1}{2})-2\left(\Psi(1)+\frac{1}{1}+\dots+\frac{1}{n}\right)\right).$$

In view of (4.8–4) we have

$$\Psi(n+\tfrac{1}{2}) = \frac{2}{2n-1}+\frac{2}{2n-3}+\dots+\frac{2}{1}+\Psi(\tfrac{1}{2}).$$

The equations (14.5–87) state that

$$\Psi(1)-\Psi(\tfrac{1}{2}) = 2\log 2.$$

Accordingly we have

$$\Psi(n+\tfrac{1}{2})-\Psi(1)-\frac{1}{1}-\dots-\frac{1}{n} = 2\left(\frac{1}{1}-\frac{1}{2}+\dots-\frac{1}{2n}\right)-2\log 2$$

and the residue of the integrand at $s = n$ turns out to be

$$\frac{\Gamma^2(n+\tfrac{1}{2})}{\Gamma^2(n+1)}z^n\left(\log\frac{z}{16}+4\left(\frac{1}{1}-\frac{1}{2}+\dots-\frac{1}{2n}\right)\right). \qquad (16.5\text{–}21)$$

As a consequence we obtain the expansion

$$2\pi \mathsf{K}'(z) = -\sum_{\nu=0}^{\infty} \frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)} z^\nu \left(\log z - \log 16 + 4\left(\frac{1}{1} - \frac{1}{2} + \ldots - \frac{1}{2\nu}\right)\right), \tag{16.5–22}$$

in accordance with (14.5–89). The expansion is valid in the region $|z| < 1$, $|\arg z| < \pi$.

In a similar way we can expand $\mathsf{K}(z)$ in a series of ascending powers of z^{-1}, for the integral on the right of (16.5–16) is equal to the sum of the residues of the integrand at $s = -n-\frac{1}{2}$, $n = 0, 1, \ldots$, provided that $|z| > 1$. The residue at $s = -n-\frac{1}{2}$ is

$$\frac{d}{ds}(s+n+\tfrac{1}{2})^2\Gamma^2(s+\tfrac{1}{2})\frac{\Gamma(-s)}{\Gamma(1+s)}(-z)^s\Bigg|_{s=-n-\frac{1}{2}}$$

$$= -\frac{d}{ds}(s+n+\tfrac{1}{2})^2\Gamma^2(s+\tfrac{1}{2})\Gamma^2(-s)\frac{\sin \pi s}{\pi}(-z)^s\Bigg|_{s=-n-\frac{1}{2}}$$

$$= -\frac{1}{\pi}\frac{d}{ds}(s-n)^2\Gamma^2(-s)\Gamma^2(s+\tfrac{1}{2})\cos \pi s\,(-z)^{-s-\frac{1}{2}}\Bigg|_{s=n}.$$

As above we deduce that this is equal to

$$\frac{1}{\pi}(-z)^{-\frac{1}{2}}\frac{\Gamma^2(n+\frac{1}{2})}{\Gamma^2(n+1)}z^{-n}\left(\log(-z) + \log 16 - 4\left(\frac{1}{1} - \frac{1}{2} + \ldots - \frac{1}{2n}\right)\right).$$

If we restrict z to $|\arg(-z)| < \pi$ we have evidently

$$\log(-z) = \log z \mp \pi i,$$

the upper or the lower sign being taken as $\operatorname{Im} z$ is positive or negative. Therefore

$$2\pi \mathsf{K}(z) = (-z)^{-\frac{1}{2}}\sum_{\nu=0}^{\infty}\frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma^2(\nu+1)}z^{-\nu} \times \left(\log z + \log 16 \mp \pi i - 4\left(\frac{1}{1} - \frac{1}{2} + \ldots - \frac{1}{2\nu}\right)\right), \tag{16.5–23}$$

valid if $|z| > 1$, $|\arg(-z)| < \pi$.

Combining (16.5–19), (16.5–22) and (16.5–23) we get

$$\mathsf{K}(z) = \mp i(-z)^{-\frac{1}{2}}\mathsf{K}\left(\frac{1}{z}\right) + (-z)^{-\frac{1}{2}}\mathsf{K}'\left(\frac{1}{z}\right).$$

Now in the region under consideration $(-z)^{-\frac{1}{2}} = \pm iz^{-\frac{1}{2}}$, the upper or the lower sign corresponding to Im $z > 0$ or Im $z < 0$. Thus we obtain the fundamental relation

$$\boxed{\mathsf{K}(z) = z^{-\frac{1}{2}}\mathsf{K}\left(\frac{1}{z}\right) \pm iz^{-\frac{1}{2}}\mathsf{K}'\left(\frac{1}{z}\right), \qquad |\arg(-z)| < \pi.} \qquad (16.5\text{–}24)$$

This result is difficult to obtain from the Legendre integrals.

16.6 – Conformal mapping

16.6.1 – The mapping of a curvilinear triangle

A curvilinear triangle in the s-plane with angles $\alpha_1\pi$, $\alpha_2\pi$, $\alpha_3\pi$, $0 < \alpha_1$, α_2, $\alpha_3 < 2$ is mapped conformally on the upper half of the z-plane by means of a function which, in accordance with (14.2–3), satisfies Schwarz's differential equation

$$[s]_z = \frac{1-\alpha_1^2}{2z^2} + \frac{1-\alpha_2^2}{2(z-1)^2} + \frac{\alpha_1^2+\alpha_2^2-\alpha_3^2-1}{2z(z-1)}. \qquad (16.6\text{–}1)$$

This equation is the differential resolvent of a hypergeometric differential equation whose parameters may be found by the method described in section 16.1.3. Since we may take

$$a-b = \alpha_3, \qquad c-a-b = \alpha_2, \qquad 1-c = \alpha_1 \qquad (16.6\text{–}2)$$

whence

$$a = \tfrac{1}{2}(1-\alpha_1-\alpha_2+\alpha_3), \quad b = \tfrac{1}{2}(1-\alpha_1-\alpha_2-\alpha_3), \quad c = 1-\alpha_1. \quad (16.6\text{–}3)$$

The quotient of two solutions which form a fundamental system of this hypergeometric equation maps the upper half of the z-plane onto the interior of a curvilinear triangle with the given angles. Different fundamental systems lead to different triangles, but all the triangles in the set have the same angles and we can pass from one triangle to another by a fractional linear transformation.

We consider more closely the function

$$s(z) = \frac{w_{01'}(z)}{w_{01}(z)}, \qquad (16.6\text{–}4)$$

where

$$\begin{aligned} w_{01}(z) &= F(a, b; c; z), \\ w_{01'}(z) &= z^{1-c}F(a-c+1, b-c+1; 2-c; z). \end{aligned} \qquad (16.6\text{–}5)$$

Since $1-c = \alpha_1 > 0$ the corresponding triangle has a vertex at $s(0) = 0$.

Since $c-a-b = \alpha_2 > 0$ the series (16.6–5) converge at $z = 1$ and from (16.3–34) we deduce that $z = 1$ corresponds to the vertex

$$s(1) = \frac{\Gamma(2-c)\Gamma(c-a)\Gamma(c-b)}{\Gamma(c)\Gamma(1-a)\Gamma(1-b)}. \tag{16.6–6}$$

The vertex corresponding to $z = \infty$ may be found from (16.1–56) applied to both functions (16.6–5). Making use of the fact that $a-b = \alpha_3 > 0$ we find after some computation

$$s(\infty) = e^{\pi i(1-c)} \frac{\Gamma(a)\Gamma(c-b)\Gamma(2-c)}{\Gamma(c)\Gamma(a-c+1)\Gamma(1-b)}. \tag{16.6–7}$$

The preceding formulas remain valid if $\alpha_2 = \alpha_3 = 0$. If $\alpha_1 = 0$ the solutions (16.6–5) coincide and we have to take the second solution obtained in section 16.4.2. This is in accordance with the fact that if $a_1 = \alpha_2 = \alpha_3 = 0$ the functions $\mathsf{K}(z)$ and $i\mathsf{K}'(z)$ fulfill our requirements.

The case that $b = 0$, i.e., $\alpha_1+\alpha_2+\alpha_3 = 1$ deserves mention. The triangle is rectilinear and we have

$$\frac{|s(\infty)-s(0)|}{|s(1)-s(0)|} = \frac{|s(\infty)|}{|s(1)|} = \frac{\Gamma(a)\Gamma(1-a)}{\Gamma(c-a)\Gamma(1+\alpha-c)} = \frac{\sin \pi(c-a)}{\sin \pi a} = \frac{\sin \pi\alpha_2}{\sin \pi\alpha_3}$$

in accordance with the sine rule in elementary trigonometry. In this case the first function (16.6–5) reduces to the constant 1. The second may be represented by one of the hypergeometric integrals of the type (16.1–42). Taking the second integral the mapping function becomes

$$z^{\alpha_1}F(1-\alpha_2, \alpha_1; 1+\alpha_1; z) = \frac{\Gamma(1+\alpha_1)}{\Gamma(\alpha_1)\Gamma(1)} z^{\alpha_1} \int_0^1 u^{\alpha_1-1}(1-zu)^{\alpha_2-1}\, du$$

which is, apart from a multiplicative constant, the same function as (10.3–10).

16.6.2 – The mapping of a regular curvilinear polygon

The problem of the mapping of the interior of the unit circle in the z-plane onto the interior of a regular curvilinear polygon can also be reduced to the solution of a hypergeometric differential equation.

Suppose we are given a polygon with centre at the origin and vertices at the nth roots of unity in the w-plane whose sides are congruent circular arcs which form the interior angle $\alpha\pi$ with each other, $0 \leq \alpha < 2$, (fig. 16.6–1). It follows from Riemann's mapping theorem (section 10.5.2) that there is a function $f(z)$ which maps the region $|z| < 1$ conformally onto the interior of the polygon and which is uniquely determined by the conditions

$$f(0) = 0, \qquad f'(0) > 0. \tag{16.6–8}$$

It is clear that the function $\overline{f(\bar{z})}$ provides a mapping of the interior of the unit circle onto the interior of the polygon as well. Since it vanishes at $z = 0$ and its derivative thereof is also $f'(0)$ we conclude that

$$f(z) = \overline{f(\bar{z})}. \tag{16.6–9}$$

This means that $f(z)$ is real for real values of z. Now $f'(z) \neq 0$ everywhere in the interior of the unit circle and it follows that $f'(x) > 0$ for $0 \leqq x < 1$. Hence $f(z)$ is monotonously uncreasing along the radius from 0 to 1 and since $f(z)$ is continuous at $z = 1$ it follows that $f(1) = 1$.

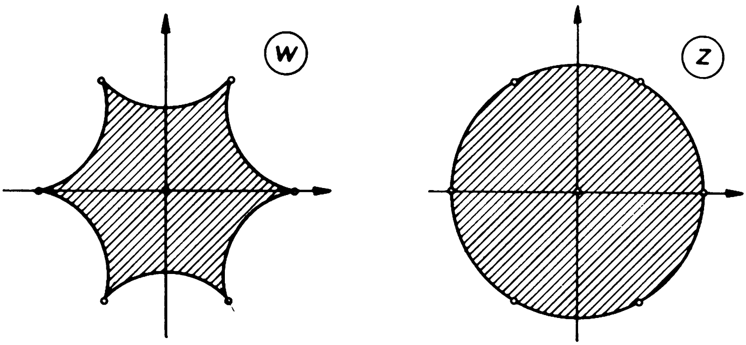


Fig. 16.6–1. The mapping of a regular curvilinear polygon

The function

$$e^{-2\pi i/n} f(e^{2\pi i/n} z)$$

maps again the interior of the unit circle onto the interior of the polygon. The derivative of this function takes the value $f'(0)$ at $z = 0$. As a consequence $f(z)$ satisfies the functional equation

$$f(e^{2\pi i/n} z) = e^{2\pi i/n} f(z) \tag{16.6–10}$$

and in particular

$$f(e^{2\pi i/n}) = e^{2\pi i/n}.$$

Thus we may infer that the n-th roots of unity

$$a_k = e^{2\pi ik/n}, \qquad k = 1, \ldots, n,$$

in the z-plane correspond to the vertices of the polygon.

Expanding both members of (16.6–10) in a power series of z we readily see that the function $f(z)$ is of the type

$$f(z) = zf_1(z^n); \tag{16.6–11}$$

where $f_1(z)$ is a power series with $f_1(0) = 1$.

It is easily verified that the invariant $R(z)$, being equal to the Schwarzian derivative

$$[f]_z = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2$$

turns out to be of the type

$$R(z) = z^{-2}f_2(z^n), \tag{16.6–12}$$

where $f_2(z)$ is an ordinary power series.

By virtue of (14.1–5) this invariant is equal to

$$\tfrac{1}{2}(1-\alpha^2)\sum_{\nu=1}^{n}\frac{1}{(z-a_\nu)^2}+\sum_{\nu=1}^{n}\frac{C_\nu}{z-a_\nu}.$$

Now

$$\sum_{\nu=1}^{n}\frac{1}{(z-a_\nu)^2} = -\frac{d}{dz}\sum_{\nu=1}^{n}\frac{1}{z-a_\nu} = -\frac{d}{dz}\sum_{\nu=1}^{n}\frac{\frac{d}{dz}(z-a_\nu)}{z-a_\nu}$$

$$= -\frac{d}{dz}\frac{\frac{d}{dz}(z^n-1)}{z^n-1} = -\frac{d}{dz}\frac{nz^{n-1}}{z^n-1} = n\left(\frac{nz^{2n-2}}{(z^n-1)^2}-\frac{(n-1)z^{n-2}}{z^n-1}\right)$$

$$= n\left(\frac{nz^{n-2}}{(z^n-1)^2}+\frac{z^{n-2}}{z^n-1}\right),$$

being again a function of the type (16.6–12). It follows that

$$\sum_{\nu=1}^{n}\frac{C_\nu}{z-a_\nu} = \frac{p(z)}{z^n-1}$$

must also be a function of this type. Since the degree of the polynomial $p(z)$ does not exceed $n-2$ it is of the form cz^{n-2}.

Now we recall that the invariant $R(z)$ satisfies the condition that it is regular at $z = \infty$ (section 14.1.1). As a consequence

$$c = -\tfrac{1}{2}n(1-\alpha^2)$$

and

$$R(z) = \tfrac{1}{2}n^2(1-\alpha^2)\frac{z^{n-2}}{(z^n-1)^2}. \tag{16.6–13}$$

The solution of the equation

$$[f]_z = R(z) \qquad (16.6\text{–}14)$$

may be reduced to the finding of a fundamental system of solutions of the second order equation

$$u'' + \tfrac{1}{4}n^2(1-\alpha^2)\frac{z^{n-2}}{(z^n-1)^2}\,u = 0, \qquad (16.6\text{–}15)$$

(eq. 15.1–28).

It is natural to take $t = z^n$ as an independent variable. If now primes denote differentiation with respect to t we readily find

$$tu'' + \left(1 - \frac{1}{n}\right)u' + \tfrac{1}{4}(1-\alpha^2)\frac{u}{(t-1)^2} = 0. \qquad (16.6\text{–}16)$$

Unfortunately this equation is not of the Fuchsian type. But we may introduce the function w by

$$u = w(t-1)^k$$

and we get

$$t(t-1)w'' + \left(t\left(2k+1-\frac{1}{n}\right) - \left(1-\frac{1}{n}\right)\right)w' +$$

$$+ \left(\frac{1}{t-1}\left(\tfrac{1}{4}(1-\alpha^2) + k(k-1)t\right) + k\left(1-\frac{1}{n}\right)\right)w = 0. \qquad (16.6\text{–}17)$$

The coefficient of w takes the desired form if $k(k-1) = -\frac{1}{4}(1-\alpha^2)$. We may take $k = \frac{1}{2}(1-\alpha)$. Our final result is the equation

$$t(1-t)w'' - \left(t\left(2-\alpha-\frac{1}{n}\right) - \left(1-\frac{1}{n}\right)\right)w' +$$

$$+\tfrac{1}{2}(1-\alpha)\left(\tfrac{1}{2}(1-\alpha) - \frac{1}{n}\right)w = 0. \qquad (16.6\text{–}18)$$

This is indeed a hypergeometric differential equation with parameters

$$a = \tfrac{1}{2}(1-\alpha) - \frac{1}{n}; \qquad b = \tfrac{1}{2}(1-\alpha), \qquad c = 1 - \frac{1}{n}.$$

Thus we have proved the following theorem

The function $f_\alpha(z)$ which effects the mapping of the interior of the unit circle in the z-plane onto the interior of a regular polygon with vertices at the n-th roots of unity and interior angle $\alpha\pi$, and satisfying the conditions

$$f_\alpha(0) = 0, \qquad f'_\alpha(0) > 0,$$

is

$$f_\alpha(z) = \frac{w_{01}(1)}{w_{01'}(1)} \frac{w_{01'}(z^n)}{w_{01}(z^n)}, \tag{16.6–19}$$

where

$$\begin{aligned} w_{01}(z) &= F\left(\tfrac{1}{2}(1-\alpha)-\frac{1}{n}, \tfrac{1}{2}(1-\alpha); 1-\frac{1}{n}; z^n\right), \\ w_{01'}(z) &= zF\left(\tfrac{1}{2}(1-\alpha)+\frac{1}{n}, \tfrac{1}{2}(1-\alpha); 1+\frac{1}{n}; z^n\right). \end{aligned} \tag{16.6–20}$$

For applications in the next section we need the value of $f'_\alpha(0)$. It is readily seen by (16.3–34) that

$$f'_\alpha(0) = \lim_{z\to 0} \frac{f_\alpha(z)}{z} = \frac{w_{01}(1)}{w_{01'}(1)} = \frac{\Gamma\left(1-\frac{1}{n}\right)\Gamma\left(\frac{1}{2}(1+\alpha)+\frac{1}{n}\right)}{\Gamma\left(1+\frac{1}{n}\right)\Gamma\left(\frac{1}{2}(1+\alpha)-\frac{1}{n}\right)}. \tag{16.6–21}$$

It is interesting to check the theorem for the case that the polygon is rectilinear. Then

$$\alpha = \frac{n-2}{n} = 1-\frac{2}{n}$$

and the first function (16.1–20) is identically equal to the constant 1. The second function may be obtained from the second hypergeometric integral (16.1–42), viz.,

$$\frac{\Gamma\left(1+\frac{1}{n}\right)}{\Gamma\left(\frac{1}{n}\right)\Gamma(1)} z\int_0^1 u^{1/n-1}(1-z^n u)^{-2/n}\,du = \int_0^z (1-t^n)^{-2/n}\,dt,$$

in accordance with (10.3–11).

16.6.3 – Upper bound for the Bloch and the Landau constants

L. Ahlfors and H. Grunsky have shown that the function (16.6–19) may serve to obtain an upper bound for the Bloch constant B defined in sections 9.9.1. By a similar method we can also obtain an upper bound for Landau's constant L.

We take $n = 3$ and denote the corresponding triangle by Δ_α. If $f_\alpha(z)$ denotes the function which effects the mapping of the open unit disc $|z| < 1$ onto the interior of the triangle Δ_α, then evidently

$$F(z) = f_{\frac{1}{3}}(\check{f}_{\frac{1}{3}}(z)), \tag{16.6–22}$$

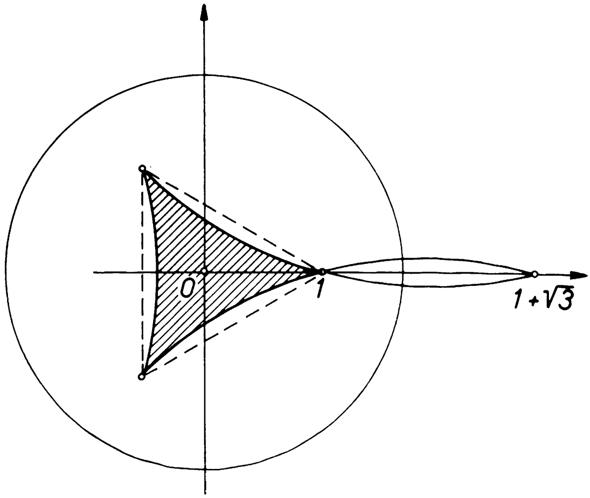


Fig. 16.6–2. The triangle $\Delta_{\frac{1}{6}}$

where $\check{f}$ denotes, as usual, the inverse of f, is a function which maps the interior of $\Delta_{\frac{1}{6}}$ conformally onto the interior of $\Delta_{\frac{1}{6}}$. The boundary circles of $\Delta_{\frac{1}{6}}$ are orthogonal to a circle about the origin whose radius R we wish to determine. Elongating the boundary circles through $z = 1$ they meet the real axis again in a point whose distance to $z = 1$ is equal to a chord of the bounding arcs of $\Delta_{\frac{1}{6}}$; these have the length $\sqrt{3}$, (fig. 16.6–2). Since the points 1 and $1+\sqrt{3}$ are inverses with respect to the circle of radius R we find that

$$R = \sqrt{1+\sqrt{3}}. \tag{16.6–23}$$

Successive reflexions in the sides of $\Delta_{\frac{1}{6}}$ and the resulting triangles lead to a net of curvilinear triangles which fills the open disc $|z| < R$ without overlapping. The images as given by $F(z)$ form a net of equilateral triangles. Since six curvilinear triangles which have a vertex in common form a full neighbourhood of this vertex, the corresponding equilateral triangles have a vertex in common and cover a neighbourhood of this vertex twice. The radius of the circumscribed circle of $\Delta_{\frac{1}{6}}$ is equal to unity and the interior corresponds univalently to a subregion of $|z| < R$. It is, moreover, the largest disc having this property. The function

$$f(z) = \frac{F(Rz)}{RF'(0)} \tag{16.6–24}$$

is holomorphic in $|z| < 1$ and satisfies the condition $f'(0) = 1$. It follows that the Bloch radius of this function is

$$B_f = \frac{1}{RF'(0)} = \frac{1}{R}\,\frac{f'_{\frac{1}{6}}(0)}{f'_{\frac{1}{3}}(0)}.$$

With the aid of (16.6–9) we now may conclude that

$$B \leqq \frac{\Gamma(\frac{1}{3})\Gamma(\frac{11}{12})}{\sqrt{1+\sqrt{3}}\,\Gamma(\frac{1}{4})} = 0.47\ldots. \tag{16.6–25}$$

An upper bound for the Landau constant L may be obtained by means of the function

$$F(z) = f_{\frac{1}{3}}(\check{f}_0(z)). \tag{16.6–26}$$

The radius of the circumscribed circle of Δ_0 is unity. The function $F(z)$ maps the interior of Δ_0 onto the interior of $\Delta_{\frac{1}{3}}$ and may be continued throughout the disc $|z| < 1$. The image of the disc as given by this function covers the plane infinitely many times. The vertices of the triangles are not within the circumscribed circle of each triangle $\Delta_{\frac{1}{3}}$ and, therefore, do not belong to the image of $|z| < 1$. Hence the circumscribed circle of $\Delta_{\frac{1}{3}}$ yields the largest disc which is covered by the image of $|z| < 1$. The function

$$f(z) = \frac{F(z)}{F'(0)} \tag{16.6–27}$$

has the property $f'(0) = 1$ and the Landau radius of f turns out to be

$$L_f = \frac{1}{F'(0)} = \frac{f_0'(0)}{f'_{\frac{1}{3}}(0)}.$$

It follows that

$$L \leqq \frac{\Gamma(\frac{5}{6})\Gamma(\frac{1}{3})}{\Gamma(\frac{1}{6})} = 0.54\ldots. \tag{16.6–28}$$

16.7 – Confluent hypergeometric functions

16.7.1 – Kummer's differential equation

A differential equation which may be derived from another differential equation by making two or more singularities of the latter tend to coincide is called a *confluent form* of the latter. The limiting process is called *confluence*. We encountered it already in section 15.8–1.

A very important type of confluent equations is obtained from the hypergeometric equation. We start with the differential equation for the function

$$F\left(a, b; c; \frac{z}{b}\right), \tag{16.7–1}$$

viz.,

$$z\left(1-\frac{z}{b}\right)w''-\left(\left(\frac{a}{b}+1+\frac{c}{b}\right)z-c\right)w'-aw = 0. \tag{16.7–2}$$

This is a Riemannian equation with singular points at $z = 0$, $z = b$, $z = \infty$. If b tends to ∞ we obtain the *confluent hypergeometric equation of Kummer*

$$\boxed{zw''-(z-c)w'-aw = 0.} \tag{16.7–3}$$

If we introduce the operator ϑ defined in section 16.1.2 we get

$$\vartheta(\vartheta+c-1)w = z(\vartheta+a)w. \tag{16.7–4}$$

The equation (16.7–3) has a regular singularity at $z = 0$, whereas the singularity at $z = \infty$ is not regular. The indicial equation at $z = 0$ is

$$\rho(\rho+c-1) = 0 \tag{16.7–5}$$

and the exponents are $\rho = 0$, $\rho = 1-c$. Hence, if c is not an integer, there is one solution regular at $z = 0$.

16.7.2 – Solution of Kummer's equation

Let us insert the series

$$w_0(z) = \sum_{\nu=0}^{\infty} c_\nu z^\nu \tag{16.7–6}$$

into the equation (16.7–3). Taking (16.1–12) into account we find

$$\vartheta(\vartheta+c-1)w_0(z) = \sum_{\nu=0}^{\infty} \nu(\nu+c-1)c_\nu z^\nu = \sum_{\nu=0}^{\infty} (\nu+c)(\nu+1)c_{\nu+1} z^{\nu+1}$$

and

$$z(\vartheta+a)w_0(z) = \sum_{\nu=0}^{\infty} (\nu+a)c_\nu z^{\nu+1}.$$

As a consequence the coefficients of the power series (16.7–6) must satisfy the relations

$$(n+c)(n+1)c_{n+1} = (n+a)c_n, \qquad n = 0, 1, 2, \ldots. \tag{16.7–7}$$

Assuming that c is neither zero, nor a negative integer, we may take

$c_0 = 1$ and the remaining coefficients are uniquely determined by

$$\frac{c_{n+1}}{c_n} = \frac{a+n}{(c+n)(1+n)}, \tag{16.7–8}$$

whence

$$c_n = \frac{a(a+1)\ldots(a+n-1)}{c(c+1)\ldots(c+n-1)n!} = \frac{(a)_n}{(c)_n n!} \tag{16.7–9}$$

and we find *Kummer's series*

$$\boxed{\Phi(a; c; z) = \sum_{\nu=0}^{\infty} \frac{(a)_\nu}{(c)_\nu} \frac{z^\nu}{\nu!}.} \tag{16.7–10}$$

Since $c_{n+1}/c_n \to 0$ as $n \to \infty$ the series (16.7–10) is convergent throughout the z-plane and represents, therefore, an integral function. A particular example is

$$\Phi(a; a; z) = \exp z. \tag{16.7–11}$$

It is clear that the series (16.7–10) arises from the series (16.7–1) by a formal limiting process.

A second solution belongs to the exponent $1-c$. Inserting the series

$$w_1(z) = \sum_{\nu=0}^{\infty} c_\nu z^{\nu+1-c} \tag{16.7–12}$$

into the equation (16.7–3) we obtain as above the recurrent relations

$$(n+2-c)(n+1)c_{n+1} = (n+a-c+1)c_n, \qquad n = 0, 1, \ldots. \tag{16.7–13}$$

Assuming that c is not an integer $\geqq 2$ we get, by taking $c_0 = 1$

$$c_n = \frac{(a-c+1)_n}{(2-c)_n n!} \tag{16.7–14}$$

and the second solution turns out to be

$$w_1(z) = z^{1-c}\Phi(a-c+1; 2-c; z). \tag{16.7–15}$$

Proceeding as in paragraph 16.4 we may also derive a fundamental system for the case that c is an integer. The results can be foreseen by a formal limiting process.

16.7.3 – Kummer's relation

In many cases the investigation of Kummer's series is much facilitated if we represent it by an integral like the hypergeometric integrals introduced in section 16.1.7. In the first equation (16.1–42) we replace z by z/b; we get

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)} F\left(a, b; c; \frac{z}{b}\right) = \int_0^1 u^{a-1}(1-u)^{c-a-1}\left(1+\frac{zu}{-b}\right)^{-b} du.$$

Assuming that passing to the limit within the sign of integration is legitimate, we find

$$\boxed{\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)} \Phi(a; c; z) = \int_0^1 u^{a-1}(1-u)^{c-a-1} e^{zu} du.} \tag{16.7–16}$$

The proof of this formula is easy. Expanding the exponential function and interchanging the order of integration and summation (which is allowed) we have

$$\int_0^1 u^{a-1}(1-u)^{c-a-1} e^{zu} du = \sum_{\nu=0}^{\infty} \frac{z^\nu}{\nu!} \int_0^1 u^{a+\nu-1}(1-u)^{c-a-1} du$$

$$= \sum_{\nu=0}^{\infty} \frac{\Gamma(a+\nu)\Gamma(c-a)}{\Gamma(c+\nu)} \frac{z^\nu}{\nu!} = \frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)} \sum_{\nu=0}^{\infty} \frac{(a)_\nu}{(c)_\nu} \frac{z^\nu}{\nu!}.$$

The result is true under the restriction $\mathrm{Re}\, a > 0$, $\mathrm{Re}\,(c-a) > 0$; in more unfavourable circumstances we consider a similar integral taken along a Jordan-Pochhammer contour.

The formula (16.7–16) enables us to derive a very important relation, due to Kummer. Replacing u by $1-u$ we get

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)} \Phi(a; c; z) = e^z \int_0^1 u^{c-a-1}(1-u)^{a-1} e^{-zu} du$$

$$= e^z \frac{\Gamma(c-a)\Gamma(a)}{\Gamma(c)} \Phi(c-a; c; -z).$$

Thus we proved the relation

$$\boxed{\Phi(a; c; z) = e^z \Phi(c-a; c; -z).} \tag{16.7–17}$$

This formula has been obtained under restrictive conditions for the parameters a and b. But it can be proved as in section 16.3.1 that $\Phi(a; c; z)$ is holomorphic with respect to a and c respectively. If we employ the integral taken along a Jordan-Pochhammer contour the restrictions can be dropped before.

An interesting specialization of Kummer's function may be derived from the integral on the right of (16.7–16) by taking $c = a+1$. Replacing, moreover, z by $-s$ we obtain

$$\frac{1}{a} \Phi(a; a+1; -s) = \int_0^1 u^{a-1} e^{-su} du = s^{-a} \int_0^s t^{a-1} e^{-t} dt.$$

Thus the incomplete gamma function defined by the integral (7.19–1) appears. Taking Kummer's formula (16.7–17) into account we finally have

$$\boxed{P(s, a) = \int_0^s t^{a-1} e^{-t} dt = \frac{s^a e^{-s}}{a} \Phi(1; a+1; s).} \tag{16.7–18}$$

16.7.4 – Contiguous functions

Contiguous Kummer series are defined like contiguous hypergeometric series. Proceeding in quite the same way as in section 16.3.2 we may find relations analogous to those listed in (16.3–14). First we have

$$a\Phi_{a+} = (\vartheta + a)\Phi \tag{16.7–19}$$

and

$$(c-1)\Phi_{c-} = (\vartheta + c - 1)\Phi. \tag{16.7–20}$$

By virtue of (16.7–4) the function Φ_{a-} satisfies the relation

$$\vartheta(\vartheta + c - 1)\Phi_{a-} = z(\vartheta + a - 1)\Phi_{a-},$$

or

$$(\vartheta + c - a - z)(\vartheta + a - 1)\Phi_{a-} = (c-a)(a-1)\Phi_{a-}.$$

Using (16.7–19) with $a-1$ written for a we get

$$(c-a)\Phi_{a-} - \vartheta\Phi + (a - c + z)\Phi = 0. \tag{16.7–21}$$

The function Φ_{c+} satisfies the relation

$$\vartheta(\vartheta + c)\Phi_{c+} = z(\vartheta + a)\Phi_{c+},$$

or

$$(\vartheta - z)(\vartheta + c)\Phi_{c+} = z(a-c)\Phi_{c+}.$$

Using (16.7–20) this becomes

$$(\vartheta - z)c\Phi = z(a-c)\Phi_{c+}. \tag{16.7–22}$$

Summing up we have

$$\begin{aligned} z\Phi' &= a(\Phi_{a+} - \Phi), \\ z\Phi' &= (c-1)(\Phi_{c-} - \Phi), \\ z\Phi' &= (c-a)\Phi_{a-} + (a - c + z)\Phi, \\ c\Phi' &= -(c-a)\Phi_{c+} + c\Phi. \end{aligned} \tag{16.7–23}$$

It is readily seen that we may obtain these relations from those listed in (16.3–14) by a formal limiting process in an obvious way. Eliminating Φ' from the relations (16.7–23) yields six relations between contiguous Kummer series.

16.7.5 – The differential equation of Whittaker

By a simple transformation we may derive from Kummer's differential equation another differential equation which plays an important part in mathematical physics. Inserting $w = W\varphi$ in the equation (16.7–2), where φ is a function to be determined appropriately, we find that W satisfies a differential equation

$$zW''+\left(2z\frac{\varphi'}{\varphi}+c-z\right)W'+\left(z\frac{\varphi''}{\varphi}+(c-z)\frac{\varphi'}{\varphi}-a\right)W = 0. \qquad (16.7\text{–}24)$$

This equation is considerably simplified if φ satisfies the condition

$$2z\frac{\varphi'}{\varphi}+c-z = 0, \qquad (16.7\text{–}25)$$

or

$$\frac{\varphi'}{\varphi} = -\frac{c}{2z}+\tfrac{1}{2}.$$

A solution of this equation is

$$\varphi(z) = z^{-\frac{1}{2}c}e^{\frac{1}{2}z}.$$

The coefficient of W becomes

$$-\tfrac{1}{4}z+\tfrac{1}{2}c-a+\frac{\tfrac{1}{2}c-\tfrac{1}{4}c^2}{z}.$$

It is common use to put

$$\tfrac{1}{2}c-a = \kappa, \qquad c = 2\mu+1. \qquad (16.7\text{–}26)$$

Then the equation (16.7–24) appears in the form

$$\boxed{W''+\left(-\tfrac{1}{4}+\frac{\kappa}{z}+\frac{\tfrac{1}{4}-\mu^2}{z^2}\right)W = 0.} \qquad (16.7\text{–}27)$$

This is *Whittaker's differential equation.*

It is clear that a fundamental system of solutions is given by the symmetric pair

$$\begin{aligned} M_{\kappa,\mu}(z) &= z^{\frac{1}{2}+\mu}e^{-\frac{1}{2}z}\Phi(\tfrac{1}{2}+\mu-\kappa;\, 2\mu+1;\, z),\\ M_{\kappa,-\mu}(z) &= z^{\frac{1}{2}-\mu}e^{-\frac{1}{2}z}\Phi(\tfrac{1}{2}-\mu-\kappa;\, -2\mu+1;\, z), \end{aligned} \qquad (16.7\text{–}28)$$

provided that 2μ is not an integer.

Kummer's relation (16.7–17) takes the elegant form

$$z^{-\frac{1}{2}-\mu}M_{\kappa,\mu}(z) = (-z)^{-\frac{1}{2}-\mu}M_{-\kappa,\mu}(-z). \qquad (16.7\text{–}29)$$

A simple transformation of Whittaker's equation leads to another important differential equation. First we replace in (16.7–27) z by $\frac{1}{2}z^2$. The modified equation is

$$W'' - \frac{W'}{z} + \left(-\tfrac{1}{4}z^2 + 2\kappa + \frac{1-4\mu^2}{z^2}\right) W = 0. \qquad (16.7\text{–}30)$$

Next we replace W by $u\varphi$ where φ satisfies the condition

$$2\frac{\varphi'}{\varphi} - \frac{1}{z} = 0.$$

We may take

$$\varphi = z^{\frac{1}{2}}.$$

Then (16.7–30) appears as

$$u'' + \left(2\kappa - \tfrac{1}{4}z^2 + \frac{1-16\mu^2}{4z^2}\right) u = 0.$$

The particular equation with $\mu = \pm\frac{1}{4}$ is *Weber's equation*

$$\boxed{u'' + (2\kappa - \tfrac{1}{4}z^2)u = 0.} \qquad (16.7\text{–}31)$$

A fundamental system of solutions is evidently

$$\begin{aligned} z^{-\frac{1}{2}}M_{\kappa,\frac{1}{4}}(\tfrac{1}{2}z^2) &= 2^{-\frac{3}{4}}ze^{-\frac{1}{4}z^2}\Phi(\tfrac{3}{4}-\kappa;\, \tfrac{3}{2};\, \tfrac{1}{2}z^2), \\ z^{-\frac{1}{2}}M_{\kappa,-\frac{1}{4}}(\tfrac{1}{2}z^2) &= 2^{-\frac{1}{4}}ze^{-\frac{1}{4}z^2}\Phi(\tfrac{1}{4}-\kappa;\, \tfrac{1}{2};\, \tfrac{1}{2}z^2). \end{aligned} \qquad (16.7\text{–}32)$$

The solutions of (16.7–31) are called *parabolic cylinder functions.* They are single-valued throughout the z-plane.

16.7.6 – Bessel's equation

If instead of (16.7–25) we impose on φ the condition

$$2z\frac{\varphi'}{\varphi} + c - z = 1,$$

then we may take φ as

$$\varphi(z) = z^{-\frac{1}{2}(c-1)}e^{\frac{1}{2}z} = z^{-\kappa}e^{\frac{1}{2}z},$$

where we have put

$$c = 2\kappa + 1. \qquad (16.7\text{–}33)$$

Inserting $w = u\varphi$ into the equation (16.7–2) the coefficient of u becomes

$$-\frac{\kappa^2}{z}-\frac{z}{4}+\kappa+\tfrac{1}{2}-a.$$

This is simplified if we take $a = \frac{1}{2}+\kappa$. The modified equation takes the form

$$z^2u''+zu'-(\tfrac{1}{4}z^2+\kappa^2)u = 0.$$

Replacing z by $2iz$ we finally have

$$z^2u''+zu'+(z^2-\kappa^2)u = 0. \tag{16.7–34}$$

This is, however, Bessel's differential equation (15.3–1). As a consequence the functions

$$\mathrm{e}^{iz}(2iz)^{-\kappa}J_\kappa(z) = (4i)^{-\kappa}\mathrm{e}^{iz}\left(\frac{z}{2}\right)^{-\kappa}J_\kappa(z)$$

and

$$\Phi(\kappa+\tfrac{1}{2};\,2\kappa+1;\,2iz),$$

both being regular at $z = 0$, differ only by a multiplicative constant. The value of the first function at $z = 0$ is

$$\frac{(4i)^{-\kappa}}{\Gamma(\kappa+1)}.$$

It follows that

$$\boxed{J_\kappa(z) = \frac{\mathrm{e}^{-iz}}{\Gamma(\kappa+1)}\left(\frac{z}{2}\right)^{\kappa}\Phi(\kappa+\tfrac{1}{2};\,2\kappa+1;\,2iz),} \tag{16.7–35}$$

provided that 2κ is not a negative integer.

16.8 – **Confluent hypergeometric polynomials**

16.8.1 – The Laguerre polynomials

The process of confluence applied to Jacobian polynomials produces an interesting class of orthogonal polynomials which play an important part in various fields of mathematics and mathematical physics. From

$$P_n^{(\alpha,\beta)}\left(1-\frac{2z}{\beta}\right) = \binom{n+\alpha}{n}F\left(-n,\,n+\alpha+\beta+1;\,\alpha+1;\,\frac{z}{\beta}\right)$$

(being a modification of (16.2–10)) arises, if $\beta \to 1$, the polynomial

$$\boxed{L_n^{(\alpha)}(z) = \binom{n+\alpha}{n}\Phi(-n;\,\alpha+1;\,z).} \tag{16.8–1}$$

The polynomials $L_n^{(0)}(z)$ briefly denoted by $L_n(z)$, were discovered by Laguerre. It is common use to call the more general confluent hypergeometric polynomials (16.8–1) *Laguerre polynomials.*

From the expression for the Kummer series on the right of (16.7–1) we find

$$L_n^{(\alpha)}(z) = \frac{(1+\alpha)_n}{n!} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{(-1)^\nu}{(1+\alpha)_\nu} z^\nu. \tag{16.8–2}$$

The first polynomials are

$$L_0^{(\alpha)}(z) = 1, \qquad L_1^{(\alpha)}(z) = \alpha+1-z,$$
$$L_2^{(\alpha)}(z) = \tfrac{1}{2}(\alpha+1)(\alpha+2)-(\alpha+2)z+\tfrac{1}{2}z^2, \ldots .$$

The degree of $L_n^{(\alpha)}(z)$ is precisely n, the coefficient of the leading term being

$$k_n = \frac{(-1)^n}{n!}. \tag{16.8–3}$$

It is clear from (16.7–1) that the Laguerre polynomials are solutions of the differential equation

$$\boxed{zw''+(\alpha+1-z)w'+nw = 0.} \tag{16.8–4}$$

16.8.2 – The Rodrigues formula

By transforming (16.8–2) we may derive a kind of a Rodrigues formula for the Laguerre polynomials. We proceed as in section 16.2.2. First we observe that

$$\frac{d^k e^{-z}}{dz^k} = (-1)^k e^{-z}, \qquad k = 0, 1, \ldots, n,$$

whence

$$(-1)^k = e^z \frac{d^k e^{-z}}{dz^k}.$$

Secondly we have

$$\frac{d^{n-k}}{dz^{n-k}} z^{\alpha+n} = (\alpha+n)\ldots(\alpha+k+1)z^{\alpha+k} = \frac{(\alpha+1)_n}{(\alpha+1)_k} z^{\alpha+k},$$

whence

$$z^k = z^{-\alpha} \frac{(\alpha+1)_k}{(\alpha+1)_n} \frac{d^{n-k}}{dz^{n-k}} z^{\alpha+n}.$$

Hence (16.8–2) becomes

$$L_n^{(\alpha)}(z) = \frac{e^z z^{-\alpha}}{n!} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{d^\nu}{dz^\nu} e^{-z} \frac{d^{n-\nu}}{dz^{n-\nu}} z^{\alpha+n}.$$

In view of Leibniz' rule for the nth derivative of a product we now have

$$L_n^{(\alpha)}(z) = \frac{1}{n!p(z)} \frac{d^n}{dz^n}(p(z)z^n), \qquad (16.8\text{–}5)$$

with

$$p(z) = e^{-z}z^\alpha. \qquad (16.8\text{–}6)$$

16.8.3 – Orthogonality

In order to obtain the relations of orthogonality we start with the function

$$u(x) = p(x)x^n, \qquad (16.8\text{–}7)$$

where $p(x)$ is the function (16.8–6) and n an integer $\geqq 0$.
Let m denote one of the numbers $0, \ldots, n$. By integrating by parts we obtain

$$\int_0^\infty x^m u^{(n)}(x)dx = (-1)^m m! \int_0^\infty u^{(n-m)}(x)dx = 0, \qquad (16.8\text{–}8)$$

if $m < n$. The integral is convergent if we assume $\alpha > -1$. In the case $m = n$ we have

$$\int_0^\infty x^n u^{(n)}(x)dx = (-1)^n n! \int_0^\infty u(x)dx = (-1)^n n! \int_0^\infty x^{\alpha+n} e^{-x} dx,$$

whence

$$\int_0^\infty x^n u^{(n)}(x)dx = (-1)^n n! \Gamma(\alpha+n+1). \qquad (16.8\text{–}9)$$

From (16.8–8), (16.8–9), (16.8–5) and (16.8–3) we easily obtain the *relations of orthogonality for the Laguerre polynomials*

$$\int_0^\infty e^{-x} x^\alpha L_m^{(\alpha)}(x) L_n^{(\alpha)}(x)dx = \begin{cases} 0, & \text{if } m \neq n. \\ \dfrac{\Gamma(\alpha+n+1)}{n!}, & \text{if } m = n. \end{cases} \qquad (16.8\text{–}10)$$

16.8.4 – Recurrent relations

The method for obtaining recurrent relations as described in section 16.2.4 is also applicable to Laguerre polynomials. In quite the same way we find by the aid of the relations of orthogonality the analogon of

(16.2–19), viz.,

$$L_{n+1}^{(\alpha)}(z) = (A_n z + B_n)L_n^{(\alpha)}(z) + C_n L_{n-1}^{(\alpha)}(z), \tag{16.8–11}$$

where A_n, B_n and C_n have the values (16.2–25) by effecting the process of confluence. In our case we have to take

$$A_n = \frac{-1}{n+1}, \qquad B_n = \frac{2n+\alpha+1}{n+1}, \qquad C_n = -\frac{n+\alpha}{n+1} \tag{16.8–12}$$

whence

$$\boxed{(n+1)L_{n+1}^{(\alpha)}(z) = ((2n+\alpha+1)-z)L_n^{(\alpha)}(z) - (n+\alpha)L_{n-1}^{(\alpha)}(z).} \tag{16.8–13}$$

It is of interest to derive this formula by a wholly different method. To this end we employ the relations listed in (16.7–22). Since

$$\Phi(-n, \alpha+1; z) = \frac{n!}{(\alpha+1)_n} L_n^{(\alpha)}(z),$$

we readily find

$$\boxed{\begin{aligned} z\frac{dL_n^{(\alpha)}(z)}{dz} &= -(\alpha+n)L_{n-1}^{(\alpha)}(z) + nL_n^{(\alpha)}(z), \\ z\frac{dL_n^{(\alpha)}(z)}{dz} &= (\alpha+n)L_n^{(\alpha-1)}(z) - \alpha L_n^{(\alpha)}(z), \\ z\frac{dL_n^{(\alpha)}(z)}{dz} &= (n+1)L_{n+1}^{(\alpha)}(z) - (n+\alpha+1-z)L_n^{(\alpha)}(z), \\ \frac{dL_n^{(\alpha)}}{dz} &= -L_n^{(\alpha+1)}(z) + L_n^{(\alpha)}(z). \end{aligned}} \tag{16.8–14}$$

Eliminating $dL_n^{(\alpha)}(z)/dz$ from the first and the third relation we immediatey get the recurrent relation (16.8–13) which is, therefore, essentially a relation between contiguous functions. In this way a great variety of relations can be found. Eliminating, for instance, $dL_n^{(\alpha)}(z)/dz$ from the first two relations we get

$$\boxed{L_n^{(\alpha-1)}(z) = L_n^{(\alpha)}(z) - L_{n-1}^{(\alpha)}(z).} \tag{16.8–15}$$

Eliminating $dL_n^{(\alpha)}(z)/dz$ from the third and the fourth relation yields

$$(\alpha+n+1)L_n^{(\alpha)}(z) = (n+1)L_{n+1}^{(\alpha)}(z) + zL_n^{(\alpha+1)}(z)$$

and a shift in the index n gives

$$zL_{n-1}^{(\alpha)}(z) = (\alpha+n)L_{n-1}^{(\alpha)}(z) - nL_n^{(\alpha)}(z).$$

Comparing this with the right hand member of the first relation (16.8–14) we see that

$$\boxed{\frac{d}{dz} L_n^{(\alpha)}(z) = -L_{n-1}^{(\alpha+1)}(z).} \tag{16.8–16}$$

16.8.5 – THE GENERATING FUNCTION

The recurrent relation (16.8–13) enables us to find a simple generating function for the Laguerre polynomials. Let

$$F(w) = \sum_{\nu=0}^{\infty} L_\nu^{(\alpha)}(z) w^\nu. \tag{16.8–17}$$

Then, in view of (16.8–13),

$$\begin{aligned}
F'(w) &= \sum_{\nu=0}^{\infty} (\nu+1) L_{\nu+1}^{(\alpha)}(z) w^\nu \\
&= \sum_{\nu=0}^{\infty} (2\nu+\alpha+1-z) L_\nu^{(\alpha)}(z) w^\nu - \sum_{\nu=0}^{\infty} (\nu+\alpha+1) L_\nu^{(\alpha)}(z) w^{\nu+1} \\
&= 2w \sum_{\nu=1}^{\infty} \nu L_\nu^{(\alpha)}(z) w^{\nu-1} + (\alpha+1-z) \sum_{\nu=0}^{\infty} L_\nu^{(\alpha)}(z) w^\nu + \\
&\quad - w^2 \sum_{\nu=0}^{\infty} \nu L_\nu^{(\alpha)}(z) w^{\nu-1} - (\alpha+1) w \sum_{\nu=0}^{\infty} L_\nu^{(\alpha)}(z) w^\nu \\
&= (2w-w^2)F'(w) + ((\alpha+1-z)-(\alpha+1)w)F(w).
\end{aligned}$$

It follows that

$$\frac{F'(w)}{F(w)} = \frac{\alpha+1-z-(\alpha+1)w}{1-2w+w^2} = \frac{\alpha+1}{1-w} - \frac{z}{(1-w)^2}.$$

Integration yields

$$F(w) = C(1-w)^{-(\alpha+1)} \exp\left(\frac{-zw}{1-w}\right).$$

Since $F(0) = L_0^{(\alpha)}(z) = 1$ the constant C is unity and we have

The Laguerre polynomial $L_n^{(\alpha)}(z)$ is the coefficient of w^n in the expansion

$$\boxed{(1-w)^{-(\alpha+1)} \exp\left(\frac{-zw}{1-w}\right) = \sum_{\nu=0}^{\infty} L_\nu^{(\alpha)}(z) w^\nu.} \tag{16.8–18}$$

Differentiating both members of (16.8–18) with respect to z yields, if

$F(w)$ stands for the left hand member of (16.8–18),

$$-\frac{w}{1-w}F(w) = \sum_{\nu=0}^{\infty} \frac{dL_\nu^{(\alpha)}(z)}{dz} w^\nu$$

and this is evidently

$$-\sum_{\nu=0}^{\infty} L_\nu^{(\alpha+1)}(z) w^{\nu+1} = \sum_{\nu=0}^{\infty} \frac{dL_\nu^{(\alpha)}(z)}{dz} w^\nu.$$

Comparing the coefficients of w^n we obtain again the relation (16.8.16).

The generating function leads to simple finite sum properties. From

$$(1-w)^{-(\alpha+1)} \exp\left(\frac{-zw}{1-w}\right) = (1-w)^{-(\alpha-\beta)}(1-w)^{-(\beta+1)} \exp\left(\frac{-zw}{1-w}\right)$$

follows

$$\sum_{\nu=0}^{\infty} L_\nu^{(\alpha)}(z) w^\nu = \sum_{\nu=0}^{\infty} (-1)^\nu \binom{-(\alpha-\beta)}{\nu} w^\nu \sum_{\nu=0}^{\infty} L_\nu^{(\beta)}(z) w^\nu$$

and, taking (16.1–21) into account, by the multiplication rule of power series

$$\boxed{L_n^{(\alpha)}(z) = \sum_{\nu=0}^{n} \frac{(\alpha-\beta)_\nu}{\nu!} L_{n-\nu}^{(\beta)}(z).} \qquad (16.8\text{–}19)$$

The specialization $\alpha-\beta = 1$ yields the interesting sum formula

$$\boxed{L_n^{(\alpha)}(z) = \sum_{\nu=0}^{n} L_\nu^{(\alpha-1)}(z).} \qquad (16.8\text{–}20)$$

From

$$(1-w)^{-(\alpha+1)} \exp\left(\frac{-uw}{1-w}\right) (1-w)^{-(\beta+1)} \exp\left(\frac{-vw}{1-w}\right)$$
$$= (1-w)^{-((\alpha+\beta+1)+1)} \exp\left(\frac{-(u+v)w}{1-w}\right)$$

follows

$$\boxed{L_n^{(\alpha+\beta+1)}(u+v) = \sum_{\nu=0}^{n} L_\nu^{(\alpha)}(u) L_{n-\nu}^{(\beta)}(v).} \qquad (16.8\text{–}21)$$

16.8.6 – General properties of orthogonal polynomials

Up to now we studied certain types of orthogonal polynomials which are closely related to hypergeometric functions. There are certain properties of orthogonal polynomials which are independent of their being hypergeometric series. On the other hand the class of orthogonal polynomials is very large and the question arises whether it is possible to characterize in a general way the class of polynomials studied before. As we shall see in the next section the generalized Rodrigues formula gives a satisfactory answer to this question. In this section we shall discuss some elementary properties.

We consider a sequence of polynomials

$$P_0(z), P_1(z), \dots,$$

with real coefficients such that the degree of $P_n(z)$ is precisely n. If there exists an interval $a < z < b$ and a function $p(z)$ which is positive on that interval we say that the functions form an orthogonal set with respect to the *weight function* $p(z)$ over the interval if

$$\int_a^b p(x)P_m(x)P_n(x)dx = 0, \qquad m \neq n. \tag{16.8–22}$$

For our purpose we shall assume that $p(x)$ is continuous throughout the interval and that the integral

$$\int_a^b p(x)dx \tag{16.8–23}$$

exists in the case that a or b or both are infinite. First we shall prove

Given a function $p(x)$ satisfying the conditions stated above we can always find a sequence of real polynomials orthogonal with respect to this function as weight function.

The proof is given by induction. Take $P_0(x) = 1$ identically. Suppose we are already in possession of an orthogonal system of n polynomials $P_0(x), \dots, P_{n-1}(x)$. A polynomial of the type

$$P_n(x) = \lambda_0 P_0(x) + \lambda_1 P_1(x) + \dots + \lambda_{n-1}P_{n-1}(x) + x^n,$$

where $\lambda_0, \dots, \lambda_{n-1}$ are real constants, is orthogonal to $P_m(x)$, $m = 0, \dots,$ $n-1$ if and only if

$$\int_a^b p(x)P_m(x)dx$$
$$= \sum_{\nu=0}^{n-1} \lambda_\nu \int_a^b p(x)P_m(x)P_\nu(x)dx + \int_a^b p(x)P_m(x)x^n\,dx$$
$$= \lambda_m \int_a^b p(x)P_m^2(x)dx + \int_a^b p(x)P_m(x)x^n\,dx = 0.$$

Since $p(x) > 0$ in the interior of the interval we have

$$\int_a^b p(x)P_m^2(x)dx > 0$$

and it follows

$$\lambda_m = -\frac{\int_a^b p(x)P_m(x)x^n\,dx}{\int_a^b p(x)P_m^2(x)dx}, \qquad m = 0, \ldots, n-1.$$

It is clear that $P_n(x)$ with these coefficients has the desired property.

The orthogonal polynomials $P_0(z)$, $P_1(z)$, . . . are linearly independent.

In fact, if

$$\lambda_0 P_0(x) + \ldots + \lambda_n P_n(x) = 0$$

identically, with certain constants $\lambda_0, \ldots, \lambda_n$, then by multiplying by $p(x)p_m(x)$ and integrating between a and b we get

$$\lambda_m \int_a^b P(x)P_m^2(x)dx = 0, \qquad m = 0, \ldots, n,$$

whence $\lambda_m = 0$.

In quite the same way as in section 3.14.3 we may prove that z^n is uniquely expressible as a linear combination of the polynomials $P_0(z), \ldots, P_n(z)$. Hence

Every polynomial $\varphi(z)$ with real coefficients and degree $\leqq n-1$ has the unique representation

$$\varphi(z) = c_0 P_0(z) + \ldots + c_{n-1} P_{n-1}(z), \tag{16.8–24}$$

where $c_0, \ldots, c_{n-1}$ are constants.

From the orthogonality follows

$$\int_a^b p(x)\varphi(x)P_n(x)dx = \sum_{\nu=0}^{n-1} c_\nu \int_a^b p(x)P_\nu(x)P_n(x)dx = 0.$$

Thus:

An arbitrary real polynomial is orthogonal to each polynomial of the sequence $P_0(z)$, $P_1(z)$, . . . of higher degree.

A very useful theorem is the following uniqueness theorem

The polynomials of a sequence orthogonal with respect to a given weight function are determined up to a multiplicative constant.

We proceed by induction. The assertion is trivial for $n = 0$, for a polynomial of zero degree is a (non-vanishing) constant. Suppose the statement is true until the degree $n-1$. Now $P_n(x)$ has degree n and is expressible as

$$P_n(x) = \sum_{\nu=0}^{n-1} c_\nu P_\nu(x) + a_n x^n.$$

Multiplying by $P_m(x)$, $m = 0, \ldots, n-1$ and integrating between a and b we get

$$0 = \int_a^b p(x)P_m(x)P_n(x)dx = c_m \int_a^b P_m^2(x)dx + a_n \int_a^b p(x)P_m(x)x^n\,dx,$$

whence

$$c_m = -a_n \frac{\int_a^b p(x)P_m(x)x^n\,dx}{\int_a^b p(x)P_m^2(x)dx}, \qquad m = 0, \ldots, n-1.$$

Hence the constants c_m are determined by $P_0(x), \ldots, P_{n-1}(x)$ up to the multiplicative constant a_n.

A remarkable property of the zeros of a polynomial belonging to an orthogonal set is the assertion

The zeros of the polynomial of an orthogonal set are all real and simple and lie in the interval $a < x < b$.

We denote by $a_1, \ldots, a_m$ the zeros of $P_n(x)$ in the interval $a < x < b$, where $P_n(x)$ changes sign if x increases from a to b. Let

$$Q(x) = \begin{cases} 1, & \text{if} \quad m = 0, \\ (x-a_1)\ldots(x-a_m), & \text{if} \quad m > 0. \end{cases}$$

This function changes sign at the same points as does $P_n(x)$. Hence $P_n(x)Q(x)$ does not change sign at any point in the interval, whence

$$\int_a^b p(x)Q(x)P_n(x)dx > 0.$$

By the fourth theorem of this section $Q(x)$ has the degree at least n. But the number of zeros of $P_n(x)$ does not exceed n and, as a consequence, $m = n$.

16.8.7 – The formula of Rodrigues

By a *function of Rodrigues* related to the weight function $p(z)$ over an interval $a < z < b$ we shall understand the function

$$\boxed{R_n(z) = \frac{1}{p(z)} \frac{d^n}{dz^n} (p(z)q^n(z)),} \qquad (16.8\text{–}25)$$

where $q(z)$ is a quadratic polynomial with zeros a and b. Either or both endpoints of the interval $a < x < b$ may be taken to be infinite. If one is infinite then $q(z)$ is a polynomial of the first degree and if both are infinite $q(z)$ is a (non-vanishing) constant.

In general $R_n(z)$ is not a polynomial. It is our aim to investigate for which weight functions $p(z)$ the expression (16.8–25) does represent a polynomial of precisely the degree n.

i) Assume that a and b are finite. Let

$$q(z) = (z-a)(z-b).$$

Without loss of generality we may take $a = -1$, $b = 1$, for these values can be obtained from given values a and b by a suitable linear transformation of the variable z. Hence we continue with

$$q(z) = 1-z^2. \tag{16.8–26}$$

A necessary condition is expressed by the fact that

$$\frac{1}{p(z)}\frac{d}{dz}(1-z^2)p(z) = \frac{p'(z)}{p(z)}(1-z^2)-2z$$

is a linear polynomial. Hence

$$\frac{p'(z)}{p(z)} = \frac{Az+B}{1-z^2} = \frac{-\alpha}{1-z} + \frac{\beta}{1+z}.$$

Integration yields

$$p(z) = c(1-z)^{\alpha}(1+z)^{\beta}. \tag{16.8–27}$$

This is, apart from an unimportant multiplicative constant, the weight function for the Jacobian polynomials. It is easy to verify that the conditions imposed on $p(z)$ is also sufficient if we assume that $\alpha > -1$, $\beta > -1$. In fact, by Leibniz's rule the function

$$\begin{aligned}
R_n(z) &= (1-z)^{-\alpha}(1+z)^{-\beta}\frac{d^n}{dz^n}((1-z)^{n+\alpha}(1+z)^{n+\beta}) \\
&= (1-z)^{-\alpha}(1+z)^{-\beta}\sum_{\nu=0}^{n}(-1)^{n-\nu}\binom{n}{\nu}\frac{d^{n-\nu}}{dz^{n-\nu}}(1-z)^{n+\alpha}\frac{d^{\nu}}{dz^{\nu}}(1+z)^{n+\beta} \\
&= n!(1-z)^{-\alpha}(1+z)^{-\beta}\sum_{\nu=0}^{n}(-1)^{n-\nu}\binom{n+\alpha}{\nu}\binom{n+\beta}{n-\nu}(1-z)^{\nu+\alpha}(1+z)^{n-\nu+\beta} \\
&= n!\sum_{\nu=0}^{n}(-1)^{n-\nu}\binom{n+\alpha}{\nu}\binom{n+\beta}{n-\nu}(1-z)^{\nu}(1+z)^{n-\nu} \\
&= (-1)^n n!\sum_{\nu=0}^{n}\binom{n+\alpha}{\nu}\binom{n+\beta}{n-\nu}(z-1)^{\nu}(z+1)^{n-\nu}
\end{aligned}$$

is a polynomial of degree n. The coefficient of z^n is $(-1)^n n!$ times the number

$$\sum_{\nu=0}^{n}\binom{n+\alpha}{\nu}\binom{n+\beta}{n-\nu}$$

and for $n > 0$ each term in this sum is positive because, by hypothesis, $\alpha > -1, \beta > -1$.

ii) Let $b = \infty$. Without loss of generality we may suppose that $a = 0$, i.e.,

$$q(z) = z. \tag{16.8–28}$$

A necessary condition for $p(z)$ is expressed by

$$\frac{p'(z)}{p(z)} = \frac{Az+B}{z} = A + \frac{B}{z}.$$

Integration yields

$$p(z) = Ce^{Az}z^{\alpha}. \tag{16.8–29}$$

Under the assumption $\alpha > -1$ this condition is also sufficient, for

$$R_n(z) = e^{-Az}z^{-\alpha}\frac{d^n}{dz^n}(e^{Az}z^{\alpha+n}) = e^{-Az}z^{-\alpha}\sum_{\nu=0}^{n}\binom{n}{\nu}\frac{d^{n-\nu}}{dz^{n-\nu}}e^{Az}\frac{d^\nu}{dz^\nu}z^{\alpha+n}$$

$$= n!e^{-Az}z^{-\alpha}\sum_{\nu=0}^{n}\frac{A^{n-\nu}}{(n-\nu)!}e^{Az}\binom{n+\alpha}{\nu}z^{\alpha+n-\nu} = n!\sum_{\nu=0}^{n}\frac{A^{n-\nu}}{(n-\nu)!}\binom{n+\alpha}{\nu}z^{n-\nu},$$

and this is again a polynomial of exactly the degree n if $\alpha > -1$.

iii) There remains the case $a = -\infty$, $b = \infty$. We take

$$q(z) = 1. \tag{16.8–30}$$

It follows that a necessary condition for $p(z)$ is expressed by

$$\frac{p'(z)}{p(z)} = Az+B,$$

whence

$$p(z) = \exp\left(\tfrac{1}{2}Az^2+Bz+C\right),$$

the multiplicative constant being absorbed in e^C. By a suitable linear transformation we can give to $p(z)$ the form

$$p(z) = \exp\left(\tfrac{1}{2}Az^2+C\right). \tag{16.8–31}$$

The condition is also sufficient if we assume that $A \neq 0$. This may be proved by induction. First $R_0(z) = 1$ is a polynomial of degree zero. If we assume that $R_{n-1}(z)$ is a polynomial $P_{n-1}(z)$ of degree $n-1$, then

$$\frac{d^{n-1}}{dz^{n-1}}\exp\left(\tfrac{1}{2}Az^2\right) = P_{n-1}(z)\exp\left(\tfrac{1}{2}Az^2\right),$$

whence

$$\frac{d^n}{dz^n}\exp\left(\tfrac{1}{2}Az^2\right) = P'_{n-1}(z)\exp\left(\tfrac{1}{2}Az^2\right)+AzP_{n-1}(z).$$

The expression $P'_{n-1}(z) + AzP_{n-1}(z)$ is a polynomial of exactly the degree n.

The orthogonality of the polynomials with respect to the weight function requires that the integral (16.8–27) exists. This imposes on the function (16.8–29) (with the limits of integration $a = 0$, $b = \infty$) the conditions that A is negative. Without loss of generality we may take $A = -1$. For the function (16.8–30) (with the limits of integration $a = -\infty$, $b = \infty$) the condition $A < 0$ must be fulfilled; without loss of generality we may take $A = -2$. Thus the three possible forms for $p(z)$ are essentially

$$p(z) = \begin{cases} (1-z)^{\alpha}(1+z)^{\beta}, & \alpha > -1, \beta > -1, \\ e^{-z}z^{\alpha}, & \alpha > -1 \\ e^{-z^2}. \end{cases} \tag{16.8–32}$$

The first two functions are weight function for the Jacobian and the Laguerre polynomials respectively. The third function is new and belongs to the so-called Hermite polynomials which constitute the subject-matter of the next section. It should be noticed that we have restricted ourselves to polynomials with real coefficients.

16.8.8 – Polynomials of Hermite

By the polynomials of Hermite we understand the polynomials defined by the Rodrigues formula

$$H_n(z) = e^{z^2} \frac{d^n}{dz^n} e^{-z^2}. \tag{16.8–33}$$

Since

$$\frac{d^n}{dz^n} e^{-z^2} = (-2)^n z^n e^{-z^2} + \dots,$$

the leading coefficient in $H_n(z)$ has the value

$$k_n = (-1)^n 2^n. \tag{16.8–34}$$

Many authors, in particular theoretical physicists, prefer to adopt the expressions $(-1)^n H_n(z)$ as Hermite polynomials. This causes only minor modifications in the formulas.

Proceeding along well-known lines we start with the function

$$u(z) = e^{-z^2} \tag{16.8–35}$$

which has the property

$$u'(z) = -2zu(z). \tag{16.8–36}$$

Differentiating both members n times we have by Leibniz's rule

$$u^{(n+1)}(z)+2zu^{(n)}(z)+2nu^{(n-1)}(z) = 0$$

and this leads us to the recurrent relation

$$\boxed{H_{n+1}(z)+2zH_n(z)+2nH_{n-1}(z) = 0.} \tag{16.8–37}$$

The first few Hermite polynomials are

$$H_0(z) = 1, \qquad H_1(z) = -2z, \qquad H_2(z) = 4z^2-2,$$
$$H_3(z) = -8z^3+12z, \qquad H_4(z) = 16z^4-48z^2+12, \ldots .$$

A generating function of the type

$$F(w) = \sum_{\nu=0}^{\infty} \frac{H_\nu(z)}{\nu!} w^\nu$$

can be obtained by the aid of (16.8–37). Differentiating with respect to w gives

$$F'(w) = \sum_{\nu=0}^{\infty} \frac{H_{\nu+1}(z)}{\nu!} w^\nu = -2z\sum_{\nu=0}^{\infty} \frac{H_\nu(z)}{\nu!} w^\nu - 2\sum_{\nu=1}^{\infty} \frac{H_{\nu-1}(z)}{(\nu-1)!} w^\nu$$
$$= -2z\sum_{\nu=0}^{\infty} \frac{H_\nu(z)}{\nu!} w^\nu - 2w\sum_{\nu=0}^{\infty} \frac{H_\nu(z)}{\nu!} w^\nu,$$

whence

$$F'(w) = -2(z+w)F(w). \tag{16.8–38}$$

Integration yields

$$F(w) = C \exp(-2zw-w^2),$$

the constant C being equal to unity, because $H_0(z) = 1$. Thus we proved

The n-th Hermite polynominal is n! times the coefficien' of w^n in the expansion of the generating function

$$\boxed{\exp(-2zw-w^n) = \sum_{\nu=0}^{\infty} \frac{H_\nu(z)}{\nu!} w^\nu.} \tag{16.8–39}$$

This formula contains much information about the Hermite polynomials and is frequently taken as a definition of these polynomials.

Replacing z by $-z$ and at the same time w by $-w$ we immediately find

$$\boxed{H_n(-z) = (-1)^n H_n(z).} \tag{16.8–40}$$

This equation expresses the fact that $H_n(z)$ contains only even or odd powers of z according as n is even or odd.

Differentiation both members of (16.8–39) with respect to z yields

$$-2w\sum_{\nu=0}^{\infty}\frac{H_\nu(z)}{\nu!}w^\nu = \sum_{\nu=0}^{\infty}\frac{H'_\nu(z)}{\nu!}w^\nu,$$

whence, on equating coefficients of equal powers of w

$$\boxed{H'_n(z) = -2nH_{n-1}(z).} \qquad (16.8\text{–}41)$$

Differentiating both members of (16.8–38) with respect to z and taking into account (16.8–41) leads to

$$-2(n+1)H_n(z)+2H_n(z)+2zH'_n(z)+2nH'_{n-1}(z) = 0,$$

or

$$2nH_n(z)-2zH'_n(z)-2nH'_{n-1}(z) = 0.$$

Again by (16.8–41)

$$H''_n(z) = -2nH'_{n-1}(z)$$

and it follows

The polynomial of Hermite $H_n(z)$ is a solution of the differential equation

$$\boxed{w''-2zw'+2nw = 0.} \qquad (16.8\text{–}42)$$

It is the only polynomial solution of degree n of this equation (apart from a multiplicative constant).

In order to prove this we proceed as follows. Let

$$H_n(z) = \sum_{\nu=0}^{n} a_\nu z^{n-\nu}$$

be a solution. Then, by inserting this into the equation we obtain the recursive relations

$$2ka_k+(n-k+2)(n-k+1)s_{k-2} = 0,$$

whence

$$a_k = -\frac{(n-k+2)(n-k+1)}{2k}a_{k-2}.$$

Taking $a_0 = (-1)^n 2^n$ we get

$$a_{2k} = \frac{n(n-1)\ldots(n-2k+1)}{k!}(-1)^{n-k}2^{n-2k} = \frac{(-1)^n n!}{(n-2k)!k!}(-1)^k 2^{n-2k},$$

the coefficients of odd index being zero since the differential equation

(16.8–42) remains unchanged if we replace z by $-z$. The coefficients of the polynomial solution are uniquely determined by a_0 and this proves the assertion. In addition we now have an explicit expression for the polynomial, viz.

$$H_n(z) = (-1)^n \sum_{\nu=0}^{[\frac{1}{2}n]} (-1)^\nu \frac{n!}{\nu!(n-2\nu)!} (2z)^{n-2\nu}. \tag{16.8–43}$$

The same result is obtained by expanding

$$\exp(-2zw-w^2) = \exp(-2zw)\exp(-w^2),$$

viz.,

$$\sum_{\nu=0}^{\infty} \frac{H_\nu(z)}{\nu!} w^\nu = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{(2z)^\nu}{\nu!} w^\nu \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{\nu!} w^{2\nu}$$

and equating coefficients of like powers of w.

Another interesting result may be obtained by means of the generating function by expanding

$$\exp(-2uw-w^2)\exp(-2vw-w^2) = \exp\left(-2\frac{u+v}{\sqrt{2}} w\sqrt{2} - (w\sqrt{2})^2\right).$$

We find

$$\sum_{\nu=0}^{\infty} \frac{H_\nu(u)}{\nu!} w^\nu \sum_{\nu=0}^{\infty} \frac{H_\nu(v)}{\nu!} w^\nu = \sum_{\nu=0}^{\infty} \frac{2^{\frac{1}{2}\nu}}{\nu!} H_\nu\left(\frac{u+v}{\sqrt{2}}\right) w^\nu.$$

We now immediately have *Runge's addition formula*

$$2^{\frac{1}{2}n} H_n\left(\frac{u+v}{\sqrt{2}}\right) = \sum_{\nu=0}^{n} \binom{n}{\nu} H_\nu(u) H_{n-\nu}(v). \tag{16.8–44}$$

The relations of orthogonality may be obtained in the usual way. If $u(x)$ is again the function (16.8–35) we have

$$\int_{-\infty}^{\infty} x^m u^{(n)}(x)dx = \int_{-\infty}^{\infty} x^m du^{(n-1)}(x) = -m\int_{-\infty}^{\infty} x^{m-1}u^{(n-1)}(x)dx$$

$$= \ldots = (-1)^m m! \int_{-\infty}^{\infty} u^{(n-m)}(x)dx$$

and this is zero if $m < n$. In the case $m = n$ we have

$$\int_{-\infty}^{\infty} x^n u^{(n)}(x)dx = (-1)^n n! \int_{-\infty}^{\infty} e^{-x^2} dx = (-1)^n n! \sqrt{\pi},$$

by (4.7–5). Because the coefficient k_n of x^n in $H_n(x)$ is $(-1)^n 2^n$, we may state

The relation of orthogonality of the Hermite polynomials are

$$\boxed{\int_{-\infty}^{\infty} e^{-x^2} H_m(x) H_n(x) dx = \begin{cases} 0, & \text{if} \quad m = n, \\ 2^n n! \sqrt{\pi}, & \text{if} \quad m = n. \end{cases}} \tag{16.8–45}$$

16.8.9 – Relations between the polynomials of Hermite and Laguerre

The Hermite polynomials belong to the class of the Laguerre polynomials, inasmuch a polynomial of Hermite can be expressed as a certain Laguerre polynomial with appropriately transformed variable. This may be found by considering the solutions of Weber's equation (16.7–31). The form of these solutions suggests the substitution

$$u = v e^{-\frac{1}{4}z^2}$$

and this leads to the differential equation

$$v'' - zv' + (2\kappa - \tfrac{1}{2})v = 0. \tag{16.8–46}$$

A solution of this equation is the even function

$$\Phi(\tfrac{1}{4} - \kappa; \tfrac{1}{2}; \tfrac{1}{2}z^2). \tag{16.8–47}$$

If we replace the variable z by $z\sqrt{2}$ the equation (16.8–46) takes the form

$$v'' - 2zv' + 2(2\kappa - \tfrac{1}{2})v = 0 \tag{16.8–48}$$

and comparing this with (16.8–42) we may conclude that $H_{2m}(z)$ is a solution of this equation, provided that $2\kappa - \frac{1}{2} = 2m$, i.e., $\frac{1}{4} - \kappa = -m$. It is clear that there exists a relation

$$H_{2m}(z) = H_{2m}(0)\Phi(-m; \tfrac{1}{2}; z^2). \tag{16.8–49}$$

On the other hand, by virtue of (16.8–1)

$$L_m^{(-\frac{1}{2})}(z^2) = \binom{m - \frac{1}{2}}{m} \Phi(-m; \tfrac{1}{2}; z^2). \tag{16.8–50}$$

The constant $H_{2m}(0)$ occurring in (16.8–49) may be evaluated by the aid of (16.8–43). The result is

$$H_{2m}(0) = (-1)^m \frac{(2m)!}{m!}. \tag{16.8–51}$$

From Legendre's duplication formula (4.6–26) we get

$$\Gamma(m + \tfrac{1}{2}) = \frac{(2m)! \sqrt{\pi}}{2^{2m} m!}, \tag{16.8–52}$$

whence

$$\binom{m-\frac{1}{2}}{m} = \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m+1)\Gamma(\frac{1}{2})} = \frac{(2m)!}{2^{2m}(m!)^2}. \tag{16.8–53}$$

This formula may be obtained by straightforward computation from the definition (2.16–19) of the binomial coefficient, of course. Thus we find

$$\boxed{H_{2m}(z) = (-1)^m 2^{2m} m!\, L_m^{(-\frac{1}{2})}(z^2).} \tag{16.8–54}$$

By shifting m to $m+1$ and differentiating both sides of the equation thus obtained we have, by virtue of (16.8–41) and (16.8–16)

$$\boxed{H_{2m+1}(z) = (-1)^{m+1} 2^{2m+1} m!\, z L_m^{(\frac{1}{2})}(z^2).} \tag{16.8–55}$$

An integral expression of the Laguerre polynomials in terms of the Hermite polynomials is readily found by the aid of an integral relation between $L_n^{(\alpha)}(z)$ and $L_n^{(\beta)}(z)$. The expansion (16.8–2) may be put in the form

$$L_n^{(\alpha)}(z) = \frac{\Gamma(1+\alpha+n)}{n!} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{1}{\Gamma(1+\alpha+\nu)} (-1)^\nu z^\nu.$$

Assuming $\alpha > \beta$ this is equivalent to

$$\begin{aligned} L_n^{(\alpha)}(z) &= \frac{\Gamma(1+\alpha+n)}{n!\Gamma(\alpha-\beta)} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{1}{\Gamma(1+\beta+\nu)} (-1)^\nu z^\nu \frac{\Gamma(1+\beta+\nu)\Gamma(\alpha-\beta)}{\Gamma(1+\alpha+\nu)} \\ &= \frac{\Gamma(1+\alpha+n)}{n!\Gamma(\alpha-\beta)} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{(-1)^\nu z^\nu}{\Gamma(1+\beta+\nu)} \int_0^1 t^{\beta+\nu}(1-t)^{\alpha-\beta-1}\, dt \\ &= \frac{\Gamma(1+\alpha+n)}{n!\Gamma(\alpha-\beta)} \int_0^1 t^\beta (1-t)^{\alpha-\beta-1} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{(-1)^\nu z^\nu t^\nu}{\Gamma(1+\beta+\nu)}\, dt \\ &= \frac{\Gamma(1+\alpha+n)}{n!\Gamma(\alpha-\beta)} z^{-\alpha} \int_0^z u^\beta (z-u)^{\alpha-\beta-1} \sum_{\nu=0}^{n} \binom{n}{\nu} \frac{(-1)^\nu u^\nu}{\Gamma(1+\beta+\nu)}\, du \end{aligned}$$

and this yields

$$\boxed{L_n^{(\alpha)}(z) = \frac{\Gamma(1+\alpha+n)}{\Gamma(1+\beta+n)} \frac{z^{-\alpha}}{\Gamma(\alpha-\beta)} \int_0^z u^\beta (z-u)^{\alpha-\beta-1} L_n^{(\beta)}(u)\, du} \tag{16.8–56}$$

provided that $\alpha > \beta$.

In this formula we replace z by z^2 and then u by t^2. If we take $\beta = -\frac{1}{2}$ then, assuming $\alpha > -\frac{1}{2}$,

$$L_n^{(\alpha)}(z^2) = \frac{2\Gamma(1+\alpha+n)}{\Gamma(n+\frac{1}{2})\Gamma(\alpha+\frac{1}{2})} z^{-2\alpha} \int_0^z (z^2-t^2)^{\alpha-\frac{1}{2}} L_n^{(-\frac{1}{2})}(t^2)\,dt \qquad (16.8\text{–}57)$$

or, by virtue of (16.8–54),

$$\boxed{2^{2n} n!(-1)^n B(n+\tfrac{1}{2}, \alpha+\tfrac{1}{2}) z^{2\alpha} L_n^{(\alpha)}(z) = 2\int_0^z (z^2-t^2)^{\alpha-\frac{1}{2}} H_{2n}(t)\,dt,}$$

the desired formula.

INDEX