John Stillwell

Classical Topology and Combinatorial Group Theory

Second Edition

Illustrated with 312 Figures by the Author

Springer

John Stillwell
Department of Mathematics
Monash University
Clayton (Victoria 3168)
Australia

Mathematics Subject Classifications (1991): 55-01, 51-01, 57-01

Library of Congress Cataloging-in-Publication Data
Stillwell, John.
Classical topology and combinatorial group theory / John Stillwell.—2nd ed.
p. cm.—(Graduate texts in mathematics; 72)
Includes bibliographical references and index.
ISBN 0-387-97970-0
1. Topology. 2. Combinatorial group theory. I. Title. II. Series.
QA611.S84 1993
514—dc20 92-40606

Printed on acid-free paper.

Production managed by Ellen Seham; manufacturing supervised by Vincent Scelta.
Composition by Asco Trade Typesetting Ltd., Hong Kong.
Printed and bound by R.R. Donnelley & Sons Co., Harrisonburg, VA.
Printed in the United States of America.

9 8 7 6 5 4 3

ISBN 0-387-97970-0 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-97970-0 Springer-Verlag Berlin Heidelberg New York SPIN 10664636

To my mother
and father

Preface to the First Edition

In recent years, many students have been introduced to topology in high school mathematics. Having met the Möbius band, the seven bridges of Königsberg, Euler's polyhedron formula, and knots, the student is led to expect that these picturesque ideas will come to full flower in university topology courses. What a disappointment "undergraduate topology" proves to be! In most institutions it is either a service course for analysts, on abstract spaces, or else an introduction to homological algebra in which the only geometric activity is the completion of commutative diagrams. Pictures are kept to a minimum, and at the end the student still does not understand the simplest topological facts, such as the reason why knots exist.

In my opinion, a well-balanced introduction to topology should stress its intuitive geometric aspect, while admitting the legitimate interest that analysts and algebraists have in the subject. At any rate, this is the aim of the present book. In support of this view, I have followed the historical development where practicable, since it clearly shows the influence of geometric thought at all stages. This is *not* to claim that topology received its main impetus from geometric recreations like the seven bridges; rather, it resulted from the *visualization* of problems from other parts of mathematics—complex analysis (Riemann), mechanics (Poincaré), and group theory (Dehn). It is these connections to other parts of mathematics which make topology an important as well as a beautiful subject.

Another outcome of the historical approach is that one learns that classical (prior to 1914) ideas are still alive, and still being worked out. In fact, many simply stated problems in 2 and 3 dimensions remain unsolved. The development of topology in directions of greater generality, complexity, and abstractness in recent decades has tended to obscure this fact.

Attention is restricted to dimensions ≤ 3 in this book for the following reasons.

(1) The subject matter is close to concrete, physical experience.
(2) There is ample scope for analytic, geometric, and algebraic ideas.
(3) A variety of interesting problems can be constructively solved.
(4) Some equally interesting problems are still open.
(5) The combinatorial viewpoint is known to be completely general.

The significance of (5) is the following. Topology is ostensibly the study of arbitrary continuous functions. In reality, however, we can comprehend and manipulate only functions which relate finite "chunks" of space in a simple combinatorial manner, and topology originally developed on this basis. It turns out that for figures built from such chunks (simplexes) of dimension ≤ 3, the combinatorial relationships reflect all relationships which are topologically possible. Continuity is therefore a concept which can (and perhaps should) be eliminated, though of course some hard foundational work is required to achieve this.

I have not taken the purely combinatorial route in this book, since it would be difficult to improve on Reidemeister's classic *Einführung in die Kombinatorische Topologie* (1932), and in any case the relationship between the continuous and the discrete is extremely interesting. I have chosen the middle course of placing one combinatorial concept—the fundamental group—on a rigorous foundation, and using others such as the Euler characteristic only descriptively. Experts will note that this means abandoning most of homology theory, but this is easily justified by the saving of space and the relative uselessness of homology theory in dimensions ≤ 3. (Furthermore, textbooks on homology theory are already plentiful, compared with those on the fundamental group.)

Another reason for the emphasis on the fundamental group is that it is a two-way street between topology and algebra. Not only does group theory help to solve topological problems, but topology is of genuine help in group theory. This has to do with the fact that there is an underlying computational basis to both combinatorial topology and combinatorial group theory. The details are too intricate to be presented in this book, but the relevance of computation can be grasped by looking at topological problems from an algorithmic point of view. This was a key concern of early topologists and in recent times we have learned of the *nonexistence* of algorithms for certain topological problems, so it seems timely for a topology text to present what is known in this department.

The book has developed from a one-semester course given to fourth year students at Monash University, expanded to two-semester length. A purely combinatorial course in surface topology and group theory, similar to the one I originally gave, can be extracted from Chapters 1 and 2 and Sections 4.3, 5.2, 5.3, and 6.1. It would then be perfectly reasonable to spend a second semester deepening the foundations with Chapters 0 and 3 and going on to 3-manifolds in Chapters 6, 7, and 8. Certainly the reader is not obliged to master Chapter 0 before reading the rest of the book. Rather, it should be skimmed once and then referred to when needed later. Students who have had a conventional first course in topology may not need 0.1–0.3 at all.

The only prerequisites are some familiarity with elementary set theory, coordinate geometry and linear algebra, ε-δ arguments as in rigorous calculus, and the group concept.

The text has been divided into numbered sections which are small enough, it is hoped, to be easily digestible. This has also made it possible to dispense with some of the ceremony which usually surrounds definitions, theorems, and proofs. Definitions are signalled simply by italicizing the terms being defined, and they and proofs are not numbered, since the section number will serve to locate them and the section title indicates their content. Unless a result already has a name (for example, the Seifert-Van Kampen theorem) I have not given it one, but have just stated it and followed with the proof, which ends with the symbol □.

Because of the emphasis on historical development, there are frequent citations of both author and date, in the form: Poincaré 1904. Since either the author or the date may be operative in the sentence, the result is sometimes grammatically curious, but I hope the reader will excuse this in the interests of brevity. The frequency of citations is also the result of trying to give credit where credit is due, which I believe is just as appropriate in a textbook as in a research paper. Among the references which I would recommend as parallel or subsequent reading are Giblin 1977 (homology theory for surfaces), Moise 1977 (foundations for combinatorial 2- and 3-manifold theory), and Rolfsen 1976 (knot theory and 3-manifolds).

Exercises have been inserted in most sections, rather than being collected at the ends of chapters, in the hope that the reader will do an exercise more readily while his mind is still on the right track. If this is not sufficient prodding, some of the results from exercises are used in proofs.

The text has been improved by the remarks of my students and from suggestions by Wilhelm Magnus and Raymond Lickorish, who read parts of earlier drafts and pointed out errors. I hope that few errors remain, but any that do are certainly my fault. I am also indebted to Anne-Marie Vandenberg for outstanding typing and layout of the original manuscript.

October 1980 JOHN C. STILLWELL

Preface to the Second Edition

There have been several big developments in topology since the first edition of this book. Most of them are too difficult to include here, or else, well written up elsewhere, so I shall merely mention below what they are and where they may be found. The main new inclusion in this edition is a proof of the unsolvability of the word problem for groups, and some of its consequences. This is made possible by a new approach to the word problem discovered by Cohen and Aanderaa around 1980. Their approach makes it feasible to prove a series of unsolvability results we previously mentioned without proof, and thus to tie up several loose ends in the first edition. A new Chapter 9 has been added to incorporate these results. It is particularly pleasing to be able to give a proof of the unsolvability of the homeomorphism problem, which has not previously appeared in a textbook.

What are the other big developments? They would have to include the proof by Freedman in 1982 of the 4-dimensional Poincaré conjecture, and the related work of Donaldson on 4-manifolds. These difficult results may be found in Freedman and Quinn's *The Topology of 4-manifolds* (Princeton University Press, 1990) and Donaldson and Kronheimer's *The Geometry of Four-Manifolds* (Oxford University Press, 1990). With Freedman's proof, only the original (3-dimensional) Poincaré conjecture remains open. In fact, the main problems of 3-dimensional topology seem to be just as stubborn as they were in 1980. There is still no algorithm for deciding when 3-manifolds are homeomorphic, or even for recognizing the 3-sphere. Since the first printing of the second edition, the latter problem has been solved by Hyam Rubinstein. However, there has been important progress in knot theory, most of which stems from the *Jones polynomial*, a new knot invariant found by Jones in 1983. For a sampling of this rapidly growing field, and its mysterious connections with physics, see Kauffman's *Knots and Physics* (World Scientific, 1991).

Recent developments in combinatorial group theory are a natural continuation of two themes in the present book—the tree structure behind free groups and the tessellation structure behind Dehn's algorithm. The main results on tree structure and its generalizations may be found in Dicks and Dunwoody's *Groups Acting on Graphs* (Cambridge University Press, 1989). Dehn's algorithm has been generalized to many other groups which act on tessellations with combinatorial properties like those discovered by Dehn in the hyperbolic plane (see *Group Theory from a Geometrical Viewpoint*, edited by Ghys, Haefliger and Verjovsky, World Scientific, 1991). Both these lines of research should be accessible to readers of the present book, though a little more preparation is advisable. I recommend Serre's *Trees* (Springer-Verlag, 1980) and Dehn's *Papers in Group Theory and Topology* (Springer-Verlag, 1987). My own *Geometry of Surfaces* (Springer-Verlag, 1992) may also serve as a source for hyperbolic geometry, and as a replacement for the very sketchy account of geometric methods given in 6.2 below.

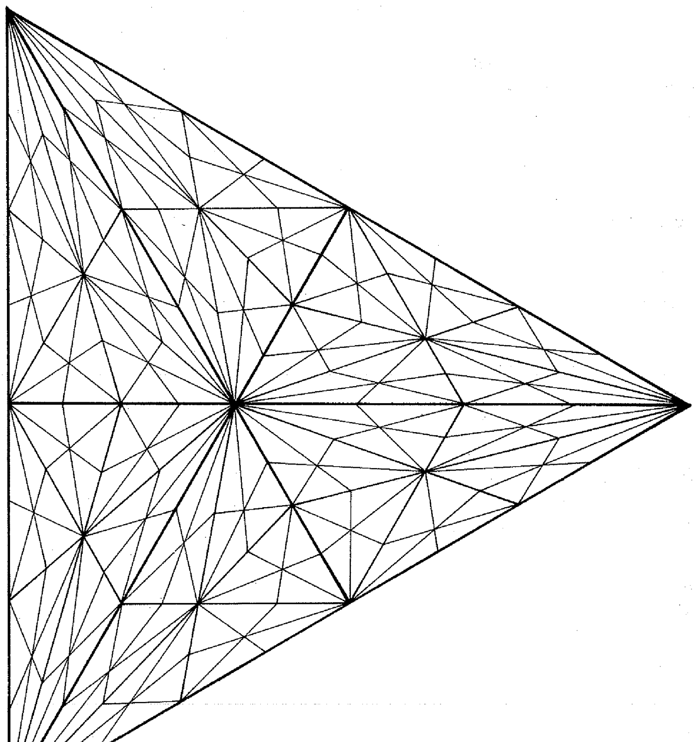
Finally, I should mention that this edition includes numerous corrections sent to me by readers. I am particularly grateful to Peter Landweber, who contributed the most thorough critique, as well as encouragement for a second edition.

Clayton, November 1992 JOHN C. STILLWELL

Contents

CHAPTER 0

Introduction and Foundations



0.1 The Fundamental Concepts and Problems of Topology

0.1.1 The Homeomorphism Problem

Topology is the branch of geometry which studies the properties of figures under arbitrary continuous transformations. Just as ordinary geometry considers two figures to be the same if each can be carried into the other by a rigid motion, topology considers two figures to be the same if each can be mapped onto the other by a one-to-one continuous function. Such figures are called topologically equivalent, or *homeomorphic*, and the problem of deciding whether two figures are homeomorphic is called the *homeomorphism problem.*

One may consider a geometric figure to be an arbitrary point set, and in fact the homeomorphism problem was first stated in this form, by Hurwitz 1897. However, this degree of generality makes the problem completely intractable, for reasons which belong more to set theory than geometry, namely the impossibility of describing or enumerating all point sets. To discuss the problem sensibly we abandon the elusive "arbitrary point set" and deal only with *finitely describable* figures, so that a solution to the homeomorphism problem can be regarded as an *algorithm* (0.4) which operates on descriptions and produces an answer to each homeomorphism question in a finite number of steps.

The most convenient building blocks for constructing figures are the simplest euclidean space elements in each dimension:

dimension 0: point
dimension 1: line segment
dimension 2: triangle
dimension 3: tetrahedron
⋮

We call the simplest space element in n-dimensional euclidean space R^n the *n-simplex* Δ^n. It is constructed by taking $n + 1$ points $P_1, \ldots, P_{n+1}$ in R^n which do not lie in the same $(n - 1)$-dimensional hyperplane, and forming their *convex hull*; that is, closing the set under the operation which fills in the line segment between any two points. In algebraic terms, we take $n + 1$ linearly independent vectors $\boldsymbol{OP}_1, \ldots, \boldsymbol{OP}_{n+1}$ (where $\boldsymbol{OP}_i$ denotes the vector from the origin O to P_i) and let Δ^n consist of the endpoints of the vectors

$$x_1\boldsymbol{OP}_1 + \cdots + x_{n+1}\boldsymbol{OP}_{n+1},$$

where $x_1 + \cdots + x_{n+1} = 1$ and $x_i \geq 0$. It is now an easy exercise (0.1.1.1 below) to show that any two n-simplexes are homeomorphic, so we are entitled to speak of *the* n-simplex Δ^n.

Each subset of $m + 1$ points from $\{P_1, \ldots, P_{n+1}\}$ similarly determines an *m-dimensional face* Δ^m of Δ^n. The union of the $(n - 1)$-dimensional faces is called the *boundary* of Δ^n, so all lower-dimensional faces lie in the boundary. We shall build figures, called *simplicial complexes*, by pasting together simplexes so that faces of a given dimension are either disjoint or coincide completely. This method of construction, which is due to Poincaré 1899, will be studied more thoroughly in 0.2. For the moment we wish to claim that all "natural" geometric figures are either simplicial complexes or homeomorphic to them, which is just as good for topological purposes.

This claim is supported by some figures which play a prominent role in this book—surfaces and knots. Surfaces may be constructed by pasting triangles together, so they are simplicial complexes of dimension 2. For example, the surface of a tetrahedron (which is homeomorphic to a sphere) is a simplicial complex of four triangles as shown in Figure 1. The torus surface (Figure 2) can be represented as a simplicial complex as shown in Figure 3. The representation is of course not unique, and from this one begins to see the *combinatorial* core of the homeomorphism problem, which remains after the point set difficulties have been set aside. Given a description of a surface as a list of triangles and their edges, how does one assess its *global* form? In particular, are the sphere and the torus topologically different? In fact we know how to solve this problem (by the classification theorem of 1.3, and 5.3.3), but not the corresponding 3-dimensional problem.

Much of the difficulty in dimension 3 is due to the existence of *knots*. We could define a knot to be any simple closed curve $\mathscr{K}$ in $\mathbb{R}^3$, but any such

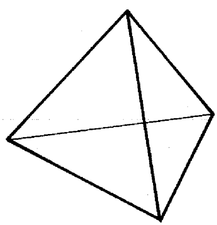


Figure 1

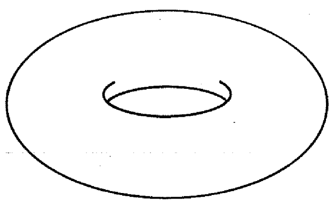


Figure 2

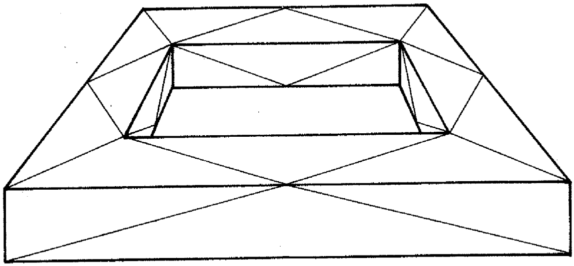


Figure 3

$\mathscr{K}$ is homeomorphic to a circle and its "knottedness" actually resides in the complement space $\mathsf{R}^3 - \mathscr{K}$. This space is not finitely describable in terms of simplexes, so we replace R^3 by, say, a cube and drill a thin tube out of it following the "knotted part" of $\mathscr{K}$ (see Figure 4).

This figure can be divided into small tetrahedra and hence is a finite simplicial complex representing the knot. The homeomorphism problem for such figures is extraordinarily difficult; Riemann was perhaps the first to think about it seriously (see Weil 1979), and it has been solved only recently (see Hemion 1979, Waldhausen 1978). The solution extends to more general "knot spaces" obtained by drilling any number of tubes out of cubes, but not as yet to all the figures which result from pasting knot spaces together.

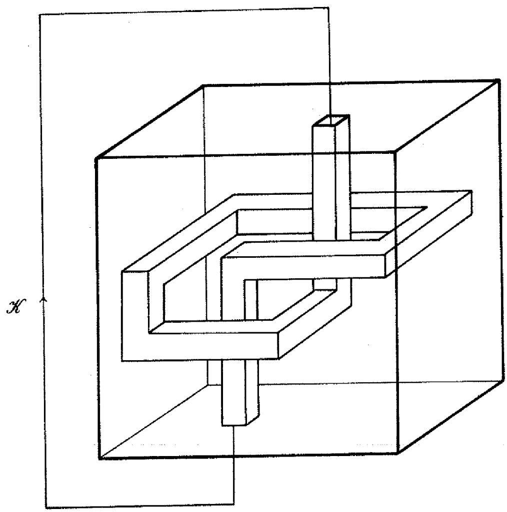


Figure 4

It seems very gratifying that the three dimensions provided by nature pose such a strong mathematical challenge. Moreover, it is known (Markov 1958) that the homeomorphism problem cannot be solved in dimensions ≥ 4, so we have every reason to concentrate our efforts in dimensions ≤ 3. This is the motivation for the present book. Our aim has been to give solutions to the main problems in dimension 2, and to select results in dimension 3 which illuminate the homeomorphism problem and seem likely to remain of interest if and when it is solved.

Like other fundamental problems in mathematics, the homeomorphism problem turns out not to be accessible directly, but requires various detours, some apparently technical and others of intrinsic interest. The first technical detour, which is typical, takes us away from the relation "is homeomorphic to" to the functions which relate homeomorphic figures. Thus we define a *homeomorphism* $f: \mathscr{A} \to \mathscr{B}$ to be a one-to-one continuous function with a continuous inverse $f^{-1}: \mathscr{B} \to \mathscr{A}$ (in particular, f is a bijection). Then to say $\mathscr{A}$ and $\mathscr{B}$ are homeomorphic is to say that there is a homeomorphism $f: \mathscr{A} \to \mathscr{B}$.

This point of view enables us to draw on general facts about continuous functions, which are reviewed in 0.1.2. We wish to avoid specific functions as far as possible, since topological properties by their nature do not reside in single functions so much as in classes of functions which are "qualitatively the same" in some sense. When we claim that there is a continuous function with particular qualitative features, it will always be straightforward to construct one by elementary means, such as piecing together finitely many linear functions. Readers should reassure themselves of this fact before proceeding too far, perhaps by working out explicit formulae for some of the examples in 0.1.3 (but not the "map of the Western Europe"!).

EXERCISE 0.1.1.1. Show that any two n-simplexes are homeomorphic.

EXERCISE 0.1.1.2. Construct a homeomorphism between the surface of a tetrahedron and the sphere.

0.1.2 Continuous Functions, Open and Closed Sets

The definition of a continuous function on R, the real line, is probably familiar. We shall phrase this definition so that it applies to any space $\mathscr{S}$ for which there is a distance function $|P - Q|$ defined for all points P, Q. If $\mathscr{S} = \mathsf{R}^n$, which is the most general case we shall ultimately need, and if

$$P = (x_1, \ldots, x_n),$$
$$Q = (y_1, \ldots, y_n),$$

we have

$$|P - Q| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

Then f is *continuous at* P if for each $\varepsilon > 0$ there is a δ such that

$$|P - Q| < \delta \quad \Rightarrow \quad |f(P) - f(Q)| < \varepsilon. \tag{*}$$

The function f is simply called *continuous* if it is continuous at each point P in its domain.

Informally, we say that a continuous function sends neighbouring points to neighbouring points. In fact, if we define the *ε-neighbourhood* of a point X to be

$$\mathcal{N}_\varepsilon(X) = \{Y \in \mathcal{S} : |X - Y| < \varepsilon\},$$

then (*) says that any neighbourhood of $f(P)$ has all sufficiently small neighbourhoods of P mapped into it by f. (An ε-neighbourhood of a point is often called a *ball* neighbourhood because this is the actual form of the above set in the "typical" space R^3. One can generalize $\mathcal{N}_\varepsilon$ to any figure in an obvious way. We later consider ε-neighbourhoods of curves, which are "strips" in R^2 and "tubes" in R^3, and ε-neighbourhoods of surfaces, which are "plates.")

A set $\mathcal{O} \subset \mathcal{S}$ in which each point X has an $\mathcal{N}_\varepsilon(X) \subset \mathcal{O}$ is called *open* (in $\mathcal{S}$). Thus any space $\mathcal{S}$ is an open subset of itself, and the empty set $\varnothing$ is open for the silly reason that it has no elements to contradict the definition. More important examples are open intervals $\{x \in \mathsf{R}: a < x < b\}$ in the line R, and cartesian products of them in higher dimensions (rectangles in R^2, "hyperrectangles" in R^n).

The complement $\mathcal{C} = \mathcal{S} - \mathcal{O}$ of an open set $\mathcal{O}$ is called *closed* (in $\mathcal{S}$). The key property of a closed set is that it contains all its *limit points*. X is a limit point of a set $\mathcal{D}$ if every $\mathcal{N}_\varepsilon(X)$ contains a point of $\mathcal{D}$ other than X itself. It is immediate that a limit point X of $\mathcal{C}$ cannot lie in the open set $\mathcal{S} - \mathcal{C}$. If X is a limit point of both $\mathcal{D}$ and $\mathcal{S} - \mathcal{D}$ then X is called a *frontier point* of $\mathcal{D}$ and $\mathcal{S} - \mathcal{D}$, and the set of frontier points is called the *frontier* (of $\mathcal{D}$ and $\mathcal{S} - \mathcal{D}$). For example, the frontier of an n-simplex Δ^n in R^n is its boundary, while the frontier of a Δ^m in R^n, $m < n$, is Δ^m itself.

For every set $\mathcal{A}$ there is a smallest closed set $\bar{\mathcal{A}}$ containing it, and called its *closure*, and a largest open set $\operatorname{int}(\mathcal{A})$ contained in it, and called its *interior*.

We now review some important properties of continuous functions, open sets, and closed sets.

(1) (Bolzano–Weierstrass theorem). *A closed set $\mathcal{C} \subset \mathsf{R}^n$ is bounded if and only if every infinite subset $\mathcal{D}$ of $\mathcal{C}$ has a limit point (in $\mathcal{C}$).*

If $\mathcal{C}$ is bounded, enclose it in a hyperrectangle and bisect repeatedly, each time choosing a half containing infinitely many points of $\mathcal{D}$. Doing this so that all edge lengths of the hyperrectangle $\to 0$ defines a point X which is a limit point of $\mathcal{D}$ by construction.

Conversely, if $\mathcal{C}$ is unbounded it contains a set $\mathcal{D} = \{P_i\}$ of points such that P_i is at distance ≥ 1 from $P_1, \ldots, P_{i-1}$ for each i, so $\mathcal{D}$ has no limit point. □

(2) *Two disjoint bounded closed sets* $\mathscr{C}_1, \mathscr{C}_2$ *have a non-zero distance* $d(\mathscr{C}_1, \mathscr{C}_2)$ *where*

$$d(\mathscr{C}_1, \mathscr{C}_2) = \inf\{|P_1 - P_2|: P_1 \in \mathscr{C}_1, P_2 \in \mathscr{C}_2\}$$

If $d(\mathscr{C}_1, \mathscr{C}_2) = 0$ choose $P_1^{(n)} \in \mathscr{C}_1$, $P_2^{(n)} \in \mathscr{C}_2$ for each n so that $|P_1^{(n)} - P_2^{(n)}| < 1/n$. If $\mathscr{C}_1, \mathscr{C}_2$ are disjoint this distance is always >0, hence the sets $\{P_1^{(i)}\}$ and $\{P_2^{(i)}\}$ are infinite and have limit points P_1, P_2 (by the Bolzano–Weierstrass Theorem) which are in $\mathscr{C}_1$, $\mathscr{C}_2$ respectively since the sets are closed. But then $|P_1 - P_2| > 0$, which contradicts the fact that P_1, P_2 are approached arbitrarily closely by $P_1^{(n)}$, $P_2^{(n)}$ which are arbitrarily close to each other. □

A bounded closed set in $\mathbf{R}^n$ is called *compact*. (By (1), an equivalent definition is that a compact set contains a limit point of each of its infinite subsets.) In many circumstances compact figures are equivalent to finite ones in the sense of 0.1.1, and this allows combinatorial arguments to be applied to rather general figures. Two propositions crucial to this "finitization" process are:

(3) *The continuous image of a compact set is compact.*

Let f be a function continuous on a compact set $\mathscr{C}$. By (1) it will suffice to show that every infinite $\mathscr{D} \subset f(\mathscr{C})$ has a limit point in $\mathscr{C}$. If not, there is an infinite set $\{f(X_i)\}$ of points in $f(\mathscr{C})$ with no limit point in $f(\mathscr{C})$. But $\{X_i\}$ has a limit point $X \in \mathscr{C}$ by (1), and every neighbourhood of $f(X)$ contains points $f(X_i)$ by the continuity of f, so $f(X)$ is a limit point of $\{f(X_i)\}$ and we have a contradiction. □

(4) *A continuous function f on a compact set* $\mathscr{C} \subset \mathbf{R}^n$ *is* uniformly continuous, *that is, for any* $\varepsilon > 0$ *there is a* $\delta > 0$ *such that*

$$|X - Y| < \delta \quad \Rightarrow \quad |f(X) - f(Y)| < \varepsilon$$

regardless of the choice of $X, Y \in \mathscr{C}$.

Suppose on the contrary that there is no such δ for some fixed ε. Then there are $X_1, X_2, \ldots \in \mathscr{C}$ such that $\mathscr{N}_\delta(X_n)$ does not map into $\mathscr{N}_\varepsilon(f(X_n))$ unless $\delta < 1/n$. Let $X \in \mathscr{C}$ be a limit point of $\{X_1, X_2, \ldots\}$, using (1). Since f is continuous there is a $\delta > 0$ such that $\mathscr{N}_\delta(X)$ maps into $\mathscr{N}_{\varepsilon/2}(f(X))$.

Now for n sufficiently large we have not only $X_n \in \mathscr{N}_\delta(X)$, but also $\mathscr{N}_{1/n}(X_n) \subset \mathscr{N}_\delta(X)$, since X_n approaches arbitrarily close to X. Thus $\mathscr{N}_{1/n}(X_n)$ maps into $\mathscr{N}_{\varepsilon/2}(f(X))$, and in particular $f(X_n) \in \mathscr{N}_{\varepsilon/2}(f(X))$. But then $\mathscr{N}_{\varepsilon/2}(f(X)) \subset \mathscr{N}_\varepsilon(f(X_n))$ and hence $\mathscr{N}_{1/n}(X_n)$ maps into $\mathscr{N}_\varepsilon(f(X_n))$, contrary to the choice of X_n. □

For example, a curve c is a continuous map of the compact interval $[0, 1]$, so by (4) we can divide $[0, 1]$ into a finite number of subintervals (of

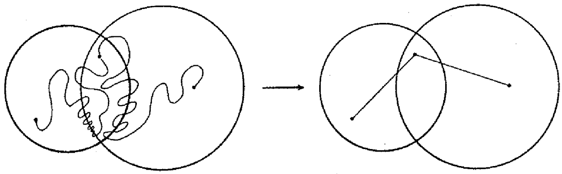


Figure 5

length $<\delta$) whose images (subarcs of c) lie in ε-neighbourhoods. If c lies in a figure with reasonable ε-neighbourhoods (say ε-balls, for ε sufficiently small), these subarcs can be deformed into line segments as in Figure 5. Thus c is equivalent to a polygonal curve, up to deformation. The notion of deformation required for this finitization process will be defined precisely in 0.1.9.

EXERCISE 0.1.2.1. If f is one-to-one consider the ordering of points on the curve $f(\mathscr{C})$ induced by the natural order on the line interval $\mathscr{C}$. Show that if $f(\mathscr{C})$ meets a closed set $\mathscr{K}$ then it has a *first* point of intersection with $\mathscr{K}$.

EXERCISE 0.1.2.2. The proofs of (1), (2), (3), (4) above use the Axiom of choice (where?). This can be avoided by giving an explicit rule for choosing a point $P(\mathscr{C})$ from a closed set $\mathscr{C} \subset \mathsf{R}^n$. Devise such a rule, starting in R^1.

EXERCISE 0.1.2.3. Construct a countable set of ball neighbourhoods in R^n, from which any open set is obtainable as the union of a subset. Deduce a rule for choosing a point from an open set.

EXERCISE 0.1.2.4. Show that a continuous one-to-one function on a bounded closed set has a continuous inverse (and hence is a homeomorphism).

EXERCISE 0.1.2.5. Show that an m-simplex is closed in any R^n, $n \geq m$.

EXERCISE 0.1.2.6. Show that $\overline{\mathscr{A}} = \mathscr{A} \cup \{\text{limit points of } \mathscr{A}\}$ and $\operatorname{int}(\mathscr{A}) = \mathscr{S} - \overline{(\mathscr{S} - \mathscr{A})}$.

EXERCISE 0.1.2.7 (intermediate-value theorem). If $f\colon [a, b] \to \mathsf{R}$ is continuous, prove that f takes every value between $f(a)$ and $f(b)$.

0.1.3 Examples of Continuous Maps

Although it is superfluous to introduce another name for functions, we often call them *maps*, to emphasize the idea of a function as an image-forming process. This is particularly appropriate in topology, which owes its existence to the fact that some visual information is preserved even by arbitrary homeomorphisms. Homeomorphisms, or *topological maps*, can be called "maps" with some justice, and we extend the usage by courtesy to other continuous functions (though the continuous function which sends everything to the same point is a poor sort of "map"!).

Interestingly, modern geography has expanded its concept of "map" to virtually coincide with the general homeomorphism concept. One now sees maps in which each country is represented by a polygon, with area proportional not to its actual area, but to some other quantity such as population. The region being mapped nevertheless remains recognizable, mainly by the boundary relations between different countries, which are topologically invariant. Western Europe, for example, is shown in Figure 6.

However, we should not push the geographic analogy too far, as this can lead to the misconception that topology is just rubber sheet geometry, in other words, that all homeomorphisms are *deformations* (defined precisely as *isotopies* in 0.1.9). Once we leave the plane most of them are not—it is quite in order to cut a figure, deform it, and then rejoin, provided that rejoining restores the neighbourhood of each point on the cut. The torus provides a good illustration of this *cut and paste* method. In Figure 7 we cut the torus along a meridian a, twist one edge of the cut through 2π relative to the other, then rejoin. A small disc neighbourhood of any point on the cut is separated into semidiscs at the first step, but reunited after the twist of 2π, so for any ε-neighbourhood on the final torus we can find a δ-neighbourhood on the initial torus which maps into it. The transformation therefore defines a continuous one-to-one function, as does its inverse, so we have a homeomorphism f. It is intuitively clear that f cannot be realized by deformation alone, in particular b cannot be deformed onto $f(b)$. In fact, when one studies homeomorphisms of the torus algebraically (6.4) the deformations are factored out as trivial.

Continuous maps which are not necessarily one-to-one are also important. For example, a *curve* is nothing but a continuous map of a line segment. If

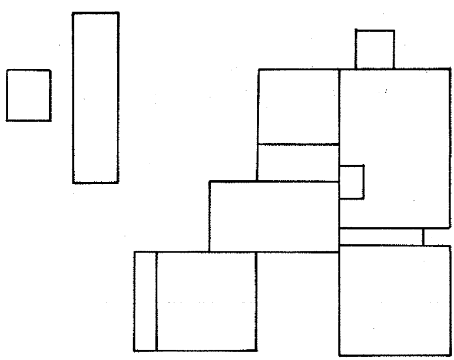


Figure 6

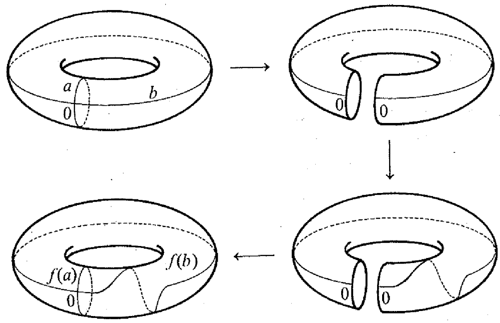


Figure 7

the endpoints have distinct images it is an *arc*, otherwise a *closed curve*, which is also the continuous map of a circle. Points on the arc or closed curve which are images of more than one point on the line segment or circle respectively are called *multiple points* or *singularities*. For example (see Figure 8), there is an obvious map of the circle $\mathbf{S}^1$ into $\mathbf{R}^2$ which realizes the figure eight. The figure eight has a double point which in this case is the image of the two points $\pi/2$, $3\pi/2$ on $\mathbf{S}^1$. We refer to a topological map of $\mathbf{S}^1$ as a *topological* $\mathbf{S}^1$, otherwise a *singular* $\mathbf{S}^1$. Similarly, one can speak of a *topological disc* and *singular disc*, etc.

An important class of many-to-one maps are *covering maps*, the paradigm of which is the covering of $\mathbf{S}^1$ by $\mathbf{R}^1$. This is defined by the function $f: \mathbf{R}^1 \to \mathbf{S}^1$ which maps successive segments of length 2π onto the circumference of the unit circle, in other words

$$f(x) = x \bmod 2\pi,$$

where the right-hand side denotes the number y, $0 \le y < 2\pi$ such that $x = y + 2n\pi$ for some integer n. Covering maps have the property of being *local homeomorphisms*, that is, their restrictions to sufficiently small neighbourhoods are homeomorphisms. In particular, the covering of $\mathbf{S}^1$ by $\mathbf{R}^1$ is a homeomorphism on any interval of length $<2\pi$. Coverings of 1- and

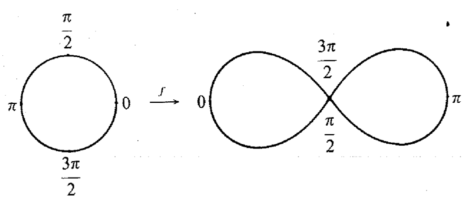


Figure 8

2-dimensional complexes will be defined precisely later (2.2.1 and 4.3.2); they turn out to have an elegant group-theoretic interpretation.

0.1.4 Identification Spaces

Every simplicial complex can be embedded in some R^n (0.2), however, it is not always necessary or natural to do this. The dimension of the ambient space R^n is usually higher than that of the embedded figure, and this leads to confusion between properties of the embedding and properties of the figure itself. The problem is that construction *inside* a given space may involve bending or intertwining parts in rather arbitrary ways, and to avoid the bias of a particular method of assembly one should simply list the parts and say which are to be made equal.

For example, the torus can be constructed from a unit square by joining opposite sides according to the plan shown in Figure 9. In other words, points on the perimeter which differ by unit vertical or horizontal translations become equal. Actually joining opposite sides in R^3 leads for example to the torus shown in Figure 10 which treats the curves a and b quite differently, whereas the original plan is completely symmetrical with respect to a and b.

The process of "saying points are equal when they're not" can be formalized by the construction of an *identification space* whose points are the sets $X = \{X_1, X_2, \ldots\}$ of points $X_1, X_2, \ldots$ which we want to be equal and

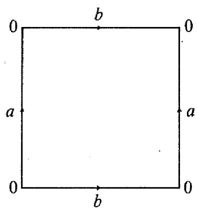


Figure 9

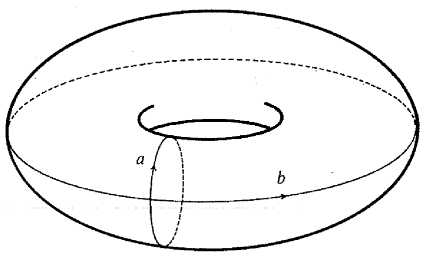


Figure 10

whose neighbourhoods $\mathcal{N}_\varepsilon(X)$ consist of the points in $\mathcal{N}_\varepsilon(X_1) \cup \mathcal{N}_\varepsilon(X_2) \cup \cdots$, for sufficiently small ε. X is called the result of *identifying* $X_1, X_2, \ldots$.

When the torus is constructed as an identification space of the square the sets X are either (i) one-element sets (interior points of the square), (ii) two-element sets (corresponding interior points of opposite sides), or (iii) a four-element set (corners). The neighbourhoods of these three types of point are respectively (i) discs, (ii) unions of two semidiscs (=discs), and (iii) the union of four quarter discs (=disc) which confirms the fact that the torus is *homogeneous*—every point has a disc neighbourhood.

A related, but more elegant, construction of the torus is the "plane mod 1." One identifies any two points in $\mathbf{R}^2$ whose x- and y-coordinates differ by integers. The homogeneity of this space is clear, but it is also clear that every point is identified with some point in the unit square, from which we recover the above representation. The map which sends $(x, y) \in \mathbf{R}^2$ to its equivalence class mod 1 is a covering of the torus by the plane, which we shall investigate further in 1.4.1 and 6.2.2.

EXERCISE 0.1.4.1. What is the identification space of $\mathbf{R}^2$ obtained by identifying points with the same y-coordinate whose x-coordinates differ by an integer?

0.1.5 The n-ball and the n-sphere

The *n-ball* is usually defined to be the set

$$\mathbf{B}^n = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_1^2 + \cdots + x_n^2 \leq 1\}$$

or any set homeomorphic to it, such as an n-simplex. The frontier of this set is the $(n-1)$-sphere

$$\mathbf{S}^{n-1} = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_1^2 + \cdots + x_n^2 = 1\}.$$

In particular, $\mathbf{B}^1$ is represented by the line segment $[-1, 1]$, and $\mathbf{S}^1$ by the unit circle in $\mathbf{R}^2$. $\mathbf{S}^0$ is then the point pair $\{-1, 1\}$. This *equatorial* pair divides $\mathbf{S}^1$ into upper and lower *hemi-1-spheres*, which are seen to be homeomorphic to $\mathbf{B}^1$ by projection onto the x_1 axis. Thus $\mathbf{S}^1$ is an identification space of two $\mathbf{B}^1$'s, obtained by identifying corresponding points on their frontier $\mathbf{S}^0$'s (see Figure 11). This construction easily generalizes to n-dimensions (try it for $n = 2$), so we have the result that $\mathbf{S}^n$ is the identification space of two $\mathbf{B}^n$'s, obtained by identifying corresponding points on their frontiers.

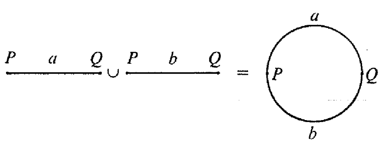


Figure 11

EXERCISE 0.1.5.1. Find a homeomorphism between Δ^n and B^n, and show that it maps the boundary of Δ^n onto the frontier of B^n in R^n.

0.1.6 Manifolds

The most attractive figures from the topological point of view are those which are homogeneous, in the sense that each point has a neighbourhood homeomorphic to the interior of a B^n (an open ball) for some fixed n. These are called the *n-dimensional manifolds*, or *n-manifolds* for short.

The simplest examples are R^n and S^n, whose homogeneity is obvious. Other examples arise as spaces whose elements are not points (at least, not in the initial interpretation) but other geometric objects or phases of mechanical systems.

A good example is given in Figure 12 which shows the system of two rigid rods free to rotate about P (which is fixed) and Q, and constrained to move in a vertical plane. The space of positions of this system is clearly 2-dimensional and homogeneous, but it comes as a surprise to find it is the torus! The reason is simply that position is uniquely determined by values $0 \leq \theta \leq 2\pi$ and $0 \leq \phi \leq 2\pi$, as is position on the torus if we interpret θ and ϕ as longitude and latitude (see Figure 13).

An example from geometry is the space of all unit tangents to the unit sphere. Using any reasonable measure of the distance between two tangents, the space is clearly homogeneous and locally 3-dimensional (for example, use two coordinates to fix the point of contact with the sphere, one for the direction of the tangent), hence a 3-manifold. However, there is no obvious coordinate system for the whole space. In fact this is a manifold we have not seen before, and it will be identified only in 8.3.4.

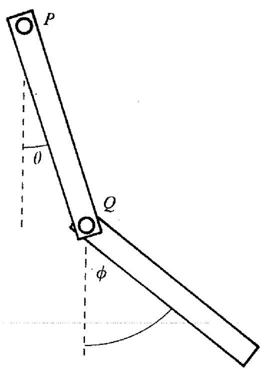


Figure 12

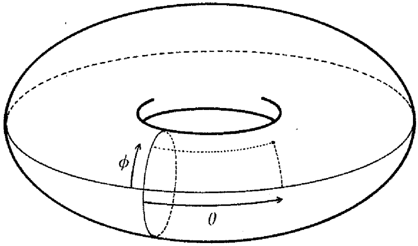


Figure 13

It is less easy to tell, in general, when a figure constructed as an identification space is a manifold, and the neighbourhoods of individual points may have to be checked, as we did for the identification space of the square in 0.1.4. The check in that case revealed a 2-manifold (the torus). On the other hand, if we identify all three sides of a triangle as in Figure 14, the result $\mathscr{C}$ is not a manifold, because a point P on one of the sides has a "book with three leaves" as neighbourhood (Figure 15) and presumably no neighbourhood homeomorphic to a disc. We shall not prove this, however, it is possible to show this complex is not a 2-manifold by computing its fundamental group (see Chapter 4) and showing that it is unequal to the group of any 2-manifold by the methods of Chapter 5.

EXERCISE 0.1.6.1: What is the dimension of the space of all straight lines through the origin in R^3? Describe this manifold as an identification space of S^2.

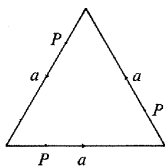


Figure 14

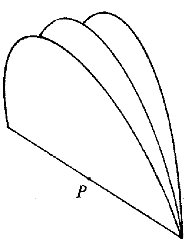


Figure 15

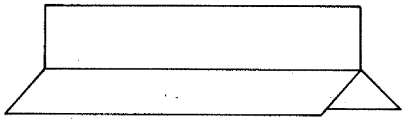


Figure 16

EXERCISE 0.1.6.2. Show that the complex $\mathscr{C}$ above may also be obtained by pasting a disc onto the figure obtained by identifying the ends of Figure 16 after a twist of $2\pi/3$.

EXERCISE 0.1.6.3. Show that the only 1-manifolds are $\mathbf{R}^1$ and $\mathbf{S}^1$.

0.1.7. Bounded Manifolds

The n-simplex does not appear to be a manifold because we cannot find open ball neighbourhoods for points on its boundary. Instead, the boundary points have "half-n-ball" neighbourhoods, homeomorphic to the open n-ball minus the open half-space determined by a hyperplane through its centre. A figure in which every point has either an open n-ball or half-n-ball neighbourhood is called a *bounded n-manifold* or *n-manifold with boundary.* If we were to prove that the open n-ball and half-n-ball were really not homeomorphic then we could define the boundary of a bounded n-manifold in a topologically invariant way as the set of points with half-n-ball neighbourhoods; it would coincide with the boundary we have already defined for the n-simplex (0.1.1), and we would also know that bounded manifolds are not manifolds.

These results are correct, however they are not as useful as they seem. In dimension 2 we can distinguish manifolds from bounded manifolds by the fundamental group (4.2.1 and 5.3.3), while in dimension 3 the problem is to distinguish manifolds from each other rather than from bounded manifolds. We shall therefore adopt the easier course of using "boundary" as a term which is useful in the discussion of simplicial complexes, without appealing to its topological invariance, just as we use genuinely nontopological terms such as "length" and "straight line." The same applies to "dimension," which is in fact intimately related to "boundary."

The nontopological definitions of these terms are as follows.

The *dimension n* of a simplicial complex is the maximum dimension among its simplexes. (Thus n exists automatically for a finite complex. For an infinite complex its existence is made part of the definition, see 0.2.1). The *boundary* $\partial\mathscr{C}$ of an n-dimensional simplicial complex $\mathscr{C}$ is the "mod 2 union" of the $(n-1)$-simplexes occurring as faces in $\mathscr{C}$. That is, one counts the number of occurrences (assumed finite, 0.2.1) of a given $(n-1)$-simplex as a face among the simplexes of $\mathscr{C}$, reduces it mod 2, and takes the union of the $(n-1)$-simplexes which are counted once. An example is given in Figure 17.

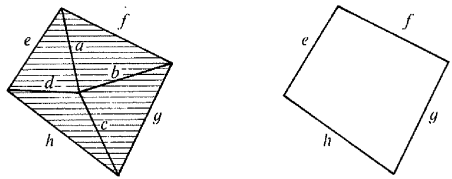


Figure 17

If $\mathscr{C}$ is the complex shown on the left, then $\partial\mathscr{C}$ is given by the figure on the right because a, b, c, d occur twice and e, f, g, h occur once.

0.1.8 Embedding Problems

Next to the homeomorphism problem, the most important type of topological problem is that of distinguishing different embeddings of one figure in another. An *embedding* of $\mathscr{C}_1$ in $\mathscr{C}_2$ is a one-to-one continuous map

$$f: \mathscr{C}_1 \to \mathscr{C}_2.$$

Given $\mathscr{C}_1$ and $\mathscr{C}_2$, the first question is whether an embedding exists, and then if there is one, how many? The latter question of course assumes that we only distinguish embeddings which differ in a topologically significant way. This will be clarified further in 0.1.9, for the moment we shall illustrate the kind of results available by looking at embeddings of S^1 in R^1, R^2, and R^3.

(1) S^1 cannot be embedded in R^1. An embedding of S^1 is equivalent to a continuous map

$$f: [0, 1] \to \mathsf{R}^1$$

which is one-to-one except that $f(0) = f(1)$. This is impossible by the intermediate-value theorem (Exercise 0.1.2.7).

(2) An embedding of S^1 in R^2 is a simple closed curve in the plane. By the Jordan–Schoenflies theorem (0.3.9) any such curve may be mapped onto the unit circle by a homeomorphism of R^2. Presumably we should not distinguish embeddings which are equivalent up to homeomorphism of R^2, hence *there is only one embedding of* S^1 *in* R^2.

(3) It is intuitively clear that there are different embeddings of S^1 in R^3, namely, different knots. We shall prove in Chapter 4 that there are infinitely many embeddings, by finding knots $\mathscr{K}_1, \mathscr{K}_2, \ldots$ such that $\mathsf{R}^3 - \mathscr{K}_i$ and $\mathsf{R}^3 - \mathscr{K}_j$ are nonhomeomorphic for $i \neq j$. Then there certainly cannot be any homeomorphism of R^3 which maps $\mathscr{K}_i$ onto $\mathscr{K}_j$.

EXERCISE 0.1.8.1. Use an embedding argument to show that R^1 is not homeomorphic to R^2.

EXERCISE 0.1.8.2. Use the Jordan–Schoenflies theorem to show that there are only finitely many ways to embed a finite graph (1-dimensional simplicial complex) in R^2. If $\mathscr{K}_n$ denotes the graph with n vertices $1, 2, \ldots, n$ and edges $\{i, j\}$ for each $i \neq j \leq n$, show that $\mathscr{K}_5$ does not embed in R^2, but that $\mathscr{K}_5$, $\mathscr{K}_6$, and $\mathscr{K}_7$ embed in the torus.

0.1.9 Homotopy and Isotopy

The homotopy concept captures the notion of *deformation* of a map. Two maps $f: \mathscr{C}_1 \to \mathscr{C}_2$ and $g: \mathscr{C}_1 \to \mathscr{C}_2$ are called *homotopic* if there is a continuous map

$$h: [0, 1] \times \mathscr{C}_1 \to \mathscr{C}_2$$

such that $h(0, x) = f(x)$ and $h(1, x) = g(x)$. We can think of h as a deformation process over the time interval $[0, 1]$, and the section $h_t(x) = h(t, x)$ at time t as the map into which f has been deformed by time t.

The most important case is where $\mathscr{C}_1 = \mathsf{S}^1$, so that f and g are closed curves in $\mathscr{C}_2$. For a picture illustrating this case see Figure 133 in 3.1.5. It turns out that the study of homotopic curves is the most important tool in the classification of manifolds of dimension ≤ 3. Not surprisingly, a manifold of small dimension is determined to a large extent by the behaviour of curves inside it; in particular we can distinguish the sphere and the torus in this way (see Figure 18). Any curve c on S^2 is *null-homotopic*, that is, homotopic to a point, whereas we can prove that the curve a on the torus is not. The property of being null-homotopic is obviously preserved by homeomorphisms, whence it follows that S^2 and the torus are not homeomorphic.

A space in which every closed curve is null-homotopic is called *simply connected*; so the difference between S^2 and the torus can also be expressed by saying that S^2 is simply connected but the torus is not.

This type of reasoning would not be very useful if each case required an ad hoc argument that certain curves are not null-homotopic. The power of the homotopy concept lies in *algebraic* properties which ultimately permit us to compute a *fundamental group* for each complex (0.5.1) and systematically reduce homotopy questions to group theory.

The group properties depend crucially on the fact that the curve is not required to be simple at any stage, and in fact the deformation may create more singularities than were present at the beginning. Only then can one introduce a *product* of closed curves, and cancel a closed curve by its inverse.

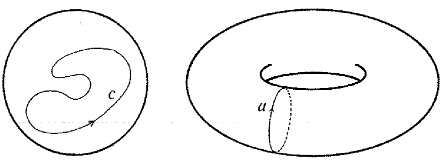


Figure 18

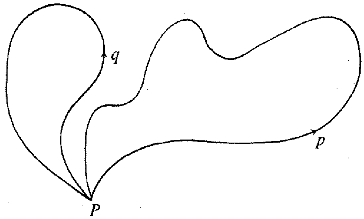


Figure 19

The natural product of two curves p, q which begin at P—their concatenation—will obviously have a multiple point at P (as shown in Figure 19) and the natural inverse p^{-1} of p will lie on top of p but with the opposite orientation (pp^{-1} is then null-homotopic). These ideas are formalized in 3.1.4–3.1.6.

If homotopy is the applied notion of deformation in topology, there is nevertheless a pure notion, which we call *isotopy*. An isotopy is a homotopy h for which every section h_t is a homeomorphism (onto its image). In particular, during an isotopy of a simple closed curve the image remains simple at every stage.

Isotopy seems to be a more natural notion of deformation, but it is not algebraically tractable. In the case of simple curves on a 2-manifold the situation is saved by a theorem of Baer 1928 (6.2.5) which says that simple curves are isotopic if and only if they are homotopic. This enables us to classify the embeddings of S^1 in a 2-manifold by computations in the fundamental group.

Isotopy is a suitable equivalence relation for classifying embeddings of S^1 in surfaces, but definitely not in R^3, since a knot can be isotopic to circle. The "knotted part" can be shrunk to nothing without acquiring a singularity at any stage. Figure 20 shows an example (Alexander 1932). A better notion in this case is that of *ambient isotopy*: two curves in R^3 are ambient isotopic if one is mapped onto the other by a homeomorphism of R^3 isotopic to the identity map. In particular, ambient isotopic curves must have homeomorphic complements, which is not the case for a knot and the circle, as we shall see in 4.2.5.

EXERCISE 0.1.9.1. Show that any homeomorphism of R^1 is isotopic either to the identity or the map $x \to -x$. What is the situation in R^2 and R^3?

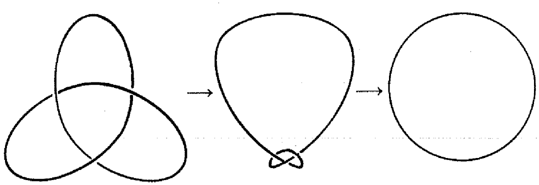


Figure 20

0.2 Simplicial Complexes

0.2.1 Definition and Basic Properties

Recalling the definition of a simplex and its faces in 0.1.1, we define an *n-dimensional simplicial complex* (n-complex) $\mathscr{K}$ to be a union of simplexes of dimension $\leq n$ satisfying the following conditions:

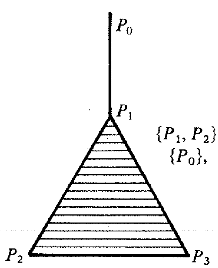
(i) Each simplex meets only finitely many others.
(ii) Two simplexes are either disjoint or their intersection is a common face.

It is best to think of cutting the n-simplexes out of R^n, then assembling the complex as an identification space, as in 0.1.4. Nevertheless it is also possible to embed the whole complex in a suitable R^m, as we shall see in 0.2.3.

Since an n-simplex is determined by its vertices, an n-complex is determined by a list of its vertices, together with those subsets of the vertices which correspond to simplexes. Since any face of a simplex is itself in the complex, it follows that any subset of an element of the list is itself in the list. In particular, the vertices are listed as the singleton subsets. It is not necessary to give coordinates for the vertices, merely distinct names, since different choices of coordinates give homeomorphic simplexes and hence homeomorphic complexes. This description, called a *schema*, is therefore combinatorial in the strictest sense of the word.

As an example we write down the schema for the 2-complex shown in Figure 21, consisting of a triangle with an attached line segment. It is a consequence of the triangulation and *Hauptvermutung* results of 0.2.5 that all homeomorphism questions for 2- and 3-manifolds reduce to combinatorial questions about schemata.

Condition (i) in the definition of simplicial complex is the *local finiteness* condition. It is automatically satisfied when there are only finitely many simplexes, in which case we call the complex *finite*. It is clear that a finite complex is compact, and similarly local finiteness implies *local compactness*,



$\{P_1, P_2, P_3\}$

$\{P_1, P_2\}$, $\{P_2, P_3\}$, $\{P_3, P_1\}$, $\{P_0, P_1\}$

$\{P_0\}$, $\{P_1\}$, $\{P_2\}$, $\{P_3\}$

Figure 21

that is, a neighbourhood with compact closure for each point. More importantly, local finiteness implies every point has a simply connected (0.1.9) neighbourhood, that is, one in which every closed curve is null-homotopic.

A simplex Δ is simply connected because it is convex (0.1.1). This allows any curve c in Δ to be contracted to one of its points P by moving each point on c along the ray from P so that its distance from P at time t, $0 \leq t \leq 1$, is a fraction $(1 - t)$ of its initial distance. With local finiteness one can find an ε-neighbourhood of any point P which contains only simplexes $\Delta_1, \ldots, \Delta_k$ containing P, and then any curve in this neighbourhood can be contracted to a point by sliding it down rays to a common point of $\Delta_1, \ldots, \Delta_k$ in the same way.

The union of the simplexes containing a given vertex P in a complex $\mathscr{C}$ is called the *neighbourhood star* of P. Typical neighbourhood stars are shown in Figure 22. The neighbourhood star is a suitable combinatorial notion of a neighbourhood, because it is homeomorphic to the closure of any sufficiently small ε-neighbourhood of P. A homeomorphism is obtained by mapping each line segment from P to the frontier of the ε-neighbourhood linearly onto its prolongation to the boundary of the simplex in which it lies.

It follows that if $\mathscr{C}$ is an n-manifold then each of its neighbourhood stars is a topological B^n.

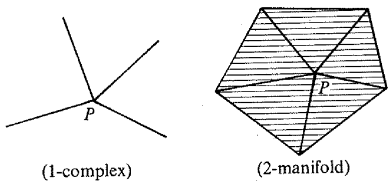


Figure 22

EXERCISE 0.2.1.1. Show that an infinite complex is not compact.

EXERCISE 0.2.1.2. Construct a figure in R^2 which is not locally simply connected.

EXERCISE 0.2.1.3. In a simplicial n-manifold, show that the faces not containing P in the neighbourhood star of P constitute a topological S^{n-1}.

0.2.2 Orientation

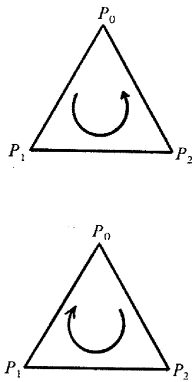
A 1-simplex Δ^1 has a natural orientation as the topological image of the unit interval $[0, 1]$. Namely, if $f: [0, 1] \to \Delta^1$ is a topological map we let $f(x) < f(y)$ if $x < y$. If $P_0 = f(0)$, $P_1 = f(1)$ we can describe the orientation combinatorially by the ordered pair (P_0, P_1) and pictorially by

$$P_0 \longrightarrow P_1.$$

In general, we interpret the ordered $(n + 1)$-tuple $(P_0, \ldots, P_{n+1})$ as an orientation of the n-simplex Δ^n with vertices $P_0, \ldots, P_{n+1}$. Orientations are equivalent if they differ by an even permutation of the vertices, so there are in fact two possible orientations, $+(P_0, \ldots, P_{n+1})$ which is just $(P_0, \ldots, P_{n+1})$, and $-(P_0, \ldots, P_{n+1})$, obtained by an odd number of exchanges of vertices.

In a 2-simplex the orientation can be indicated by a circular arrow as shown in Figure 23. An orientation of an n-simplex induces an orientation in each face, simply by omitting the vertices not in that face.

An *orientation of an n-complex* is an assignment of orientations to its simplexes. The orientation is *coherent* if n-simplexes which share an $(n - 1)$-dimensional face induce opposite orientations in that face. An example of what a coherent orientation for a 2-manifold looks like is given in Figure 24. Intuitively, one can slide a circular arrow all over the surface and match it



$$(P_0, P_1, P_2) = (P_1, P_2, P_0) = (P_2, P_0, P_1)$$

$$(P_0, P_2, P_1) = (P_2, P_1, P_0) = (P_1, P_0, P_2)$$

Figure 23

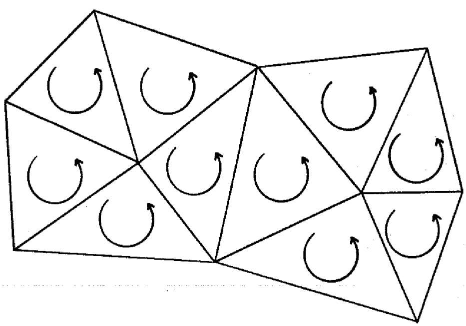


Figure 24

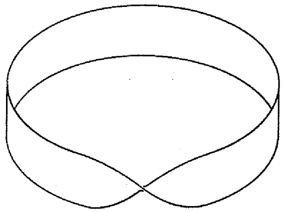


Figure 25

with the circular arrow drawn in each triangle. A complex is called *orientable* if it has a coherent orientation.

The classic *nonorientable* figure is the Möbius band (Figure 25). The reader is invited to triangulate this surface and see why it cannot be oriented coherently.

0.2.3 Realization in Euclidean Space

Any n-complex can be embedded in R^{2n+1}.

To motivate the proof, first consider how to embed a 1-complex in R^3. A *topological* embedding is certainly possible if we simply bend the edges to avoid collisions, but a *rectilinear* embedding is also possible if we place the vertices on a suitable twisted curve. There are many curves with the property that no four points on them are coplanar, so chords meet only when they have a common endpoint, and hence can serve as edges for the 1-complex. One such curve is given by the parametric equations

$$x = t, \qquad y = t^2, \qquad z = t^3$$

for if distinct points t_1, t_2, t_3, t_4 lie on the plane

$$ax + by + cz = d,$$

they are four distinct roots of the cubic equation

$$at + bt^2 + ct^3 = d$$

which is impossible.

The argument readily generalizes to embed an n-complex in R^{2n+1}. We put vertices on the curve

$$x_1 = t, \qquad x_2 = t^2, \qquad \ldots, \qquad x_{2n+1} = t^{2n+1}$$

and then no $2n + 2$ distinct vertices lie in a common hyperplane, as this would imply an equation of degree $2n + 1$ with $2n + 2$ distinct roots. It follows that two n-simplexes (each determined by $n + 1$ vertices) meet only if they have vertices in common. Since the simplex determined by the common vertices is itself in the complex, we have an embedding. □

The above proof was found by Leigh Samphier. Other proofs use only linear algebra (one using the above curve may be found in Giblin 1977), but they are slightly longer. In any case, the result that an mth degree equation has $\leq m$ roots may be proved using the mean-value theorem of calculus, and hence is quite elementary.

The dimension $2n + 1$ cannot be lowered. We saw this for $n = 1$ in Exercise 0.1.8.2. Van Kampen 1932 proved the generalization of this fact for the "complete n-complex" on $2n + 3$ vertices.

EXERCISE 0.2.3.1. Show that one turn of the helix $x = \cos t$, $y = \sin t$, $z = t$ also has the property that no four points are coplanar.

0.2.4 Cell Complexes

Viewing a figure as a simplicial complex is one way to assemble it from *cells*, in this case simplexes. Taking a cell to be any figure homeomorphic to a simplex, we can also consider more complicated methods of assembly, perhaps involving identification of the boundary of a cell with itself. For example, the construction of the torus by identifying sides of the square may be viewed as a 2-dimensional cell structure with one 0-cell (the vertex 0), two 1-cells (the edges a and b) and one 2-cell (the square) as shown in Figure 26. In general, a cell complex is constructed by first assembling the 0-cells; then attaching the 1-cells by identifying their boundaries with 0-cells to form the 1-*skeleton*; then attaching the 2-cells by mapping their boundaries onto the 1-skeleton to form the 2-*skeleton*; and so on. These stages for the above cell structure for the torus are shown in Figure 27. If the attaching maps are

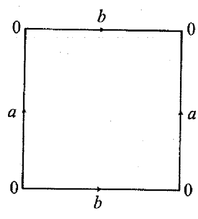


Figure 26

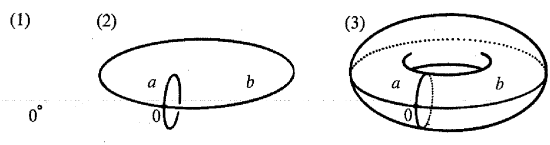


Figure 27

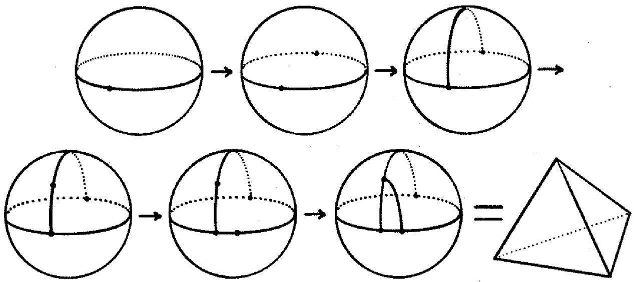


Figure 28

sufficiently simple, as they will be in all the cases we consider, it is possible to reduce a cell decomposition to a simplicial decomposition by *elementary subdivision*. An elementary subdivision of a 1-cell is the introduction of a new interior 0-cell, an elementary subdivision of a 2-cell is the introduction of an interior 1-cell connecting 0-cells, and in general one m-cell is divided into two by the introduction of a new interior $(m - 1)$-cell spanning an $(m - 2)$-sphere in its boundary.

For example, the cell decomposition of $\mathbf{S}^2$ into two hemispheres can be made simplicial by the series of elementary subdivisions of 1-cells and 2-cells shown in Figure 28. Conversely, one can view the initial cell decomposition as the result of amalgamating certain cells in a simplicial decomposition (reverse the arrows). Since all the cell decompositions we use can be viewed in this way, it will not be necessary to make our definitions of cell complex and elementary subdivision any more formal, since in the last resort one can always view cells and the dividing cells inside them as unions of simplexes in a simplicial decomposition. The point of considering cell complexes at all is to minimize the number of cells, which usually helps to shorten computations.

EXERCISE 0.2.4.1. Obtain the two decompositions of the torus in Figure 29 by elementary subdivision of the square cell structure. Which of them is simplicial?

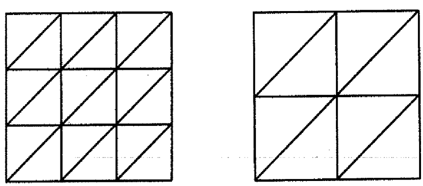


Figure 29

EXERCISE 0.2.4.2. The *barycentric subdivision* of a simplex Δ^n is obtained by introducing a new vertex at the centre of mass (the *barycentre*) of each face, and then introducing all simplexes of dimension $\leq n$ determined by the enlarged set of vertices. Why is this a subdivision? (*Hint*: Generalize the theorem that the medians of a triangle are concurrent.)

Show that by repeating barycentric subdivision a sufficient number of times in a finite n-complex, the diameter of all simplexes may be made less than a given $\varepsilon > 0$.

EXERCISE 0.2.4.3. Making the obvious interpretation of barycentric subdivision for arbitrary 1-cells, not necessarily straight, show that the second barycentric subdivision of a 1-dimensional cell complex is simplicial.

EXERCISE 0.2.4.4. Show that the boundary and orientability character of a simplicial complex are invariant under elementary subdivision.

0.2.5 Triangulation and *Hauptvermutung*

Our definition of a manifold in 0.1.6 depended on the notions of neighbourhood and homeomorphism, and it is by no means clear that every n-manifold is a simplicial complex. However, this is true for $n \leq 3$. For $n = 1$ it is clear, since the only 1-manifolds are R^1 and S^1; for $n = 2$ it was proved by Rado 1924; and for $n = 3$ by Moise 1952. A simplicial decomposition of a manifold is also called a *triangulation*, and proofs that 2- and 3-manifolds possess triangulations may be found in Moise 1977.

We shall bypass these theorems by confining our attention to figures which *are* simplicial complexes. As pointed out in 0.1.1, we shall certainly not miss any reasonable figures with this approach. It is also possible to give purely combinatorial criteria for 2- and 3-complexes to be manifolds. For 2-manifolds these are given in 1.3.1, and for 3-manifolds in 8.2..1 and 8.2.2.

Finally, one can give a combinatorial definition of homeomorphism using the notion of elementary subdivision. Two simplicial complexes are certainly homeomorphic if they possess isomorphic schemata (schemata which are identical up to renaming of vertices). More generally, they are homeomorphic if their schemata become isomorphic after finite sequences of elementary subdivisions, in other words, if they have a common simplicial refinement. We say that two complexes are *combinatorially homeomorphic* if this is the case. We might naively expect a common simplicial refinement to follow from superimposing the two simplicial decompositions of the manifold, if indeed the two manifolds are the same. However, one must bear in mind that in mapping one decomposition onto the other rectilinearity may be lost, so that two edges, for example, may intersect in infinitely many points. (The superimposition error has a distinguished history, being first committed by Riemann 1851 in discussing the connectivity of surfaces.)

The *Hauptvermutung* (main conjecture) of Steinitz 1908 states that homeomorphic manifolds are combinatorially homeomorphic. It is known to be correct for manifolds of dimension ≤ 3, in fact it is a rather easy consequence of the triangulation theorems. We shall derive the *Hauptvermutung* for triangulated 2-manifolds as a consequence of the classification theorem in 1.3.7 and 5.3.3.

With the proofs of triangulation and *Hauptvermutung* we are entitled to say that the homeomorphism problems for 2- and 3-manifolds are purely combinatorial questions. To answer them, however, we need combinatorial tools from group theory, and it turns out to be easier to develop these tools directly, without appeal to *Hauptvermutung*. This is the route we shall take in this book, particularly for 3-manifolds. The theory of 2-manifolds under elementary subdivisions is presented in Chapter 1, but before it can be completed we need the group theory of Chapters 2 and 3, which also serves for higher dimensions.

0.3 The Jordan Curve Theorem

0.3.1 Connectedness and Separation

The statement, as a theorem, that every simple closed curve in $\mathbf{R}^2$ separates it into two regions (Jordan 1887) was important in the history of topology as the first moment when an "obvious" fact was seen to require proof. As is well-known, Jordan's own proof was faulty, and this has only added to the theorem's reputation for subtlety. The first rigorous proof was given by Veblen 1905, and a variety of lengthy proofs have been reproduced in textbooks. A very short and transparent proof is given in Moise 1977, and we reproduce it below, slightly modified. Little use will actually be made of the theorem, but it is an excellent example of the process of reducing general topology to combinatorial topology.

The first step is to reduce the general notion of connectedness to one in terms of polygonal curves. This reduces questions about general curves to questions about polygonal curves, for which the separation properties are easily proved.

The key proposition is the following:

Let $P, Q \in \mathcal{O}$, an open set in $\mathbf{R}^n$. Then the following statements are equivalent.

(i) *P, Q are the endpoints of a polygonal arc $\subset \mathcal{O}$.*
(ii) *P, Q are the endpoints of an arc $\subset \mathcal{O}$.*
(iii) *P, Q lie in an open set $\mathcal{O}' \subset \mathcal{O}$ which is not the union of two disjoint non-empty open sets.*

(iii) $\Rightarrow$ (i). Consider the set of all points R which are connected to P by a finite chain of open balls $\mathcal{B}_1, \ldots, \mathcal{B}_k \subset \mathcal{O}$. That is

$$P \in \mathcal{B}_1, \qquad R \in \mathcal{B}_k, \quad \text{and} \quad \mathcal{B}_i \cap \mathcal{B}_{i+1} \neq \varnothing.$$

These points R obviously constitute an open set $\mathcal{O}_P \subset \mathcal{O}$. If $\mathcal{O}_P \neq \mathcal{O}$, then $\mathcal{O} - \mathcal{O}_P$ is also open, because any ball $\subset \mathcal{O}$ which is partly in $\mathcal{O}_P$ is entirely in $\mathcal{O}_P$, hence any $S \in \mathcal{O} - \mathcal{O}_P$ has its ball neighbourhoods in $\mathcal{O} - \mathcal{O}_P$.

Then if $Q \notin \mathcal{O}_P$ the set $\mathcal{O}'$ decomposes into disjoint nonempty open sets $\mathcal{O}' \cap \mathcal{O}_P$ and $\mathcal{O}' \cap (\mathcal{O} - \mathcal{O}_P)$, which is a contradiction. Thus Q is connected to P by a finite chain of open balls, and hence by a polygonal arc.

(i) $\Rightarrow$ (ii) is trivial.

(ii) $\Rightarrow$ (iii). Let a be an arc connecting P and Q, and let $\mathcal{O}'$ be an open set $\supset a$, obtained as the union of ball neighbourhoods in $\mathcal{O}$ of all the points in a. If $\mathcal{O}'$ decomposes into disjoint open sets $\mathcal{O}'', \mathcal{O}''', \ldots$, let X be the first point of a not in $\mathcal{O}''$ (Exercise 0.1.2.1). Then X lies on the frontier of $\mathcal{O}''$ and cannot belong to any open set disjoint from $\mathcal{O}''$, so we have a contradiction. □

In general topology an open set $\mathcal{O}$ is called *connected* if it is not a disjoint union of nonempty open sets. This is also expressed by saying $\mathcal{O}$ has only one *component*, the component containing a given point P being the $\mathcal{O}_P$ constructed above. Thus we have just proved that a connected open set $\mathcal{O} \subset \mathbb{R}^n$ has the stronger property of being *arc connected*, that is, any two points in $\mathcal{O}$ are the endpoints of an arc in $\mathcal{O}$; and furthermore the arc can be assumed polygonal.

A set $\mathcal{S}$ contained in a set $\mathcal{D}$ *separates* points $P, Q \in \mathcal{D} - \mathcal{S}$ if any arc from P to Q in $\mathcal{D}$ meets $\mathcal{S}$. If $\mathcal{D} - \mathcal{S}$ is open (as it will be if $\mathcal{D}$ is open and $\mathcal{S}$ is a closed set, such as a curve), then an equivalent statement (by the above proposition) is that P and Q lie in distinct components of $\mathcal{D} - \mathcal{S}$.

From now on we refer to a simple closed curve in $\mathbb{R}^2$ as a *Jordan curve*.

EXERCISE 0.3.1.1. Show that $\mathcal{O}_P = \{Q \in \mathcal{O}: P, Q \text{ are the endpoints of an arc } \subset \mathcal{O}\} = \{Q \in \mathcal{O}: P, Q \text{ are the endpoints of a polygonal arc } \subset \mathcal{O}\}$.

0.3.2 The Polygonal Jordan Curve Theorem

A polygonal Jordan curve p separates $\mathbb{R}^2$ into two components.

The open set $\mathbb{R}^2 - p$ has at most two components, determined by the components of $\mathcal{N} - p$, where $\mathcal{N}$ is a strip neighbourhood of p in $\mathbb{R}^2$. For any point $P \in \mathbb{R}^2$ is connected to one "side" of $\mathcal{N}$ by a line segment, and any point in $\mathcal{N} - p$ is connected to either P_1 or Q_1 by a polygonal arc in $\mathcal{N} - p$ (see Figure 30).

We now prove that $\mathbb{R}^2 - p$ has at least two components.

Consider a family of parallel lines l in a direction different from that of any segment of p. Intuitively, P is outside p if it lies on an unbounded segment of an $l - p$, or in general if one crosses p an even number of times in order to reach P from an unbounded segment of an $l - p$ (see Figure 31). (Touching a vertex as shown does *not* count as a crossing.) The points $P \in \mathbb{R}^2 - p$

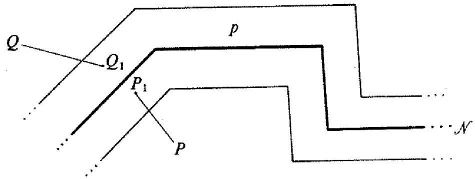


Figure 30

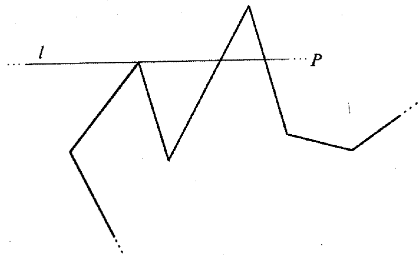


Figure 31

with this property obviously constitute an open set $\mathcal{O}$, and the points with the contrary property constitute an open set $\mathcal{I}$ (the "inside"). Since $\mathcal{O}$ and $\mathcal{I}$ are disjoint by definition, $\mathsf{R}^2 - p$ has at least two components, and therefore exactly two. □

We define a *polygon* $\mathcal{P}$ to be a region in R^2 consisting of a polygonal curve p and its inside. The next section deals with separation in polygons.

EXERCISE 0.3.2.1. Show that the polygon $\mathcal{P}$ determined by a polygonal Jordan curve p may be triangulated, by first dividing it into convex polygons. Deduce that

$$p = \partial\mathcal{P} = \text{frontier of } \mathcal{P}$$

and that the inside of p is the interior of $\mathcal{P}$.

EXERCISE 0.3.2.2. Show that a polygonal arc does not separate R^2.

EXERCISE 0.3.2.3. Show that a semidisc (half 2-ball, cf. 0.1.7) may be separated by an arc.

0.3.3 θ-graphs

A figure $\mathcal{T}$ consisting of a polygonal Jordan curve p and a simple polygonal arc p_3 connecting points Q, S on p, and elsewhere lying in the interior of the polygon $\mathcal{P}$ determined by p, is called a *θ-graph.*

If $\mathcal{T}$ is a θ-graph and p_1, p_2 denote the arcs into which p is divided by Q, S, then p_3 separates an interior point P_1 of p_1 from an interior point P_2 of p_2 in $\mathcal{P}$.

As in 0.3.2, the components of $\mathsf{R}^2 - \mathscr{T}$ are determined by the components of $\mathscr{N} - \mathscr{T}$, where $\mathscr{N}$ is a strip neighbourhood of $\mathscr{T}$ in R^2. The latter components are

(i) a strip $\mathscr{N}_3$ around the "outside" of p
(ii) a strip $\mathscr{N}_1$ commencing on the "inside" of p_1
(iii) a strip $\mathscr{N}_2$ commencing on the "inside" of p_2.

Strips (ii) and (iii) continue up the sides of p_3 (see Figure 32) and either close into separate strips or (somehow!) join into one. In fact there must be three separate strips by 0.3.2, since they are pairwise separated from each other by polygonal Jordan curves $p_i \cup p_j$.

Now extend p_3 to the outer frontier of $\mathscr{N}_3$ by transverse segments at each end, to become p'_3 (see Figure 33). Then $(\mathscr{P} \cup \mathscr{N}_3) - p'_3$ consists of two

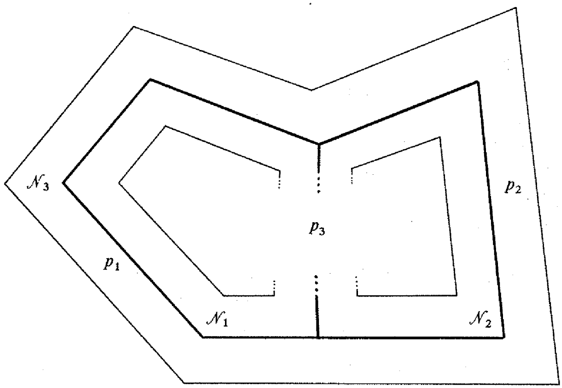


Figure 32

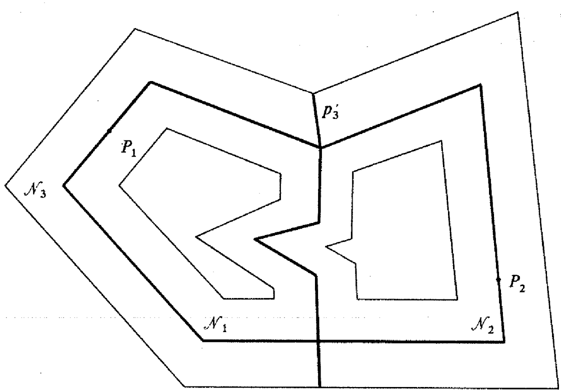


Figure 33

components, determined by $\mathcal{N}_1$ and $\mathcal{N}_2$, which contain P_1 and P_2 respectively. Thus there is no arc from P_1 to P_2 in $(\mathcal{P} \cup \mathcal{N}_3) - p'_3$, and *a fortiori* none in $\mathcal{P} - p_3$.

In other words, p_3 separates P_1 from P_2 in $\mathcal{P}$. □

0.3.4 Arcs Across a Polygon

If P, Q, R, S are points in cyclic order on the boundary p of a polygon $\mathcal{P}$, and a is a simple arc from P to R which elsewhere lies in $\text{int}(\mathcal{P})$, *then a separates Q from S in $\mathcal{P}$.*

Since p is polygonal, points $Q', S' \in \text{int}(\mathcal{P})$ sufficiently close to Q, S respectively can be connected to them by line segments in $\text{int}(\mathcal{P})$. Furthermore, these line segments will miss a if they are sufficiently short, since the closed sets Q, S, a are nonzero distances apart by 0.1.2(2). Thus if Q, S are not separated by a, neither are Q', S' and they then lie in the same component of the open set $\mathcal{P} - (p \cup a)$. It follows from 0.3.1 that Q', S' are connected by a polygonal arc in $\mathcal{P} - (p \cup a)$, so Q, S are connected by a polygonal arc p_3 in $\mathcal{P} - a$, which meets p only at Q, S. We can assume p_3 is simple, since loops can be omitted, so we have a θ-graph (see Figure 34).

Then, by 0.3.3, p_3 separates P from R in $\mathcal{P}$, contrary to the existence of the arc a connecting them. □

Corollary. *If a_1, a_2 are two simple arcs from P to R in* $\text{int}(\mathcal{P})$, *disjoint except at P, R, and if a_1 is the first arc encountered on an arc p from Q to S in* $\text{int}(\mathcal{P})$, *then a_2 is the last encountered.*

Suppose on the contrary that the first and last points encountered (which exist by exercise 0.1.2.1) X, Y both lie on a_1 as in Figure 35. Let p_1 be the subarc of p from Q to X; let a be the subarc of a_1 from X to Y; and let p_2 be the subarc of p from Y to S. Then $p_1 \cup a \cup p_2$ is an arc in $\mathcal{P} - a_2$ connecting Q to S, contrary to the fact that a_2 separates these points. □

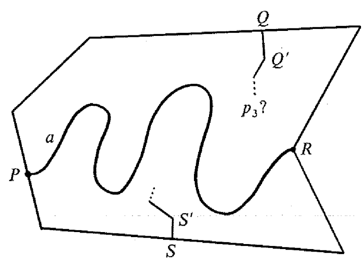


Figure 34

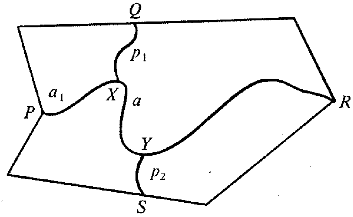


Figure 35

0.3.5 The Jordan Separation Theorem

If c is a Jordan curve in R^2, *then* $\mathsf{R}^2 - c$ *is not connected.*

Since c is compact by 0.1.2(3), we can assume it lies in the interior of a square $ABCD \subset \mathsf{R}^2$. Let $A'D'$ be the left-most vertical in the square which meets c, and $B'C'$ the rightmost. They exist, otherwise there would be vertical disjoint from c but at zero distance from it, contrary to 0.1.2(2). Now choose points P, R where c meets $A'D'$, $B'C'$ respectively, and construct the hexagon $APDCRB$ (see Figure 36). This polygon contains c in its interior except at P and R.

We now define a point Z which intuitively lies "inside" c, and prove that c separates it from $\partial(ABCD)$.

Let c_1, c_2 be the two arcs into which c is divided by P, R, where c_1 is the first encountered on a vertical from $Q \in \operatorname{int}(A'B')$ to $S \in \operatorname{int}(D'C')$. Let X be the last point at which this vertical meets c_1, and Y the first point below X at which it meets c_2. X, Y exist by the corollary in 0.3.4. Then let Z be any point in $\operatorname{int}(XY)$ (see Figure 37).

Now if Z lies in the same component of $\mathsf{R}^2 - c$ as any point on $\partial(ABCD)$ there is a polygonal arc connecting them by 0.3.1, and hence a polygonal

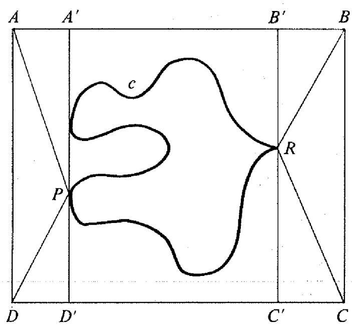


Figure 36

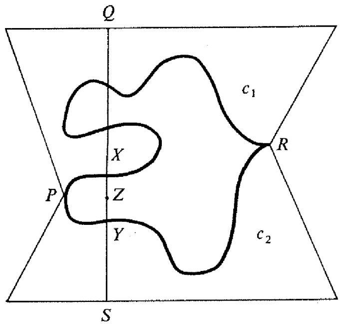


Figure 37

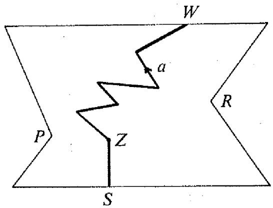


Figure 38

simple arc a between Z and a point $W \neq P, R$ on the boundary of the hexagon. P, R divide this boundary into upper and lower parts. If, say, W is on the upper part we construct the θ-graph shown in Figure 38 by uniting a with ZS. By hypothesis a does not meet c, and ZS does not meet c_1, so $a \cup ZS$ does not meet c_1. But by 0.3.3, $a \cup ZS$ separates P from R, contrary to the existence of the arc c_1 connecting them.

The argument is completely analogous when W is on the lower part. □

0.3.6 Arcs in a Polygon

Let P, Q, R, S be points in cyclic order on the boundary of a polygon $\mathscr{P}$ and let a_1, a_2 be disjoint simple arcs which lie in $\text{int}(\mathscr{P})$ *except that a_1 begins at P and a_2 ends at R. Then Q and S are not separated by $a_1 \cup a_2$ in $\mathscr{P}$.*

Since Q, S, a_1, a_2 are disjoint closed sets, there is some minimum distance $\delta > 0$ between them, by 0.1.2(2). We now pave R^2 with rectangular "bricks" of diameter $<\delta/2$ in the pattern shown in Figure 39. This paving has the property that any finite (arc) connected union $\mathscr{A}$ of bricks has a boundary consisting of disjoint Jordan curves, and by paving so that $\partial\mathscr{P}$ does not pass

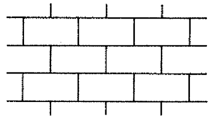


Figure 39

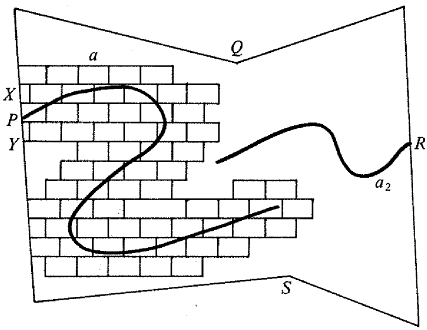


Figure 40

through a corner or touch (as distinct from cross) an edge of a brick, the same is true for $\mathscr{A} \cap \mathscr{P}$.

Now let $\mathscr{A}_1$ consist of the bricks which meet a_1, or meet bricks which meet a_1. Then $a_1 \subset \text{int}(\mathscr{A}_1)$ but $\mathscr{A}_1$ does not meet a_2, since the distance between a_1, a_2 is $\geq \delta$ (see Figure 40). The Jordan curves which constitute $\partial(\mathscr{A}_1 \cap \mathscr{P})$ do not meet a_1 except for a segment XY containing P in $\partial\mathscr{P}$. The boundary arc a of $\mathscr{A}_1 \cap \mathscr{P}$ complementary to XY therefore runs in $\mathscr{P}$ from the point X between Q and P to the point Y between P and S. Thus if p_1 is the subarc of $\partial\mathscr{P}$ from Q to X and if p_2 is the subarc of $\partial\mathscr{P}$ from Y to S, then the arc $p_1 \cup a \cup p_2$ connects Q to S in $\mathscr{P} - (a_1 \cup a_2)$. □

By moving p_1, p_2 into $\text{int}(\mathscr{P})$ by a sufficiently small distance, and removing any loops, we can connect Q, S by a *simple* polygonal arc which is in $\text{int}(\mathscr{P}) - (a_1 \cup a_2)$ except at its endpoints.

0.3.7 No Simple Arc Separates R^2

If a is a simple arc in R^2, *then* $\mathsf{R}^2 - a$ *has only one unbounded component because a is bounded. We show that* $\mathsf{R}^2 - a$ *has no bounded component.*

If there is a bounded component $\mathscr{B}$ of $\mathsf{R}^2 - a$ its frontier is a closed subset of a. The minimal subarc of a containing the frontier of $\mathscr{B}$ therefore has its endpoints on the frontier. This subarc will then also separate R^2, so we may as well assume that the endpoints T, T' of a are on the frontier of $\mathscr{B}$ to begin with.

Since T, T' are limit points of the connected set $\mathscr{B}$ they cannot be separated by any curve disjoint from $\mathscr{B}$. We now derive a contradiction by constructing such a curve.

We use the method of 0.3.5 to enclose a in a polygon $\mathscr{P}$ whose boundary meets a at exactly two points P, R, not necessarily the endpoints T, T' of a. By the concluding remark of 0.3.6 there is a polygonal arc p from Q to S in $\text{int}(\mathscr{P})$ which misses the subarcs a_1, a_2 of a from P to T and R to T' respectively (see Figure 41).

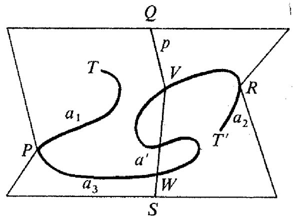


Figure 41

Let a_3 be the subarc of a from P to R, and let V and W be the first and last points of p on a_3. Let p_1 be the subarc of p from Q to V; let a' be the subarc of a_3 from V to W; and let p_2 be the subarc of p from W to S. Then by 0.3.4 $p_1 \cup a' \cup p_2$ separates P from R in $\mathscr{P}$, and hence T from T', since T is connected to P and T' to R. However, $p_1 \cup a' \cup p_2$ does not meet $\mathscr{B}$, since p_1, p_2 lie in the unbounded component of R^2 and $a' \subset a$, so we have a contradiction. □

The above theorem depends crucially on the fact that a simple arc is the topological image of a *closed* interval, and hence has endpoints. The topological image of an open interval can obviously separate R^2—for example, an infinite straight line.

EXERCISE 0.3.7.1. Give an example of the topological image of an open interval which is bounded and separates R^2.

EXERCISE 0.3.7.2. Show that an open disc cannot be separated by a simple arc and deduce that the boundary of a bounded 2-manifold is topologically invariant (cf. 0.1.7 and Exercise 0.3.2.3.).

EXERCISE 0.3.7.3. Where does the above proof assume that a is simple?

0.3.8 The Jordan Curve Theorem

If c is a Jordan curve in R^2, then $\mathsf{R}^2 - c$ has exactly two components.

Since we know from 0.3.5 that $\mathsf{R}^2 - c$ has at least two components, one of which is unbounded, it will suffice to show that there is only one bounded component.

Enclose c in a polygon $\mathscr{P}$ as in 0.3.5, so that c meets $\partial\mathscr{P}$ at exactly two points P and R, and let c_1, c_2, Q, S, $p = QS$, X, Y, Z also be as in 0.3.5 (see Figure 42). In addition, let U be the first point at which p meets c (on c_1, by definition of c_1) and V the last (on c_2, by 0.3.4). Let $p_1 = QU$; let a_1 be the subarc of c_1 from U to X; let $p_2 = XY$; let a_2 be the subarc of c_2 from Y to V; and let $p_3 = VS$. Then if there is any bounded component $\mathscr{B}'$ of $\mathsf{R}^2 - c$ other than the one $\mathscr{B}$ containing Z, $a = p_1 \cup a_1 \cup p_2 \cup a_2 \cup p_3$ does not meet it. However, a separates P from R by 0.3.4, so the frontier of $\mathscr{B}'$ cannot contain both P and R. But then the frontier of $\mathscr{B}'$ lies in an *arc* of c, which is impossible by 0.3.7. □

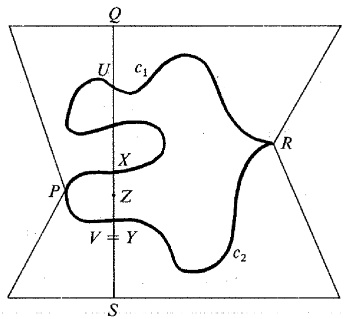


Figure 42

EXERCISE 0.3.8.1. Show that a simple closed curve on S^2 separates it into two components, and that a simple curve in R^2 which goes to infinity at both ends (make a suitable definition of this) separates R^2 into two components.

0.3.9 The Jordan–Schoenflies Theorem

To strengthen the Jordan curve theorem we might ask whether the inside of a Jordan curve c is in fact a topological disc. This was already suggested by the Riemann mapping theorem of Riemann 1851, though not actually proved until Schoenflies 1906, 1908. In fact one has the even stronger result that a Jordan curve can be mapped onto a circle by a homeomorphism of the whole of $\mathbf{R}^2$, and any homeomorphism of $\mathbf{R}^2$ is isotopic to either the identity or a reflection. Thus there is only one way to embed an $\mathbf{S}^1$ in $\mathbf{R}^2$ from the topological point of view.

The proofs of these theorems can be obtained with machinery similar to that used above, but rather than take up more space we simply refer the reader to Moise 1977.

In higher dimensions this extension of the Jordan curve theorem breaks down, in particular for an $\mathbf{S}^2$ in $\mathbf{R}^3$. It remains true that the sphere separates the space into two components, but neither need be homeomorphic to the components obtained with the standard embedding. Thus there are topologically distinct embeddings of $\mathbf{S}^2$ in $\mathbf{R}^3$. Some of these will be studied in 4.2.6.

0.4 Algorithms

0.4.1 Algorithmic Problems

Strictly speaking, it is more logical to define the notion of "algorithm" before we do anything else, but to understand the purpose of algorithms one needs to know the kind of problems they are intended to solve. Typical algorithmic problems are

(i) Decide whether a natural number n is prime.
(ii) Decide whether an algebraic function f is integrable in terms of elementary functions
(iii) Decide whether two schemata Σ_1, Σ_2 define homeomorphic simplicial complexes.

Each of these problems consists of an infinite set of questions which can be effectively enumerated as finite expressions (*words*) in some finite alphabet. For example (i) is

$$\{\text{Is 1 prime?, Is 2 prime?, Is 3 prime?}, \ldots\}$$

and its questions, like those of the other two, can be expressed in the alphabet of the ordinary typewriter keyboard. The purpose of an algorithm is to answer the questions in a systematic, mechanical way.

An algorithm is therefore a computer, the first general definition of which was given by Turing 1936 (and independently by Post 1936). The Turing formulation, now known as the *Turing machine*, illustrated in Figure 43, involves a finite alphabet $\mathscr{A} = \{\text{blank}, S_1, \ldots, S_m\}$, a finite set of internal states $\mathscr{Q} = \{q_1, \ldots, q_n\}$, a *read/write head* and a *tape*. The tape is infinite and divided into squares, each of which can carry a single alphabet symbol. Only finitely many squares are nonblank at any time, and the initial tape

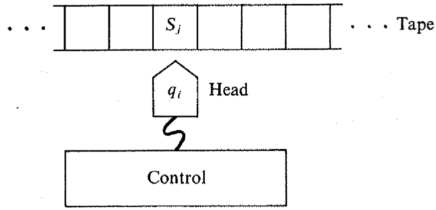


Figure 43

expression (*input*) represents the question being asked. The read/write head begins on the leftmost symbol of the question, and it is directed to perform a sequence of *atomic acts* at unit time intervals by the internal control, which is initially in state q_1. An atomic act is uniquely determined by the pair

$$(q_i, S_j) = (\text{current internal state, scanned symbol})$$

and it is of three possible types:

(a) Replace S_j by S_k, move one square to the right, go into state q_l, or
(b) Replace S_j by S_k, move one square to the left, go into state q_l, or
(c) Halt.

The set $\mathscr{R}$ of responses to the (q_i, S_j) situations possible for a given $\mathscr{A}$ and $\mathscr{Q}$ therefore completely determines the behaviour of the machine, and we may identify a machine M with a list of such responses, or more precisely with a triple $(\mathscr{A}, \mathscr{Q}, \mathscr{R})$ where $\mathscr{R}: \mathscr{A} \times \mathscr{Q} \to \mathscr{A} \times \{\text{left, right}\} \times \mathscr{Q} \cup \{\text{halt}\}$ is the response function.

M answers the input question after a finite number of atomic acts by halting on some specified expression, say 1 for "yes" and 0 for "no." *M solves* a problem by correctly answering all questions in it.

Experience and certain theoretical arguments (Turing 1936) suggest that Turing machines precisely capture the notion of algorithm, but there is no question of proof since we are trying to formalize an informal notion. Thus when we claim that a problem is unsolvable, the statement actually proved is that no Turing machine solves the problem.

The formal notion was slow to appear because until the twentieth century it was taken for granted that algorithms existed, the only problem was to find them. As Hilbert put it: "We must know! We shall know!" If one expects to find an algorithm there is no need to define the class of all algorithms—this is necessary only if *nonexistence* is to be proved.

Perhaps the first to claim nonexistence of an algorithm was Tietze (Tietze 1908, p. 80), who said of finitely presented groups: "*Die Frage, ob zwei Gruppen isomorph sein,* [*ist*] *nicht allgemein lösbar*" (The question whether two groups are isomorphic is not generally solvable). This problem, which arose from the homeomorphism problem (see 0.5.1) was eventually proved unsolvable by Rabin 1958.

The main objective of the present book is to find algorithms to solve topological problems, so the formal theory of unsolvability is left to the final chapter. However, along the way we shall point out where algorithms are unknown, or known to be nonexistent, and also indicate the reasons why unsolvability occurs in topology.

0.4.2 Recursively Enumerable Sets

The Turing machine concept formalizes all computational notions, including the notion of effective enumeration we used to define algorithmic problems. The most convenient procedure is to subsume all notions under that of the *partial recursive function* (p.r. function): viewing the input to a machine M as the argument x of a function, the expression on the tape when (and if) M halts is taken as the function value $\phi_M(x)$. Since M need not halt for all inputs, ϕ_M is generally only a "partial" function.

An algorithm may then be defined as a p.r. function ϕ_M whose domain is a set of questions Q, and such that

$$\phi_M(Q) = \begin{cases} 1 & \text{if the answer to } Q \text{ is "yes,"} \\ 0 & \text{if the answer to } Q \text{ is "no."} \end{cases}$$

Thus a *problem* is just the domain of a p.r. function.

This glib definition does not seem to be what we originally had in mind, so let us see why the domain of a p.r. function ϕ_M can indeed be effectively enumerated. Observe first of all that the words in the alphabet of M can be effectively enumerated as $w_1, w_2, \ldots$; first list the one-letter words, then the two-letter words, and so on. We can similarly enumerate all computations of M; first the one-step computations on input words of length 1, then the two-step computations on input words of length ≤ 2, and so on; and whenever a w_n is found to lead to a halting computation, we place it on another list $\mathscr{L}$. Then $\mathscr{L}$ is an effective enumeration of the domain of ϕ_M. The domain of a p.r. function is called a recursively enumerable set (r.e. set).

Although the complement of an r.e. set with respect to the set of all words is enumerable, it need not be recursively enumerable. This remarkable fact can even be illustrated by a natural example—the homeomorphism problem for finite complexes.

The set of all pairs (Σ_i, Σ_j) of combinatorially homeomorphic schemata can be recursively enumerated by a machine which systematically tries all elementary subdivisions, halting when isomorphic refinements of Σ_i, Σ_j are obtained. However, we cannot recursively enumerate the complement of this set, as this would yield a recursive enumeration of *non*-combinatorially homeomorphic pairs. Such an enumeration is unknown, and with good reason—by enumerating both the set and its complement until we found a given pair (Σ_i, Σ_j) we could decide whether Σ_i, Σ_j were combinatorially homeomorphic, whereas this problem is known to be unsolvable (Markov 1958).

0.4.3 The Diagonal Argument

The diagonal argument was first used by du Bois-Reymond 1874 and Cantor 1891 to show that certain collections of objects could not be enumerated. Despite the negative conclusions drawn from the argument, it is in fact highly constructive, and this makes it equally suitable for proving nonexistence of *effective* enumerations.

For example, Cantor proves that one cannot enumerate all sets $\mathscr{S}$ of natural numbers. Namely, any enumeration $\mathscr{S}_0, \mathscr{S}_1, \mathscr{S}_2, \ldots$ fails to include the "diagonal set"

$$\mathscr{D} = \{n : n \notin \mathscr{S}_n\}$$

because $\mathscr{D}$ differs from the set $\mathscr{S}_n$ with respect to the number n ($n \in \mathscr{S}_n \Rightarrow n \notin \mathscr{D}, n \notin \mathscr{S}_n \Rightarrow n \in \mathscr{D}$).

Cantor needs $\mathscr{D}$ only as a counterexample to an assumed enumeration of all $\mathscr{S}$, but its nature becomes more interesting when we have a specific list of sets $\mathscr{S}_0, \mathscr{S}_1, \mathscr{S}_2, \ldots$. In particular, we can obtain an effective enumeration of the r.e. sets by effectively enumerating the descriptions of Turing machines, and $\mathscr{D}$ is then a specific non-r.e. set. The surprise is that

$$\mathscr{K} = \text{complement of } \mathscr{D} = \{n : n \in \mathscr{S}_n\}$$

is r.e.! One performs a giant computation which looks at each step of each Turing machine computation, and whenever a number n is found in $\mathscr{S}_n$, it is placed in $\mathscr{K}$.

Thus the effectivized diagonal argument yields an r.e. set $\mathscr{K}$ of natural numbers whose complement is *not* r.e.. Then there is no algorithm for deciding membership of $\mathscr{K}$, since any algorithm would immediately yield a recursive enumeration of the complement of $\mathscr{K}$.

$\mathscr{K}$ is the direct source of all known unsolvability results in mathematics. Such results are obtained by showing that Turing machines can be simulated by various mathematical systems, and then showing that solutions of certain problems in these systems would imply algorithms for deciding membership of $\mathscr{K}$. The first, and most direct simulation was obtained by Post 1947 by means of finitely presented semigroups. Post was able to conclude from this that the word problem for semigroups (see 0.5.7) was unsolvable. Novikov 1955 did the same for groups, and this paved the way for the unsolvability of the homeomorphism problem proved by Markov 1958. See Chapter 9.

All these problems inherit the asymmetric character of $\mathscr{K}$—the "yes" answers can be effectively enumerated, but not the "no" answers. The process of enumerating the "yes" cases is sometimes called a *semidecision procedure*, and it invariably consists of searching systematically through computations until one with the desired result is found. The enumeration of homeomorphic pairs (Σ_i, Σ_j) in 0.4.2 is an example, and others will be mentioned in 0.5.8.

0.5 Combinatorial Group Theory

0.5.1 The Fundamental Group

In 0.1.9 we sketched the ideas of *product* and *inverse* for closed curves in a complex $\mathscr{C}$. Given a fixed origin P for the curves (the choice of which is arbitrary if $\mathscr{C}$ is arc connected), we call curves p, p' *equivalent* if there is a homotopy between them which leaves P fixed. The equivalence class of p is denoted by $[p]$. Then the natural product $\cdot$ for curves (concatenation) extends in a well-defined way to equivalence classes by

$$[p_1]\cdot[p_2] = [p_1p_2]$$

and this product inherits the obvious associativity of concatenation. There is an identity element 1, represented by the "point path" P, and most importantly

$$[p]\cdot[p^{-1}] = [p\cdot p^{-1}] = 1$$

so that $[p^{-1}]$ is the inverse $[p]^{-1}$ of $[p]$. The homotopy between pp^{-1} and P which proves this is suggested by Figure 44.

Thus we have a group $\pi_1(\mathscr{C})$ of equivalence classes of closed paths in $\mathscr{C}$, called the *fundamental group* (Poincaré 1895). A rigorous construction of π_1 is given in 3.1. It is clear that the fundamental group is invariant under homeomorphisms, so one way to prove $\mathscr{C}_1$ and $\mathscr{C}_2$ are nonhomeomorphic is to prove $\pi_1(\mathscr{C}_1) \neq \pi_1(\mathscr{C}_2)$.

This idea is made feasible by the fact that we can read off a "finite presentation" of $\pi_1(\mathscr{C})$ from a finite cell decomposition of $\mathscr{C}$. The method is to use the finitization process sketched in 0.1.2 to deform all paths onto the 1-skeleton of $\mathscr{C}$. All paths in the 1-skeleton are homotopic to products of finitely many *generating paths* $a_1, \ldots, a_m$; hence $a_1, \ldots, a_m$ generate all

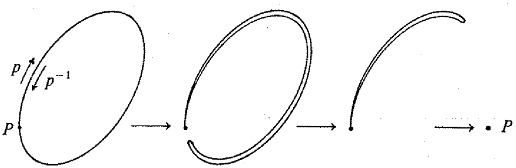


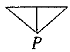
Figure 44

paths in $\mathscr{C}$, up to homotopy. Any homotopy between paths then reduces to a series of *elementary deformations* across the cells $\Delta_1, \ldots, \Delta_n$ in the 2-skeleton of $\mathscr{C}$. If we let $r_j(a_i)$ denote the null homotopic path which runs from P to Δ_j, round $\partial\Delta_j$ and back to P, then the relations

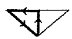
$$r_j(a_i) = 1$$

completely determine $\pi_1(\mathscr{C})$. The details of this construction are carried out rigorously in 3.2–3.4 and 4.1.

EXERCISE 0.5.1.1. Show that any closed edge path with origin P in the 1-complex



is homotopic to a product of $a_1^{\pm 1}$, $a_2^{\pm 1}$, where

$a_1 =$  , $a_2 =$ 

0.5.2 Generators, Words, and Relations

The above sketch of the fundamental group is intended to suggest that the proper way to view groups in combinatorial topology is in terms of "generators" and "relations." We shall now drop the topological interpretation and discuss generators and relations in purely combinatorial terms.

A *generator* is a letter a_i, and it has a formal inverse a_i^{-1}. A *word* is any finite sequence

$$a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_k}^{\varepsilon_k}$$

of generators or their inverses, so that each $\varepsilon_j = \pm 1$, where a_i^{+1} denotes a_i.

The *product* $w_1 w_2$ of words w_1, w_2 is the concatenation of the corresponding sequences, in other words, the result of writing w_1, then w_2. Since concatenation is trivially associative, this is an associative product.

We abbreviate the product $a_i a_i \cdots a_i$ (n factors) by a_i^n. The empty word is denoted by 1, so that $1w = w1 = w$ for any word w.

A *relation* is an equation $r = 1$ where r is a word (called a *relator* in this context), and words w, w' are called *equivalent* with respect to relations $r_j = 1$ if w is convertible to w' by a finite sequence of operations of the following types

(i) insertion or deletion of a subword r_j,
(ii) insertion or deletion of a subword $a_i a_i^{-1}$ or $a_i^{-1} a_i$.

The relations $a_i a_i^{-1} = a_i^{-1} a_i = 1$ implicit in (ii) are called *trivial relations*, and words equivalent by trivial relations alone are called *freely* equivalent.

The equivalence class of w (with respect to a fixed set of relations) is denoted by $[w]$, and we extend the product operation to equivalence classes by setting

$$[w_1] \cdot [w_2] = [w_1 w_2].$$

This product is well-defined, for if w_1' is equivalent to w_1, then $w_1' w_2$ is equivalent to $w_1 w_2$, since the operations which convert w_1 to w_1' are indifferent to the presence of w_2 concatenated on the right. Thus the product is independent of the choice of representative for the first factor, and similarly for the second factor.

0.5.3 Group Presentations

The structure $\langle a_1, a_2, \ldots; r_1, r_2 \ldots \rangle$ of equivalence classes of words in the a_i with respect to the relations r_j, under the product operation, is a group G.

The product in G inherits associativity from the associativity of concatenation:

$$[w_1]([w_2][w_3]) = [w_1][w_2 w_3] = [w_1 w_2 w_3] = ([w_1][w_2])[w_3].$$

The identity element is $[1]$ because

$$[1][w] = [1w] = [w] = [w][1]$$

and the inverse of $[w]$ exists, because if

$$w = a_{i_1}^{\varepsilon_1} \cdots a_{i_k}^{\varepsilon_k}$$

and we set $w^{-1} = a_{i_k}^{-\varepsilon_k} \cdots a_{i_1}^{-\varepsilon_1}$ it is clear that ww^{-1} is freely equivalent to 1, so

$$[w][w^{-1}] = [ww^{-1}] = [1]$$

and we can let $[w]^{-1} = [w^{-1}]$. □

We usually drop the equivalence class brackets and simply speak of the element w of G. This has the same advantages as speaking of the "rational number $\frac{1}{2}$" when we really mean the rational number $\{\frac{1}{2}, \frac{2}{4}, \frac{3}{6}, \ldots\}$.

The expression $\langle a_1, a_2, \ldots; r_1, r_2, \ldots \rangle$ is called a *presentation* of G. Of course, a group G has many presentations, but we do not distinguish a presentation from G itself except to point out properties of G which are evident from some presentations but not from others. For example, G is *finitely presented* if it has a presentation in which the sets $\{a_i\}$ and $\{r_j\}$ are finite. Some finite presentations of well-known groups are

$$\mathbf{Z} = \langle a; - \rangle$$

$$\mathbf{Z}_2 = \langle a; a^2 \rangle$$

$$\mathbf{Z}_2 \times \mathbf{Z}_3 = \langle a, b; a^2, b^3, aba^{-1}b^{-1} \rangle.$$

When a relation $r_j = 1$ is written more naturally as some other equation $r'_j = r''_j$ we sometimes write this equation in the presentation in place of r_j. For example, $\mathbf{Z}_2 \times \mathbf{Z}_3$ is better expressed as $\langle a, b; a^2, b^3, ab = ba\rangle$.

The theory of groups in terms of generators and relations is largely self-contained, however it is sometimes useful to interpret relations more conventionally in terms of normal subgroups and quotients. We now review these notions.

EXERCISE 0.5.3.1. Does the trivial group $\{1\}$ have a presentation? Does every group have a presentation?

0.5.4 Coset Decomposition, Normal Subgroups

If H is a subgroup of G the sets

$$Hg = \{hg: h \in H\}$$

for $g \in G$ are called *right cosets* of G modulo H. They constitute a partition of G, called the *right coset decomposition*, because if Hg_1, Hg_2 have a common element

$$h_1 g_1 = h_2 g_2$$

then

$$g_2 g_1^{-1} = h_2^{-1} h_1 \in H$$

in which case

$$H = Hg_2 g_1^{-1}$$

and

$$Hg_1 = Hg_2.$$

Thus cosets are either equal or disjoint.

H is called *normal* if

$$gHg^{-1} = H \quad \text{for each } g \in G,$$

where gH and gHg^{-1} are defined in the obvious analogy to Hg. Normal subgroups are characterised by the following proposition.

Any normal subgroup N of G is of the form $N(v_1, v_2, \ldots)$ consisting of elements expressible by words

$$(*) \qquad \prod_{k=1}^{l} g_k v_{j_k}^{\varepsilon_k} g_k^{-1} \qquad g_k, v_j \in G$$

and called the normal subgroup generated by $v_1, v_2, \ldots \in G$.

$N(v_1, v_2, \ldots)$ is a subgroup because products and inverses of words of the form (*) are again of this form. To show normality, let

$$x = \Pi g_k v_{j_k}^{\varepsilon_k} g_k^{-1}$$

so

$$\begin{aligned} gxg^{-1} &= g(\Pi g_k v_{j_k}^{\varepsilon_k} g_k^{-1})g^{-1} \\ &= \Pi g(g_k v_{j_k}^{\varepsilon_k} g_k^{-1})g^{-1} \\ &= \Pi(gg_k)v_{j_k}^{\varepsilon_k}(gg_k)^{-1} \in N(v_1, v_2, \ldots). \end{aligned}$$

Hence $gNg^{-1} \subseteq N$ and by repeating the argument with g^{-1} in place of G we get $g^{-1}Ng \subseteq N$. But

$$g^{-1}Ng \subseteq N \quad \Rightarrow \quad Ng \subseteq gN \quad \Rightarrow \quad N \subseteq gNg^{-1}$$

which is the reverse containment, hence $gNg^{-1} = N$.

Conversely, any normal subgroup N of G is of the form $N(v_1, v_2, \ldots)$. Namely, let $v_1, v_2, \ldots$ be all the elements of N. Then $v_j^{\varepsilon_j} \in N \Rightarrow g_k v_j^{\varepsilon_j} g_k^{-1} \in N$ (by normality) $\Rightarrow \Pi g_k v_j^{\varepsilon_j} g_k^{-1} \in N$ since N is a subgroup, and any element $v_j \in N$ is a trivial product of this form. □

The proposition may be interpreted as saying that the operations required to generate a normal subgroup N of G from an arbitrary set $\{v_1, v_2, \ldots\} \subseteq G$ are inverses, products, and conjugates by arbitrary elements of G. Thus when we speak of generating a normal subgroup we include the operation of conjugation, in contrast to generating a group which requires only inverses and products.

0.5.5 Quotient Groups and Homomorphisms

The cosets of G modulo a normal subgroup N are made into a group G/N by setting

$$Ng_1 \cdot Ng_2 = Ng_1g_2.$$

The group properties are inherited from G and we only have to show that the product is well-defined. Any representative of Ng_1 has the form xg_1 for some $x \in N$, and if we use xg_1 instead of g_1 the product is

$$\begin{aligned} N(xg_1)g_2 &= (Nx)g_1g_2 \\ &= Ng_1g_2 \end{aligned}$$

since $Nx = N$ because $x \in N$. If we use another representative xg_2 of Ng_2 we get

$$\begin{aligned} Ng_1(xg_2) &= g_1Nxg_2 && \text{by normality} \\ &= g_1Ng_2 && \text{since } Nx = N \\ &= Ng_1g_2 && \text{by normality.} \end{aligned}$$

Thus the product is independent of the choice of representatives. □

$\phi: G \to G/N$ defined by $\phi(g) = Ng$ is an example of a *homomorphism*, that is,

$$\phi(g_1 g_2) = \phi(g_1)\phi(g_2)$$
$$\phi(1) = 1.$$

It is called the *canonical homomorphism* of G onto G/N. The *kernel* of ϕ, $\ker(\phi) = \{g: \phi(g) = 1\}$, is equal to N.

Conversely, if ϕ is any homomorphism of G onto a group G', then $\ker(\phi)$ *is a normal subgroup N of G and $G' = G/N$.*

If $x \in \ker(\phi)$, then $gxg^{-1} \in \ker(\phi)$ also, because

$$\phi(gxg^{-1}) = \phi(g)1\phi(g^{-1}) = \phi(gg^{-1}) = \phi(1) = 1.$$

Thus $gNg^{-1} \subseteq N$, and by the argument of 0.5.4, N is normal. Now

$$\begin{aligned}\phi(x_1) = \phi(x_2) \quad &\Leftrightarrow \quad \phi(x_1)\phi(x_2)^{-1} = 1 \\ &\Leftrightarrow \quad \phi(x_1 x_2^{-1}) = 1 \\ &\Leftrightarrow \quad x_1 x_2^{-1} \in N \\ &\Leftrightarrow \quad x_1 \in N x_2\end{aligned}$$

if and only if x_1, x_2 are in the same coset of G mod N. Thus distinct elements of G' correspond to distinct cosets of G modulo N, in fact $\phi(x)$ corresponds to Nx, so ϕ can be identified with the canonical homomorphism: $G \to G/N$, and hence G' with G/N. □

This proposition yields the standard test for ϕ to be an *isomorphism* (one-to-one homomorphism), namely $\ker(\phi) = \{1\}$. We have constructed an isomorphism between G' and G/N in effect by setting N, the kernel of ϕ, equal to 1.

Unless we are interested in the isomorphism itself we do not distinguish between corresponding elements in isomorphic groups. For example, if there is an isomorphism of G *into* G' (a *monomorphism* or *embedding*) we are likely to say G *is* a subgroup of G'. The kind of isomorphism most likely to interest us is one from a group onto itself, called an *automorphism* (see for example 7.1). The automorphisms of a group G themselves constitute a group, under composition, called the *automorphism group* of G.

0.5.6 Dyck's Theorem (Dyck 1882)

$G = \langle a_1, a_2, \ldots; r_1, r_2, \ldots \rangle$ is the quotient of $F = \langle a_1, a_2, \ldots; — \rangle$ (called the free group *on generators $a_1, a_2, \ldots$) by its normal subgroup $N(r_1, r_2, \ldots)$.*

Since the function $\phi: F \to G$ which sends an element of F to its equivalence class in G is clearly a homomorphism, with kernel equal to the set of words equivalent to 1, it will suffice to show that the kernel equals $N(r_1, r_2, \ldots)$.

Certainly, any $\Pi g_k r_{j_k}^{\varepsilon_k} g_k^{-1} \in N$ is equal to 1 in G, since each $r_j = 1$. Conversely, suppose a word $w = 1$ in G. We shall show that each insertion or deletion of $r_j^{\pm 1}$ in w can be accomplished by multiplying w by $g_k^{-1} r_j^{\pm 1} g_k$ for some g_k.

Note firstly that deletion of $r_j^{\pm 1}$ can always be accomplished by insertion of $r_j^{\mp 1}$ next to it, followed by cancellation (which is valid in F). Thus it remains to deal with insertions.

Let $w = uv \to ur_j v$ be the insertion of r_j between the factors u, v of w. We can obtain the same result by multiplying w by $v^{-1} r_j v$, since $ur_j v$ is freely equivalent to $uv \cdot v^{-1} r_j v$.

Repetition of this process for each insertion in the sequence required to convert w to 1 gives a word

$$w \Pi g_k^{-1} r_{j_k}^{\varepsilon_k} g_k$$

which is freely equivalent to 1, and therefore

$$w = \Pi g_k r_{j_k}^{-\varepsilon_k} g_k^{-1} \quad \text{in } F$$

so that $w \in N(r_1, r_2, \ldots)$. □

Dyck's 1882 paper is the beginning of combinatorial group theory as a subject, and the first to recognize the fundamental role of free groups. Dyck viewed free groups as the most general groups, since any other group is obtainable by imposing relations on them. The explanation of relations in terms of normal subgroups and quotients suggests a reconstruction of combinatorial group theory in more conventional algebraic terms. This can indeed be done, including the definition of free groups themselves, but it proves to be an object lesson in the impotence of abstract algebra. All substantial theorems in combinatorial group theory still require honest toil with words and relations, and the best labour-saving device turns out to be the topological interpretation of 0.5.1, rather than algebra.

Exercise 0.5.6.1. If G is any group show that the result of adding relations $v_1 = 1$, $v_2 = 1, \ldots$, to G is $G/N(v_1, v_2, \ldots)$.

0.5.7 The Word Problem and Cayley Diagrams

When a group G arises as a fundamental group, as in 0.5.1, null-homotopic paths correspond to words w which equal 1 in G. Thus the problem of deciding null-homotopy (contractibility to a point) is reduced to deciding whether a given word $w = 1$ in G. Even though we can compute a presentation of G, this problem is not trivial, and its fundamental importance for topology and group theory was first recognized by Dehn 1910, who called it the *word problem*.

Early topologists, such as Poincaré, Tietze, and Reidemeister, frequently commented on the difficulty of group-theoretic problems in topology, on occasion (Reidemeister) saying that the fundamental group seemed merely to translate hard topological problems into hard group-theoretic problems. This pessimism was vindicated when Novikov 1955 proved that the word problem (for specific, finitely presented G) was unsolvable. Novikov's proof is based on the idea of Post 1947 of simulating Turing machines by systems of generators and relations. A word corresponds (roughly) to the tape expression on a Turing machine M, and the relations permit the word to be changed to reflect the atomic acts of M. (The technical difficulty, which is absent in the semigroup case, is the presence of relations $a_i a_i^{-1} = a_i^{-1} a_i = 1$ which do not correspond to acts of M. See Chapter 9.)

Solution of the word problem for G is equivalent to the construction of a figure $\mathscr{C}_G$ called the *Cayley diagram* of G, introduced for finite groups by Cayley 1878 and for infinite groups by Dehn 1910. If G is generated by $a_1, a_2, \ldots$, then $\mathscr{C}_G$ is a graph with a vertex P_g for each distinct $g \in G$ and an oriented edge labelled a_i from P_g to P_{ga_i} for each generator a_i. It follows that each vertex has exactly one outgoing, and one incoming, edge for each generator. Examples (labelling each vertex g instead of P_g for simplicity) are given in Figure 45. The last example is constructed by noting that there are six distinct elements $g = 1, b, b^2, a, ab, ab^2$, then multiplying each of these by a, b and using the defining relations to reduce each product to one of the six forms already chosen.

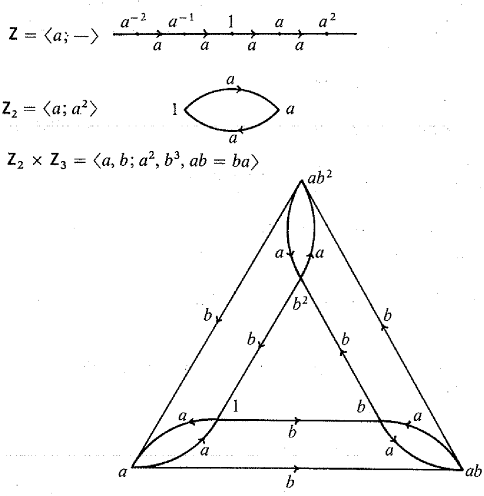


Figure 45

Each word

$$w = a_{i_1}^{\varepsilon_1} \cdots a_{i_k}^{\varepsilon_k}$$

determines a path from P_1 to P_w by following the labels $a_{i_1}, \ldots, a_{i_k}$ in succession, with or against the arrow according as the exponent ε is $+1$ or -1. It follows that $w = 1$ if and only if the path is closed.

Thus if $\mathscr{C}_G$ can be effectively constructed we have a solution of the word problem for G.

Conversely, if the word problem for G can be solved, we can construct $\mathscr{C}_G$.

Effectively list the words of G as $w_1, w_2, \ldots$ and as each w_j appears, use the solution of the word problem to decide whether $w_j =$ any w_i earlier on the list (see if $w_i w_j^{-1} = 1$). If not, put w_j on a *second list*. The second list is then an effective enumeration of the distinct elements of G, which we use as labels for the vertices of $\mathscr{C}_G$.

As each vertex P_{w_j} is constructed, we again use the solution of the word problem to find which of the words $w_j a_i$ is equivalent to a w_k already on the second list (if an equivalent is not found, one will be found later by repeated checking as the second list grows). For each such word we construct an oriented edge labelled a_i from P_{w_j} to $P_{w_j a_i} = P_{w_k}$. This is an effective process which eventually gives each vertex and edge in $\mathscr{C}_G$. □

Since G has many different presentations, $\mathscr{C}_G$ is not unique. However, if there is a solution to the word problem for one finite presentation of G there is a solution for any other finite presentation of G, hence the effective constructibility of $\mathscr{C}_G$ does not depend on the presentation chosen.

EXERCISE 0.5.7.1. Prove the last remark.

EXERCISE 0.5.7.2. Show that $\{w: w = 1 \text{ in } G\}$ is r.e. when G is finitely presented.

EXERCISE 0.5.7.3. Sketch the Cayley diagram of the free group $F_2 = \langle a, b; - \rangle$.

EXERCISE 0.5.7.4. Describe the Cayley diagrams of the *free abelian groups* $\mathsf{Z} \times \mathsf{Z} \times \ldots \times \mathsf{Z} = \langle a_1, \ldots, a_m; a_i a_j = a_j a_i \ (i, j \le m)\rangle$ as figures in R^m.

EXERCISE 0.5.7.5. Figure 46 shows the Cayley diagram of a group. Why is this group nonabelian?

Show that the group is the group of symmetries of an equilateral triangle.

0.5.8 Tietze Transformations

Tietze transformations are simply the obvious ways of transforming a finite presentation $\langle a_1, \ldots, a_m; r_1, \ldots, r_n\rangle$.

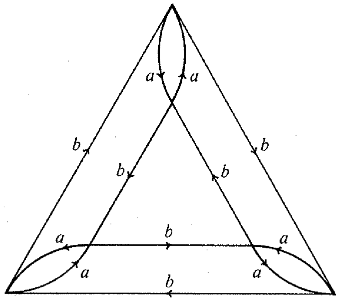


Figure 46

T_1: Add a relation $r_{n+1}(=1)$ which is a consequence of $r_1, \ldots, r_n$. (That is, r_{n+1} is equivalent to 1 with respect to the relations $r_1 = \cdots = r_n = 1$. We write this $r_1, \ldots, r_n \Rightarrow r_{n+1}$.)

T_2: Add a generator a_{m+1} together with a relation

$$a_{m+1} = w(a_1, \ldots, a_m)$$

which defines it as a word in the old generators.

The inverse transformations, which we denote by T_1^{-1}, T_2^{-1}, can also be applied when meaningful.

Tietze's Theorem. *Any two finite presentations of a group G are convertible into each other by a finite sequence of Tietze transformations.*

Suppose G has presentations $\langle a_1, \ldots, a_m; r_1, \ldots, r_n\rangle$ and $\langle a'_1, \ldots, a'_{m'}; r'_1, \ldots, r'_{n'}\rangle$, which we abbreviate to $\langle a_i; r_j(a_i)\rangle$ and $\langle a'_i; r'_j(a'_i)\rangle$. We use the notation $w(x_i)$ to express the fact that w is a word in the letters x_i, and denote the result of substituting a word χ_i for x_i in $w(x_i)$ by $w(\chi_i)$.

Since both presentations denote the same group, there are words α'_i in $a_1, a_2, \ldots$ representing the a'_i and hence satisfying the relations $r'_j(\alpha'_i)$. Then the $r_j(a_i) \Rightarrow$ the $r'_j(\alpha'_i)$ since *all* relations in the a_i are consequences of the $r_j(a_i)$. Similarly there are words α_i in $a'_1, a'_2, \ldots$, representing the a_i, and the $r'_j(a'_i) \Rightarrow$ the $r_j(\alpha_i)$.

We can therefore make the following modifications of the group presentation by Tietze transformations:

$$\begin{aligned}
&\langle a_i; r_j(a_i)\rangle \\
&\to \langle a_i; r_j(a_i), r'_j(\alpha'_i)\rangle \quad \text{by} \quad T_1 \text{ since the } r_j(a_i) \Rightarrow \text{the } r'_j(\alpha'_i) \\
&\to \langle a_i, a'_i; r_j(a_i), r'_j(\alpha'_i), a'_i = \alpha'_i\rangle \quad \text{by } T_2 \\
&\to \langle a_i, a'_i; r_j(a_i), r'_j(a'_i), r_j(\alpha'_i), a'_i = \alpha'_i\rangle \quad \text{by } T_1 \\
&\to \langle a_i, a'_i; r_j(a_i), r'_j(a'_i), a'_i = \alpha'_i\rangle \quad \text{by } T_1^{-1} \\
&\to \langle a_i, a'_i; r_j(a_i), r'_j(a'_i), a'_i = \alpha'_i, a_i = \alpha_i\rangle \quad \text{by } T_1 \qquad (*)
\end{aligned}$$

since the relations $a_i = \alpha_i$ are true in the group and hence consequences of the relations already present. But (*) is symmetric with respect to primed and unprimed symbols, so it could equally well be obtained from $\langle a_i'; r_j'(a_i')\rangle$. By reversing the latter derivation we obtain

$$\langle a_i; r_j(a_i)\rangle \to (*) \to \langle a_i'; r_j'(a_i')\rangle. \qquad \square$$

Since we can effectively enumerate all consequences of a given finite set of relations, and hence all possible sequences of Tietze transformations which can be applied to a given presentation, Tietze's theorem shows that we can effectively enumerate all finite presentations of a given group. Thus the problem of deciding when two presentations are the same, the *isomorphism problem* of Tietze 1908, is similar to the word problem—in both cases we can effectively enumerate the pairs of equal objects, and the difficulty is to find the pairs of unequal objects. It actually follows from basic results of recursive function theory (see Rogers 1967) that the two problems are of the same degree of unsolvability, that is, a solution of one would effectively yield a solution of the other. (In particular, the isomorphism problem is unsolvable.) In individual cases, however, the isomorphism problem is usually harder to solve than the word problem.

On the positive side, the Tietze theorem is often a slick way to prove existence of algorithms or semidecision procedures. For example, if G has a property that can be recognized from one of its presentations we can eventually verify this property by enumerating all the presentations of G. Examples of such properties are:

(i) being abelian (all generators commute)
(ii) being finite (all relations of the form $a_i a_j = a_k$)
(iii) being a specific finite group (relations given by multiplication table)
(iv) being free (no relations).

EXERCISE 0.5.8.1. Show that $\langle a, b; abab^{-1}\rangle = \langle c, d; c^2d^2\rangle$.

EXERCISE 0.5.8.2. Suppose that *infinitely many* consequence relations or new generators can be added in a transformation of type T_1 or T_2 respectively. Deduce that any two presentations of the same group are then convertible to each other by a finite sequence of Tietze transformations.

EXERCISE 0.5.8.3. Give an algorithm for finding α_i and α_i' from two presentations $\langle a_i; r_j(a_i)\rangle$ and $\langle a_i'; r_j'(a_i')\rangle$ of the same group. (This gives a "uniform" solution to Exercise 0.5.7.1.)

EXERCISE 0.5.8.4. If G has a finite presentation, show that in any presentation

$$G = \langle a_1, \ldots, a_n; r_1, r_2, \ldots\rangle$$

all but a finite number of relations are superfluous.

0.5.9 Coset Enumeration

As a final example of the way finiteness can be discovered by systematically enumerating words, consider the case of a subgroup H of a finitely presented group G. If the set of cosets Hg for $g \in G$ is finite, H is said to be of *finite index* in G. In this case there is a finite set $\{g_1, \ldots, g_k\}$ of *coset representatives* such that

$$G = Hg_1 \cup \cdots \cup Hg_k.$$

We now show how to find such a set, if one exists.

$G = Hg_1 \cup \cdots \cup Hg_k$ if and only if the set $\{Hg_1, \ldots, Hg_k\}$ is closed under right multiplication by the generators of G and their inverses. That is

$$Hg_j a_i = \text{some } Hg_{j'},\ Hg_j a_i^{-1} = \text{some } Hg_{j''}$$

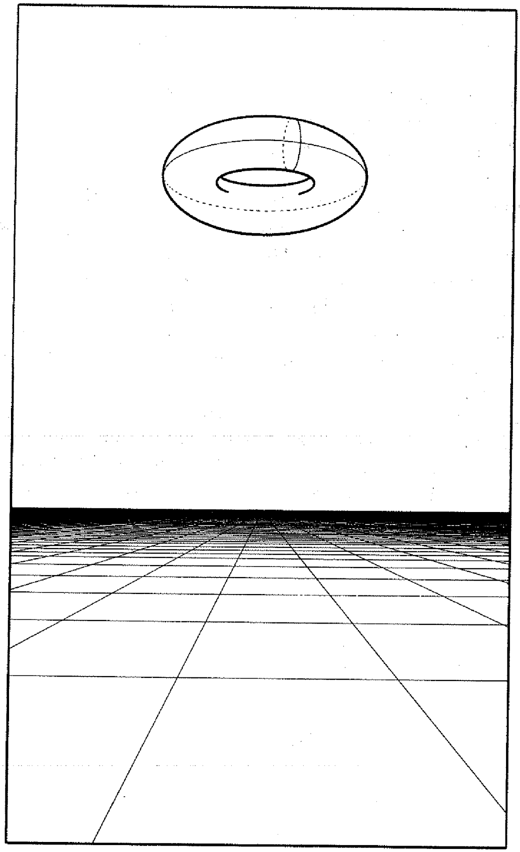
for each generator $a_1, \ldots, a_m$ of G. Now assuming H is effectively enumerable, we can verify the equality of two cosets by enumerating their members, along with an enumeration of equal words in G, until we find a common element.

It therefore suffices to enumerate all the finite sets $\{g_1, \ldots, g_k\}$ in G, and for each one try to verify that $\{Hg_1, \ldots, Hg_k\}$ is closed under right multiplication by looking for equal pairs $Hg_j a_i$, $Hg_{j'}$ and $Hg_j a_i^{-1}$, $Hg_{j''}$. Eventually such a verification will succeed. □

A more practical version of the above idea is known as the *Todd–Coxeter coset enumeration method* (Todd, Coxeter 1936).

CHAPTER 1

Complex Analysis and Surface Topology



1.1 Riemann Surfaces

1.1.1 Introduction

Topology may have had its tentative beginnings in isolated thoughts of Descartes, Leibniz, and Euler, but it was Riemann who brought the subject into the mainstream of mathematics with his inaugural dissertation in Göttingen in 1851. His introduction of the Riemann surface in that year showed the indispensable rôle of topology in questions of analysis, and thus ensured the future cultivation of the subject by the mathematical community, if only for the service of analysis. In fact, of course, Riemann surfaces were quickly seen to be of interest in themselves, and were the source of two ideas of profound significance in later topology—*connectivity* and *covering spaces.*

It hardly does Riemann justice to present only the topological aspects of his theory, however, limitations of space aside, it may be worthwhile to avoid the heavy burden of analysis found in texts on Riemann surfaces. The next section therefore presents a purely topological notion of Riemann surface, the *branched covering of the sphere.* Just a few words of motivation may be of value before we start.

In complex function theory it is convenient to treat the value ∞ as just another number, as far as possible, and one therefore completes the complex number plane by a *point at infinity.* The completed plane may be viewed as a sphere, since stereographic projection from the north pole N of a sphere resting on the plane at the origin O establishes a continuous one-to-one correspondence between the finite points P' of the plane and the points $P \neq N$ on the sphere. The point N itself is naturally reckoned to correspond to ∞ (Figure 47).

A complex function $w(z)$ can then be viewed as a map of the sphere onto itself, but of course the map need not be one-to-one, even for algebraic functions such as z^2. In a natural sense, $w(z) = z^2$ maps the sphere *twice* onto itself except at O and ∞, since any other value of w is the square of two distinct values $+\sqrt{w}$ and $-\sqrt{w}$. In fact, if we divide the z-sphere into hemispheres

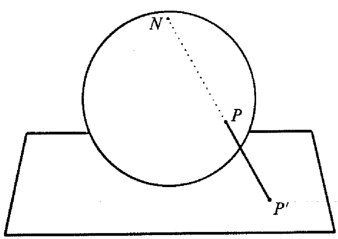


Figure 47

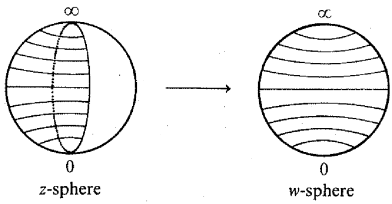


Figure 48

by any meridian (say, the one corresponding to the imaginary axis) then both are mapped onto the whole z-sphere by squaring (Figure 48).

If we were to place the z-sphere so that each point lay above its image on the w-sphere the result would be what we call a 2-*sheeted* cover of the w-sphere with branch points O and ∞. We cover the w-sphere with two spheres, each slit along a meridian, and identify the edges according to the labels a and b shown in Figure 49. The slit spheres are the *sheets*. The points O and ∞ have the property that a small circuit around them on the covering surface is not closed—it passes from one sheet to the other as if on a ramp—nevertheless each of these points has a disc neighbourhood, the perimeter of which is obtained by making *two* circuits around the branch point (Figure 50).

Thus the covering surface is a genuine surface from the topological point of view; unfortunately our psychological need to force the identified edges together in ordinary space, causing a line of intersection, tends to obscure this fact.

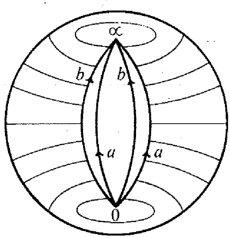


Figure 49

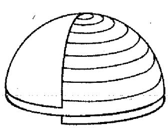


Figure 50

The "two-valued function" $z = \sqrt{w}$ can be viewed as a single-valued function if its domain is taken to be the covering surface instead of the w-sphere—the values in each hemisphere of the z-sphere occur as w moves over each sheet. The general purpose of Riemann surfaces in function theory is to provide domains on which all algebraic functions become single-valued.

1.1.2 Branched Coverings of the 2-sphere

It is easy to see that the Riemann surface for $w = z^2$, and in fact any example with only two branch points, is topologically a sphere. The interest in the theory stems from the fact that topologically different surfaces occur when there are more branch points.

In general a *branched covering* of the 2-sphere is determined by a finite set of branch points $P_1, \ldots, P_m$, a sheet number n, and specification of the way the sheets join up around the branch points. To do this, each branch point P_i is associated with a permutation

$$\pi_i = (k_1, k_2, \ldots, k_n)$$

of the integers $1, 2, \ldots, n$, the interpretation of which is that an anticlockwise circuit of P_i starting on sheet j ends on sheet k_j. These permutations must satisfy a consistency condition which is obtained as follows.

Take a point $P \neq P_1, \ldots, P_m$ and not on the same great circle as any pair P_i, P_j (when P_i, P_j are antipodal choose the great circle joining them arbitrarily) and connect it to $P_1, \ldots, P_m$ by great circle arcs $a_1, \ldots, a_m$ (Figure 51).

We cut the sheets along these arcs and identify their edges according to the permutations π_i in order to form the covering. Then since P is not a branch point, any circuit around P must begin and end on the same sheet. Assuming the subscripts are chosen so that $P_1, \ldots, P_m$ is the order of branch points around P, the permutation determined by a circuit around P is $\pi_1\pi_2 \cdots \pi_m$, and hence the consistency condition is

$$\pi_1\pi_2 \cdots \pi_m = 1$$

where 1 denotes the identity permutation.

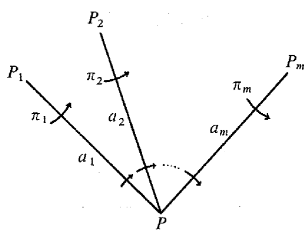


Figure 51

(This equation says that $\pi_2, \ldots, \pi_m$ are arbitrary, which corresponds to the fact that loops around $P_2, \ldots, P_m$ are not related by any nontrivial homotopies on $\mathsf{S}^2 - \{P_1, \ldots, P_m\}$. In the language of Chapter 4, such loops freely generate the fundamental group of $\mathsf{S}^2 - \{P_1, \ldots, P_m\}$.)

If the surface is to be connected we need $\pi_1, \ldots, \pi_m$ to generate a permutation which sends i to j for each $i, j \leq n$—in other words, to generate a *transitive* permutation group—since this says that any sheet is connected to any other. The group generated by $\pi_1, \ldots, \pi_m$ is called the *monodromy group* of the covering. This term was originally introduced by Hermite 1851 to denote the group of substitutions of the corresponding algebraic function induced by circuits around its branch points.

Now to see that the construction does produce non-spherical surfaces it suffices to look at 2-sheeted coverings with an even number of branch points (> 2). We shall place all branch points on a meridian and permute the sheets along arcs which connect the branch points in pairs. Thus with four branch points we have Figure 52.

To see the topological form of this surface more clearly we peel off the outer sphere and place it opposite the inner sphere (Figure 53). Then when the identified edges are pasted together we get a surface homeomorphic to the torus (Figure 54). It is easy to see how to construct any surface of the form in Figure 55 by this method. The more surprising fact is that any Riemann surface is homeomorphic to one of these forms—a result which was proved by Clifford 1877—indeed it is far from clear that any arbitrary Riemann surface can even be embedded in R^3.

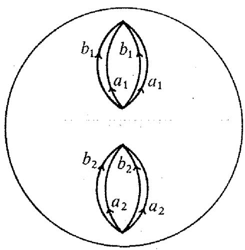


Figure 52

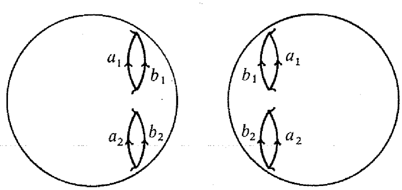


Figure 53

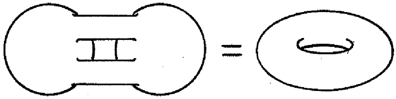


Figure 54

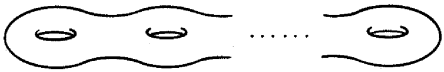


Figure 55

We shall deduce Clifford's result as a corollary of the general classification of surfaces in 1.3, a result first proved rigorously by Dehn and Heegaard 1907. Before doing so, it is of interest to look at some earlier attempts to classify surfaces.

EXERCISE 1.1.2.1. Show that the Riemann surface for

$$w^2 = (1 - z^2)(1 - k^2z^2) \qquad k^2 \neq 0, 1$$

is topologically a torus.

EXERCISE 1.1.2.2. Is there a 2-sheeted cover with an odd number of branch points?

1.1.3 Connectivity and Genus

The property which distinguishes a torus topologically from the sphere is the presence of a nonseparating closed curve a (Figure 56). Any points P, Q outside a on the torus can be connected by an arc which does not meet a, whereas any closed curve separates the sphere by the Jordan curve theorem. In general the *connectivity* of a surface can be measured by the maximum number of disjoint closed curves which can be drawn on the surface without separating it. This number is called the *genus* of the surface (German: *Geschlecht*), a term introduced by Clebsch. Surfaces of different genus

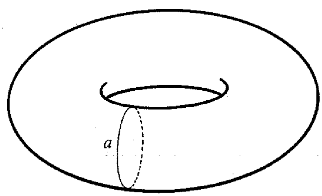


Figure 56

Figure 57

(Figure 57) are necessarily nonhomeomorphic and Riemann satisfied himself that the converse is also true. A somewhat unsatisfactory proof was offered by Jordan 1866a.

A similar result had already been obtained by Möbius 1863; in fact Möbius was the first to classify surfaces into normal forms. Despite his discovery of the famous one-sided surface which bears his name, Möbius's classification deals only with two-sided surfaces. He takes the surface to be smoothly embedded in R^3—an assumption which now strikes us as rather restrictive—and slices it into thin pieces by a family of parallel planes. He then argues that if the planes are suitably inclined and sufficiently close together the pieces will be three possible forms of perforated sphere (Figure 58), which he denotes (a), (ab), and (abc) respectively, where a, b, c are the boundary curves. The surface is then represented as a formal sum of such terms, for example Figure 59 is $(a) + (ab) + (c) + (bcd) + (d)$. By introducing the symbol $(abcd\ldots)$ for the sphere with perforations bounded by $a, b, c, d, \ldots$ and

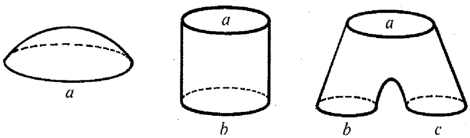


Figure 58

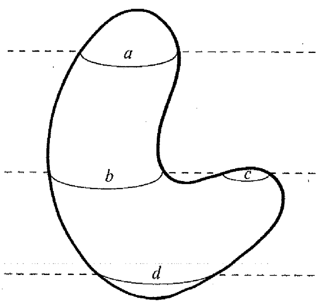


Figure 59

computation rules which reflect permissible ways of pasting or cutting perforated spheres, Möbius is able to bring any surface into the form

$$(a_1 a_2 \cdots a_n) + (a_1 a_2 \cdots a_n)$$

which is the normal form surface of genus $n - 1$ shown in Figure 60.

The normal form for Riemann surfaces obtained by Clifford 1877 is similar to the Möbius form, but obtained by joining together two copies of the perforated disc (Figure 61), along corresponding boundaries. The final variation of the surface, due to Klein 1882a, is the sphere with handles, Figure 62. This picturesque term has now become standard.

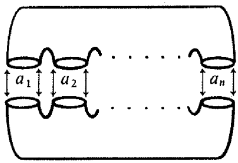


Figure 60

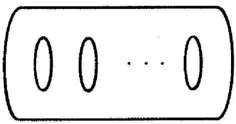


Figure 61

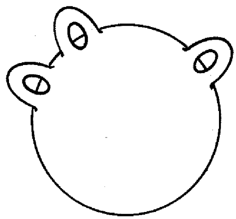


Figure 62

Exercise 1.1.3.1. Möbius's rules are

$$(xa_1a_2\ldots) + (xb_1b_2\ldots) = (a_1a_2\ldots b_1b_2\ldots),$$

where $x, a_1, a_2, \ldots, b_1, b_2, \ldots$ are all different and elements may be permuted inside their brackets.

(1) Interpret the transformations LHS (left-hand side) → RHS and RHS → LHS geometrically.
(2) Use the rules to show that the surface in Figure 63 has genus 4.

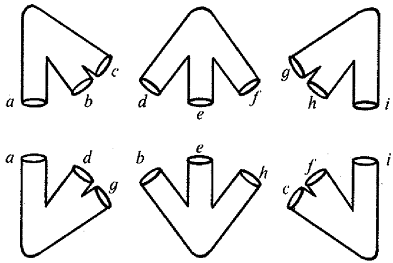


Figure 63

1.1.4 Branched Coverings of Higher Dimension

Potential theory in three dimensions gives rise to branching phenomena like those encountered in complex function theory (indeed the latter is related to 2-dimensional potential theory, as is well known). Instead of branch points one has branch curves, and if R^3 is completed by a point at infinity the result is the 3-sphere, S^3. Appell 1887 gave an example from potential theory of a covering of R^3 branched over a circle, and Sommerfeld 1897, also working in potential theory, introduced the term "Riemann spaces" for branched coverings of R^3.

Branched coverings of R^3 (or S^3) were first studied for their own sake in Heegaard 1898, where they were used as a means of constructing 3-dimensional manifolds. Proceeding by analogy with the 2-dimensional case, Heegaard takes a point P not on the branch curves, and connects it to each branch curve by a conical surface. Since the branch curves may be linked or knotted it is not generally possible to prevent these cones intersecting themselves or each other. The best one can do is to position P and the branch curves so that, when viewed from P, no crossing point of the curves is more than double, in which case the intersection lines on the cones will also be double (Figure 64). For each piece of a cone, a curve which pierces it and

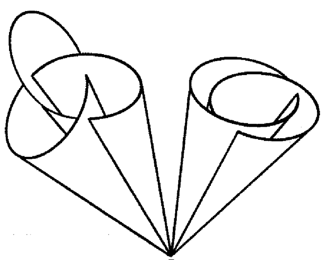


Figure 64

loops around its edge defines a permutation of the sheets, where each sheet is a copy of R^3 or S^3.

The typical 3-dimensional sheet minus its branch curves is far more complicated topologically than the 2-sphere minus a finite set of points, and one cannot formulate consistency conditions on the permutations until the relations between loops in the complement of a linked system of curves are known. The problem in effect is to determine the fundamental group of such a space.

This problem was solved by Wirtinger around 1904 and it was the first significant result in the mathematical theory of knots (see 4.2.3). Knowing the relations which the permutations had to satisfy, Wirtinger was able to give a procedure for finding all n-sheeted covers over a given system of branch curves. Namely, enumerate all finite sets of permutations on n letters, check whether they satisfy the consistency relations when associated with cone pieces, then see if they generate a transitive permutation group. The first exposition of Wirtinger's method, and its analogy with the determination of Riemann surfaces, is in Tietze 1908.

(Wirtinger was led to branched coverings in studying the singularities of functions of two complex variables. In fact, he arrived at a 3-sheeted covering branched over the trefoil knot, with permutations (12), (23), (31) associated with the three pieces of the cone from, of all things, Cardan's formula for the solution of cubic equations! An account of this example appears in Brauner 1928.)

While it is possible to enumerate the branched covers of the 3-sphere, there is no obvious method for deciding when two are the same, as one can do for Riemann surfaces by computing the genus (see 1.3.7, 1.3.8).

We shall see in 1.3 that the spheres with handles include all orientable closed surfaces, so branched covers of the 2-sphere (indeed, 2-sheeted covers alone) are a completely general method of constructing orientable surfaces. Branched covers of the 3-sphere attained a similar significance when Alexander 1919b proved that they include all orientable closed 3-manifolds. (He also proved the generalization to n dimensions.) An interesting recent result, found independently by Hilden 1974 and Montesinos 1976, is that 3-sheeted covers suffice and that the branch set can be a single (knotted) curve.

1.2 Nonorientable Surfaces

1.2.1 The Möbius Band

This surface appears to be the independent discovery of Listing and Möbius, both of whom mentioned it in unpublished manuscripts in 1858. Its paradoxical properties—one side and one edge—never fail to astonish those

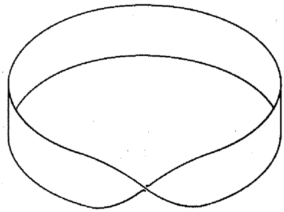


Figure 65

meeting it for the first time; nevertheless it is hard to see why such a simple surface (Figure 65) was not discovered until 1858. Twisted bands occur frequently as decorative borders in Roman mosaics and searches have been made for Möbius bands among them, without success (of course any band with an odd number of half twists is topologically a Möbius band). One might also look at the history of belt-driven machinery, since at some stage it was realized that a belt with a half twist wears evenly on "both" sides.

The property of one-sidedness seems to assume that the surface is embedded in ordinary space, however, an intrinsic counterpart to this property was found by Klein 1876. Klein imagines a small oriented circle (the *indicatrix*) placed on the surface, then transported round an arbitrary closed curve. If there is a curve which brings the indicatrix back with its orientation reversed then the surface is called *nonorientable*. The Möbius band has this property, see Figure 66, as does any surface containing a Möbius band. Conversely, any nonorientable surface contains a Möbius band, namely, a strip neighbourhood of a curve which reverses the indicatrix.

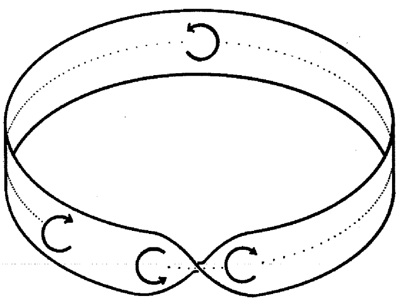


Figure 66

1.2.2 The Projective Plane

The plane of projective geometry is constructed by adding a line at infinity to the ordinary plane, with the property that each ordinary line has *one* point on the line at infinity. This surface can be realized topologically by resting a hemisphere on the plane and considering projection from the centre of the sphere (Figure 67). This establishes a continuous one-to-one correspondence between the finite points P' of the plane and the interior points P of the hemisphere. Lines in the plane correspond to great semicircles on the hemisphere, so the requirement that each line have exactly one point at infinity forces diametrically opposite points on the boundary of the hemisphere (which represents the line at infinity) to be identified.

Then by projecting the hemisphere vertically onto the plane we obtain the projective plane in the form of a disc with boundary divided into two halves identified as in Figure 68. This will be called the canonical polygon (in this case a 2-gon) for the projective plane.

Returning to the construction using the hemisphere, we can derive an elegant realization of the projective plane due to Klein and Schläfli 1874. Namely, identify all diametrically opposite points of the sphere. Since every point above the equator is identified with one below, we can omit the points above the equator, in which case we have precisely the original construction. Klein and Schläfli's construction shows the complete homogeneity of the projective plane, in particular the line at infinity (the equator) is no different from any other line. It also exhibits the sphere as a 2-sheeted *unbranched* cover of the projective plane (see 1.4).

It is intuitively plausible that the projective plane cannot be embedded in $\mathbf{R}^3$, though a rigorous proof of this was not available until the Alexander

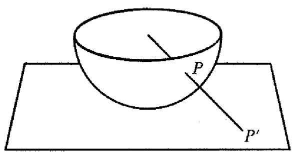


Figure 67

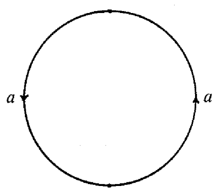


Figure 68

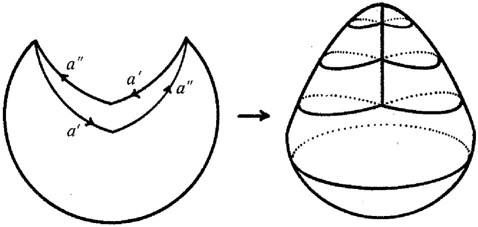


Figure 69

duality theorem appeared in Alexander 1923a. The best one can do in R^3 is to divide the boundary of the canonical polygon into four equal arcs a', a'', a', a'', then join the identified arcs along a single line of intersection (Figure 69).

The top part of this surface is called a *crosscap,* and we often describe the surface as a *sphere with crosscap,* meaning that a disc has been removed from the sphere and replaced by a crosscap.

Exercise 1.2.2.1. Show that if a disc is removed from the projective plane the result is a Möbius band. (In other words, a crosscap is a Möbius band.)

1.2.3 The Klein Bottle

This surface, introduced by Klein 1882a, is the easiest to visualize among the closed nonorientable surfaces. However, like the others, it cannot be realized in R^3 without intersecting itself. One begins with a cylinder whose ends are identified with opposite orientations, and joins them together as in Figure 70. If we slit the cylinder along a line b parallel to the axis we obtain the Klein bottle as a rectangle with the edge identifications of Figure 71. This is a more impartial representation of the surface, since it is easily converted to other

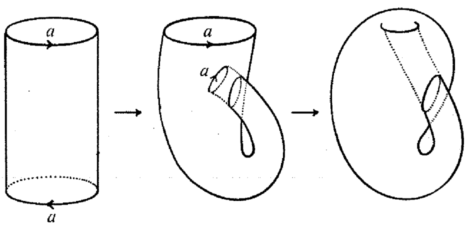


Figure 70

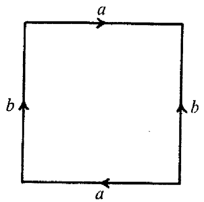


Figure 71

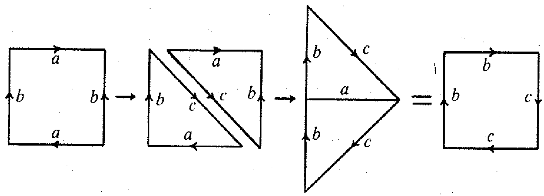


Figure 72

forms which look quite different when we attempt to realize them in $\mathbf{R}^3$. For example, see Figure 72. This is called the *two-crosscaps* form and it is taken as the canonical polygon for the Klein bottle. The physical realization of the two crosscaps is achieved as in Figure 73. If we now paste along a' we get the crosscaps in Figure 74, in the same way the crosscap was produced for the projective plane.

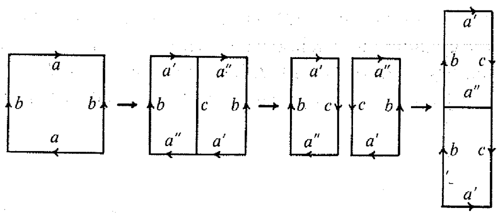


Figure 73

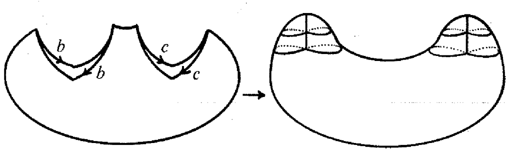


Figure 74

EXERCISE 1.2.3.1 (Hilbert and Cohn–Vossen 1932). Show that the Klein bottle can be constructed by diametric point identification of a (centrally symmetric) torus.

EXERCISE 1.2.3.2 (Dyck 1888). Show that the surface in Figure 75 is also a Klein bottle.

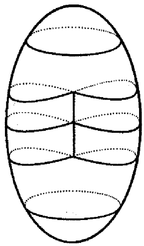


Figure 75

EXERCISE 1.2.3.3. Find a curve on the Klein bottle which separates it into two Möbius bands.

EXERCISE 1.2.3.4. Show that the Klein bottle minus a disc is a "nonorientable handle" (Figure 76) and that this figure can be deformed isotopically in R^3 into that shown in Figure 77.

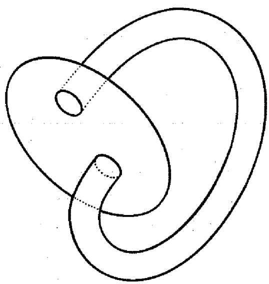


Figure 76

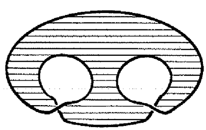


Figure 77

1.2.4 Dyck's Classification of Nonorientable Surfaces

Dyck 1888 gave a classification of nonorientable surfaces analogous to the classification of orientable surfaces, in which crosscaps take the place of handles. His proof is not really satisfactory; however, it introduces the important result

$$\text{crosscap} + \text{handle} = 3 \text{ crosscaps}.$$

Recalling that a crosscap is just a Möbius band (Exercise 1.2.2.1). Dyck's result can be explained intuitively by attaching a handle to a Möbius band, then dragging one end of it round the band to make a nonorientable handle (Figure 78). The nonorientable handle is just a perforated Klein bottle, joined to the Möbius band along its boundary; in other words, we have two crosscaps joined to a third.

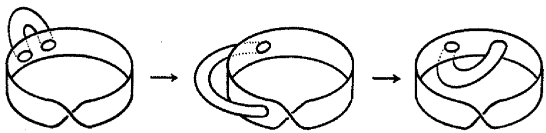


Figure 78

It follows that for any surface on which handles and crosscaps both appear we can remove the handles in favour of crosscaps. The hard part is to show that these are the only features a surface can have.

The formal version of Dyck's theorem, which we prove in the next section, is that any closed nonorientable surface can be represented as a polygon with edges identified as in Figure 79. The successive pairs a_i, a_i with the same orientation are interpreted as crosscaps, as is understandable from the construction of the two-crosscap form of the Klein bottle in 1.2.3.

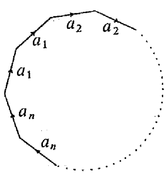


Figure 79

1.3 The Classification Theorem for Surfaces

1.3.1 Combinatorial Definition of a Surface

The precise definition and classification of surfaces from a combinatorial point of view was first given by Dehn and Heegaard 1907. They define a *closed surface* to be a finite 2-dimensional simplicial complex in which each edge is incident with two triangles and the set of triangles incident with a given vertex P can be ordered $\Delta_1, \Delta_2, \ldots, \Delta_k$ so that Δ_i has exactly one edge e_i in common with Δ_{i+1}, Δ_k has exactly one edge e_k in common with Δ_1, and these are the only common edges. Such a neighbourhood complex, the combinatorial equivalent of a disc, is called an *umbrella* by Lefschetz 1975. (See Figure 80.)

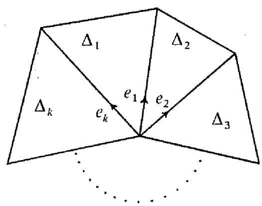


Figure 80

This definition obviously includes all the surfaces we have considered so far. In particular, the branch points on a triangulated Riemann surface have umbrella neighbourhoods, so Riemann surfaces will be included in our classification.

1.3.2 Schemata

It will be convenient to build surfaces from polygons other than triangles, so we now go to an alternative definition. A (finite) closed surface is a (finite) set of polygons with oriented edges identified in pairs. Such a system is called a *schema*. This definition is equivalent to the former, for if the polygons are given sufficiently fine simplicial decompositions, which are compatible on identified edges, then each edge in the decomposition will be either an interior edge on a polygon, hence incident with two triangles, or else a pair of subedges identified by the schema, hence also incident with two triangles. Similarly, vertices P' introduced by the decomposition will automatically have umbrella neighbourhoods when they lie in the interior of polygons, and they will get them as the result of pasting "half-umbrella" neighbourhoods when they lie on a polygon edge (Figure 81). Vertices P in the original schema also get umbrella neighbourhoods since the "corners" of the schema which come together at P can be arranged in a cyclic sequence in which each has one edge in common with its successor (Figure 82). This is because each edge in the schema is identified with exactly one other.

A portion of the boundary of a polygon with the form in Figure 83, which we shall represent symbolically by $aba^{-1}b^{-1}$ (reading labels and orientations clockwise), will be called a *handle*. The reason for this becomes clear when we cut it from the rest of the polygon and paste the identified edges (Figure 84).

Similarly, a portion like that in Figure 85, which we shall represent symbolically by aa or a^2, will be called a *crosscap* (cf. 1.2.2).

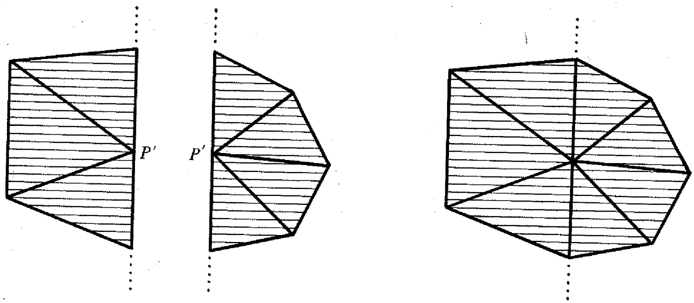


Figure 81

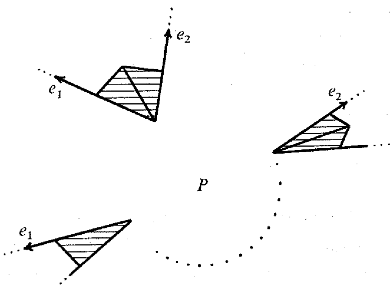


Figure 82

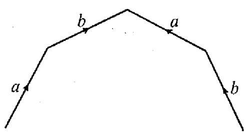


Figure 83

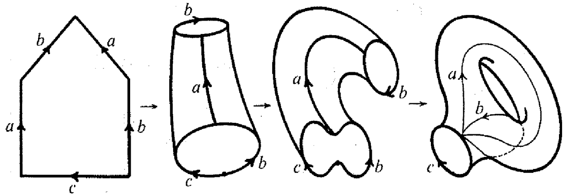


Figure 84

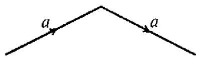


Figure 85

1.3.3 Reduction to a Single Polygon with a Single Vertex

Assuming the polygons in the schema define a connected surface, it will be possible to amalgamate them all into a single polygon by a sequence of pastings along identified edges of separate polygons. We say the resulting polygon has a single vertex if its edge identifications bring all vertices into coincidence. For example, the standard schemata for the torus and Klein bottle have this property (Figure 86). On the other hand, the schema for the sphere has two distinct vertices P, Q (Figure 87). This schema is exceptional in having only a "cancelling pair" of edges, aa^{-1}. In any other schema with

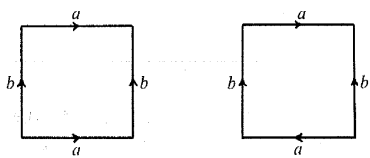


Figure 86

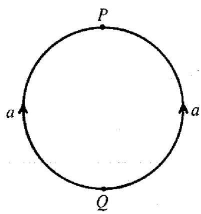


Figure 87

more than one pair of identified edges and more than one vertex it is possible to reduce the number of vertices as follows.

Divide the apparent vertices of the polygon into equivalence classes of vertices which are identified with each other, then, assuming there are ≥ 2 equivalence classes, consider an edge a whose endpoints P, Q belong to different classes. Then the construction in Figure 88 reduces the number of vertices in the equivalence class of Q by 1 (assuming there are at least 2) and by repeating we can reduce the class of Q to one member, at which time Q will be the endpoint of a cancelling pair, which can be closed up, eliminating this class entirely (Figure 89). We similarly eliminate other equivalence classes until only one remains.

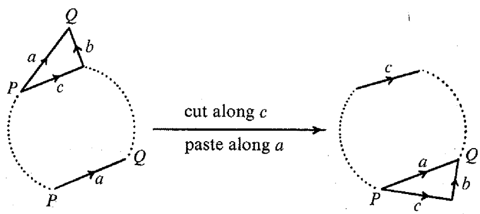


Figure 88

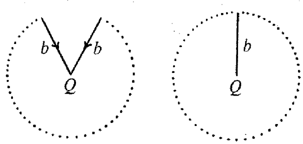


Figure 89

When we have only one vertex, it continues to be the only one when the polygon is cut from vertex to vertex and pasted along some identified edges. All the constructions which follow involve only operations of this type, so we can assume there is only one vertex throughout the remainder of the construction.

1.3.4 Crosscap Normalization

All identified pairs of like-oriented edges can be replaced by adjacent pairs (crosscaps) by the operation in Figure 90. Notice that any adjacent pairs already present (on the dotted lines) remain adjacent after this operation, so all pairs of like-oriented edges can be replaced by adjacent pairs if the operation is repeated.

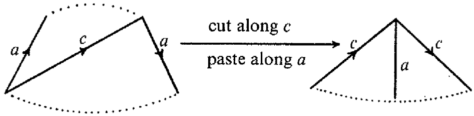


Figure 90

1.3.5 Handle Normalization

If any pairs of oppositely oriented edges remain after 1.3.4 they must occur as "crossed pairs"

$$\ldots a \ldots b \ldots a^{-1} \ldots b^{-1} \ldots$$

in the boundary of the polygon. For if, say, $\ldots a \ldots a^{-1} \ldots$ is not separated by any other pair of oppositely oriented edges we have Figure 91, where each edge in α is identified with another edge in α, and each edge in β is identified with another edge in β. This is because all like-oriented edges were made adjacent in 1.3.4. But then it is impossible for the two ends of a to be identified, contrary to 1.3.3.

We now replace two crossed pairs by a handle as in Figure 92. It is easily verified that any other oppositely oriented pairs in the boundary remain oppositely oriented under this operation, so if all pairs in the original polygon are oppositely oriented the result of repeating the operation as long as possible is a sphere with n handles (Figure 93).

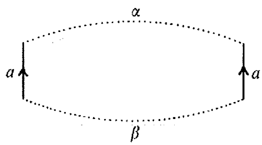


Figure 91

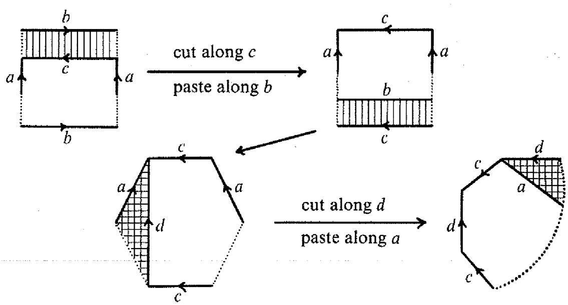


Figure 92

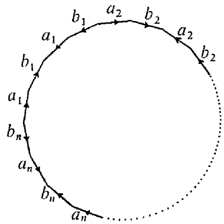


Figure 93

1.3.6 Metamorphosis of Handles into Crosscaps

In the general case, crosscap normalization followed by handle normalization yields a boundary consisting of both crosscaps and handles. The boundary must then contain a sequence

$$\ldots aabcb^{-1}c^{-1}\ldots$$

which we convert to three crosscaps as follows. First do as shown in Figure 94, then replace the three like-oriented pairs by adjacent pairs as in 1.3.4.

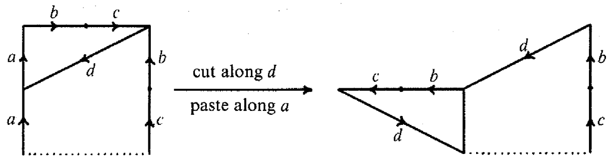


Figure 94

This will not disturb the dotted part of the boundary, and the handle will not reappear if crosscaps are normalized in the right order. One order which works is: b, c, d.

It follows that if the original polygon contains any like-oriented edges in its boundary it can be converted into a sphere with n crosscaps (Figure 95).

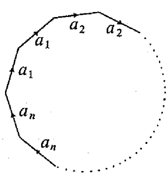


Figure 95

1.3.7 The Normal Forms for Classification

The procedure of 1.3.3–1.3.6, which is due to Brahana 1921, reduces any finite closed surface to one of the forms in Figure 96, called the sphere with n handles, sphere with n crosscaps and sphere respectively, and denoted symbolically by $a_1b_1a_1^{-1}b_1^{-1}\cdots a_nb_na_n^{-1}b_n^{-1}$, $a_1^2a_2^2\cdots a_n^2$ and aa^{-1}. We shall later (4.2.1 and 5.3.3) prove rigorously that these surfaces are topologically distinct. The proof develops the group theory which one can sense lurking behind the above construction and its symbolism. In the meantime we shall use the classical invariants for distinguishing surfaces, without attempting more than an intuitive proof of their topological invariance.

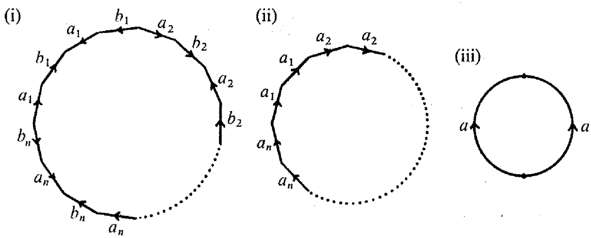


Figure 96

1.3.8 Euler Characteristic and Orientability

The Euler polyhedron formula (Euler 1752)

$$V - E + F = 2$$

where V, E, F are the numbers of vertices, edges, and faces, is valid for any schema which represents the sphere. One proves this by observing that the quantity $V - E + F$ is invariant under *elementary subdivisions*

(a) division of an edge into two by a new vertex
(b) division of a face into two by a new edge
and their inverses.

(Namely, in (a) V and E both increase by 1, in (b) E and F both increase by 1.) It follows that any schema equivalent to the sphere schema (iii) by these operations has the same value of $V - E + F$, namely $2 - 1 + 1 = 2$.

The value of $V - E + F$, which we call the *Euler characteristic*, for the schemata (i) and (ii) is

(i) $V - E + F = 1 - 2n + 1 = 2 - 2n$
(ii) $V - E + F = 1 - n + 1 = 2 - n$

so none of these surfaces is equivalent to the sphere under elementary subdivisions. In fact the Euler characteristic distinguishes the individual schemata in (i) by their different handle numbers n, and the individual schemata in (ii) by their different crosscap numbers n.

To distinguish the schemata in (i) from those in (ii) we need a means of computing the orientability character of the surface from a schema. The method is to triangulate the schema, then orient each triangle with a circular arrow. The oriented triangle induces orientations in its edges, as shown in Figure 97. The orientation of the whole surface is called *coherent* if each edge receives opposite orientations from the two triangles incident with it, and the surface is called *orientable* if it has a coherent orientation. A coherent orientation carries over to a triangulation obtained by elementary subdivision in the obvious way, for example see Figure 98, and thus the orientability character, like the Euler characteristic, is an invariant of schemata under elementary subdivisions.

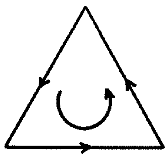


Figure 97

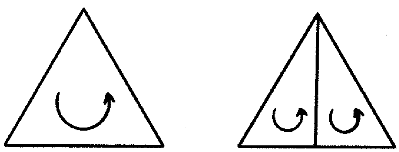


Figure 98

Now it is not difficult to find coherent orientations for the spheres with handles, and likewise one finds triangulations of the spheres with crosscaps which cannot be coherently oriented. The sphere itself is of course orientable. Thus all the normal forms can be distinguished by Euler characteristic and orientability character, as far as equivalence under elementary subdivisions is concerned. Since the normal form of any schema is obtained using only elementary subdivisions, it follows in particular that the normal form of any schema can be read directly from its Euler characteristic and orientability character.

To prove the topological invariance of these combinatorial invariants one needs a way to move from one triangulation of the surface to another. The difficulty is that triangulations are not necessarily "straight," so if two triangulations are superimposed we may have edges which intersect at infinitely many points (think of $x \sin 1/x$ and the x-axis). It seems reasonable to expect that the intersections could be reduced to a finite number by deformation of one of the triangulations, but this is awkward to prove. (It is done in Kerekjarto 1923.) When it is done, however, one gets a decomposition of the surface obtainable from both the given triangulations by elementary subdivision, and hence with the same Euler characteristic and orientability character as both. This proves that Euler characteristic and orientability character are independent of the triangulation initially chosen.

EXERCISE 1.3.8.1. Identify the surfaces with schemata $a_1 a_2 \cdots a_n a_1^{-1} a_2^{-1} \cdots a_n^{-1}$ and $a_1 a_2 \cdots a_{n-1} a_n a_1^{-1} a_2^{-1} \cdots a_{n-1}^{-1} a_n$.

EXERCISE 1.3.8.2. Let $\mathscr{F}$ be an actual euclidean polyhedral surface. Define the *curvature* $\kappa(P)$ of $\mathscr{F}$ at a vertex P to be (2π minus the sum of the face angles incident with P). Then show that

$$\sum_{\text{vertices } P} \kappa(P) = 2\pi \times (\text{Euler characteristic of } \mathscr{F}).$$

(For polyhedra homeomorphic to the sphere, this is Descartes's version of the Euler polyhedron formula; known in 1639 but not published until the 1850s.)

1.3.9 Bounded Surfaces

A bounded surface is a finite 2-dimensional simplicial complex in which each edge is incident with one or two triangles and the triangles incident with a vertex form either an umbrella or a *half-umbrella*, that is, a sequence $\Delta_1, \Delta_2, \ldots, \Delta_k$ in which each Δ_i has exactly one edge in common with Δ_{i+1}, and these are the only common edges. It is easily seen that the *free edges*—those incident with only one triangle—form one or more closed curves, which we call the boundary curves of the surface.

The corresponding notion of schema is one in which each edge of a polygon is identified with at most one other, and it follows by simplicial decomposition of the schema that the free polygon edges likewise form one or more closed curves. In fact, by making the simplicial decomposition sufficiently fine one obtains a thin annulus from the triangles incident with a given boundary curve (it cannot be a Möbius band, since it has two edges). Starting with these annuli, and amalgamating disjoint pieces of the schema along identified edges, one obtains a schema consisting of a single polygon with "holes" bounded by the boundary curves, and an outer boundary consisting of edges identified in pairs.

We can then repeat the construction of 1.3.3–1.3.7, taking care that all cuts in the polygon avoid the holes, and obtain one of the normal forms in Figure 99, called the sphere with handles and holes, sphere with crosscaps and holes, and sphere with holes respectively.

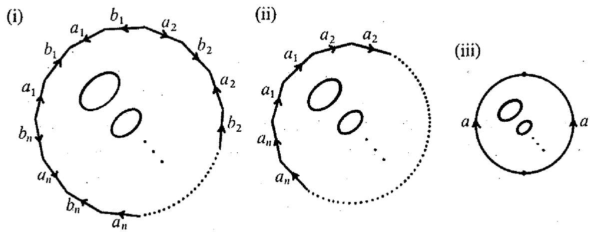


Figure 99

Thus any bounded surface can be regarded as a closed surface with perforations. An interesting property of the perforated surfaces is that they are all embeddable in $\mathbf{R}^3$. To see this, take a canonical polygon *without* holes and perforate it in the neighbourhood of its single vertex by cutting off the corners. The corners cut off form a disc, like slices of a pie, and the portions of edges which remain can be physically pasted together in $\mathbf{R}^3$ to form bands attached to the body of the polygon. Examples are shown in Figure 100.

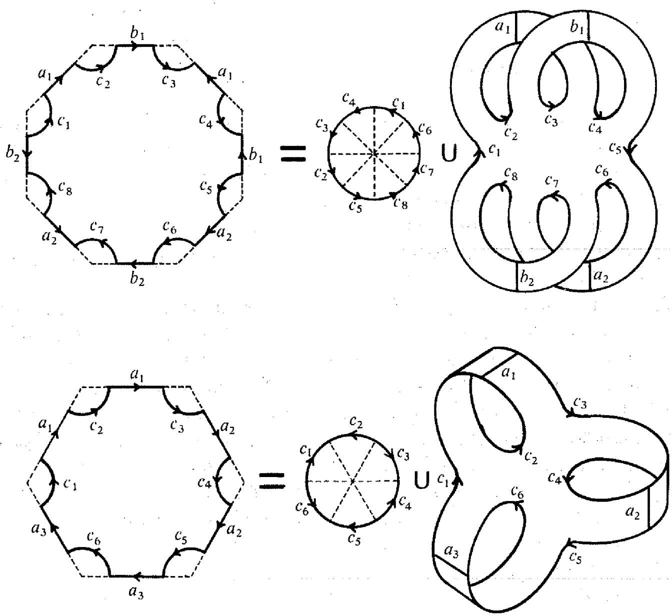


Figure 100

As these examples show, handles yield double bands, crosscaps yield Möbius bands, and of course any extra perforations yield single untwisted bands. These forms of the perforated surfaces were used by Dehn and Heegaard 1907 in their proof of the classification theorem.

If the perforated surface is thought of as lying in the hyperplane $x = 0$ of (x, y, z, t)-space, one can close it by a cone of line segments from any point P off $x = 0$ to the boundary curve of the surface, thus realizing the closed surface without self-intersections in R^4.

EXERCISE 1.3.9.1. Show that the normal forms of bounded surfaces are distinguished from each other by Euler characteristic, orientability character, and number of boundary curves.

EXERCISE 1.3.9.2. By suitable cuts in the perforated polygons show that the normal forms can be expressed

(i) $w_1c_1w_1^{-1} \cdots w_mc_mw_m^{-1}a_1b_1a_1^{-1}b_1^{-1} \cdots a_nb_na_n^{-1}b_n^{-1}$,
(ii) $w_1c_1w_1^{-1} \cdots w_mc_mw_m^{-1}a_1^2 \cdots a_n^2$,
(iii) $w_1c_1w_1^{-1} \cdots w_mc_mw_m^{-1}$,

where $c_1, \ldots, c_m$ are the boundary curves.

EXERCISE 1.3.9.3. If $\mathscr{F}$ denotes the "standard" sphere with handles in R^3, with one perforation (Figure 101), show that $\mathscr{F}$ can be isotopically deformed into its disc-with-bands form (Figure 102). (*Hint*: Stretch the perforation so that it sends a "tentacle" into each of the handles.) Deduce that $\mathscr{F}$ can be turned inside-out in R^3.

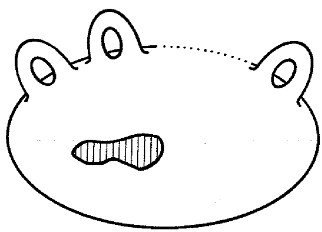


Figure 101

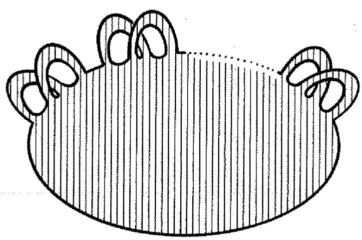


Figure 102

EXERCISE 1.3.9.4. Show how the models of nonorientable surfaces with self-intersections in R^3 can be viewed as embeddings in R^4 if a fourth coordinate is introduced by applying colour, with varying intensity, to the surface.

1.4 Covering Surfaces

1.4.1 The Universal Covering Surface

An interesting consequence of the representation of a surface $\mathscr{F}$ by a polygon was discovered by Schwarz in 1882. Schwarz observed that if one takes infinitely many copies of $\mathscr{P}$ and joins them to each other (rather than to themselves) along the identified edges then the result is a simply-connected surface $\tilde{\mathscr{F}}$ which is an *unbranched* covering of $\mathscr{F}$. In fact, if $\mathscr{F}$ is not the sphere or projective plane then $\tilde{\mathscr{F}}$ is homeomorphic to the plane.

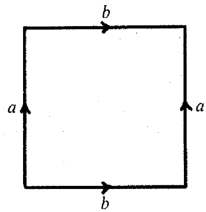


Figure 103

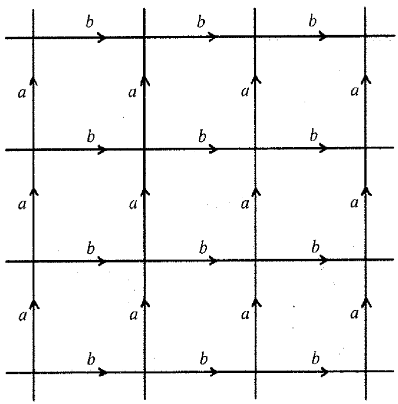


Figure 104

$\tilde{\mathscr{F}}$ is called the *universal covering surface* of $\mathscr{F}$, and its construction is most easily seen in the case of the torus. In this case it is possible to imagine the plane covering the surface quite literally (Hilbert and Cohn–Vossen 1932). We take the canonical polygon for the torus (Figure 103) then paste infinitely many copies together along the like-labelled edges to form a plane (Figure 104). Now to cover the torus with the plane we first roll up the plane into an infinite cylinder with circumference a (Figure 105) then wrap the cylinder round a torus of axis b like an infinite snake swallowing its tail infinitely often (Figure 106). The result is an infinite-sheeted covering of the torus without any branch points (Figure 107). Each sheet is a copy of the canonical polygon for the torus, namely one of the squares from which we originally assembled the plane.

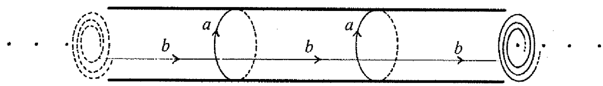


Figure 105

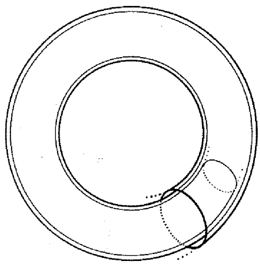


Figure 106

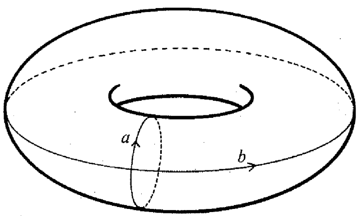


Figure 107

Exercise 1.4.1.1. Associate each sheet in the obvious way with an ordered pair $\langle m, n\rangle$ of integers $m, n \in \mathbb{Z}$. Then describe the permutations of the sheets induced by crossing the lines a, b on the torus as permutations of $\mathbb{Z} \times \mathbb{Z}$.

1.4.2 The Universal Cover of an Orientable Surface of Genus > 1

It will suffice to construct the universal cover of the orientable surface $\mathscr{F}_2$ of genus 2, whose canonical polygon $\mathscr{P}_2$ is the octagon (Figure 108), which folds up into the closed surface as shown in Figure 109. $\mathscr{F}_2$ is not only sufficient to illustrate the difficulties of the general case, it is also sufficient in a technical sense, because any orientable surface of higher genus is a (finite-sheeted) cover of $\mathscr{F}_2$ and the universal covering surface covers any other covering surface—hence the name "universal." These technical points are explored in the exercises below.

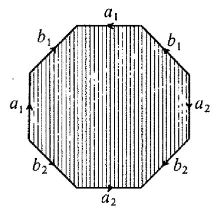


Figure 108

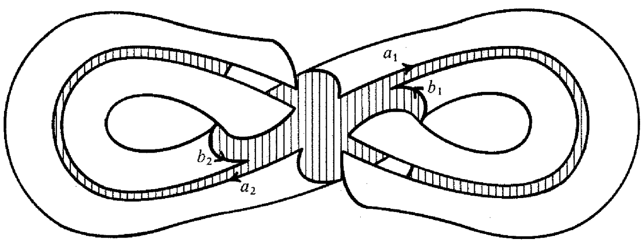


Figure 109

We now construct the universal cover of $\mathscr{F}_2$.

Since the eight corners of the octagon meet at a single point on $\mathscr{F}_2$, eight octagons must meet at each vertex of the tessellation which represents the universal covering surface. Furthermore, the labelling of edges at a vertex must be the same as on $\mathscr{F}_2$, namely Figure 110.

It is certainly possible to produce this cycle of edges by putting eight copies of $\mathscr{P}_2$ together, since the eight wedges at the vertex correspond to the different corners of $\mathscr{P}_2$. In fact, assuming we cannot flip the $\mathscr{P}_2$'s over, this is the *only* way eight $\mathscr{P}_2$'s can be joined at a vertex. Because each label a_i or b_i occurs only twice in the boundary of $\mathscr{P}_2$, once with the interior of $\mathscr{P}_2$ on its left and once with the interior of $\mathscr{P}_2$ on its right; so there is exactly one way to attach one $\mathscr{P}_2$ to another along a given edge.

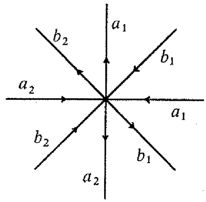


Figure 110

Thus if we begin with a single $\mathscr{P}_2$, with vertices on a circle $\mathscr{C}_1$ say, there is a topologically unique way to complete the neighbourhood of each vertex of $\mathscr{P}_2$ with further (noncongruent!) copies of $\mathscr{P}_2$, whose new vertices we can assume to lie on a larger circle $\mathscr{C}_2$ concentric with $\mathscr{C}_1$ (Figure 111). (We interpolate vertices on $\mathscr{C}_2$, too close together to show in the diagram, so as to make each polygon an octagon.) We then proceed to complete the neighbourhood of all vertices on $\mathscr{C}_2$ similarly, with the new vertices lying on a circle $\mathscr{C}_3$, and so on. By choosing the radius of $\mathscr{C}_n$ to tend to ∞ as $n \to \infty$ we obtain a tessellation with the required vertex neighbourhoods covering the whole plane.

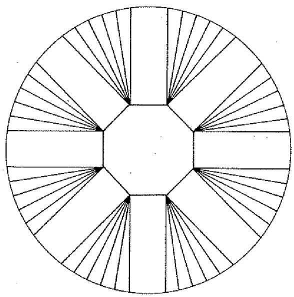


Figure 111

The covering of $\mathscr{F}_2$ by the plane is then defined by mapping each octagon in the tessellation onto $\mathscr{P}_2$ by a continuous function which is one-to-one except where it is required to identify boundary points, and making the functions for adjacent polygons agree on the common edges. □

Naturally we cannot use regular octagons in the euclidean plane for this construction, but in any case we need only the topological properties of the tessellation, and irregular octagons in the euclidean plane are the most direct means of demonstrating its existence. In the noneuclidean plane of hyperbolic geometry there are regular octagons with corner angles of size

$\pi/4$ (in fact of any size $<$ the value for euclidean regular octagons) so it is possible for these regular octagons to meet eight at each vertex, and hence to tessellate the hyperbolic plane. This observation was made by Poincaré after he learned of the universal covering surface in a letter from Klein (Klein 1882b), and he developed it into a powerful tool with this model of the hyperbolic plane as the open unit disc with circles orthogonal to the boundary as "straight lines."

The work of Poincaré and Klein in this period was devoted to automorphic functions and Fuchsian groups, to which we briefly turn in the next section, but it contained implicitly some fundamental results of surface topology.

EXERCISE 1.4.2.1. The n-sheeted cyclic cover of $\mathscr{F}_2$ is defined by taking n coaxial copies of $\mathscr{F}_2$, cutting them through one of the handles (Figure 112), then identifying the boundaries of the cuts cyclically (ith on the left with $(i + 1)$th on the right, nth on the left with first on the right). By computing the Euler characteristic of the cover, show that it can be an orientable surface of arbitrary genus > 1, if n is suitably chosen.

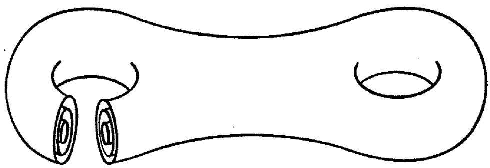


Figure 112

EXERCISE 1.4.2.2. Give an alternative description of the n-sheeted cyclic cover of $\mathscr{F}_2$ as n copies of $\mathscr{P}_2$ joined together, and hence show how it is covered by the universal covering surface of $\mathscr{F}_2$.

1.4.3 Fuchsian Groups

The monodromy group was the first group to appear in topology, as a means of specifying the way the sheets of a Riemann surface permute around the branch points. We can also speak of the monodromy group of an unbranched covering, the generating permutations now being the permutations of the sheets induced by crossing one of the canonical curves on the surface, that is, one of the edges of its canonical polygon. This permutation can be viewed as an automorphism of the tessellation of the covering surface. For example, the two generating automorphisms of the rectangular tessellation of the plane which covers the torus are simply vertical and horizontal translations of lengths a and b respectively (Figure 113), and the monodromy group is therefore the free abelian group of rank 2. The automorphisms of the universal covers of surfaces of higher genus can also be viewed as (noneuclidean)

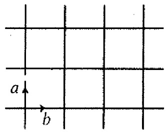


Figure 113

translations if one uses regular tessellations of the hyperbolic plane. Otherwise, one can define an automorphism to be a one-to-one continuous map of the tessellation onto itself which preserves labels and orientations.

Such groups of automorphisms were first studied in complex function theory, notably by Poincaré and Klein from 1882 onwards after their initial discovery by Fuchs. Thus the terms *automorphic functions* and *Fuchsian groups*. Space does not permit us to give an account of this vast and interesting theory (see Magnus 1974, or Fricke and Klein 1897, 1912) except to say that it arises naturally from algebraic functions so a connection with surface topology is to be expected. This connection was emphasized by Klein from the beginning, though in the then immature state of topology one did not pursue theorems about surfaces for their own sake, but only as a tool of function theory. Consequently, the first group-theoretic results about surfaces appear only as special cases of results on Fuchsian groups.

Since our interest is only in this special case, we shall immediately specialize the methods of Fuchsian groups to deal just with the tessellations obtained from universal coverings of orientable surfaces.

Let $\mathscr{F}_n$ denote the orientable surface of genus n whose canonical polygon is $\mathscr{P}_n$ (Figure 114). The universal cover of $\mathscr{F}_n$ is obtained by tessellating the plane with copies of the $4n$-gon $\mathscr{P}_n$ so that $4n$ of them meet at each vertex. As we observed for $n = 2$ in 1.4.2, there is a unique way of doing this. In anticipation of Chapters 3 and 4 we shall denote the automorphism group of the universal cover by $\pi_1(\mathscr{F}_n)$. We now show that

$$\pi_1(\mathscr{F}_n) = \langle a_1, b_1, \ldots, a_n, b_n; a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1} \rangle.$$

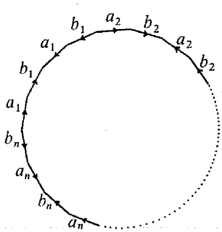


Figure 114

To discuss automorphisms of the tessellation we introduce the notion of an *edge path.* An edge path is a finite sequence of oriented edges of the tessellation $d_1, d_2, \ldots, d_k$ such that

$$\text{final point of } d_i = \text{initial point of } d_{i+1}.$$

It can be uniquely determined relative to its initial vertex by the corresponding sequence of edge labels a_i or b_i, provided with exponents $+1$ or -1 to indicate whether the edge is traversed according to, or against, its given orientation. This is because each vertex has exactly one incoming and one outgoing edge for each label. We shall consider edge paths in this relative sense so they can be identified with elements in the free group generated by the a_i's and b_i's.

If we fix a vertex P of the tessellation, an automorphism is determined by the vertex P' to which P is sent. For any vertex Q is determined relative to P by an edge path q from P to Q, and since any automorphism preserves labels and orientations the image Q' of Q is found at the end of the same edge path q from P'. But P' in turn is determined by an edge path p' from P, so the automorphism group is naturally isomorphic to the group of edge paths p' modulo closed paths.

This group is generated by the single edge paths $a_1, b_1, \ldots, a_n, b_n$ which are automatically subject to the relations

$$a_i a_i^{-1} = a_i^{-1} a_i = 1 \quad \text{and} \quad b_i b_i^{-1} = b_i^{-1} b_i = 1 \tag{1}$$

saying that a path out and back along an edge returns to its starting point, and also to the relation

$$a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1} = 1 \tag{2}$$

the left-hand side of which, r_n, is a circuit round $\mathscr{P}_n$. To show that (2) is in fact the defining relation of the group we have to show that any closed edge path is equivalent to the trivial path under the relations (1) and (2).

But it is clear that any closed path can be contracted to its initial point by a finite sequence of operations.

(a) Pulling out portions of the form pp^{-1} (Figure 115), where the inverse p^{-1} is p written backwards with all exponents reversed. This is obtainable from the relations (1).

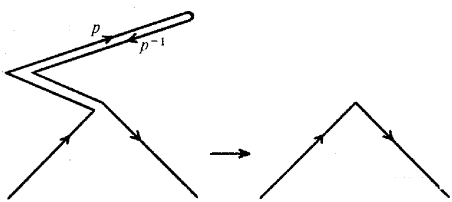


Figure 115

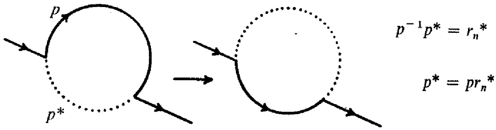


Figure 116

(b) Pulling a portion of the path from one side of a polygon to the other (Figure 116). This means replacing the portion p by pr_n^*, where r_n^* is a cyclic permutation of r_n, and hence it is obtainable by application of (2). □

This result was obtained independently by Poincaré 1882 and Klein 1882c as a special case of the presentation of Fuchsian groups. Its first explicit application to surface topology was made by Poincaré 1904, in a study of curves on surfaces. We lay the foundations for this study in Chapters 3 and 4, and carry it out in Chapter 6. In the meantime, Exercise 1.4.3.2 will serve to explain the connection.

EXERCISE 1.4.3.1. Show that the edge complex of the tessellation forming the universal cover of $\mathscr{F}_n$ can also be interpreted as the Cayley diagram of $\pi_1(\mathscr{F}_n)$.

EXERCISE 1.4.3.2. Show that a path on $\mathscr{F}_n$ covered by an edge path p in the tessellation contracts to a point on $\mathscr{F}_n$ just in case p is closed.

1.4.4 The 2-sheeted Cover of a Nonorientable Surface

Every nonorientable surface has an orientable surface as a 2-sheeted cover.

The most intuitive way to see this is to take the perforated form of the sphere with n crosscaps, namely the disc with n Möbius strips attached (Figure 117). This surface has the 2-sheeted cover $\mathscr{F}$ shown in Figure 118, which is evidently orientable because of its "two-sidedness." Then to cover the closed surface one attaches two discs along the boundary curves of $\mathscr{F}$, covering the single disc needed to close the nonorientable surface. □

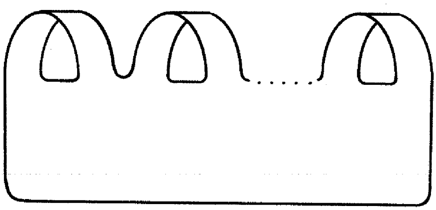


Figure 117

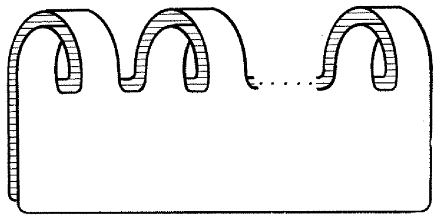


Figure 118

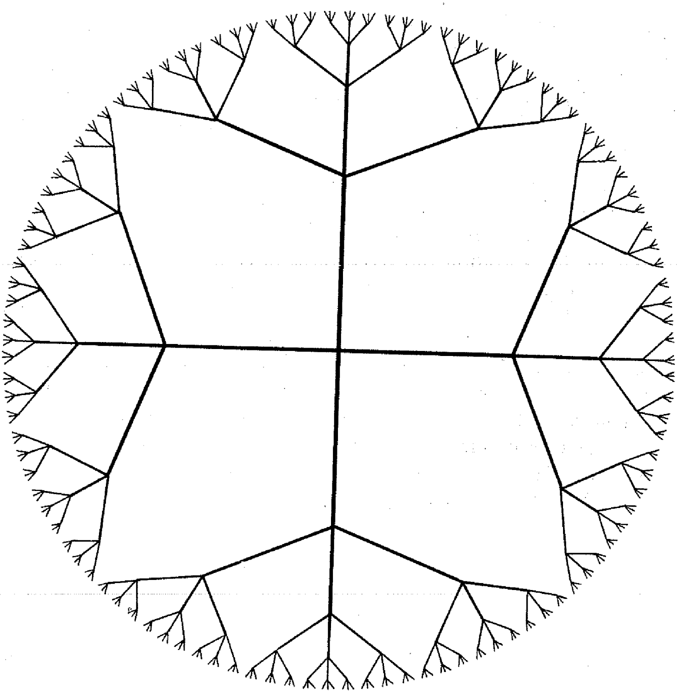
It follows that the universal covers of the orientable surfaces also cover the nonorientable surfaces. The universal covering surfaces are therefore the sphere (for the sphere itself, and the projective plane) and the plane. All the covers we have discussed in this section are unbranched, and indeed this will be the predominant type from now on. Unbranched covers came to the fore in the 1920s when Reidemeister discovered important group-theoretic applications (see 4.3) and since then the word "covering" has been taken to refer to them.

The combinatorial definition of an unbranched covering will be given in 4.3. The idea is that the covering surface must project continuously onto the underlying surface in a way which is *locally one-to-one*. That is, the projection is one-to-one when restricted to sufficiently small neighbourhoods on the covering surface. It is precisely in the neighbourhoods of branch points where this condition fails for Riemann surfaces—if a small disc is removed around each branch point one obtains an unbranched covering of the perforated sphere by a perforated orientable surface.

Exercise 1.4.4.1. Show that any nonorientable surface can be obtained by diametric point identification of either a sphere with handles or a torus with handles.

CHAPTER 2

Graphs and Free Groups



2.1 Realization of Free Groups by Graphs

2.1.1 Introduction

Free groups first appeared in mathematics as subgroups of the modular group in complex function theory. When Dyck 1882 pointed out the fundamental role of free groups in combinatorial group theory, as the most general groups from the point of view of generators and relations, his picture of them remained the function-theoretic one—a tessellation of the unit disc by curvilinear triangles whose sides were circular arcs orthogonal to the disc boundary. The first mathematician to study free groups in their own right and discover significant theorems about them was Jakob Nielsen (see Nielsen 1918, 1919, 1921), in fact the term "free group" did not appear until Nielsen 1924a. Nielsen's technique is partly geometric (based on the length of words and "cancellation," see exercise 2.2.4.1 below), however, it suppresses the natural geometric structure of a free group by imposing a "linear" appearance on the elements as strings of letters.

The appropriate geometric framework for describing free groups, the "two-dimensional" one of graphs, was first exploited in unpublished work of Max Dehn. According to Magnus and Moufang 1954, Dehn used this method to obtain the first proof that subgroups of free groups are free. When Schreier 1927 published his algebraic proof of the theorem he concluded the paper by describing the graph-theoretic interpretation of his construction. Schreier had been influenced by ideas of Reidemeister, who published the first full treatment of free groups on a graph-theoretic basis in Reidemeister 1932.

Reidemeister's treatment has influenced ours, however we have used the graph-theoretic framework even in the elementary stages to explain the notion of reduced word and to solve the word problem. This not only serves to unify the exposition, but vividly illustrates the dual view of a group as fundamental group of a space and automorphism group of a covering space, which will be a continuing theme in the chapters to follow.

It should be emphasized that the fundamental group is needed only in a combinatorial sense for the results on free groups below. Indeed the proofs could be viewed as mere translations of arguments about letters, words, and cancellation if it were not for the fact that the graph-theoretic form is more natural. However, since the topological invariance of certain groups will be needed later, the next chapter contains a topologically invariant construction of the fundamental group and its computation for graphs.

2.1.2 Graphs, Paths, and Trees

We shall interpret a graph in its broadest sense as an arbitrary collection of points (vertices) joined by lines (edges). Thus we admit examples such as those in Figure 119, in which more than one line may connect a given pair

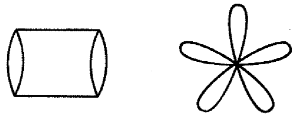


Figure 119

of points, and the two endpoints of a line may coincide. In the latter case, the two directions in which a line may be traversed cannot be distinguished by the order of the endpoints, so it is necessary to adjoin a notion of orientation in order to fully describe the concept of path in a graph.

Accordingly, the formalization of these concepts is as follows: A *graph* consists of two sets $\{P_i\}$ and $\{e_j\}$ of elements called *vertices* and *edges* respectively, subject to certain *incidence relations.* Each edge e_j is a pair $\{e_j^{+1}, e_j^{-1}\}$ of *oriented edges* called the *positive* and *negative* orientation of e_j, and e_j is incident with two vertices X_j and Y_j called respectively the *initial point* and *final point* of e_j^{+1} (and referred to collectively as *endpoints*). The vertices X_j and Y_j can also be described as the final point and initial point, respectively, of e_j^{-1}. When we write an edge without a superscript the orientation is understood to be positive.

A path p in a graph $\mathscr{G}$ is a finite sequence $d_1, d_2, \ldots, d_m$ of oriented edges of $\mathscr{G}$ such that

$$\text{final point of } d_i = \text{initial point of } d_{i+1}$$

for $i = 1, 2, \ldots, m-1$. We write $p = d_1 d_2 \cdots d_m$ and also write $p_3 = p_1 p_2$ if $p_1 = d_1 \cdots d_m$, $p_2 = d_1' \cdots d_n'$ and $p_3 = d_1 \cdots d_m d_1' \cdots d_n'$. The number of oriented edges in the sequence is called the *length* of the path. A path $p = d_1 d_2 \cdots d_m$ is *closed* if

$$\text{final point of } d_m = \text{initial point of } d_1$$

and *reduced* if no two successive oriented edges are opposite orientations of the same edge. Such a subpath $e_j^{+1} e_j^{-1}$ or $e_j^{-1} e_j^{+1}$ is called a *spur*. For convenience we also admit a single vertex to be a closed path (which is therefore reduced). The *inverse* of a path $p = d_1 d_2 \cdots d_m$ is the path $p^{-1} = d_m^{-1} \cdots d_2^{-1} d_1^{-1}$, where d_i^{-1} denotes the result of reversing the exponent of d_i.

A graph is *connected* if there is a path between any two of its vertices. A *tree* is a connected graph containing no reduced closed paths other than vertices. If one looks at a typical tree, such as Figure 120, then the proposition which follows is obvious; however we give a careful proof, in view of its fundamental importance.

Path Uniqueness Property of Trees. *Any two vertices in a tree are connected by a unique reduced path.*

If there are vertices P, Q connected by different reduced paths p_1, p_2 in a tree $\mathscr{T}$ we can assume that one of these paths is a path of minimal length

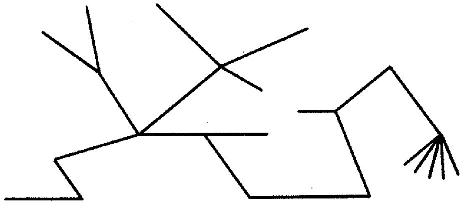


Figure 120

between P and Q, and that no other pair of vertices are connected by a nonunique reduced path of smaller length.

Since p_1, p_2 contain no spurs by hypothesis, whereas the closed path $p_1p_2^{-1}$ must, since $\mathcal{T}$ is a tree, the spur can only occur as the last oriented edge in p_1 and the first in p_2^{-1}, that is, if the last oriented edge in both p_1 and p_2 is the same; or similarly if the first oriented edge in both p_1, p_2 is the same. In the former case we can omit this last edge d from both paths p_1, p_2, obtaining shorter nonunique reduced paths p_1', p_2' from P to the initial point of d, contrary to hypothesis. In the latter case, omission of the first edge of both p_1, p_2 leads to a similar contradiction. □

The converse of this proposition is immediate, hence the path uniqueness property can also be used to define trees. Two important equivalents of this property can be derived with the help of the notion of *path equivalence.* Paths p, p' are called *equivalent* if p' results from p by a finite number of insertions or removals of spurs between successive oriented edges or at the endpoints. In particular, the reduced form of p is equivalent to p, so since path uniqueness says that two paths p, p' between the same endpoints have the same reduced form we have

(1) Paths in a tree with the same initial and final point are equivalent. In particular
(2) Any closed path is equivalent to its initial vertex.

But (2) in turn implies that the graph is a tree, because if the reduced form of the closed path is not a single vertex we have a contradiction to the definition of a tree. Thus all three properties are equivalent.

2.1.3 The Cayley Diagram of a Free Group

Given a free group F with free generators $a_1, a_2, \ldots$ we can construct a tree $\mathcal{T}$ which is the Cayley diagram of F. The edges of $\mathcal{T}$ are assigned orientations and labels $a_1, a_2, \ldots$ so that each a_i occurs exactly twice at a given vertex, once on an incoming edge and once on an outgoing edge. The typical vertex of $\mathcal{T}$ will be imagined to look like Figure 121, though this picture cannot be

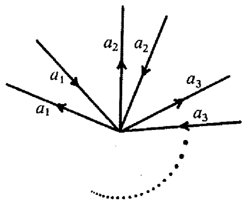


Figure 121

taken quite literally if there is more than a countable infinity of generators.

An explicit construction of $\mathcal{T}$, which at the same time shows it to be a tree, is as follows:

Step 1. Draw the typical vertex and its neighbouring edges (Figure 121), with the free endpoints lying on a circle $\mathscr{C}_1$.

Step $k + 1$. Assuming all the free endpoints of the graph constructed up to the stage k lie on a circle $\mathscr{C}_k$, attach edges to each free vertex so as to complete its neighbourhood star to the form (Figure 121), but so that all new free endpoints lie on a circle $\mathscr{C}_{k+1}$ outside $\mathscr{C}_k$. Figure 122 shows the construction when the vertex on $\mathscr{C}_k$ is the initial point of an outgoing a_i, the construction is similar for an incoming a_i.

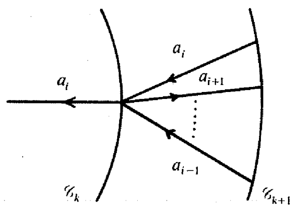


Figure 122

Assuming that there is no nontrivial reduced closed path inside $\mathscr{C}_k$ (which is certainly true for $k = 1$), the same is true inside $\mathscr{C}_{k+1}$, since any path inside $\mathscr{C}_{k+1}$ is separated by $\mathscr{C}_k$ into paths inside $\mathscr{C}_k$ and single edges or spurs between $\mathscr{C}_k$ and $\mathscr{C}_{k+1}$. Thus it follows by induction on k that no $\mathscr{C}_k$ contains a nontrivial reduced closed path, and hence the graph $\mathcal{T}$ obtained by uniting all graphs within the $\mathscr{C}_k$'s is a tree, since any path in $\mathcal{T}$ must lie in some $\mathscr{C}_k$. □

The paths in $\mathcal{T}$ which emanate from some fixed vertex correspond naturally to elements of F. The word corresponding to a given path is read by taken the labels on successive edges in the path and giving them exponents $+1$ or -1 according as the edge is traversed with the assigned orientation or its opposite. Conversely, there is exactly one path from a given vertex corresponding to a given word, since there is exactly one edge at each vertex for each of the generators a_i^{+1}, a_i^{-1}. A product p_1p_2 of paths p_1, p_2 corresponds to the product of words read from p_1, p_2 respectively. In particular, the paths which correspond to the trivial products $a_i^{+1}a_i^{-1}$ and $a_i^{-1}a_i^{+1}$ are exactly the spurs, so the notion of equivalence of words in a free group (0.5.2) agrees with that for paths (2.1.2).

The paths which correspond to the identity element 1 are the closed paths, since it is exactly these which reduce to a single vertex. (These include the paths of the form pp^{-1}, so the group element corresponding to p^{-1} is the inverse of the element corresponding to p.) It follows that once a particular vertex is chosen to represent 1, all the other vertices represent distinct elements of F, so $\mathscr{T}$ is indeed its Cayley diagram.

Exercise 2.1.3.1. If the cardinality of the set $\{a_1, a_2, \ldots\}$ is too large for the graph to be actually embedded in the plane, how should "inside $\mathscr{C}_k$" and "outside $\mathscr{C}_k$" be interpreted so as to retain the property that $\mathscr{C}_k$ separates paths inside $\mathscr{C}_{k+1}$ into paths inside $\mathscr{C}_k$ and single edges and spurs?

Exercise 2.1.3.2. In the noneuclidean geometry of the hyperbolic plane (see also 6.2) equilateral triangles exist with (equal) angles of arbitrary size $< \pi/3$. In particular, there is an equilateral triangle Δ whose angles are $\pi/4$, so the hyperbolic plane can be paved with copies of Δ, eight of which surround each vertex in the tessellation. If we select any vertex P_0, take alternate edges emanating from P_0, ending at P_1, P_2, P_3, P_4 say, again take alternate edges emanating from these P_i (starting with those which lead back to P_0) then repeat the process at the free endpoints of the new edges, etc., then the resulting graph is a tree. Prove this, and deduce that there is a pair of rigid motions of the hyperbolic plane which generate the free group F_2 on two generators.

Generalize the construction to obtain the free group on n generators F_n.

2.1.4 Solution of the Word Problem for Free Groups

The construction of the Cayley diagram in 2.1.3 is effective (relative to the generating set, at any rate), hence it yields an algorithm for the solution of the word problem. Namely, trace the path in $\mathscr{T}$ corresponding to a given word in F and see if it is closed. More generally, one can compute a normal form equivalent of a given word w, the *reduced word* $\rho(w)$, by constructing the path p corresponding to w and finding its reduced form by removing spurs. Because of the uniqueness of reduced paths (2.1.2), the result is independent of the order in which spurs are removed. The corresponding algebraic process, cancellation of terms $a_ia_i^{-1}$ or $a_i^{-1}a_i$, therefore leads to a unique reduced word regardless of the order of operations, and $w = 1$ in F if and only if $\rho(w) \equiv 1$.

Thus simple cancellation (in any order) is an algebraic algorithm for the solution of the word problem in F. This confirms the commonsense impression that one decides whether a given element equals 1 in F simply by cancelling as much as possible.

EXERCISE 2.1.4.1. If paths are taken to emanate from the centre of the circle $\mathscr{C}_1$ in 2.1.3, show that the reduced form of a path p has length equal to the index k of the circle $\mathscr{C}_k$ containing p's final point.

2.1.5 Spanning Trees

A *spanning tree* $\mathscr{T}$ of a graph $\mathscr{G}$ is a tree contained in $\mathscr{G}$ which includes all the vertices of $\mathscr{G}$.

Every connected graph $\mathscr{G}$ contains a spanning tree.

To give a constructive proof we shall assume that $\mathscr{G}$ has at most a countable infinity of edges $\mathscr{C}_1, \mathscr{C}_2, \ldots$.

Step 1. Select any vertex P_o, and for each vertex P_i of $\mathscr{G}$ which is one edge distant from P_o choose one edge (say, the one with least index) between P_o and P_i to put in $\mathscr{T}$. The result is a star we call $\mathscr{T}_1$.

Step $k + 1$. Let $\mathscr{T}_k$ be the tree constructed up to the end of step k. For each vertex P_i of $\mathscr{G}$ which is not in $\mathscr{T}_k$ and which is one edge distant from a vertex of $\mathscr{T}_k$ choose the edge of least index connecting P_i to $\mathscr{T}_k$ and add this edge to $\mathscr{T}_k$. The resulting graph $\mathscr{T}_{k+1}$ is also a tree, by the argument used in 2.1.3.

The tree $\mathscr{T}$ is the union of the $\mathscr{T}_k$ for $k = 1, 2, 3, \ldots$. $\mathscr{T}_k$ contains all vertices connected to P_o by a path of length $\leq k$, so every vertex of $\mathscr{G}$ is in some $\mathscr{T}_k$ (by connectedness) and hence in $\mathscr{T}$. □

The above proof can be repeated verbatim for any graph whose edges are indexed by a well-ordered set, hence the result is true for an arbitrary graph, assuming the axiom of choice.

EXERCISE 2.1.5.1. (For readers familiar with the axiom of choice). Prove that the existence of a spanning tree in an arbitrary graph implies the axiom of choice.

EXERCISE 2.1.5.2. Show that the two-dimensional lattice graph (Figure 123) has a spanning tree homeomorphic to the real line (that is, every vertex meets two edges).

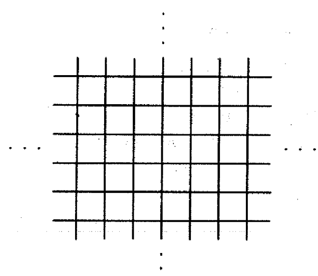


Figure 123

2.1.6 The Fundamental Group of a Graph

The fundamental group $\pi_1(\mathscr{G})$ of a graph $\mathscr{G}$ is defined combinatorially in terms of the relation of path equivalence given in 2.1.2: we choose a vertex P of $\mathscr{G}$ and consider the equivalence classes of closed paths which begin and end at P. If $[p]$ denotes the equivalence class of such a path p, we define the product of equivalence classes by

$$[p_1]\cdot[p_2] = [p_1p_2].$$

This product is well-defined on equivalence classes, since changing the representative of one factor by insertion or removal of spurs merely changes the representative of the product by insertion or removal of spurs.

It is clear that $[p]^{-1} = [p^{-1}]$ and that the identity element 1 is the class of paths whose reduced form is P itself. Thus $\pi_1(\mathscr{G})$ is indeed a group.

EXERCISE 2.1.6.1. Show that we are entitled to omit mention of P in the notation for the fundamental group of a connected graph by proving that choice of any other vertex P' as the origin for closed paths leads to an isomorphic group.

2.1.7 Generators for the Fundamental Group

We use a spanning tree $\mathscr{T}$ of $\mathscr{G}$ to find a canonical equivalent of each closed path p from P.

For each vertex P_l of $\mathscr{G}$ we construct an *approach path* w_l, namely the unique reduced path in $\mathscr{T}$ from P to P_l. Then for each edge $e_i = P_jP_k$ of $\mathscr{G}$ consider the closed path

$$a_i = w_je_iw_k^{-1}$$

Any closed path p from P is equivalent to a product of the a_i's or their inverses, in fact if

$$p = e_{i_1}^{\varepsilon_1}e_{i_2}^{\varepsilon_2}\cdots e_{i_n}^{\varepsilon_n} \quad \text{(where each } \varepsilon_j = \pm 1\text{)}$$

is such a path then p is equivalent to

$$a_{i_1}^{\varepsilon_1}a_{i_2}^{\varepsilon_2}\cdots a_{i_n}^{\varepsilon_n}$$

because the approach paths to and from $e_{i_1}^{\varepsilon_1}$, $e_{i_n}^{\varepsilon_n}$ respectively are just P itself ($=1$) and the approach paths between successive edges cancel.

Notice that if e_i is in $\mathscr{T}$ then a_i is a closed path in the tree $\mathscr{T}$ and hence equivalent to P. We can therefore omit these a_i and take the generators for $\pi_1(\mathscr{G})$ to be just the $[a_i]$ which correspond to edges e_i not in $\mathscr{T}$. □

EXERCISE 2.1.7.1. Show that the number of edges not in a spanning tree $\mathscr{T}$ of $\mathscr{G}$ is independent of the choice of $\mathscr{T}$.

2.1.8 Freeness of the Generators

To prove that the generators $[a_i]$ just obtained for $\pi_1(\mathscr{G})$ are free we first simplify the graph by shrinking the spanning tree $\mathscr{T}$ to a single vertex P. Only the edges e_i not in $\mathscr{T}$ then remain, as loops attached to P, and we call the resulting graph a *bouquet of circles*, $\mathscr{B}$ (Figure 124).

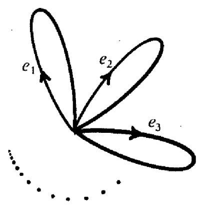


Figure 124

A given product $p(a_i)$ of a_i's in $\mathscr{G}$ becomes the corresponding product $p(e_i)$ of e_i's in $\mathscr{B}$, and $p(e_i) = 1$ in $\mathscr{B}$ if and only if $p(a_i) = 1$ in $\mathscr{G}$, for any spur $e_i e_i^{-1}$ or $e_i^{-1} e_i$ in $p(e_i)$ corresponds to $a_i a_i^{-1}$ or $a_i^{-1} a_i$ in $p(a_i)$, which can likewise be cancelled, and any spur in $\mathscr{G}$ maps to either the vertex of $\mathscr{B}$, or to a spur in $\mathscr{B}$. But the *only* spurs in $\mathscr{B}$ are of the form $e_i e_i^{-1}$ or $e_i^{-1} e_i$, so a product $p(e_i)$ equals 1 only if $p(e_i) = 1$ in the free group generated by $e_1, e_2, \ldots$. In other words, $[e_1], [e_2], \ldots$ are free generators for $\pi_1(\mathscr{B})$, and hence $[a_1], [a_2], \ldots$ are free generators for $\pi_1(\mathscr{G})$. □

The shrinking process we have just used is an instance of *collapsing*, a method of trimming unnecessary fat from a space without changing its fundamental group (for a general definition, see 3.3). The simpler form of the collapsed space makes the fundamental group easier to survey.

2.1.9 The Tree as the Universal Covering Graph of the Bouquet of Circles

In the above we have found realizations of the free group F in terms of two graphs which represent opposite extremes in structure—the tree and the bouquet of circles. The Cayley diagram of F is a tree, while the bouquet realizes F as the group of equivalence classes of closed paths.

The geometric relationship between these two can be grasped if one observes that the neighbourhood of the single vertex in the bouquet looks like a typical vertex in the Cayley diagram $\mathscr{T}$ (Figure 125). Relative to a given vertex of $\mathscr{T}$, we shall let $\tilde{e}_i^{+1}$ and $\tilde{e}_i^{-1}$ respectively denote the outgoing and incoming edges labelled e_i, and they will be said to *cover* the e_i^{+1} and e_i^{-1} respectively in the bouquet $\mathscr{B}$. Thus for a given vertex $\tilde{P}$ in $\mathscr{T}$ there is a unique

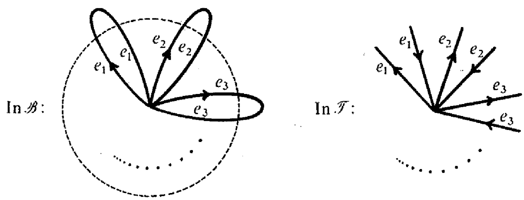


Figure 125

$\tilde{e}_i^{\pm 1}$ covering a given $e_i^{\pm 1}$ in $\mathscr{B}$. In turn, an $e_j^{\pm 1}$ which follows $e_i^{\pm 1}$ in a path p in $\mathscr{B}$ is covered by a unique $\tilde{e}_j^{\pm 1}$ relative to the final point of the $\tilde{e}_i^{\pm 1}$ covering $e_i^{\pm 1}$. Continuing in this way we obtain a unique path $\tilde{p}$ from $\tilde{P}$ in $\mathscr{T}$ which will be said to cover the path p in $\mathscr{B}$. Conversely, a given path $\tilde{p}$ emanating from $\tilde{P}$ covers a unique path p in $\mathscr{B}$.

In short, $\mathscr{T}$ exhibits all the paths of $\mathscr{B}$ in "unrolled" form. $\mathscr{T}$ is called the *universal covering graph* of $\mathscr{B}$, and it may be compared with the universal covering surface of 1.4. A universal covering is always simply connected, and therefore easier to survey, provided it can be effectively constructed at all. However its construction is equivalent to solving the word problem for the fundamental group (in this case via the Cayley diagram), so it is possible only when the solution to the word problem exists.

The paths $\tilde{p}$ and p are of course just Cayley diagram and fundamental group realizations of the same group element from F. We can also interpret $\tilde{p}$ more abstractly as the "rigid motion" of $\mathscr{T}$ which sends a given vertex $\tilde{P}^{(i)}$ at the end of the path $\tilde{p}_i$ from $\tilde{P}$ to the end of the path $\tilde{p}\tilde{p}_i$ from $\tilde{P}$. F can therefore also be realized as a group of *motions* or automorphisms of $\mathscr{T}$, and in this context we describe it as the covering motion group (German: *Deckbewegungsgruppe*). The "motion" terminology is an extrapolation from the situation with the universal covering surface, where the automorphisms can be realized by genuine rigid motions in the sense of euclidean or non-euclidean geometry. (Exercise 2.1.3.2 shows that this interpretation is also possible for finitely generated free groups.)

The most elementary, but nevertheless instructive, example of a universal covering graph is the covering of the circle by the line (Figure 126). In this case the motions of $\mathscr{T}$ are translations by an integer multiple of the generating translation, which conforms with the fact that F_1 is isomorphic to the additive group of integers.

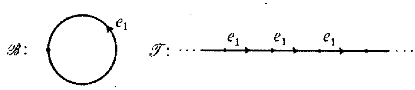


Figure 126

EXERCISE 2.1.9.1. Show that an automorphism of $\mathscr{T}$ in the above sense can be characterized as a one-to-one map of $\mathscr{T}$ onto itself which

(i) preserves endpoints (combinatorial equivalent of continuity),
(ii) preserves labels (including orientation).

2.2 Realization of Subgroups

2.2.1 Covering Graphs

A graph $\tilde{\mathscr{G}}$ is said to *cover* a graph $\mathscr{G}$ if there is a map $\phi: \tilde{\mathscr{G}} \to \mathscr{G}$ (called the *covering map* or *projection*) from the vertices and oriented edges of $\tilde{\mathscr{G}}$ onto the vertices and oriented edges, respectively, of $\mathscr{G}$ with the following properties:

(1) ϕ preserves endpoints, that is, if an oriented edge $\tilde{e}_j$ in $\tilde{\mathscr{G}}$ has initial point $\tilde{X}_j$ and final point $\tilde{Y}_j$, then $\phi(\tilde{e}_j)$ has initial point $\phi(\tilde{X}_j)$ and final point $\phi(\tilde{Y}_j)$.
(2) $(\phi(\tilde{e}_j))^{-1} = \phi(\tilde{e}_j^{-1})$.
(3) If $\phi(\tilde{P}_i) = P_i$ and $\tilde{e}_{i_1}, \tilde{e}_{i_2}, \ldots$ are the oriented edges with initial point $\tilde{P}_i$, and if $e_{i_1}, e_{i_2}, \ldots$ are the oriented edges with initial point P_i, then ϕ maps the collection $\{\tilde{e}_{i_k}\}$ one-to-one onto the collection $\{e_{i_k}\}$.

Condition (3) is a "local homeomorphism" condition which says that the neighbourhoods of corresponding vertices look alike.

When we speak of a graph $\tilde{\mathscr{G}}$ covering a graph $\mathscr{G}$ we have in mind a particular covering map $\phi: \tilde{\mathscr{G}} \to \mathscr{G}$, but in practice this map can be adequately represented by labelling and orienting the edges in $\tilde{\mathscr{G}}$, as we have done with the universal covering graph in 2.1.9. Each edge labelled e_i in $\tilde{\mathscr{G}}$ is mapped to the single edge labelled e_i in $\mathscr{G}$, with preservation of endpoints and orientation. Another example is shown in Figure 127. This $\tilde{\mathscr{G}}$ is called the *universal abelian cover* of $\mathscr{G}$, for reasons which will become more apparent later.

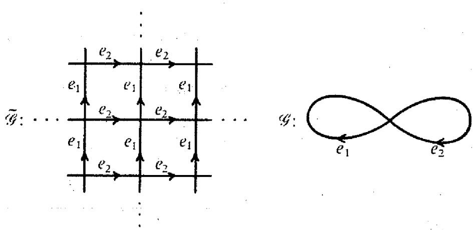


Figure 127

The local homeomorphism condition (3) implies that a path p in $\mathscr{G}$ is covered by a unique path $\tilde{p}$ in $\tilde{\mathscr{G}}$ starting from a given vertex $\tilde{P}$. The proof is by "lifting" the successive oriented edges $e_i, e_j, \ldots$ of p to covering edges exactly as in 2.1.9. If $\mathscr{G}$ is a connected graph and p is a path between an arbitrary pair of vertices P_i, P_j then the paths $\tilde{p}$ which cover p set up a one-to-one correspondence between the vertices $\tilde{P}_i^{(1)}$, $\tilde{P}_i^{(2)}, \ldots$ and $\tilde{P}_j^{(1)}$, $\tilde{P}_j^{(2)}, \ldots$ in $\tilde{\mathscr{G}}$ which cover P_i and P_j respectively. Thus the number of vertices in $\mathscr{G}$ which cover a given vertex in $\mathscr{G}$ (and similarly the number of edges which cover a given edge) is a constant, called the *sheet number* of the covering (another carryover from the theory of covering surfaces).

It will be convenient to use the following notational convention in discussing coverings: $\tilde{X}$ will denote any element in $\tilde{\mathscr{G}}$ which covers the particular element X of $\mathscr{G}$. Particular instances of $\tilde{X}$ will be distinguished, if they have to be, as $\tilde{X}^{(1)}, \tilde{X}^{(2)}, \ldots$.

2.2.2 The Subgroup Property

The definition of covering in 2.2.1 is geometrically motivated, in particular by the example of covering surfaces, however it is also significant from the viewpoint of the fundamental group.

Condition (3) implies that the paths covering spurs in $\mathscr{G}$ are exactly the spurs in $\tilde{\mathscr{G}}$, therefore equivalent paths in $\mathscr{G}$ are covered by equivalent paths in $\tilde{\mathscr{G}}$ and conversely. If we choose a vertex P in $\mathscr{G}$ as the initial point for closed paths p, we then have a one-to-one correspondence between the elements $[p]$ of $\pi_1(\mathscr{G})$ and the equivalence classes $[\tilde{p}]$ of (not necessarily closed) covering paths $\tilde{p}$ in $\tilde{\mathscr{G}}$ which emanate from some fixed vertex $\tilde{P}^{(0)}$ covering P.

Conditions (1) and (2) respectively say that ϕ sends products to products and inverses to inverses, so the correspondence between path classes is in fact a monomorphism

$$\phi_*: \pi_1(\tilde{\mathscr{G}}) \to \pi_1(\mathscr{G})$$

when restricted to the closed path classes in $\tilde{\mathscr{G}}$. In other words, $\pi_1(\tilde{\mathscr{G}})$ *is isomorphic to a subgroup of* $\pi_1(\mathscr{G})$. We shall not distinguish between $\pi_1(\tilde{\mathscr{G}})$ and the image of ϕ_*.

The classes of closed paths p in $\mathscr{G}$ which lift to nonclosed paths from $\tilde{P}^{(0)}$ in $\tilde{\mathscr{G}}$ can be classified according to the final point $\tilde{P}^{(j)}$ of the covering path $\tilde{p}$. This classification is in fact the *right coset decomposition* of $\pi_1(\mathscr{G})$ modulo $\pi_1(\tilde{\mathscr{G}})$. It follows that the number of cosets, by definition the *index* of $\pi_1(\tilde{\mathscr{G}})$ in $\pi_1(\mathscr{G})$, is the sheet number of the covering.

If $[p]$, $[p']$ are in the same coset we have $gp = p'$ for some $[g] \in \pi_1(\tilde{\mathscr{G}})$. But then $\tilde{p}' = \tilde{g}\tilde{p}$, where $\tilde{g}$ runs from $\tilde{P}^{(0)}$ to $\tilde{P}^{(0)}$ by hypothesis; hence $\tilde{p}'$ and $\tilde{p}$ have the same initial point $\tilde{P}^{(0)}$, and the same final point $\tilde{P}^{(j)}$ which must cover P since p, p' have final point P.

Conversely, if $\tilde{p}$, $\tilde{p}'$ run from $\tilde{P}^{(0)}$ to the same vertex $\tilde{P}^{(j)}$ covering P, then p and p' end at P, so $[p]$, $[p']$ are elements of $\pi_1(\mathscr{G})$, and $[p'][p]^{-1}$ is the projection of the closed path $\tilde{p}'\tilde{p}^{-1}$, hence an element of $\pi_1(\tilde{\mathscr{G}})$. Thus $[p'][p]^{-1} = [g] \in \pi_1(\tilde{\mathscr{G}})$, or $[p'] = [g][p]$, which means $[p]$, $[p']$ are in the same right coset. □

Consider the example of the universal abelian cover in 2.2.1 (Figure 128). The closed paths from $\tilde{P}^{(0)}$ in $\tilde{\mathscr{G}}$ are exactly those for which the sum of the exponents on both e_1 and e_2 is zero. Such terms do indeed constitute a subgroup of the free group F_2 generated by e_1, e_2, known as the *commutator subgroup* K_2. Thus $K_2 = \pi_1(\tilde{\mathscr{G}})$ and since any spanning tree for $\tilde{\mathscr{G}}$ omits

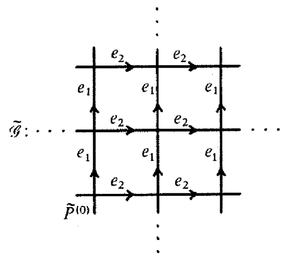


Figure 128

infinitely many edges, 2.1.7 and 2.1.8 tell us that the commutator subgroup of F_2 is an infinitely generated free group (an unpublished result of Artin from the 1920s). The vertices of $\tilde{\mathscr{G}}$ are determined by paths $e_1^m e_2^n$ as $\langle m, n\rangle$ runs through all ordered pairs of integers, which confirms the fact that the elements $e_1^m e_2^n$ are a set of right coset representatives for F_2 modulo K_2.

EXERCISE 2.2.2.1. Prove that K_2 is the normal subgroup generated by the commutator $e_1 e_2 e_1^{-1} e_2^{-1}$. (The geometric equivalent of this statement is that any closed path in $\tilde{\mathscr{G}}$ is equivalent to a product of paths of the form shown in Figure 129, or their inverses.)

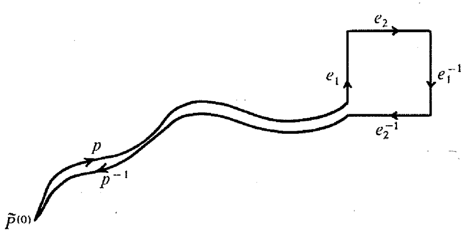


Figure 129

2.2.3 Realization of an Arbitrary Subgroup of a Free Group

Given a free group F, we realize it as $\pi_1(\mathcal{G})$ where $\mathcal{G}$ is a bouquet of circles with vertex P, as in 2.1.8. Then if G is a subgroup of F, 2.2.2 tells us that a realization of G as $\pi_1(\tilde{\mathcal{G}})$ where $\tilde{\mathcal{G}}$ covers $\mathcal{G}$ must have a vertex $\tilde{P}^{(j)}$ covering P for each right coset of F modulo G, one of which, $\tilde{P}^{(0)}$, corresponds to G itself.

Since $\tilde{\mathcal{G}}$ is a covering, each vertex $\tilde{P}^{(j)}$ will have exactly one out-going and one incoming edge labelled e_i for each generating circle e_i in $\mathcal{G}$. But this means a connected $\tilde{\mathcal{G}}$ is uniquely determined, because the outgoing edge e_i from the vertex corresponding to the coset $G[p]$ must end at the vertex corresponding to the coset $G[pe_i]$.

The $\tilde{\mathcal{G}}$ we have just described is indeed such that $\pi_1(\tilde{\mathcal{G}}) = G$.

A path $\tilde{p}$ from $\tilde{P}^{(0)}$ in $\tilde{\mathcal{G}}$ which covers p in $\mathcal{G}$ leads to the vertex $\tilde{P}^{(j)}$ corresponding to the coset $G[p]$. Thus $\tilde{p}$ is closed just in case $G[p] = G$, that is, if $[p] \in G$, and 2.2.2 then tells us that $\pi_1(\tilde{\mathcal{G}})$ is isomorphic to G. □

Just as we speak of a "covering $\tilde{\mathcal{G}}$" when a covering map $\phi: \tilde{\mathcal{G}} \to \mathcal{G}$ is actually meant, we shall also speak of a subgroup G of F being *realized by* $\tilde{\mathcal{G}}$ when we really mean that $\phi: \tilde{\mathcal{G}} \to \mathcal{G}$ induces a monomorphism $\phi_*: \pi_1(\tilde{\mathcal{G}}) \to \pi_1(\mathcal{G})$, where $\pi_1(\tilde{\mathcal{G}}) = G$ and $\pi_1(\mathcal{G}) = F$.

The example in Figure 130 is the covering $\phi: \tilde{\mathcal{G}} \to \mathcal{G}$ which realizes the subgroup F_1 (generator e_1) of F_2 (generators e_1, e_2): It is clear that the powers of e_1 are exactly the closed paths in $\mathcal{G}$ covered by closed paths from $\tilde{P}^{(0)}$ in $\tilde{\mathcal{G}}$. This $\tilde{\mathcal{G}}$ has no nontrivial automorphisms, so a covering graph need not be at all "homogeneous," as the universal cover and universal abelian cover may have tended to suggest. We shall see in 2.2.7 that a *normal* subgroup G yields a cover whose automorphism group is F/G.

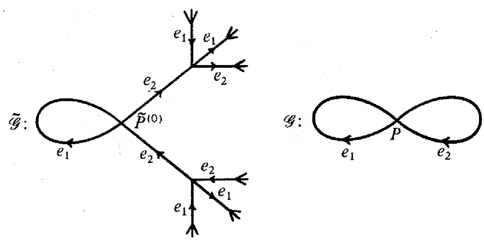


Figure 130

EXERCISE 2.2.3.1. Generalize the construction of $\tilde{\mathcal{G}}$ to the case where F is realized by an arbitrary graph $\mathcal{G}$.

2.2.4 The Nielsen–Schreier Theorem

Every subgroup of a free group is free.

This follows immediately from 2.2.3. Given any subgroup G of a free group F, we realize G as $\pi_1(\tilde{\mathscr{G}})$ for some graph $\tilde{\mathscr{G}}$. G is then a free group, since the fundamental group of any graph is free (2.1.8). □

The above proof is so slick it seems almost like magic. Perhaps the best way to explain how the result falls out is that when a group is realized as the fundamental group of a space, the notion of subgroup is exactly what is realized by the notion of covering space. This general fact (given the appropriate general notion of covering) was first observed by Reidemeister 1928, and we shall apply it to groups which are not necessarily free in Chapter 4.

Nevertheless, the cleverness of the covering space proof has some inbuilt disadvantages. Firstly, it requires the coset decomposition of F modulo G. It is not clear how to effectively construct this when G is given, say, by a set of generators. Without it, the proof does not supply a set of free generators for G. Secondly, it obscures the fact which one feels to be the intuitively correct basis of the Nielsen–Schreier theorem, namely, that elements $g_1, g_2, \ldots$ of a free group F generate a free subgroup because they cannot cancel except for trivial reasons.

These disadvantages are absent from Nielsen's proof (Nielsen 1921), which tackles the problem head-on in a way which is delightfully free of abstract technicalities. The reader does not even have to know what a group is, since the problem is posed as one of computation with products of non-commuting factors $a_1, \ldots, a_m$, each one of which has an inverse a_i^{-1} satisfying $a_i a_i^{-1} = a_i^{-1} a_i = 1$. The following exercise breaks down Nielsen's proof into simple steps.

Exercise 2.2.4.1. Consider the free group F and a subgroup G generated by elements $u_1, \ldots, u_n$ of F. For convenience we shall assume that whenever a word w is a member of a set of generators so is w^{-1}. A product of generators will be called *proper* if no generators w, w^{-1} occur as adjacent terms. A transformation of u_i into a proper product

$$u_i' = u_i u_j \quad \text{or} \quad u_i' = u_j u_i$$

is called a *Nielsen transformation*.

(1) If $u_1, \ldots, u_n$ generate G and u_i' results from u_i by a Nielsen transformation, prove that $u_1, \ldots, u_{i-1}, u_i', u_{i+1}, \ldots, u_n$ also generate G. The length $l(w)$ of a word w is the number of letters (with exponent $+1$ or -1) in the reduced word $\rho(w)$. A Nielsen transformation is called *length-reducing* when

$$l(u_i') < l(u_i).$$

(2) Describe how a finite sequence of length-reducing Nielsen transformations can be used to obtain a set of generators $v_1, \ldots, v_p$ $(p \leq n)$ for G with the property

$$l(v_i v_j) \geq l(v_i) \quad \text{for a proper product } v_i v_j \qquad (*)$$

so that no generator cancels more than half of another. The possibility remains that a generator v_i of even length may have both left and right halves cancelled in a proper product $v_k v_i v_j$. Let $v_i = l_i r_i$ be the decomposition into halves and suppose that

$$v_k = x_i l_i^{-1}, \qquad v_j = r_i^{-1} y_i.$$

(3) Show that the l_i^{-1} in v_k can be replaced by r_i, and the r_i^{-1} in v_j by l_i, by Nielsen transformations. The latter transformations, which obviously preserve length, may be called cancellation-reducing, because

(4) If $v_k' = x_i r_i$ then v_k' does not cancel the left half of v_i. Why? (Similarly $v_j' = l_i y_i$ does not cancel the right half.)

(5) Order the words in the letters of G so that each has only finitely many predecessors, and do not admit a cancellation-reducing Nielsen transformation unless the word removed (l_i^{-1} or r_i^{-1}) is replaced by a word (r_i or l_i respectively) earlier in the ordering.

(6) Show that by suitably interweaving finite sequences of cancellation-reducing and length-reducing Nielsen transformations one can obtain a set of generators $w_1, \ldots, w_q$ ($q \leq p$) for G with the properties (*) and

$$l(w_k w_i w_j) > l(w_k) + l(w_j) - l(w_i) \quad \text{for a proper product } w_k w_i w_j. \qquad (**)$$

(7) Deduce that G is freely generated by $w_1, \ldots, w_q$.

EXERCISE 2.2.4.2. Assuming a well-ordering of the words in the letters of G, extend the above argument to the infinitely generated case.

EXERCISE 2.2.4.3 (Nielsen 1921). If $G = F$ show that the generators $w_1, \ldots, w_q$ found in 2.2.4.1 must all be single letters. Deduce that q is independent of the initial choice of generators $u_1, \ldots, u_n$ (q is called the *rank* of F. Its invariance can also be proved by linear algebra if F is first "abelianized," see 5.3.2).

EXERCISE 2.2.4.4. Give an algorithm which decides whether the generators of a subgroup generate freely.

EXERCISE 2.2.4.5 (Nielsen 1921). The *generalized word problem* for a free group F is to decide, given words $u_1, \ldots, u_n$ and w, whether w is in the subgroup G of F generated by $u_1, \ldots, u_n$. Derive an algorithm for the generalized word problem using the generators $w_1, \ldots, w_q$ found for G in 2.2.4.1. (It follows that the elements of F can be effectively divided into cosets modulo G.)

2.2.5 The Schreier Index Formula

Suppose that F has rank r_F and G has index i in F. Then the rank of G is given by

$$r_G = ir_F - i + 1.$$

Since F has rank r_F it is realized as π_1 of a bouquet $\mathscr{G}$ of r_F circles $e_1, \ldots, e_{r_F}$. This means that there are r_F outgoing edges (labelled $e_1, \ldots, e_{r_F}$) from each vertex in the covering graph $\tilde{\mathscr{G}}$ which realizes G. Since G has index i in F there are i vertices in $\tilde{\mathscr{G}}$ and hence ir_F edges.

Now a spanning tree for a graph with i vertices has $(i - 1)$ edges (since the first edge takes two vertices and each subsequent edge takes one more) hence there are

$$ir_F - (i - 1) = ir_F - i + 1$$

edges of $\tilde{\mathscr{G}}$ not in the spanning tree. By 2.1.7 and 2.1.8 this is the number, r_G, of free generators of $\pi_1(\tilde{\mathscr{G}}) = G$. □

The Schreier formula can be rewritten $i = (r_G - 1)/(r_F - 1)$. Then if r_F is given, the fact that i must be an integer > 0 excludes certain values of r_G from being the ranks of subgroups of finite index in F (in particular, all values $r_G < r_F$). A subgroup G whose rank is one of the excluded values must therefore be of infinite index, so the covering graph method for finding free generators of G involves an unnecessary detour into the infinite when compared with the Nielsen method, quite apart from the problem of finding coset representatives in the first place. This should be kept in mind when reading the next section.

Exercise 2.2.5.1. To what extent can the Schreier formula be considered valid for infinite values of r_F, r_G, or i?

2.2.6 Schreier Transversals

The proof of the Nielsen–Schreier theorem in Schreier 1927 is a little more searching than the one given in 2.2.4. Schreier also finds free generators for the subgroup G by means of a special system of coset representatives. His method refines the use of coset representatives for determining subgroups in Reidemeister 1927, but it begs to be interpreted in terms of spanning trees. In fact, the method is simply an algebraic translation of the method used for finding generators of $\pi_1(\tilde{\mathscr{G}})$ in 2.1.7, as Schreier himself points out.

Let F again be realized as $\pi_1(\mathscr{G})$, where $\mathscr{G}$ is a bouquet of circles, and let $\tilde{\mathscr{G}}$ be the covering which realizes the subgroup G of F. As we saw in 2.2.2, the vertices of $\tilde{\mathscr{G}}$ correspond to the right cosets of G in F. Thus if we choose a spanning tree $\mathscr{T}$ of $\tilde{\mathscr{G}}$ the coset corresponding to a given vertex $\tilde{P}^{(j)}$ can be associated with the unique reduced path $\tilde{p}$ in $\mathscr{T}$ from $\tilde{P}^{(o)}$ to $\tilde{P}^{(j)}$, and if p is the (closed) path in $\mathscr{G}$ covered by $\tilde{p}$, its equivalence class $[p]$ is a representative of the coset in question. Because of the fact that an initial segment of a reduced path in $\mathscr{T}$ from $\tilde{P}^{(o)}$ is itself such a path (ending at a different vertex), the system of coset representatives $[p]$, taken as reduced words, has the property that any initial segment of a member of the system is another member of the system. Such a system of coset representatives is called a *Schreier transversal.*

Conversely, any Schreier transversal corresponds to a spanning tree $\mathscr{T}$ of $\tilde{\mathscr{G}}$.

$\mathscr{T}$ is found by lifting the reduced form, p say, of each coset representative $[p]$ to its covering path $\tilde{p}$ from $\tilde{P}^{(o)}$ in $\tilde{\mathscr{G}}$. The paths $\tilde{p}$ must constitute a tree, because any nontrivial closed path would involve two different initial segments $\tilde{p}_1$ and $\tilde{p}_2$ ending at the same vertex $\tilde{P}^{(j)}$, and then the initial segments p_1 and p_2 would give two different representatives of the same coset. The tree spans because each vertex corresponds to a distinct coset. □

Now let us adapt the construction of free generators from 2.1.7 to the situation where a Schreier transversal is known. We shall assume reduced words are used throughout, so that equivalence class brackets [ ] can be dropped. The coset representative of an element x of F will be denoted $\bar{x}$, and the elements of F which are themselves coset representatives will be denoted $w_1, w_2, \ldots$ (so $\bar{w}_i = w_i$). The free generators of F will be $e_1, e_2, \ldots$ so that each vertex of $\tilde{\mathscr{G}}$ has outgoing edges labelled $e_1, e_2, \ldots$. If the edge $\tilde{e}_i$ from $\tilde{P}^{(j)}$ to $\tilde{P}^{(k)}$ is not in the spanning tree $\mathscr{T}$ determined by the Schreier transversal then it yields a generator

$$a_{ij} = \tilde{w}_j \tilde{e}_i \tilde{w}_k^{-1},$$

where $\tilde{w}_j$ is the unique reduced path in $\mathscr{T}$ from $\tilde{P}^{(o)}$ to $\tilde{P}^{(j)}$, and hence covering some element w_j of the Schreier transversal, and $\tilde{w}_k$ is the unique reduced path in $\mathscr{T}$ from $\tilde{P}^{(o)}$ to $\tilde{P}^{(k)}$, the final point of $\tilde{w}_j \tilde{e}_i$, hence covering the element $(\overline{w_j e_i})$ of the Schreier transversal. Thus a_{ij} can be expressed

$$a_{ij} = w_j e_i (\overline{w_j e_i})^{-1}$$

as an element of F.

We therefore obtain all the free generators of G by letting w_j run through $w_1, w_2, \ldots$ and e_i through $e_1, e_2, \ldots$. In doing so, of course, we produce expressions $w_j e_i (\overline{w_j e_i})^{-1}$ corresponding to edges $\tilde{e}_i$ in the spanning tree $\mathscr{T}$. Such an expression represents a closed path in $\mathscr{T}$ and therefore has reduced form 1, so it may be immediately discarded.

EXERCISE 2.2.6.1. Use the Schreier method to find free generators for the commutator subgroup K_2 of F_2.

2.2.7 Normal Subgroups and Cayley Diagrams

If F is realized as π_1 of a bouquet of circles $\mathscr{G}$, the covering $\tilde{\mathscr{G}}$ which realizes a normal subgroup G of F is the Cayley diagram of F/G, and F/G is also the covering motion group. Thus any group H can be realized as a covering motion group.

We saw in 0.5 that any H with generators $e_1, e_2, \ldots$ has the form F/G where F is the free group generated by $e_1, e_2, \ldots$ and G is a normal subgroup of F. Now it is immediate from the definition and elementary properties of Cayley diagrams that

(1) The Cayley diagram of F/G is a covering $\tilde{\mathscr{G}}$ of the bouquet $\mathscr{G}$.
(2) The subgroup of F realized by the covering is G (closed paths in the Cayley diagram are just the elements of G).

(3) The automorphism group of the Cayley diagram of any group H is H itself.

On the other hand, we know from 2.2.3 that the covering $\tilde{\mathscr{G}}$ of $\mathscr{G}$ which realizes a given subgroup G of F is unique, hence if we construct the covering $\tilde{\mathscr{G}}$ for a *normal* subgroup G we must get the Cayley diagram of $H = F/G$, so H is also the covering motion group. □

This theorem is illustrated by the covering $\tilde{\mathscr{G}}$ in 2.2.2 which realizes the commutator subgroup K_2 of F_2. It is obvious that $\tilde{\mathscr{G}}$ is the Cayley diagram of the free abelian group on two generators, which is indeed equal to F_2/K_2.

The interpretation of Cayley diagrams brought to light by the theorem suggests we should regard graph coverings in general (at least when $\mathscr{G}$ is a bouquet of circles) as generalized Cayley diagrams. In fact, some authors call coverings of the bouquet of circles *Schreier coset diagrams*, since they were first used in Schreier 1927. Exercise 2.2.7.1 below yields a geometric characterization of the coset diagrams which are Cayley diagrams.

In 2.2.4 we pointed out that the nonconstructiveness of the Schreier method for finding free generators via a set of coset representatives could be overcome by the Nielsen method when the subgroup G was finitely generated. When G is also normal (but $\neq \{1\}$) then F/G is in fact finite (see Exercise 2.2.7.2), and we can proceed more directly to make Schreier's method effective. Namely, if F/G is defined by relators we use the method of 0.5.7 to effectively construct the Cayley diagram $\tilde{\mathscr{G}}$ of F/G, then construct a tree $\mathscr{T}$ spanning $\tilde{\mathscr{G}}$ by the method of 2.1.5. The Schreier generators $w_j e_i (\overline{w_j e_i})^{-1}$ can then be read from the edges of $\tilde{\mathscr{G}}$ which are not in $\mathscr{T}$.

This is the situation where Schreier's method is most often useful. Of course, it can also be applied to subgroups G of infinite index when the coset diagram is apparent and easy to survey, as is the case with the commutator subgroup.

EXERCISE 2.2.7.1. A covering $\tilde{\mathscr{G}}$ of $\mathscr{G}$ is called *regular* if the paths in $\tilde{\mathscr{G}}$ which cover a given closed path p in $\mathscr{G}$ are either all closed or all nonclosed. Show that this property is equivalent to the normality of the subgroup realized by the covering.

EXERCISE 2.2.7.2. Show that a Cayley diagram $\tilde{\mathscr{G}}$ is finite if and only if $\pi_1(\tilde{\mathscr{G}})$ is finitely generated.

EXERCISE 2.2.7.3. Identify the normal subgroup G of F_2 realized by the covering in Figure 131 and the quotient F_2/G. Give a set of free generators for G.

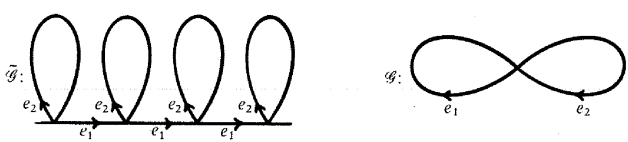
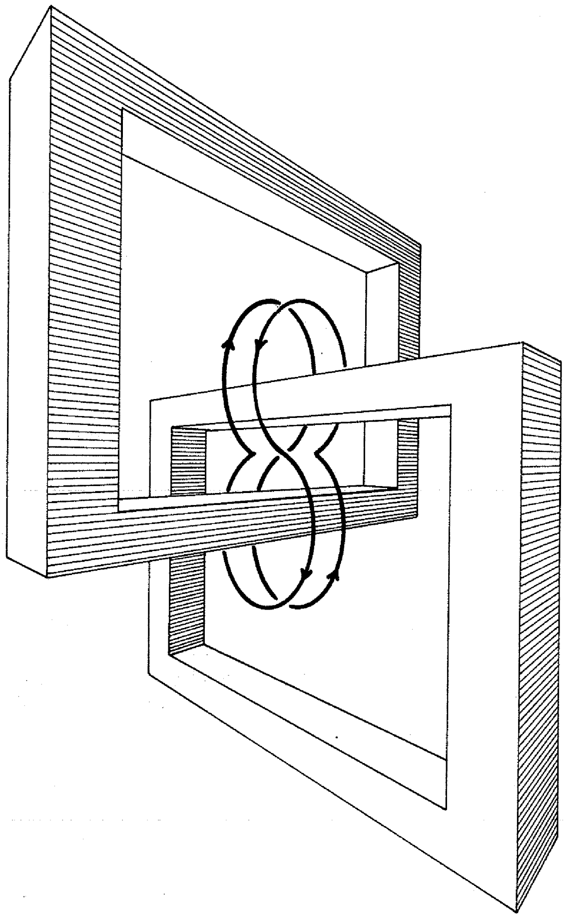


Figure 131

CHAPTER 3

Foundations for the Fundamental Group



3.1 The Fundamental Group

3.1.1 Introduction

The fundamental group was introduced by Poincaré 1892 (though anticipated to some extent by the study of curves on surfaces in Jordan 1866b). Poincaré defined the group in function-theoretic terms by considering analytic continuation of a many-valued function ϕ around a closed path p in a manifold. Since the value obtained after completing p may differ from the initial value, p may be considered to define a transformation of ϕ, and since any path p' which is deformable into p defines the same transformation, the group of transformations of the "most general" function ϕ is naturally isomorphic to the group of equivalence classes of closed paths, where "equivalent" means mutually deformable.

The inconvenience of this definition for actual computation is obvious, and Poincaré quickly moved to a combinatorial notion of the fundamental group (Poincaré 1895), in which all paths are polygonal and deformations result from pulling a polygonal path from one side of a cell to the other (the simplest case, where the cell is 1-dimensional, being insertion or removal of a spur). Such a definition makes the computation of generators and relations routine, but it is open to the objection that the group is not obviously a topological invariant. Since the topologists of the time pinned their hopes on the *Hauptvermutung*, they could be satisfied with a proof that the fundamental group was invariant under *combinatorial* homeomorphisms, which was supplied by Tietze 1908.

A new approach opened up with the proof of Alexander 1915 that the Betti and torsion numbers (see Chapter 5) are topological invariants. Alexander's proof is based on the simple observation that uniform continuity allows us to divide a curve p (or the continuous image of any simplex) in a complex $\mathscr{C}$ into a finite number of pieces which individually lie in arbitrarily small regions of $\mathscr{C}$. Such a region $\mathscr{R}$ can therefore be treated as a ball, and the piece of p in $\mathscr{R}$ deformed into a "straight" segment. In other words, there is no loss of generality in replacing p by a polygon p', and the numbers and other objects computed from polygonal paths are therefore the same as those defined, in a topologically invariant way, from arbitrary continuous paths. This applied in particular to the fundamental group, so the groups which had earlier been computed on a combinatorial basis were now placed on a topologically secure footing. This was first done in the textbook Veblen 1922.

The method of computing generators and relations from a simplicial decomposition, while routine (see 4.1.6), is clumsy, and in practice many *ad hoc* arguments were used to find simpler presentations. Seifert 1931 proved a theorem which greatly simplified the process by showing that the fundamental group of a complex which is the union of suitable sets $\mathscr{A}$ and $\mathscr{B}$

with a connected intersection arises in a canonical way from the groups of $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{A} \cap \mathscr{B}$. Seifert's proof assumed that these groups were already realized combinatorially by a simplicial decomposition of $\mathscr{A} \cup \mathscr{B}$, but a more general proof by Van Kampen 1933 led to a complete emancipation from simplicial decompositions.

The following remarks concern the exposition in this chapter. In defining the fundamental group and proving that it is a group (3.1.2–3.1.6) we have tried to avoid unnatural functions as far as possible. This involved defining a path to be an equivalence class of maps rather than a single one (an approach also adopted by de Rham 1969) and allowing arbitrary intervals and rectangles as the domains of functions. The only group which has to be derived from first principles is that of the circle (3.2); all the other fundamental groups we need are then obtained by combining the simple technique of deformation retraction (3.3) with the Seifert–Van Kampen theorem (3.4).

3.1.2 Paths

The intuitive idea of a path p is a curve with orientation indicated by arrows, for example Figure 132. However, this view is inadequate in some respects: for example, a picture of a circle with an arrow on it does not distinguish between the path which runs once round the circle and the path which runs round twice.

Figure 132

We arrive at a formal definition by considering the idea of a "journey" along the path. A *journey* in the complex $\mathscr{C}$ is a continuous map $f: \mathscr{I} \to \mathscr{C}$ where $\mathscr{I}$ is a closed interval. We think of $\mathscr{I}$ as a time interval, so $f(t)$ is the position at time t. The image of f is the curve itself while the orientation is induced by the natural (time) order on $\mathscr{I}$. In particular, if $\mathscr{I} = [a, b]$ then $f(a)$ = initial point of p, $f(b)$ = final point of p.

Another journey $f': \mathscr{I}' \to \mathscr{C}$ may be held to cover the same path if $f' = f\phi$ where $\phi: \mathscr{I}' \to \mathscr{I}$ is a continuous order-preserving bijection. (We can think of ϕ as a "time-warp.") The relation between f and f' which holds when there is such a ϕ is clearly an equivalence relation, so we can give the following

Definition. A *path* p is an equivalence class of continuous functions from closed intervals into $\mathscr{C}$, where $f: \mathscr{I} \to \mathscr{C}$ and $f': \mathscr{I}' \to \mathscr{C}$ are called equivalent if $f' = f\phi$ for some continuous order-preserving bijection $\phi: \mathscr{I}' \to \mathscr{I}$.

The *inverse* (or oppositely oriented) path to p, p^{-1} is defined to be the equivalence class of $f\psi: \mathscr{I} \to \mathscr{C}$ where ψ is a continuous *order-reversing* bijection. In intuitive terms, f and $f\psi$ are journeys along the same curve in opposite directions.

3.1.3 Notation

In what follows it will be convenient to have a notation which reflects the fact that one interval or function is a "*continuation*" of another. We call $\mathscr{I}_2$ a *continuation* of $\mathscr{I}_1$ if $\mathscr{I}_1 = [a, b]$ and $\mathscr{I}_2 = [b, c]$ and just in this case use $\mathscr{I}_{12}$ to denote $\mathscr{I}_1 \cup \mathscr{I}_2$. In analogous circumstances $\mathscr{I}_{23} = \mathscr{I}_2 \cup \mathscr{I}_3$, $\mathscr{I}_{123} = \mathscr{I}_1 \cup \mathscr{I}_2 \cup \mathscr{I}_3$, and so on.

A function $f_2: \mathscr{I}_2 \to \mathscr{C}$ is a *continuation* of $f_1: \mathscr{I}_1 \to \mathscr{C}$ provided

(i) $\mathscr{I}_1 = [a, b]$ and $\mathscr{I}_2 = [b, c]$ and
(ii) $f_1(b) = f_2(b)$

and just in this case we use $f_{12}: \mathscr{I}_{12} \to \mathscr{C}$ to denote $f_1 \cup f_2$. The idea extends similarly to f_{23}, f_{123}, and so on.

We shall also consider functions of two variables $h_1: \mathscr{J}_1 \times \mathscr{I}_1 \to \mathscr{C}$ where $\mathscr{J}_1$, $\mathscr{I}_1$ are intervals. In this case h_1 can be continued to an h_2 in two ways—by continuation of $\mathscr{J}_1$ to $\mathscr{J}_2$ or continuation of $\mathscr{I}_1$ to $\mathscr{I}_2$. In either case we use h_{12} to denote $h_1 \cup h_2$ (though of course its domain is $\mathscr{J}_{12} \times \mathscr{I}_1$ in the first case, $\mathscr{J}_1 \times \mathscr{I}_{12}$ in the second).

3.1.4 Products of Paths

The trouble involved in finding a natural formalization of the notion of path is worthwhile because it gives a natural notion of the *product* p_1p_2 which results from successive journeys along paths p_1, p_2 where

$$\text{final point of } p_1 = \text{initial point of } p_2.$$

Namely, choose journeys $f_1: \mathscr{I}_1 \to \mathscr{C}$, $f_2: \mathscr{I}_2 \to \mathscr{C}$ along p_1, p_2 where $\mathscr{I}_2$ is a continuation of $\mathscr{I}_1$; then

$$f_{12}: \mathscr{I}_{12} \to \mathscr{C}$$

is a journey along p_1, p_2 in succession and hence a natural representative of p_1p_2.

This product is well defined on equivalence classes, for if $f'_1: \mathscr{I}'_1 \to \mathscr{C}$, $f'_2: \mathscr{I}'_2 \to \mathscr{C}$ are other journeys along p_1, p_2 where $\mathscr{I}'_2$ is a continuation of $\mathscr{I}'_1$ and $f'_1 = f_1\phi_1, f'_2 = f_2\phi_2$, then $f'_{12}: \mathscr{I}'_{12} \to \mathscr{C}$ equals $f_{12}\phi_{12}$, and hence is equivalent to f_{12}.

The product is associative too, for if p_1, p_2, p_3 are paths such that

$$\text{final point of } p_1 = \text{initial point of } p_2,$$

$$\text{final point of } p_2 = \text{initial point of } p_3,$$

we can choose journeys along them, $f_1: \mathscr{I}_1 \to \mathscr{C}$, $f_2: \mathscr{I}_2 \to \mathscr{C}$, $f_3: \mathscr{I}_3 \to \mathscr{C}$ such that each is a continuation of its predecessor. But then $p_1(p_2 p_3)$ and $(p_1 p_2)p_3$ are both represented by the journey $f_{123}: \mathscr{I}_{123} \to \mathscr{C}$.

3.1.5 Homotopy

We are not really interested in individual paths, only in paths which differ in a topologically interesting way. Paths will be considered equivalent if one can be deformed into the other within $\mathscr{C}$, an idea which is formalized by the notion of *homotopy*.

A homotopy between journeys $f: \mathscr{I} \to \mathscr{C}$ and $g: \mathscr{I} \to \mathscr{C}$ is a continuous function $h: \mathscr{J} \times \mathscr{I} \to \mathscr{C}$ where $\mathscr{J} = [c, d]$, $h(c, t) = f(t)$ and $h(d, t) = g(t)$. If one now thinks of $\mathscr{J}$ as a time interval, the functions $h_x(t) = h(x, t)$ for $x \in \mathscr{J}$ represent a series of journeys in a process of deformation from f to g, for example Figure 133. As the example shows, the image curve may retrace previous positions and even cross itself under deformation. Accordingly, the image of $\mathscr{J} \times \mathscr{I}$ under h, called the *deformation rectangle*, is in general a "crumpled" or *singular* rectangle, with various possible types of singularity arising from failure of h to be one-to-one. Just as in the definition of path it is important to think dynamically—in terms of the map rather than its image.

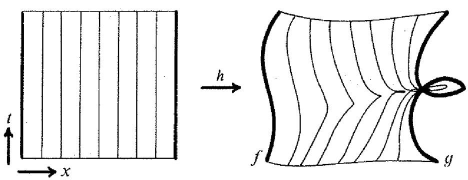


Figure 133

By a slight abuse of language, we also use the word "homotopy" to denote the *relation* "there is a homotopy between f and g." We now wish to show that this relation is well-defined on paths and that it is in fact an equivalence relation. These results are obtained by exploiting the arbitrariness of $\mathscr{I}$ and $\mathscr{J}$ respectively.

Suppose $h: \mathscr{J} \times \mathscr{I} \to \mathscr{C}$ is a homotopy between journeys $f: \mathscr{I} \to \mathscr{C}$ and $g: \mathscr{J} \to \mathscr{C}$. The bijections ϕ which establish a correspondence between the journeys $f' = f\phi$ equivalent to f and the journeys $g' = g\phi$ equivalent to g also provide homotopies between them, namely $h'(x, t) = h(x, \phi(t))$, so any journey equivalent to f is homotopic to a journey equivalent to g.

Similar application of ϕ to the x variable enables us to convert a homotopy between f and g over the interval $\mathscr{J}$ to one over any other interval $\mathscr{J}'$. In particular, if

$$h_1: \mathscr{J}_1 \times \mathscr{I} \to \mathscr{C} \text{ is a homotopy between } f_0 \text{ and } f_1$$

$$h_2: \mathscr{J}_2 \times \mathscr{I} \to \mathscr{C} \text{ is a homotopy between } f_1 \text{ and } f_2$$

we can assume $\mathscr{I}_2$ is a continuation of $\mathscr{I}_1$ and hence obtain the homotopy

$$h_{12}: \mathscr{I}_{12} \times \mathscr{I} \to \mathscr{C} \text{ between } f_0 \text{ and } f_2.$$

Thus the homotopy relation is transitive, and since it is obviously reflexive and symmetric, it is an equivalence relation, which from now on we denote by $\sim$. The $\sim$ equivalence class of a path p is denoted by $[p]$, and it will be called the *path class* of p.

EXERCISE 3.1.5.1. If p is homotopic to a point show that p is the image of the boundary of a disc $\mathscr{D}$ which maps continuously into $\mathscr{C}$ (the image of $\mathscr{D}$ is called a *singular disc*).

EXERCISE 3.1.5.2. Construct a complex $\mathscr{C}$ and a simple curve p which bounds a singular disc in $\mathscr{C}$ but not a topological disc.

3.1.6 The Group Properties

It is easy to see that the product operation on paths induces one on path classes, in other words we can define

$$[p_1] \cdot [p_2] \text{ to be } [p_1 p_2]$$

Namely, suppose $f_1 \sim g_1$ and $f_2 \sim g_2$ are homotopic pairs of journeys along paths p_1, p_2 respectively, where

$$\text{final point of } p_1 = \text{initial point of } p_2.$$

We can assume

$$h_1: \mathscr{I} \times \mathscr{I}_1 \to \mathscr{C} \text{ is a homotopy between } f_1 \text{ and } g_1$$

$$h_2: \mathscr{I} \times \mathscr{I}_2 \to \mathscr{C} \text{ is a homotopy between } f_2 \text{ and } g_2$$

and that $\mathscr{I}_2$ is a continuation of $\mathscr{I}_1$. Then

$$h_{12}: \mathscr{I} \times \mathscr{I}_{12} \to \mathscr{C}$$

is a homotopy between f_{12} and g_{12}.

Associativity for the product of path classes then follows immediately from associativity of the product of paths.

However, the path classes have a decidedly more pleasant algebraic structure under product than do paths. The classes of *closed* paths emanating from a fixed point P form a *group*, called the *fundamental group* of $\mathscr{C}$, $\pi_1(\mathscr{C})$. (It will turn out that, up to isomorphism, the group does not depend on P when $\mathscr{C}$ is arc connected.) P is called a *basepoint*.

The identity element of the group is the class of the "point path" P represented by a function with constant value P. (The paths in this class are also called *null-homotopic*.)

To see how this works, let $f_1: \mathscr{I}_1 \to \mathscr{C}$ be a journey along an arbitrary closed path p_1 from P and let $f_2: \mathscr{I}_2 \to \mathscr{C}$ be a constant journey along p_2

(= the point path P) where $\mathscr{I}_2$ is a continuation of $\mathscr{I}_1$, of unit length for simplicity. Since p_1p_2 is represented by $f_{12}: \mathscr{I}_{12} \to \mathscr{C}$ we are looking for a homotopy which "stretches" f_1 until its domain is all of $\mathscr{I}_{12}$ and at the same time "compresses" f_2 until its domain is just the final point of $\mathscr{I}_{12}$.

Let ϕ_1^x be the linear function whose inverse leaves the initial point of $\mathscr{I}_1$ fixed and increases the final point by x, ϕ_2^x the linear function whose inverse leaves the final point of $\mathscr{I}_2$ fixed and increases the initial point by x. Then

$$f_1^x = f_1\phi_1^x \text{ is the "stretched" } f_1$$

$$f_2^x = f_2\phi_2^x \text{ is the "compressed" } f_2$$

and $h(x, t) = f_{12}^x(t)$ for x from 0 to 1 is the required homotopy.

(Intuitively, one deforms the journey across p_1p_2 in the time interval $\mathscr{I}_{12}$ so that the time spent on p_1 increases and the time spent on p_2 tends to 0. This is possible without discontinuity only when p_2 is the point path, since at the end one must be everywhere on p_2 at the same instant.)

The above argument shows that $[P]$ is an identity for multiplication on the right, and a similar argument shows that it is also an identity for multiplication on the left.

Finally we show that $[p]^{-1}$, the inverse of $[p]$, exists and equals $[p^{-1}]$. Choose a journey $f_1: \mathscr{I}_1 \to \mathscr{C}$ along p so that $\mathscr{I}_1 = [0, 1]$ and let the journey along p^{-1} be $f_2: \mathscr{I}_2 \to \mathscr{C}$, where $\mathscr{I}_2 = [1, 2]$ and

$$f_2(1 + t) = f_1(1 - t).$$

To show $[p][p^{-1}] = [pp^{-1}] = [P]$ we seek a homotopy between the journey $f_{12}: \mathscr{I}_{12} \to \mathscr{G}$ along pp^{-1} and the journey with constant value P.

Let

$$f_{12}^x(t) = \begin{cases} f_1(t) & 0 \le t \le 1 - x, \\ f_1(1 - x) = f_2(1 + x) & 1 - x \le t \le 1 + x, \\ f_2(t) & 1 + x \le t \le 2. \end{cases}$$

(This describes a journey which stops x time units before reaching the end of p, waits for $2x$ units, then retraces p back to the beginning.) Then $h(x, t) = f_{12}^x(t)$ for $0 \le x \le 1$ is the required homotopy.

3.1.7 Independence from the Basepoint P and Topological Invariance

If $\mathscr{C}$ is an arc connected complex then the group of closed path classes emanating from P is isomorphic to the group of closed path classes emanating from any other point Q, hence we are entitled to speak of the fundamental group of $\mathscr{C}$, and omit mention of the basepoint.

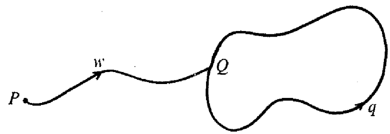


Figure 134

Choose a path w between P and Q and associate each closed path class $[q]$ emanating from Q with the closed path class $[wqw^{-1}]$ emanating from P (Figure 134). Since any path class $[p]$ from P can be written in this form, namely

$$[w(w^{-1}pw)w^{-1}]$$

we have a one-to-one correspondence Φ between the closed path classes based at Q and P respectively. Φ is also a group homomorphism, since

$$\begin{aligned}\Phi([q_1][q_2]) &= [wq_1q_2w^{-1}]\\ &= [wq_1w^{-1}wq_2w^{-1}]\\ &= [wq_1w^{-1}][wq_2w^{-1}]\\ &= \Phi([q_1])\Phi([q_2])\end{aligned}$$

and hence Φ is an isomorphism, as required. □

Since P is irrelevant to the structure of $\pi_1(\mathscr{C})$ we now denote the identity element simply by 1.

The topological invariance of $\pi_1(\mathscr{C})$ is an immediate consequence of the arbitrariness of the continuous functions which define paths. A homeomorphism $\psi: \mathscr{C} \to \mathscr{C}'$ maps a closed path p based at P to a closed path $\psi(p)$ based at $\psi(P)$, and since ψ is one-to-one and continuous it induces a one-to-one correspondence ψ_* between the path classes based at P and $\psi(P)$ respectively. Since ψ also sends products to products, it follows that ψ_* is an isomorphism between $\pi_1(\mathscr{C})$ and $\pi_1(\mathscr{C}')$.

From now on we shall deal only with arc connected complexes, and we shall also describe them by the equivalent term: *path-connected*.

3.2 The Fundamental Group of the Circle

3.2.1 The Rôle of Compactness

The combinatorial notion of path equivalence used in 2.1 and 2.2 seems a drastic simplification of the general notion of homotopy for paths. In fact this is not so, and the underlying reason is compactness. A path p, being the continuous image of a closed interval, is compact and therefore can be decomposed into finitely many subpaths $p_1, \ldots, p_n$ lying in "small" regions of the complex in question. In a "small" region, a path p_i can be tightened by an obvious homotopy to the "straight line" p_i' between its endpoints, and one can then operate combinatorially with the "straight" paths p_i' to compute certain discrete objects, such as the element corresponding to p in the (combinatorial) fundamental group.

Then in order to prove the invariance of these objects under homotopy, one makes a further application of compactness to decompose a given homotopy into "small" homotopies. Because of the discrete nature of the objects, small homotopies cannot change them, and hence they are unchanged under all homotopies.

In the present section we carry out this program for the fundamental group of the circle (see also 3.4.2, 3.4.3). We expect of course that any closed path p in the circle can be tightened by a homotopy into the product of n copies of the standard circular path e_1, for some integer n. This will show that e_1 generates the group. However, the technical difficulty in the tightening process is to know in advance what the value of n is. Consider for example the problem of tightening the portion of p between two neighbouring points P_1 and P_2 on the circle (Figure 135). The tightened form is not necessarily the short arc a_1, it could be the complementary (and oppositely oriented) arc a_2, or a complete circuit of the circle, e_1, *plus* a_1, and so on.

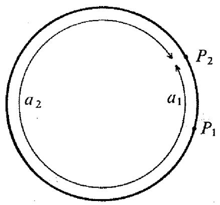


Figure 135

To overcome this problem we shall perform the tightening in two stages. Stage 1 will divide the path into finitely many subpaths which are so small that the tightened form of each is just the short arc between its endpoints, and Stage 2 will reduce the resulting product of arcs to e_1^n by removal of spurs.

3.2.2 Tightening a Path

Any closed path is homotopic to a power of e_1.

Stage 1. Let $f: \mathscr{I} \to \mathbf{S}^1$ be a journey along a closed path p in the unit circle $\mathbf{S}^1$. Since f is a continuous function and $\mathscr{I}$ is a closed interval, f is uniformly continuous and we can partition $\mathscr{I}$ into subintervals

$\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_m$ which are so small that $d(f(t), f(t')) < \pi$ for $t, t' \in \mathscr{I}_k$, where the distance function $d(\ ,\ )$ is the length of the shorter arc between the two points.

Suppose $\mathscr{I}_k = [t_k, t_{k+1}]$ and let $f_k: \mathscr{I}_k \to \mathbf{S}^1$ denote the subjourney of f restricted to $\mathscr{I}_k$. Since the image of f_k is confined to an arc $\mathscr{A}_k$ which is less than half the circle we can view $\mathscr{A}_k$ as an interval. Then it is evident from a glance at its graph (typified by Figure 136) that f_k can be deformed into the function $f'_k: \mathscr{I}_k \to \mathbf{S}^1$ which runs linearly between $f(t_k)$ and $f(t_{k+1})$. An explicit homotopy which does this is obtained by letting

$$h_k(x, t) = f_k^x(t),$$

where $f_k^x(t)$ results from $f_k(t)$ by shrinking the distance between $f_k(t)$ and $f'_k(t)$ by a factor $1 - x$ (x running from 0 to 1).

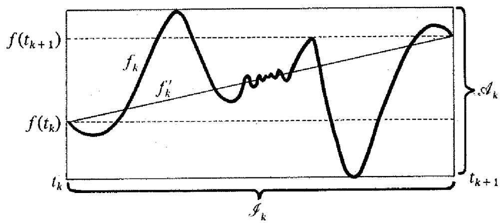


Figure 136

Stage 2. The homotopy $h_{12\cdots m}$ deforms p into the product of m circular arcs $a_1, \ldots, a_m$, each less than a semicircle. If $a_1, \ldots, a_m$ all have the same orientation then it is clear from the fact that $a_1 a_2 \cdots a_m$ is closed that it must be e_1^n for some integer n. If the orientations vary, simplify the product by cancellation as follows: let $a_1, a_2, \ldots, a_i$ have the same orientation and suppose that a_{i+1} has the opposite orientation. If a_{i+1} is shorter than a_i we have

$$a_i = a'_i a_{i+1}^{-1} \quad \text{and} \quad a_1 a_2 \cdots a_i a_{i+1} \sim a_1 a_2 \cdots a_{i-1} a'_i$$

if not

$$a_{i+1} = a_i^{-1} a'_{i+1} \quad \text{and} \quad a_1 a_2 \cdots a_i a_{i+1} \sim a_1 a_2 \cdots a_{i-1} a'_{i+1}.$$

In either case we get a homotopic path with a smaller number of factors. If all factors now have the same orientation we are finished; if not, we repeat the process until they do. □

We have now shown that e_1 generates $\pi_1(\mathbf{S}^1)$, however it is not clear that it does so freely. We have to prove that p is homotopic to e_1^n for only one value of n, but as yet it is not even clear that the value of n just computed is independent of the partition of $\mathscr{I}$ into subintervals.

EXERCISE 3.2.2.1. Show that any closed path from the vertex of a bouquet of circles $e_1, e_2, \ldots$ is homotopic to a finite product of paths e_i^{+1} or e_i^{-1}. (Actually, this is not generally true. For example, there is a single closed path which traverses every circle in the family

$$\left(x - \frac{1}{n}\right)^2 + y^2 = \frac{1}{n^2} \qquad (n = 1, 2, 3, \ldots)$$

in Figure 137. This space, sometimes called the *Hawaiian earring*, was shown to have a nonfree fundamental group by Griffiths 1956. One must therefore make an assumption which excludes this pathology—for example, let every circle have length greater than one—then find a suitable notion of "small" subpath.)

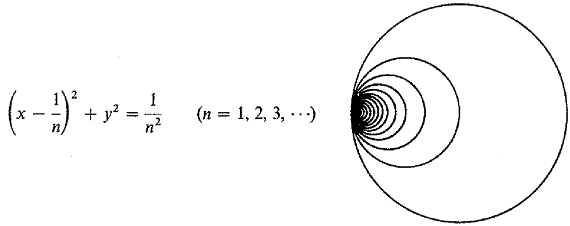


Figure 137

3.2.3 Brouwer Degree

The number n determined from $f: \mathscr{I} \to \mathbf{S}^1$ by means of the partition $\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_m$ is independent of the partition chosen and hence can be regarded as a function of f, $n(f)$, called its Brouwer degree.

To show the invariance of n under different partitions it will suffice to show that n does not change when a refinement of $\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_m$ is made, since any two partitions have a common refinement. In turn, it will suffice to look at the effect of a single subdivision of $\mathscr{I}_k = [t_k, t_{k+1}]$ into $[t_k, t^*]$ and $[t^*, t_{k+1}]$. Figure 138 compares the new situation to the original (3.2.2).

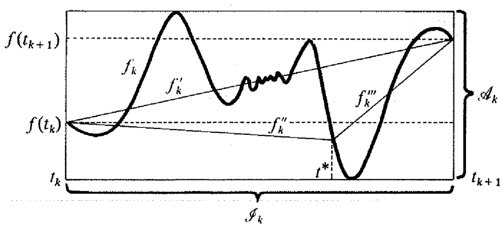


Figure 138

The tightening homotopy of Stage 1 now produces two linear functions f''_k and f'''_k in place of f'_k, and obviously the sum of the lengths of the corresponding subarcs of $\mathscr{A}_k$ (signed according to orientation) equals the length of the arc corresponding to f'_k. But n arises from Stage 2 simply as the sum of the arc lengths divided by 2π, so it is unchanged and we are justified in writing it as $n(f)$. □

The degree of a map was introduced by Brouwer 1912a, who showed that this invariant is meaningful for a mapping of any manifold into itself. In this case we are essentially dealing with a map of the manifold $\mathbf{S}^1$ into itself, since f is required to give the endpoints of $\mathscr{I}$ the same image.

EXERCISE 3.2.3.1. Show that the product of terms $e_i^{\pm 1}$ obtained according to Exercise 3.2.2.1 is independent of the decomposition of p into small subpaths, and that it is a reduced word.

3.2.4 Invariance of the Brouwer Degree under Homotopy

We first show that

An arbitrary homotopy $f \sim g$ can be decomposed into a sequence of homotopies $f \sim f_1 \sim f_2 \sim \cdots \sim f_p \sim g$ which are "small" in the sense that, for each i, $D(f_i, f_{i+1}) < \varepsilon$, where the distance D between functions defined on the interval $\mathscr{I}$ is the usual sup *norm*:

$$D(f_i, f_j) = \sup_{t \in \mathscr{I}} d(f_i(t), f_j(t)).$$

A homotopy $h: \mathscr{I} \times \mathscr{I} \to \mathbf{S}^1$ is a uniformly continuous function, so for any $\varepsilon > 0$ there is a $\delta > 0$ such that

$$|(x, t) - (x', t')| < \delta \quad \text{implies} \quad d(h(x, t), h(x', t')) < \varepsilon,$$

where $|\quad|$ denotes the usual distance in the plane. Then if we divide $\mathscr{I}$ by points $x_1, \ldots, x_p$ into subintervals of length $< \delta$ the functions $f_i(t) = h(x_i, t)$ have the property that $D(f_i, f_{i+1}) < \varepsilon$. □

Now in order to show that $n(f) = n(g)$ when $f \sim g$ it will suffice to find an ε such that $D(f, g) < \varepsilon$ implies $n(f) = n(g)$. In fact, $\varepsilon = \pi/4$ suffices.

Given $f: \mathscr{I} \to \mathbf{S}^1$ we find a partition $\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_m$ of $\mathscr{I}$ such that $d(f(t), f(t')) < \pi/2$ for any $t, t' \in \mathscr{I}_k$ and use it to determine $n(f)$ as in 3.2.2. This partition can also be used to determine $n(g)$ for any g such that $D(f, g) < \pi/4$ because in any $\mathscr{I}_k$ we have

$$\begin{aligned} d(g(t), g(t')) &\leq d(g(t), f(t)) + d(f(t), f(t')) + d(f(t'), g(t')) \\ &< \frac{\pi}{4} + \frac{\pi}{2} + \frac{\pi}{4} = \pi. \end{aligned}$$

Now $n(f) = (1/2\pi)(a_1 + a_2 + \cdots + a_m)$, where a_k = (signed) length of short arc between $f(t_k)$ and $f(t_{k+1})$ and $n(g) = (1/2\pi)(b_1 + b_2 + \cdots + b_m)$, where b_k = (signed) length of short arc between $g(t_k)$ and $g(t_{k+1})$, and the condition $D(f, g) < \pi/4$ imposes the constraints

$$|a_i - b_i| < \frac{\pi}{2} \quad \text{for each } i,$$

and for each k, $|(a_1 + \cdots + a_k) - (b_1 + \cdots + b_k)|$ differs from a multiple of 2π by $< \pi/4$. It then follows easily by induction on k that in fact

$$|(a_1 + \cdots + a_k) - (b_1 + \cdots + b_k)| < \frac{\pi}{4} \quad \text{for each } k$$

and in particular

$$|n(f) - n(g)| < \frac{1}{2\pi} \cdot \frac{\pi}{4} = \frac{1}{8}.$$

But since $n(f)$ and $n(g)$ are integers, they must be equal. □

We have now proved that paths $e_1^{n_1}$ and $e_1^{n_2}$ cannot be homotopic when $n_1 \neq n_2$ and hence that e_1 freely generates the fundamental group of the circle $\mathbf{S}^1$. Thus $\pi_1(\mathbf{S}^1)$ is the infinite cyclic group.

EXERCISE 3.2.4.1. Show that the product of terms $e_i^{\pm 1}$ associated with a closed path p in the bouquet of circles in Exercises 3.2.2.1 and 3.2.3.1 is invariant under homotopy. Deduce that the e_i's freely generate the fundamental group of the bouquet.

3.3 Deformation Retracts

3.3.1 Retracts

A complex $\mathscr{R} \subseteq \mathscr{C}$ is called a *retract* of $\mathscr{C}$ if there is a continuous map $\rho: \mathscr{C} \to \mathscr{R}$ (called the *retraction*) such that $\rho(P) = P$ for each $P \in \mathscr{R}$.

A retraction ρ induces a homomorphism ρ_* of $\pi_1(\mathscr{C})$ onto $\pi_1(\mathscr{R})$. (Borsuk 1933).

Since ρ is continuous, it sends each closed path in $\mathscr{C}$ to a closed path in $\mathscr{R}$, and it also sends homotopic paths to homotopic paths, since the continuous image of a deformation rectangle (3.1.5) is again a deformation rectangle. Thus ρ induces a map

$$\rho_*: \pi_1(\mathscr{C}) \to \pi_1(\mathscr{R})$$

which is a homomorphism because ρ obviously sends products to products and inverses to inverses. Finally, since ρ is the identity on $\mathscr{R}$, every closed path in $\mathscr{R}$ is the image of a closed path (namely itself) in $\mathscr{C}$, and hence ρ_* is onto. □

A trivial example of a retraction is the map which sends every point of $\mathscr{C}$ to some particular point $P \in \mathscr{C}$. Retractions of $\mathscr{C}$ onto more complicated subspaces $\mathscr{R}$ are constrained by the group-theoretic proposition above $-\mathscr{R}$ cannot be so complicated that $\pi_1(\mathscr{R})$ is not a homomorphic image of $\pi_1(\mathscr{C})$. For example, the disc $\mathscr{D}$ has $\pi_1(\mathscr{D}) = \{1\}$ (we shall prove this below, however it is easy to see directly that each closed path is null homotopic), while its boundary circle $\mathbf{S}^1$ has $\pi_1(\mathbf{S}^1)$ = infinite cyclic group. There is no homomorphism of the trivial group onto the infinite cyclic group, and hence no retraction of the disc onto its boundary circle.

EXERCISE 3.3.1.1. Suppose that there is a continuous map $\phi: \mathscr{D} \to \mathscr{D}$ of the disc into itself with no fixed points, that is $\phi(P) \neq P$ for each $P \in \mathscr{D}$. Use the pair $\langle P, \phi(P) \rangle$ to define a point P' on the boundary circle $\mathbf{S}^1$ such that $\rho(P) = P'$ is a retraction, thus proving the nonexistence of ϕ. (This is the famous fixed-point theorem of Brouwer 1912a. In this paper Brouwer actually proved that any continuous map of the n-dimensional ball onto itself has a fixed point.)

3.3.2 Deformation Retracts and Collapsing

A complex $\mathscr{R} \subseteq \mathscr{C}$ is a *deformation retract* of $\mathscr{C}$ if there is a retraction $\rho: \mathscr{C} \to \mathscr{R}$ which is homotopic to the identity map: $\mathscr{C} \to \mathscr{C}$ in the following sense: there is a continuous function

$$h: [0, 1] \times \mathscr{C} \to \mathscr{C}$$

such that

(i) For all $P \in \mathscr{C}$,

$$h(0, P) = P \quad \text{and} \quad h(1, P) = \rho(P).$$

(ii) Whenever $P \in \mathscr{R}$, $h(t, P) = P$ for all $t \in [0, 1]$.

A deformation retraction ρ induces an isomorphism

$$\rho_*: \pi_1(\mathscr{C}) \to \pi_1(\mathscr{R}). \qquad \text{(Borsuk 1933)}$$

We have a homomorphism $\rho_*: \pi_1(\mathscr{C}) \to \pi_1(\mathscr{R})$ as in 3.3.1. The homotopy h now guarantees that ρ maps any closed path in $\mathscr{C}$ based at a point in $\mathscr{R}$ to a homotopic path in $\mathscr{R}$, so ρ_* does not change path classes, and hence is an isomorphism. □

As examples of deformation retracts we mention

(1) The centre point of a disc is a deformation retract of the disc. Thus $\pi_1(\text{disc}) = \{1\}$.
(2) The circle is a deformation retract of the annulus, and the solid torus. Thus π_1 of both these objects is the infinite cyclic group.

It is not hard to give explicit analytic formulae for h in these examples. In the next example (which is important for computing knot groups, see 4.2.4) we shall be content with a series of pictures (Figure 139).

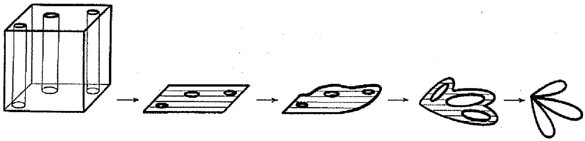


Figure 139

(3) A cube with vertical holes drilled through it has a deformation retraction to a bouquet of circles.

If $\mathscr{C}$ is a subcomplex of a complex $\mathscr{D}$ then a deformation retraction $\rho: \mathscr{C} \to R$ extends to a map ρ of $\mathscr{D}$ by letting ρ be the identity map outside $\mathscr{C}$. This extended map ρ induces a continuous map ρ' of $\mathscr{D}$ onto the identification space $\mathscr{D}'$ obtained by identifying all elements of $\mathscr{D}$ with the same ρ-image. We call $\rho': \mathscr{D} \to \mathscr{D}'$ a *collapse* of $\mathscr{D}$ onto $\mathscr{D}'$, and the proof of the theorem above generalises easily to show that the induced map of fundamental groups, $\rho'_*: \mathscr{D} \to \mathscr{D}'$ is an isomorphism.

An example of a collapse is that of an arbitrary graph to a bouquet of circles (cf. 2.1.8 and Exercise 3.3.2.1.)

The more complicated collapses we need (such as example (3) above) can be given a combinatorial form. An *elementary collapse* of a simplex Σ across a face Γ is a deformation retraction of Σ onto (boundary of Σ)-Γ. For example, a 1-simplex collapses to a point, a 2-simplex to a V-shaped pair of 1-simplexes, a 3-simplex to a "cone" made of three 2-simplexes, and so on. An elementary collapse can be made across any boundary face Γ in a simplicial complex $\mathscr{C}$, and a collapse in general is a finite sequence of elementary collapses. In the examples we shall use it is only a matter of patience to find a suitable simplicial decomposition and the right sequence of elementary collapses.

Deformation retraction is a good tool for computing fundamental groups because it simplifies the space without changing the group. When combined with the Seifert–Van Kampen theorem of the next section, which permits the fundamental group to be computed when complexes with known fundamental groups are glued together, we shall have a method powerful enough to compute the fundamental groups of all finite complexes.

EXERCISE 3.3.2.1. Show that any vertex of a tree is a deformation retract of the whole tree, and hence that any graph has a bouquet of circles as a collapse. Deduce from Exercise 3.2.4.1 that the fundamental group of any graph is free.

EXERCISE 3.3.2.2. Use collapsing to show that the bouquet of two circles is a deformation retract of the perforated torus, Figure 140, and hence deduce that the fundamental group of the perforated torus is the free group of rank 2.

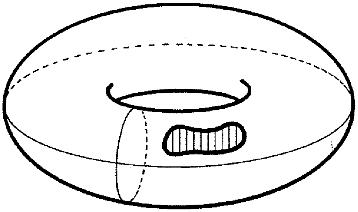


Figure 140

3.4 The Seifert–Van Kampen Theorem

3.4.1 Introduction

So far we know the fundamental groups only of spaces which possess deformation retracts of a particularly simple form, namely a point or a bouquet of circles, for which the group is either $\{1\}$ or a free group. However, many other spaces can be obtained simply by glueing such spaces together. Consider the decomposition of the torus surface in Figure 141. $\mathscr{A}$ has the bouquet of two circles as an obvious deformation retract and hence $\pi_1(\mathscr{A}) = \langle a_1, a_2 ; — \rangle$ (free group of rank 2), while $\mathscr{B}$ is a topological disc and hence $\pi_1(\mathscr{B}) = \{1\}$. $\mathscr{A}$ and $\mathscr{B}$ meet along a simple closed curve $\mathscr{C}$ (Figure 142)

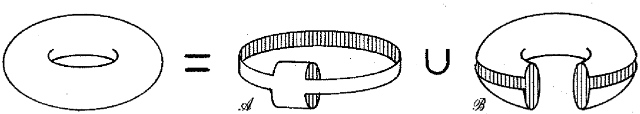


Figure 141

Figure 142

which represents the element $a_1a_2a_1^{-1}a_2^{-1}$ in $\pi_1(\mathscr{A})$, while in $\pi_1(\mathscr{B})$ it $= 1$. It therefore seems that attaching $\mathscr{B}$ to $\mathscr{A}$ will add the relation

$$a_1a_2a_1^{-1}a_2^{-1} = 1$$

so that $\pi_1(\mathscr{A} \cup \mathscr{B}) = \langle a_1, a_2; a_1a_2a_1^{-1}a_2^{-1}\rangle = \langle a_1a_2; a_1a_2 = a_2a_1\rangle$.

More generally, we should be able to realize an arbitrary finitely presented group $\langle a_1, \ldots, a_m; r_1, \ldots, r_n\rangle$ by taking a bouquet $\mathscr{A}$ of m circles to represent $a_1, \ldots, a_m$ and for each relator r_i, taking a disc $\mathscr{B}_i$ and identifying its boundary with the path in $\mathscr{A}$ which realizes the word r_i. This was pointed out by Dehn 1910. However, while it is clear that the relations $r_i = 1$ hold in the resulting complex, it is not clear that *only* they hold (or more precisely, that any relation is a consequence of them). To prove this type of result we shall use the following theorem (Seifert–Van Kampen theorem):

Suppose $\mathscr{C}$ is a space which can be expressed as the union of path-connected open sets $\mathscr{A}$, $\mathscr{B}$ such that $\mathscr{A} \cap \mathscr{B}$ is path-connected and such that $\pi_1(\mathscr{A})$ and $\pi_1(\mathscr{B})$ have respective presentations

$$\langle a_1, \ldots, a_m; r_1, \ldots, r_n\rangle$$

$$\langle b_1, \ldots, b_p; s_1, \ldots, s_q\rangle$$

while $\pi_1(\mathscr{A} \cap \mathscr{B})$ is finitely generated. Then $\pi_1(\mathscr{C})$ has the presentation

$$\langle a_1, \ldots, a_m, b_1, \ldots, b_p; r_1, \ldots, r_n, s_1, \ldots, s_q, u_1 = v_1, \ldots, u_t = v_t\rangle,$$

where u_i, v_i $(i = 1, \ldots, t)$ are expressions for the generators of $\pi_1(\mathscr{A} \cap \mathscr{B})$ in terms of the generators of $\pi_1(\mathscr{A})$, $\pi_1(\mathscr{B})$ respectively.

The assumption that $\mathscr{A}$ and $\mathscr{B}$ are open, so that $\mathscr{A} \cap \mathscr{B}$ is also, is not what we originally had in mind, but it greatly facilitates the proof. In practice it can easily be accommodated by attaching small neighbourhoods to the boundaries of closed components of $\mathscr{C}$. For example, the components $\mathscr{A}$, $\mathscr{B}$ of the torus above can be enlarged to open sets $\mathscr{A}'$, $\mathscr{B}'$ by adding points of the torus distant $< \varepsilon$ from their boundaries. $\mathscr{A}' \cap \mathscr{B}'$ is then an open annulus instead of a single curve, however deformation retraction shows that none of the fundamental groups has changed, so our previous computation will be a consequence of the Seifert–Van Kampen theorem.

3.4.2 Generators

Given $\mathscr{C} = \mathscr{A} \cup \mathscr{B}$, with the above assumptions, we choose the base point P for closed paths in $\mathscr{C}$ to lie in $\mathscr{A} \cap \mathscr{B}$. Then any closed path p from P is homotopic to a product of the paths a_i and b_i (or their inverses) which generate $\pi_1(\mathscr{A})$ and $\pi_1(\mathscr{B})$ respectively.

Let $f: \mathscr{I} \to \mathscr{C}$ be a journey along the path p. Since $\mathscr{A}$, $\mathscr{B}$ are open and $\mathscr{I}$ is compact we can find a partition $\mathscr{I} = \mathscr{I}_1 \cup \mathscr{I}_2 \cup \cdots \cup \mathscr{I}_k$ such that the subjourney f_j on each $\mathscr{I}_j = [t_j, t_{j+1}]$ lies entirely within $\mathscr{A}$ or entirely within $\mathscr{B}$. Let p_j be the path represented by this subjourney. (We get the partition by uniform continuity, after observing that there is an $\varepsilon > 0$ such that $\mathscr{N}_\varepsilon(X) \subset \mathscr{A}$ or $\mathscr{N}_\varepsilon(X) \subset \mathscr{B}$ for each $X \in p$. If there is no such ε, points $X_n \in p$ with $\mathscr{N}_{1/n}(X_n) \not\subset \mathscr{A}, \mathscr{B}$ have a limit $X \in p$, by compactness, but $X \notin \mathscr{A}, \mathscr{B}$.)

We now construct an approach path w_j from P to the initial point $f_j(t_j)$ of each p_j (w_j^{-1} will then be a path from the final point of p_{j-1} to P) in such a way that if $f_j(t_j)$ lies in $\mathscr{A} \cap \mathscr{B}$ then so does w_j. This is possible because $\mathscr{A} \cap \mathscr{B}$ is path-connected. Otherwise, if $f_j(t_j)$ lies in $\mathscr{A}$ so does w_j, and if $f_j(t_j)$ lies in $\mathscr{B}$ so does w_j, similarly using the path-connectedness of $\mathscr{A}$ and $\mathscr{B}$. The result is that the closed path $w_j p_j w_{j+1}^{-1}$ either lies entirely in $\mathscr{A}$ or entirely in $\mathscr{B}$, and hence can be expressed in terms of the generators a_i or b_i respectively. The same is true of the closed paths $p_1 w_2^{-1}$ and $w_k p_k$.

But

$$p \sim p_1 w_2^{-1} \cdot w_2 p_2 w_3^{-1} \cdot \cdots \cdot w_{k-1} p_{k-1} w_k^{-1} \cdot w_k p_k$$

and hence p is homotopic to a product of a_i's and b_i's. □

3.4.3 Relations

We can now assume that any path p in $\mathscr{A} \cup \mathscr{B}$ is written as a product of $a_1, \ldots, a_m, b_1, \ldots, b_p$. The relations

$$r_1 = \cdots = r_n = s_1 = \cdots = s_q = 1$$

must hold, since they are already true in $\mathscr{A}, \mathscr{B}$; likewise the relations

$$u_1 = v_1, \quad \ldots, \quad u_t = v_t$$

since they are true in $\mathscr{A} \cap \mathscr{B}$. Our task is to show that any other relation follows from these. This means showing that if p is any null-homotopic path in $\mathscr{A} \cup \mathscr{B}$, then $p = 1$ can be derived by means of the above relations.

Since p is null-homotopic there is a continuous map $h: \mathscr{R} \to \mathscr{C}$, where $\mathscr{R}$ is a rectangle and h maps the top edge of $\mathscr{R}$ onto p, the other three edges into the basepoint P. It will be convenient to discuss the image of $\mathscr{R}$ under h in terms of a diagram of $\mathscr{R}$ itself. One need only bear in mind that the diagram of $\mathscr{R}$ in $\mathscr{C}$ is "crumpled" to the extent that the top edge becomes p and the other three edges collapse to the point P, then the distinction between $\mathscr{R}$ and its image may be suppressed.

By compactness, there is a subdivision of $\mathscr{R}$ by vertical and horizontal lines into subrectangles $\mathscr{R}_{ij}$ so small that (the image of) each $\mathscr{R}_{ij}$ lies entirely in $\mathscr{A}$ or entirely in $\mathscr{B}$. This suggests a way of contracting p to the point P in $\mathscr{A} \cup \mathscr{B}$ by "small" steps which individually take place within $\mathscr{A}$ or $\mathscr{B}$, namely pulling portions of p across (the image of) one $\mathscr{R}_{ij}$ at a time (Figure

143). To describe the process in more detail we introduce the notation for the vertices and edges of $\mathscr{R}_{ij}$ in Figure 144. The initial position is the top edge of $\mathscr{R}$, $p = c_{0l}c_{1l} \cdots c_{k-1,l}$, and we want to reach the final position $c_{00}c_{10} \cdots c_{k-1,0}$ which is just P. It will suffice to pull p downwards across the typical horizontal strip shown in Figure 145. The typical step illustrated in Figure 143 is represented by the equation

$$d_{ij}c_{i,j+1} = c_{ij}d_{i+1,j} \tag{1}$$

and special steps at the ends, as in Figure 146 (which are valid because the vertical sides of $\mathscr{R}$ collapse to the point P in $\mathscr{C}$), are represented by the equations

$$c_{0,j+1} = c_{0j}d_{1j} \quad \text{and} \quad d_{k-1,j}c_{k-1,j+1} = c_{k-1,j}. \tag{2}$$

The problem is that the c_{ij}, d_{ij} are not closed paths, so Equations (1), (2) cannot be considered as relations in $\pi_1(\mathscr{A})$ or $\pi_1(\mathscr{B})$. The solution is to take an approach path w_{ij} from P to each vertex P_{ij}. We can then define *closed* paths

$$\gamma_{ij} = w_{ij}c_{ij}w_{i+1,j}^{-1}$$
$$\delta_{ij} = w_{ij}d_{ij}w_{i,j+1}^{-1}$$

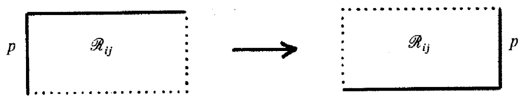


Figure 143

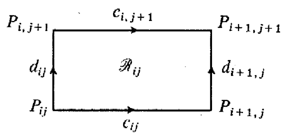


Figure 144

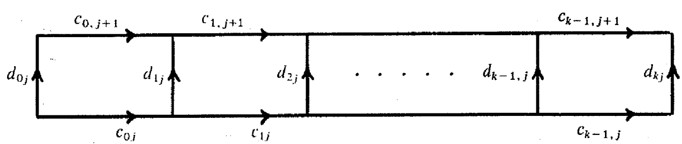


Figure 145

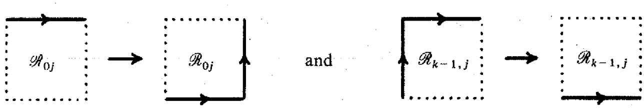


Figure 146

which behave formally like c_{ij}, d_{ij} (as will be seen). The crucial point is to construct w_{ij} so that it lies in $\mathscr{A} \cap \mathscr{B}$ if P_{ij} does, which is possible because $\mathscr{A} \cap \mathscr{B}$ is path-connected, and otherwise let w_{ij} lie in $\mathscr{A}$ if P_{ij} does, in $\mathscr{B}$ if P_{ij} does. Then all four closed paths γ_{ij}, δ_{ij}, $\gamma_{i,j+1}$, $\delta_{i+1,j}$ associated with $\mathscr{R}_{ij}$ lie in $\mathscr{A}$, or else all in $\mathscr{B}$, so it is to be expected that the relations analogous to (1), (2)

$$\delta_{ij}\gamma_{i,j+1} = \gamma_{ij}\delta_{i+1,j} \tag{1'}$$

$$\gamma_{0,j+1} = \gamma_{0j}\delta_{1j} \quad \text{and} \quad \delta_{k-1,j}\gamma_{k-1,j+1} = \gamma_{k-1,j} \tag{2'}$$

are valid in whichever of $\pi_1(\mathscr{A})$, $\pi_1(\mathscr{B})$ their elements belong to. We verify (1′) by performing homotopies on $\delta_{ij}\gamma_{i,j+1}\delta_{i+1,j}^{-1}\gamma_{ij}^{-1}$, which for definiteness we shall assume to lie in $\mathscr{A}$. A homotopy in $\mathscr{A}$ will be denoted by $\sim_{\mathscr{A}}$.

$$\begin{aligned}
&\delta_{ij}\gamma_{i,j+1}\delta_{i+1,j}^{-1}\gamma_{ij}^{-1} \\
&\quad = w_{ij}d_{ij}w_{i,j+1}^{-1} \cdot w_{i,j+1}c_{i,j+1}w_{i+1,j+1}^{-1} \\
&\qquad \cdot (w_{i+1,j}d_{i+1,j}w_{i+1,j+1}^{-1})^{-1} \cdot (w_{ij}c_{ij}w_{i+1,j}^{-1})^{-1} \\
&\quad \sim_{\mathscr{A}} w_{ij}d_{ij}c_{i,j+1}d_{i+1,j}^{-1}c_{ij}^{-1}w_{ij}^{-1} \text{ (since the cancelled paths lie in } \mathscr{A}) \\
&\quad \sim_{\mathscr{A}} w_{ij}w_{ij}^{-1} \quad \text{(contracting the perimeter } d_{ij}c_{i,j+1}d_{i+1,j}^{-1}c_{ij}^{-1} \\
&\qquad\qquad\qquad\qquad \text{of } \mathscr{R}_{ij}, \text{ which lies in } \mathscr{A}, \text{ to a point)} \\
&\quad \sim_{\mathscr{A}} 1.
\end{aligned}$$

Thus $\delta_{ij}\gamma_{i,j+1}\delta_{i+1,j}^{-1}\gamma_{ij}^{-1} = 1$ is valid in $\pi_1(\mathscr{A})$ and hence a consequence of the relations of $\pi_1(\mathscr{A})$. Similarly for the relations (2′), which are in fact degenerate cases of (1′), since

$$\delta_{0j} = \delta_{k,j} = 1$$

as a result of the collapse of the vertical sides of $\mathscr{R}$ to the point P. (Incidentally, $\gamma_{00} = \gamma_{10} = \cdots = \gamma_{k-1,0} = 1$ for the same reason.)

We now observe that

$$p = \gamma_{0l}\gamma_{1l}\cdots\gamma_{k-1,l}$$

(assuming that we have chosen w_{0l} and w_{kl} to coincide with the point path P) so we can reduce this expression to

$$p = \gamma_{00}\gamma_{10}\cdots\gamma_{k-1,0} = 1$$

by manipulating the γ's and δ's as we previously manipulated the c's and d's, using the relations (1′) and (2′) which are valid in $\pi_1(\mathscr{A})$ or $\pi_1(\mathscr{B})$. The relations $u_i = v_i$ will enter when we wish to apply a relation of $\pi_1(\mathscr{A})$ to an expression which, as a result of previous manipulations, is expressed in terms of the generators of $\pi_1(\mathscr{B})$ (or vice versa). This will only happen when the curve in question lies in $\mathscr{A} \cap \mathscr{B}$, in which case expressions for it in terms of either set of generators exist, and are intertranslatable via the relations $u_i = v_i$. □

3.4.4 Realization of Finitely Presented Groups by Surface Complexes

We are now able to justify the claim that the group

$$\langle a_1, \ldots, a_m; r_1, \ldots, r_n \rangle$$

can be realized by taking a bouquet of m circles for the generators $a_1, \ldots, a_m$ and identifying each path r_i with the boundary of a disc $\mathscr{D}_i$.

First we confirm independently that if

$$\mathscr{A} = \text{bouquet of } m \text{ circles } a_1, \ldots, a_m$$

then $\pi_1(\mathscr{A}) = \langle a_1, \ldots, a_m; - \rangle$

The method of proof may be illustrated with $m = 2$ (Figure 147). Since a_1, a_2 are not open sets we enlarge them to

$$\mathscr{A}_1 = \mathscr{A} - \{x_2\} \quad \text{and} \quad \mathscr{A}_2 = \mathscr{A} - \{x_1\}$$

respectively. Then a_1 is a deformation retraction of $\mathscr{A}_1$, so $\pi_1(\mathscr{A}_1) = \langle a_1; - \rangle$ and a_2 is a deformation retraction of $\mathscr{A}_2$, so $\pi_1(\mathscr{A}_2) = \langle a_2; - \rangle$. Also, $\mathscr{A}_1 \cap \mathscr{A}_2 = \mathscr{A} - \{x_1, x_2\}$ which is a tree, so it has a deformation retraction to a point and $\pi_1(\mathscr{A}_1 \cap \mathscr{A}_2)$ has only the trivial generator 1. Thus by the Seifert–Van Kampen theorem

$$\pi_1(\mathscr{A}) = \pi_1(\mathscr{A}_1 \cup \mathscr{A}_2) = \langle a_1, a_2; - \rangle$$

Similarly, if $\mathscr{A}$ is the bouquet of m circles, then $\pi_1(\mathscr{A})$ is the free group of rank m.

Now if r_1 is a path in the bouquet $\mathscr{A}$ and we want to identify it with the boundary of a disc $\mathscr{D}_1$, the result can be obtained by first adding a neighbourhood $\mathscr{N}_1$ of r_1 to form the open set $\mathscr{A} \cup \mathscr{N}_1$, then adding the interior $\text{int}(\mathscr{D}_1)$ of $\mathscr{D}_1$. $\mathscr{N}_1$ will consist of points distant $< \varepsilon$ from the boundary of $\mathscr{D}_1$ together with "whiskers"—small half-open segments from the circles which meet r_1 at the vertex of the bouquet (Figure 148). We take the basepoint for

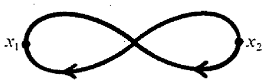


Figure 147

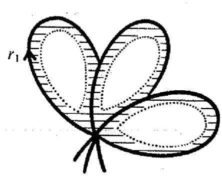


Figure 148

the fundamental group to lie in the intersection $(\mathscr{A} \cup \mathscr{N}_1) \cap \text{int}(\mathscr{D}_1)$, which is the open annulus of points distant > 0 but $< \varepsilon$ from the boundary of $\mathscr{D}_1$.

$\mathscr{A}$ is a deformation retract of the open set $\mathscr{A} \cup \mathscr{N}_1$, so

$$\pi_1(\mathscr{A} \cup \mathscr{N}_1) = \pi_1(\mathscr{A}) = \langle a_1, \ldots, a_m; -\rangle.$$

$\text{int}(\mathscr{D}_1)$ has a deformation retraction to a point, so

$$\pi_1(\text{int}(\mathscr{D}_1)) = \{1\}$$

and $\pi_1((\mathscr{A} \cup \mathscr{N}_1) \cap \text{int}(\mathscr{D}_1))$ is the infinite cyclic group generated by the centerline of the annulus, which is equal to the element r_1 in $\pi_1(\mathscr{A} \cup \mathscr{N}_1)$ and 1 in $\pi_1(\text{int}(\mathscr{D}_1))$. The openness of the various sets permits the Seifert–Van Kampen theorem to be applied and we have

$$\begin{aligned}\pi_1(\mathscr{A} \cup \mathscr{D}_1) &= \pi_1((\mathscr{A} \cup \mathscr{N}_1) \cup \text{int}(\mathscr{D}_1)) \\ &= \langle a_1, \ldots, a_m; r_1 = 1\rangle.\end{aligned}$$

Successive attachment of discs $\mathscr{D}_i$ to the paths r_i for $i \geq 2$ adds the relations $r_i = 1$ by a similar argument (using the obvious neighbourhood $\mathscr{N}_i$ of r_i in $\mathscr{A} \cup \mathscr{D}_1 \cup \cdots \cup \mathscr{D}_{i-1}$). □

Of course the discs $\mathscr{D}_i$ must all be disjoint and free of self-intersections except for points on their boundaries which are identified with the same point of $\mathscr{A}$. As a result, it is not usually possible to embed the complex which realizes a given group in ordinary space. The simplest example is the cyclic group of order 2, $\langle a_1; a_1^2\rangle$. The process of sewing the boundary of $\mathscr{D}_1$ twice around a_1 (Figure 149) cannot be completed in $\mathbf{R}^3$ since the result is a projective plane.

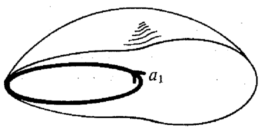


Figure 149

EXERCISE 3.4.4.1. Give a criterion for a relator r_1 to yield a closed surface when the above construction is applied.

EXERCISE 3.4.4.2. Show that $\pi_1(\mathbf{S}^2) = \{1\}$. Generalise the argument to $\mathbf{S}^{n+1}$.

3.4.5 Free Products

The Seifert–Van Kampen theorem gives a natural interpretation to the *free product* construction for finitely presented groups. Given

$$G = \langle a_1, \ldots, a_m; r_1, \ldots, r_n\rangle$$
$$H = \langle b_1, \ldots, b_p; s_1, \ldots, s_q\rangle$$

then

$$G * H = \langle a_1, \ldots, a_m, b_1, \ldots, b_p; r_1, \ldots, r_n, s_1, \ldots, s_q \rangle$$

is called the *free product* of G and H.

The free product is well defined, because if we choose another presentation $\langle a'_1, \ldots, a'_{m'}; r'_1, \ldots, r'_{n'} \rangle$ for G, there is a finite sequence of Tietze transformations (0.5.8) which convert it to $\langle a_1, \ldots, a_m; r_1, \ldots, r_n \rangle$, and the same transformations will convert $\langle a'_1, \ldots, a'_{m'}, b_1, \ldots, b_p; r'_1, \ldots, r'_{n'}; s_1, \ldots, s_q \rangle$ to $\langle a_1, \ldots, a_m, b_1, \ldots, b_p; r_1, \ldots, r_n, s_1, \ldots, s_q \rangle$ if one takes care not to give different generators the same name at any stage. Thus $G * H$ does not depend on the presentation of G, nor on the presentation of H (similarly).

Using the infinite Tietze transformations of Exercise 0.5.8.2 one can extend this argument to define free products of infinitely presented groups, and even the free product $*_i G_i$ of an infinite collection $\{G_i\}$ of groups, by taking the union of presentations of the G_i on disjoint sets of generators.

$G * H$ has a natural realization in terms of surface complexes $\mathscr{A}$, $\mathscr{B}$ such that $\pi_1(\mathscr{A}) = G$, $\pi_1(\mathscr{B}) = H$. Namely, one identifies a vertex of $\mathscr{A}$ with a vertex of $\mathscr{B}$. (To apply the Seifert–Van Kampen theorem enlarge $\mathscr{A}$, $\mathscr{B}$ to open sets $\mathscr{A}'$, $\mathscr{B}'$ by adding a small simply connected neighbourhood of the vertex.) This yields the following group-theoretic result:

There are natural embeddings of G and H in $G * H$

Let $G * H$ be realized as $\pi_1(\mathscr{A} \cup \mathscr{B})$, where $\mathscr{A}$, $\mathscr{B}$ are complexes, with a single common point P, such that $\pi_1(\mathscr{A}) = G$, $\pi_1(\mathscr{B}) = H$.

Consider a path p in $\mathscr{A}$ for which there is a homotopy $h: \mathscr{R} \to \mathscr{A} \cup \mathscr{B}$ which sends the top edge of the rectangle $\mathscr{R}$ to p and the other three edges to the point P. If we compose h with the retraction

$$\rho: \mathscr{A} \cup \mathscr{B} \to \mathscr{A}$$

which is the identity on $\mathscr{A}$ and maps all $\mathscr{B}$ to P then we get a homotopy $\rho h: \mathscr{R} \to \mathscr{A}$ which again sends the top edge to p and the other three edges to P. In other words, if $[p] \in \pi_1(\mathscr{A})$ and $[p] = 1$ in $\pi_1(\mathscr{A} \cup \mathscr{B})$, then, in fact $[p] = 1$ in $\pi_1(\mathscr{A})$, so the homotopy classes of paths in $\mathscr{A}$ represent a natural embedding of $\pi_1(\mathscr{A})$ in $\pi_1(\mathscr{A} \cup \mathscr{B})$.

In group-theoretic terms, the subgroup of elements of $G * H$ generated by the generators of G is isomorphic to G. Similarly for the subgroup generated by the generators of H. □

EXERCISE 3.4.5.1. Give a combinatorial proof that G, H embed in $G * H$ by finding a group-theoretic equivalent of the retraction map ρ.

EXERCISE 3.4.5.2. Let $G_i = \langle a_{i1}, a_{i2}, \ldots; r_{i1}, r_{i2}, \ldots \rangle$ be realized by a surface complex $\mathscr{A}_i$. Let $\mathscr{C}$ be the complex formed by attaching each $\mathscr{A}_i$ by an edge e_i to a new vertex P (Figure 150). Prove a special case of an infinite Seifert–Van Kampen theorem to show that

$$\pi_1(\mathscr{C}) = \langle a_{11}, a_{12}, \ldots, a_{21}, a_{22}, \ldots, \ldots; r_{11}, r_{12}, \ldots, r_{21}, r_{22}, \ldots, \ldots \rangle.$$

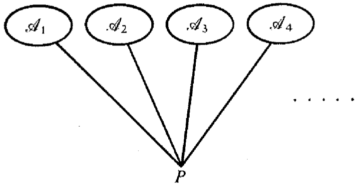


Figure 150

(The edges e_i are not strictly necessary, but they make it impossible for a path to enter infinitely many $\mathscr{A}_i$.)

3.5 Direct Products

3.5.1 Product Spaces

The *cartesian product* (or simply, *product*) of two spaces $\mathscr{C}_1$ and $\mathscr{C}_2$, $\mathscr{C}_1 \times \mathscr{C}_2$ is defined as a topological space by letting its points be the ordered pairs $\langle x_1, x_2 \rangle$ such that $x_1 \in \mathscr{C}_1$ and $x_2 \in \mathscr{C}_2$ and letting the neighbourhoods of $\langle x_1, x_2 \rangle$ be unions of the sets

$$\mathscr{N}_1 \times \mathscr{N}_2 = \{\langle y_1, y_2 \rangle : y_1 \in \mathscr{N}_1 \text{ and } y_2 \in \mathscr{N}_2\},$$

where $\mathscr{N}_1$ is a neighbourhood of x_1 and $\mathscr{N}_2$ a neighbourhood of x_2.

The best known example of a product is the torus, which can be viewed as $\mathbf{S}^1 \times \mathbf{S}^1$. If we let θ, ϕ be the angular coordinates on the two circles, the ordered pair $\langle \theta, \phi \rangle$ fixes a point on the torus by "longitude" and "latitude." With small arcs as a basis for the neighbourhoods on $\mathbf{S}^1$ we get small "square" patches, homeomorphic to the disc, as neighbourhoods on the torus (Figure 151). Finitely many such patches exhaust the whole torus surface and, when divided by their diagonals, yield a triangulation which reaffirms the identity of the torus as a simplicial complex.

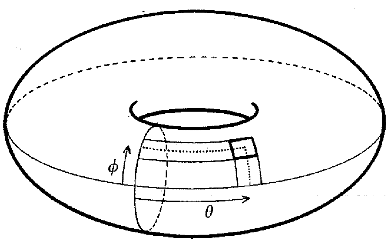


Figure 151

A similar argument shows that the product of any two simplicial complexes is a simplicial complex.

More importantly, the product $\mathcal{M}_1 \times \mathcal{M}_2$ of manifolds $\mathcal{M}_1, \mathcal{M}_2$ is itself a manifold. By hypothesis, each point in $\mathcal{M}_1$ has a neighbourhood homeomorphic to the open m-dimensional ball, which in turn is homeomorphic to R^m, while each point in $\mathcal{M}_2$ has a neighbourhood homeomorphic to R^n. Each point in $\mathcal{M}_1 \times \mathcal{M}_2$ therefore has a neighbourhood homeomorphic to $\mathsf{R}^m \times \mathsf{R}^n = \mathsf{R}^{m+n}$, which is homeomorphic to the $(m + n)$-dimensional ball. Thus $\mathcal{M}_1 \times \mathcal{M}_2$ is an $(m + n)$-dimensional manifold.

3.5.2 The Direct Product of Groups

The direct product $G \times H$ of groups G, H is defined by componentwise multiplication:

$$\langle g_1, h_1 \rangle \cdot \langle g_2, h_2 \rangle = \langle g_1 g_2, h_1 h_2 \rangle$$

on the set of ordered pairs $\langle g, h \rangle$ such that $g \in G, h \in H$. A convenient fact is:

$$\pi_1(\mathscr{C}_1 \times \mathscr{C}_2) = \pi_1(\mathscr{C}_1) \times \pi_1(\mathscr{C}_2).$$

It suffices to observe that a path p in $\mathscr{C}_1 \times \mathscr{C}_2$ can be viewed as a pair $\langle p_1, p_2 \rangle$, where $p_1 \subset \mathscr{C}_1$ and $p_2 \subset \mathscr{C}_2$. This is because a continuous map $f: \mathscr{I} \to \mathscr{C}_1 \times \mathscr{C}_2$ sends each $t \in \mathscr{I}$ to an ordered pair $\langle f_1(t), f_2(t) \rangle$, where $f_1: \mathscr{I} \to \mathscr{C}_1$ and $f_2: \mathscr{I} \to \mathscr{C}_2$ are also continuous.

Thus each journey f in $\mathscr{C}_1 \times \mathscr{C}_2$ decomposes into a pair of journeys in the two factors, paths and path classes decompose in turn, and products are formed componentwise. The elements of $\pi_1(\mathscr{C}_1 \times \mathscr{C}_2)$ are therefore ordered pairs $\langle [p_1], [p_2] \rangle$, where $[p_1] \in \pi_1(\mathscr{C}_1)$ and $[p_2] \in \pi_1(\mathscr{C}_2)$, with componentwise multiplication, which means that the group is precisely $\pi_1(\mathscr{C}_1) \times \pi_1(\mathscr{C}_2)$. □

This theorem gives an independent computation of the fundamental group of the torus, since

$$\begin{aligned} \pi_1(\mathsf{S}^1 \times \mathsf{S}^1) &= \pi_1(\mathsf{S}^1) \times \pi_1(\mathsf{S}^1) \\ &= (\text{infinite cyclic}) \times (\text{infinite cyclic}) \\ &= \text{free abelian of rank 2}. \end{aligned}$$

The theorem is not generally useful for computing fundamental groups, since most complexes are not nontrivial products, however it provides a very quick computation for those that are.

For example, $\mathsf{S}^1 \times \mathsf{S}^2, \mathsf{S}^1 \times \mathsf{S}^3, \ldots$ all have infinite cyclic fundamental group, like S^1, because

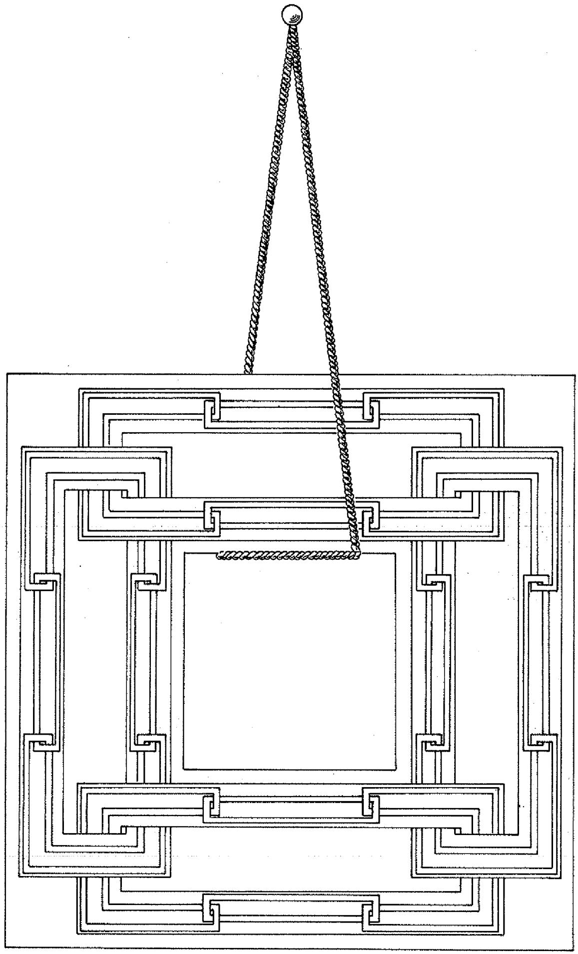
$$\pi_1(\mathsf{S}^2) = \pi_1(\mathsf{S}^3) = \cdots = \{1\}.$$

The fact that the 4-dimensional manifold $\mathbf{S}^1 \times \mathbf{S}^3$ has infinite cyclic fundamental group will be useful later in constructing 4-dimensional manifolds with prescribed fundamental groups.

EXERCISE 3.5.2.1. Show that $G \times H$ results from $G * H$ by adding relations $gh = hg$ for each generator $g \in G$, $h \in H$.

CHAPTER 4

Fundamental Groups of Complexes



4.1 Poincaré's Method for Computing Presentations

4.1.1 Introduction

What we describe as Poincaré's method for computing the fundamental group was given by Poincaré 1895 only in terms of a few special examples. These examples were mainly 3-dimensional manifolds obtained from a solid cube by identifying its faces in various ways. However, they clearly exposed the fact that one finds generators from the 1-dimensional cells of the complex, and relations from the 2-dimensional cells. Let us take Poincaré's example of the cube in which each face is identified with its opposite by translation (Figure 152).

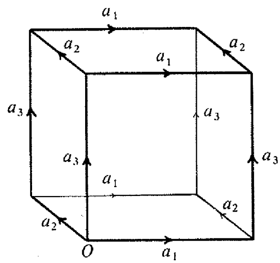


Figure 152

It is easily checked that all eight vertices become identified with 0, the neighbourhood of which is filled in by eight "corner chips" of the cube to become a topological ball. Similarly, each edge becomes identified with the three other like-labelled edges, and the neighbourhood of a point on the edge is filled in by four "edge chips" to become a topological ball also. It is obvious that points in the faces and interior of the cube also have 3-dimensional ball neighbourhoods, so this complex is indeed a 3-dimensional manifold.

Poincaré then assumes that any closed curve can be deformed onto the edges, so the curves a_1, a_2, a_3 (which, one notes, are *closed*) are generators of the fundamental group. Likewise, any deformation is based on elementary deformations which consist of pulling the curve from one side of a face to the other. Formally, this means equating the perimeter of each face to the identity, that is,

$$a_1a_2a_1^{-1}a_2^{-1} = 1, \qquad a_2a_3a_2^{-1}a_3^{-1} = 1, \qquad a_3a_1a_3^{-1}a_1^{-1} = 1.$$

These relations simply say that all generators commute with each other, so we have the free abelian group of rank 3.

Tietze 1908 gave a more formal discussion of Poincaré's method, generalising it to n dimensions and allowing for more than one cell in the complex. Tietze takes care of the possible multiplicity of vertices by using an approach path w_i from the chosen basepoint P to each vertex P_i. Each edge P_iP_j is then associated with the closed path $w_iP_iP_jw_j^{-1}$, and these paths suffice to generate the group. Tietze observes that in this general setting it is the 2-dimensional cells alone which determine the relations of the group.

The purpose of this section is to prove, using the Seifert–Van Kampen theorem, that the generators and relations obtained by the Poincaré method indeed determine the fundamental group as we have defined it in 3.1.6. This result is mainly of theoretical interest, showing that any finite simplicial complex has a finitely presented fundamental group. It would be tedious in practice to actually use a simplicial decomposition to compute a presentation, and the reader who wishes to see examples of direct computation may wish to skip ahead to Section 4.2.

EXERCISE 4.1.1.1. Show that the manifold discussed above is homeomorphic to $\mathbf{S}^1 \times \mathbf{S}^1 \times \mathbf{S}^1$.

4.1.2 1-complexes (Graphs)

We already know (from Exercise 3.3.2.1) that any graph has a collapse to a bouquet of circles, via a deformation retraction of a spanning tree, and hence that its fundamental group is free. However, in order to find explicit generators in the original graph, we sketch a fresh approach from first principles.

Given a graph $\mathscr{G}$, construct a spanning tree $\mathscr{T}$, and for each edge $e_i = P_jP_k$ construct a path

$$a_i = w_je_iw_k^{-1},$$

where w_j is the unique reduced path in $\mathscr{T}$ from P to P_j, and w_k is the unique reduced path in $\mathscr{T}$ from P to P_k, where P is the chosen basepoint. We can now show by an argument like that in 3.2 that any path p in $\mathscr{G}$ is homotopic to a product of e_i's and their inverses, whence any closed path is homotopic to the corresponding product of a_i's (cf. 2.1.7). The a_i's therefore generate $\pi_1(\mathscr{G})$, and we obtain a subset which freely generates by taking only those a_i's which correspond to e_i's not in $\mathscr{T}$ (the remaining a_i's can be set equal to 1 since they are in fact closed paths in $\mathscr{T}$). □

4.1.3 2-complexes (Surface Complexes)

We now study the result of attaching discs $\mathscr{D}$ to a graph $\mathscr{G}$. The construction is like 3.4.4 except that the boundary of $\mathscr{D}$ is no longer an element of $\pi_1(\mathscr{G})$ necessarily, but may be an arbitrary closed path given by a product of edges

$$r = e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n}.$$

This adds the relation

$$a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_n}^{\varepsilon_n} = 1$$

to the group.

The only departure from the construction in 3.4.4 is to take an approach path w from P to r, namely the unique reduced path $\mathscr{G}$ from P to the initial point of r. As in 3.4.4, we first attach a neighbourhood $\mathscr{N}$ of r to $\mathscr{G}$, then add the interior of $\mathscr{D}$, int($\mathscr{D}$). The intersection of these two open sets is an open annulus whose centreline is homotopic to r, so if we take the basepoint to lie in the annulus the relation $r = 1$ will be added to the group.

When we shift the basepoint back to its original position P, at the initial point of w, r will be replaced by wrw^{-1}. But

$$\begin{aligned} wrw^{-1} &= we_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n} w^{-1} \\ &= a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_n}^{\varepsilon_n} \end{aligned}$$

since $e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n}$ is closed and w is the unique reduced path in $\mathscr{T}$ from P to its initial point. Hence with P as basepoint the relation added to the group is

$$a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_n}^{\varepsilon_n} = 1. \qquad \square$$

The argument is the same (except for suitable reinterpretation of the "neighbourhood of r") when $\mathscr{G}$ is already a surface complex instead of a graph, hence successive attachment of discs adds the relations which correspond to their boundary paths in the above manner.

4.1.4 The n-sphere for $n \geq 2$

We know that the n-sphere S^n is obtained from two n-balls B^n by identifying their boundaries, which are S^{n-1}. But B^n has a deformation retraction to a point, so $\pi_1(\mathsf{B}^n) = \{1\}$, and since S^{n-1} is path-connected for $n \geq 2$, the Seifert–Van Kampen theorem gives $\pi_1(\mathsf{S}^n) = \{1\}$. (Expand the two B^n's slightly to open sets which intersect in a neighbourhood of S^{n-1}.)

4.1.5 Attaching an n-ball to a Complex for $n \geq 3$

When the boundary S^{n-1} *of the n-ball* B^n *is identified with part of a complex* $\mathscr{C}$,

$$\pi_1(\mathscr{C} \cup \mathsf{B}^n) = \pi_1(\mathscr{C}).$$

The construction is exactly the same as the special case $\mathsf{B}^2 = \mathscr{D}$ treated in 4.1.3. We find open sets whose union is $\mathscr{C} \cup \mathsf{B}^n$, which have deformation retractions to $\mathscr{C}$ and a point respectively, and whose intersection is a neighbourhood $\mathscr{N}$ of $w \cup \mathsf{S}^{n-1}$, where w is an approach path to the boundary S^{n-1} of B^{n-1}. But then $\mathscr{N}$ has a deformation retraction to S^{n-1} which, unlike S^1, has *trivial* fundamental group, so the Seifert–Van Kampen theorem gives

$$\pi_1(\mathscr{C} \cup \mathsf{B}^n) = \pi_1(\mathscr{C}). \qquad \square$$

4.1.6 The Method

Any simplicial complex can be constructed by first assembling the 1-simplexes (giving the 1-*skeleton*), then attaching the 2-simplexes (giving the 2-*skeleton*), then attaching the 3-simplexes, and so on. The result of 4.1.5 shows that the fundamental group is already determined by the 2-skeleton. By 4.1.2 and 4.1.3, the manner of determination is the following: construct a spanning tree $\mathscr{T}$ of the 1-skeleton by the method of 2.1.5, and for each vertex P_l construct an approach path w_l from the basepoint P to P_l, namely the unique reduced path in $\mathscr{T}$. For each edge $e_i = P_jP_k$ there is then a generator

$$a_i = w_j e_i w_k^{-1}$$

which can be set $= 1$ when e_i is in $\mathscr{T}$. And for each face in the 2-skeleton, with boundary $e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n}$ we have a relation

$$a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_n}^{\varepsilon_n} = 1.$$

Furthermore, the same is true for cell decompositions which are not necessarily simplicial, since we made no assumption about the way the boundary of B^n was mapped continuously onto the complex. In particular, if one can find a one-cell decomposition with a single vertex—as in the Poincaré example—then the generators are just the edges and the relators are the perimeters of the faces.

EXERCISE 4.1.6.1. If $\mathscr{M}_1$ and $\mathscr{M}_2$ are n-dimensional manifolds, the *connected sum* $\mathscr{M}_1 \# \mathscr{M}_2$ is constructed by removing a small open n-ball $\mathscr{E}_i$ from each $\mathscr{M}_i$ and then identifying $(\mathscr{M}_1 - \mathscr{E}_1)$ and $(\mathscr{M}_2 - \mathscr{E}_2)$ along their boundary $(n-1)$-spheres. When $n \geq 3$ show that

$$\pi_1(\mathscr{M}_1 \# \mathscr{M}_2) = \pi_1(\mathscr{M}_1) * \pi_1(\mathscr{M}_2).$$

4.1.7 Infinite Complexes

If $\mathscr{C}$ is an infinite simplicial complex then the above construction is also valid for obtaining a presentation of $\pi_1(\mathscr{C})$.

Let G be the group whose presentation is obtained by applying the construction of 4.1.6 to the complex $\mathscr{C}$. We first show that there are connected finite complexes

$$\mathscr{C}_1 \subset \mathscr{C}_2 \subset \mathscr{C}_3 \subset \cdots$$

such that

$$\mathscr{C} = \bigcup_n \mathscr{C}_n$$

and that there are "nested" presentations of $\pi_1(\mathscr{C}_1)$, $\pi_1(\mathscr{C}_2), \ldots$ whose "union" is the presentation of G.

Let $\mathscr{T}$ be a spanning tree of the 1-skeleton of $\mathscr{C}$, obtained, as in 2.1.5, as the union of

$$\mathscr{T}_1 \subset \mathscr{T}_2 \subset \mathscr{T}_3 \subset \cdots,$$

where any vertex one edge away from $\mathscr{T}_n$ is in $\mathscr{T}_{n+1}$. Then since $\mathscr{C}$ is locally finite, $\mathscr{T}_n$ is finite and hence so is

$$\mathscr{C}_n = \text{union of simplexes of } \mathscr{C} \text{ incident with } \mathscr{T}_n.$$

$\mathscr{T}_{n+1}$ is a spanning tree of the 1-skeleton of $\mathscr{C}_n$, so $\mathscr{C}_n$ is connected, and $\bigcup_n \mathscr{C}_n = \mathscr{C}$.

Taking P as the vertex of $\mathscr{T}_1$, the approach paths in $\mathscr{T}$ to vertices of $\mathscr{C}_n$ are in $\mathscr{T}_{n+1}$, so the generators a_i of $\pi_1(\mathscr{C}_n)$ obtained by using $\mathscr{T}_{n+1}$ as its spanning tree are among the generators of G. These a_i correspond to edges e_i in the 1-skeleton of $\mathscr{C}_n$, and since any face in $\mathscr{C}_n$ bounded by these e_i is *a fortiori* in $\mathscr{C}$, the relations of $\pi_1(\mathscr{C}_n)$ are among those of G. Conversely, any edge or face of $\mathscr{C}$ is in some $\mathscr{C}_n$, so each generator and relation of G eventually appears in some $\pi_1(\mathscr{C}_n)$.

Now a closed path p based at P is compact and hence must lie in some $\mathscr{C}_n$. It can then be expressed in terms of the generators of $\pi_1(\mathscr{C}_n)$ and hence in the generators of G. Thus the generators of G generate $\pi_1(\mathscr{C})$. If p is also null-homotopic then its deformation rectangle, being compact, also lies in some $\mathscr{C}_m$. In that case $p = 1$ in $\pi_1(\mathscr{C}_m)$ and hence in G, so the relations of G imply all relations of $\pi_1(\mathscr{C})$. The relations of G are certainly true in $\pi_1(\mathscr{C})$, therefore

$$G = \pi_1(\mathscr{C}). \qquad \square$$

4.2 Examples

4.2.1 Finite Surfaces

The classification theorem of 1.3 can be interpreted as giving the following constructions for closed finite surfaces:

(i) *Orientable surface of genus n.* Take a bouquet of circles $a_1, b_1, \ldots, a_n, b_n$ and attach a disc with boundary identified with the path $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1}$.

(ii) *Nonorientable surface of genus n.* Take a bouquet of circles $a_1, \ldots, a_n$ and attach a disc with boundary identified with the path $a_1^2 \cdots a_n^2$.

(iii) *Sphere.* Attach two discs to a circle a.

It is then an immediate consequence of 3.4.4 that their fundamental groups are

(i) $\langle a_1, b_1, \ldots, a_n, b_n; a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1} \rangle$
(ii) $\langle a_1, \ldots, a_n; a_1^2 \cdots a_n^2 \rangle$
(iii) $\{1\}$.

The topological distinctness of these surfaces can now be proved rigorously by showing that the groups are all different. This is not difficult to do, but we shall postpone it until it can be placed in its true setting (homology theory and abelianization, Chapter 5, in particular 5.3.3).

We also saw in 1.3 that any bounded finite surface is homeomorphic to a disc with strips attached—double strips corresponding to handles, Möbius strips to crosscaps, and single strips for any extra perforations. Such a surface (Figure 153) obviously has a deformation retraction onto a bouquet of circles passing through the strips, hence its fundamental group is free, of rank equal to the number of strips. A more general result, which does not depend on normal forms, is given in the following exercise.

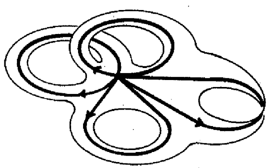


Figure 153

EXERCISE 4.2.1.1. Let $\mathscr{C}$ be a connected subcomplex of a triangulated finite surface $\mathscr{F}$. If $\mathscr{C} \neq \mathscr{F}$ show that $\mathscr{C}$ collapses onto a graph, so that $\pi_1(\mathscr{C})$ is free.

4.2.2 Infinite Surfaces

The fundamental group of an infinite surface is free (Johansson 1931).

To find free generators for $\pi_1(\mathscr{F})$, where $\mathscr{F}$ is an infinite surface, we shall refine the idea of 4.1.7 so as to get finite bounded surfaces

$$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \mathscr{F}_3 \subset \cdots$$

such that

$$\mathscr{F} = \bigcup_n \mathscr{F}_n$$

and such that the free generators of $\pi_1(\mathscr{F}_{n+1})$ include those of $\pi_1(\mathscr{F}_n)$.

First, enumerate the triangles $\Delta_1, \Delta_2, \Delta_3, \ldots$ in a simplicial decomposition of $\mathscr{F}$ in such a way that Δ_{n+1} is incident with one of $\Delta_1, \ldots, \Delta_n$ for each n. This can be done by first enumerating the triangles in the neighbourhood star of P cyclically, then the triangles one edge distant from P, then those two edges distant from P, and so on. If we now take a small disc neighbourhood $\mathscr{N}(P_k)$ of each vertex P_k in the triangulation and let

$$\Delta_n' = \Delta_n \cup \overline{\mathscr{N}(P_{n1})} \cup \overline{\mathscr{N}(P_{n2})} \cup \overline{\mathscr{N}(P_{n3})},$$

where P_{n1}, P_{n2}, P_{n3} are the vertices of Δ_n, then

$$\Delta_1' \cup \cdots \cup \Delta_n' = \mathscr{C}_n'$$

is a bounded surface for each n.

In general, if a union $\mathscr{U}$ of triangles Δ_j is connected then the union $\mathscr{U}'$ of the corresponding Δ_j' is a surface. $\mathscr{U}'$ is a subcomplex of the surface $\mathscr{F}$, so any of its edges is incident with at most two surface pieces. A vertex P_k of $\mathscr{U}$ has a disc neighbourhood, namely $\mathscr{N}(P_k)$ while a vertex P_k' not of $\mathscr{U}$ has a disc or semidisc neighbourhood according as it meets two or one of the triangles Δ_j (Figure 154).

It follows that $\mathscr{F}$ is a union $\bigcup_n \mathscr{F}_n$ of finite surfaces

$$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \mathscr{F}_3 \subset \cdots$$

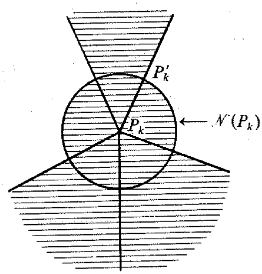


Figure 154

however, to ensure that the free generators of $\pi_1(\mathscr{F}_n)$ remain free in $\pi_1(\mathscr{F}_{n+1})$ we have to control the construction as follows, to prevent the appearance of curves which contract in $\mathscr{F}$ but not in $\mathscr{F}_n$.

Step 1. Set $\mathscr{F}_1 = \Delta'_1$. Because $\mathscr{F} - \Delta'_1$ is an infinite surface it contains no disc spanning the boundary of Δ'_1.

Step $n + 1$. Assume inductively that $\mathscr{F}_n$ has been constructed as a connected union of elements Δ'_j which include all of $\Delta'_1, \ldots, \Delta'_n$ and so that there is no disc in $\mathscr{F} - \mathscr{F}_n$ spanning any of its boundary curves. If Δ'_{n+1} is already contained in $\mathscr{F}_n$, then $\mathscr{F}_{n+1} = \mathscr{F}_n$. Otherwise, $\mathscr{F}_{n+1}$ is constructed by first attaching Δ'_{n+1} to $\mathscr{F}_n$ (they have a common point by hypothesis on the enumeration of $\Delta_1, \Delta_2, \ldots$) then, if any of the new boundary curves which result is spanned by a disc $\mathscr{D}$ in $\mathscr{F} - (\mathscr{F}_n \cup \Delta'_{n+1})$, attaching $\mathscr{D}$ also.

$\mathscr{D}$ is attached by taking the union with the minimal set of Δ'_j which suffice to cover $\mathscr{D}$, and it then follows by induction that $\mathscr{F}_{n+1}$ has properties analogous to those assumed for $\mathscr{F}_n$. Since $\Delta'_n \subset \mathscr{F}_n$ we have

$$\mathscr{F} = \bigcup_n \mathscr{F}_n.$$

To prove the claim about $\pi_1(\mathscr{F}_{n+1})$ we examine the six possible ways that Δ'_{n+1} can be attached to $\mathscr{F}_n$ (Figure 155). The polygon represents an arbitrary boundary curve of $\mathscr{F}_n$, with bumps corresponding to the $\mathscr{N}(P_k)$ neighbourhoods, and the shaded triangle is Δ_{n+1}. The case of all three edges of Δ_{n+1} lying on $\mathscr{F}_n$ is excluded by the induction hypothesis.

In Cases (1), (2), (3) the new boundary curve cannot be spanned by a disc in $\mathscr{F} - (\mathscr{F}_n \cup \Delta'_{n+1})$, otherwise the old one would be spanned by a disc in $\mathscr{F} - \mathscr{F}_n$. So in these cases the construction of $\mathscr{F}_{n+1}$ is complete and the

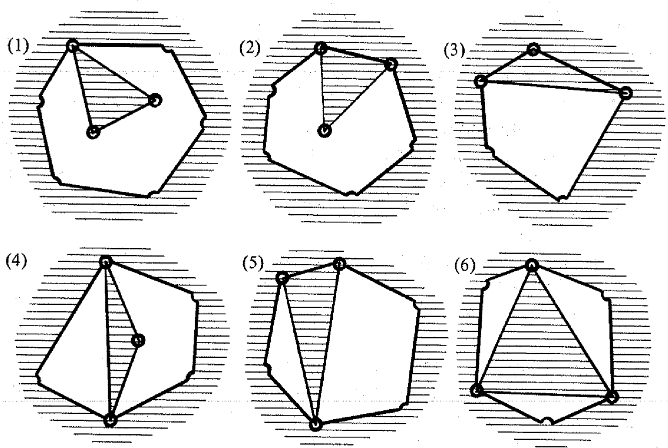


Figure 155

free generators of $\pi_1(\mathscr{F}_n)$ are also free generators for $\pi_1(\mathscr{F}_{n+1})$, for in each case there is a deformation retraction of $\mathscr{F}_{n+1}$ onto $\mathscr{F}_n$.

In Cases (4), (5), (6) the old boundary curve is replaced by two or three new ones, some of which may be spanned by discs in $\mathscr{F} - (\mathscr{F}_n \cup \Delta'_{n+1})$. When these discs are added the result is replacement of the old boundary curve by (a) one, (b) two, or (c) three new ones. In Case (a) the free generators of $\pi_1(\mathscr{F}_n)$ are also free generators for $\pi_1(\mathscr{F}_{n+1})$ by the same argument as above.

In Case (b) we get one new generator. Recall that free generators for $\pi_1(\mathscr{F}_n)$ are found by collapsing $\mathscr{F}_n$ onto a bouquet of circles $\mathscr{B}_n$. It is clear that $\mathscr{F}_{n+1}$ can be collapsed onto $\mathscr{B}_n$ plus one new edge (corresponding to Δ'_{n+1}), hence the free generators of $\pi_1(\mathscr{F}_n)$ will serve as free generators for $\pi_1(\mathscr{F}_{n+1})$ in conjunction with one new generator which passes through Δ'_{n+1}. In Case (c), which can only arise from (6), we get two new generators passing through Δ'_{n+1}.

This completes the proof of the claim.

It follows that there are nested *free* presentations of $\pi_1(\mathscr{F}_1)$, $\pi_1(\mathscr{F}_2)$, $\pi_1(\mathscr{F}_3), \ldots$, and hence a free presentation of $\pi_1(\mathscr{F})$, by the concluding argument of 4.1.7. □

EXERCISE 4.2.2.1. Construct surfaces to realize the free groups on $1, 2, 3, \ldots$ and a countable infinity of generators.

4.2.3 Wirtinger Presentation of Knot Groups

A *knot* $\mathscr{K}$ is a simple closed polygonal curve in R^3. $\mathscr{K}$ of course need not be an actual polygon, but only the image of one under a homeomorphism of all R^3 (this is to exclude wild embeddings, see 4.2.6). During the nineteenth century the study of knots and their classification was pursued on an experimental basis, but with the advent of the fundamental group decisive results could be obtained for the first time. The key observation is that when $\mathscr{K}$ is a *trivial knot* (isotopic to the circle in R^3) the group of its complement is infinite cyclic. Thus if we can show that the *knot group* $\pi_1(\mathsf{R}^3 - \mathscr{K})$ is not infinite cyclic for a particular knot $\mathscr{K}$ we have a topologically sound proof that $\mathscr{K}$ is not trivial.

The first method for computing knot groups was introduced by Wirtinger around 1904 in his lectures in Vienna, but not given wide circulation until its publication in Tietze 1908. We begin with intuitive explanation of the method.

Any knot $\mathscr{K}$ can be given by a projection on the plane with no multiple points which are more than double, and with indication being given, at each double point, which branch of $\mathscr{K}$ is uppermost. Figure 156 shows a projection of the trefoil knot. The double points are called *crossings*. If we break the lower branch at each crossing we obtain a finite set of arcs α_i,

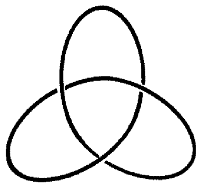


Figure 156

and it is intuitively clear that $\pi_1(\mathrm{R}^3 - \mathscr{K})$ is generated by loops a_i which pass around these arcs. Thus we have as many generators as there are crossings. We choose an orientation for the knot $\mathscr{K}$, and then orient the generators a_i around the arcs α_i by the right-hand screw convention (Figure 157). It is also convenient to order the subscripts to follow a circuit round $\mathscr{K}$, so that the lower arc α_i into a crossing is followed by the arc α_{i+1} out of the crossing. Referring to Figure 158, we see that for the crossing of type (1) the curve $a_i a_j^{-1} a_{i+1}^{-1} a_j$ contracts to a point, hence we have the relation

$$a_i a_j^{-1} a_{i+1}^{-1} a_j = 1 \quad \text{or} \quad a_j a_i = a_{i+1} a_j.$$

For the crossing of type (2) we have

$$a_i a_j a_{i+1}^{-1} a_j^{-1} = 1 \quad \text{or} \quad a_i a_j = a_j a_{i+1}.$$

(The other two possibilities correspond to the opposite orientation of $\mathscr{K}$, and hence give relations of one of these two forms.)

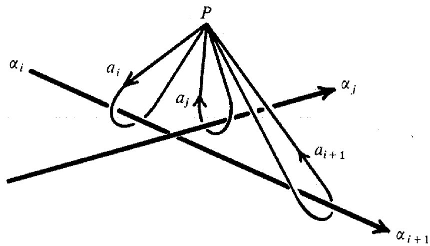


Figure 157

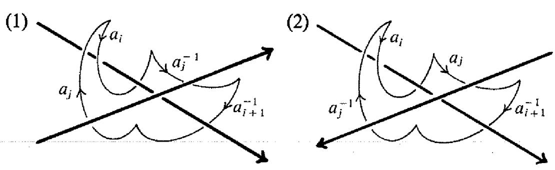


Figure 158

4.2.4 Proof of the Wirtinger Presentation

We now give a rigorous derivation of the presentation using the Seifert–Van Kampen theorem.

We can assume that each arc α_i of the knot $\mathscr{K}$ lies in the plane $z = 1$ except for a vertical segment at each end which goes down to $z = 0$. The final point of α_i can then be joined to the initial point of α_{i+1} by a segment β_j in the plane $z = 0$ passing under the upper arc of the crossing α_j, to complete the knot $\mathscr{K}$. We now remove from R^3 the "tunnel" neighbourhood $\mathscr{N}$ of $\mathscr{K}$ swept out by a cube of side ε which travels with its midpoint on $\mathscr{K}$ and faces parallel to the axes, where ε is small enough to ensure that $\mathsf{R}^3 - \mathscr{N}$ is a deformation retract of $\mathsf{R}^3 - \mathscr{K}$ (Figure 159). $\mathsf{R}^3 - \mathscr{N}$ can now be expressed as the union of open sets $\mathscr{A}$ and $\mathscr{B}$ which reflect the generators and relations, respectively, of $\pi_1(\mathsf{R}^3 - \mathscr{N})$.

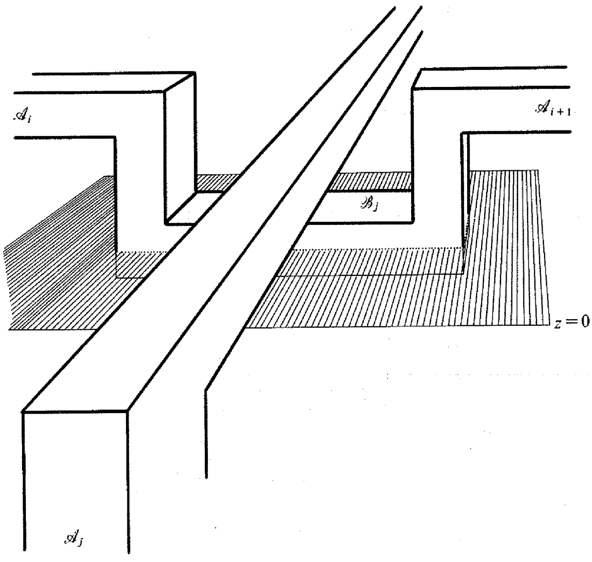


Figure 159

$\mathscr{A} = \{z > 0\} - \mathscr{N}$ has a deformation retraction onto a bouquet of circles $a_1, \ldots, a_n$, where a_i is a loop passing under the tunnel $\mathscr{A}_i$ containing α_i. The reader may become more convinced of this by first deforming $\mathscr{A}$ so that the "hollows" $\mathscr{B}_j$ containing the β_j are pressed down to $z = 0$, then pulling the tunnels $\mathscr{A}_i$ into parallel with each other, as in Figure 160 (cf. the cube with holes in 3.3.2).

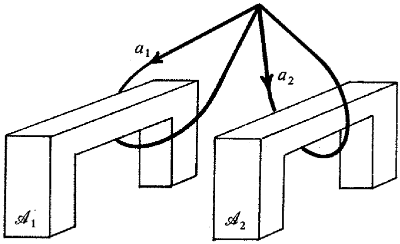


Figure 160

Hence $\pi_1(\mathscr{A}) = \langle a_1, \ldots, a_n; -\rangle$.

$\mathscr{B} = \{z < \varepsilon/2\} - \mathscr{N}$ is an open half-space with "trenches" containing the segments $\mathscr{B}_j$ dug out of it. It is clearly simply connected, so $\pi_1(\mathscr{B}) = \{1\}$.

$\mathscr{A} \cap \mathscr{B} = \{0 < z < \varepsilon/2\} - \mathscr{N}$ is an infinite plate with n holes in it (the upper halves of the trenches), hence

$$\pi_1(\mathscr{A} \cap \mathscr{B}) = \text{free group of rank } n.$$

The typical generator of $\pi_1(\mathscr{A} \cap \mathscr{B})$, a circuit round a trench (Figure 161), has the form $a_i a_j^{-1} a_{i+1}^{-1} a_j$ in $\pi_1(\mathscr{A})$ (or $a_i a_j a_{i+1}^{-1} a_j^{-1}$ for the second type of crossing) and 1 in $\pi_1(\mathscr{B})$. Thus the Seifert–Van Kampen theorem gives precisely the Wirtinger relations for

$$\pi_1(\mathscr{A} \cup \mathscr{B}) = \pi_1(\mathsf{R}^3 - \mathscr{N}) = \pi_1(\mathsf{R}^3 - \mathscr{K}). \qquad \square$$

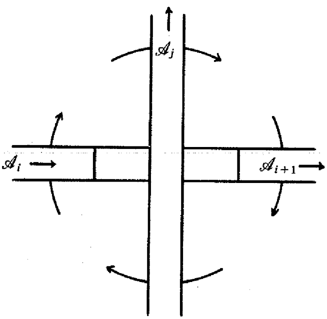


Figure 161

Exercise 4.2.4.1. Show that any one Wirtinger relation is a consequence of the remainder. (*Suggestion*: Instead of using $\mathscr{B}$ to seal the tunnels $\mathscr{B}_j$ at the bottom of $\mathscr{A}$, use a separate open set $\mathscr{C}_j$ to seal each $\mathscr{B}_j$, where $\mathscr{C}_j$ is an open cube with $\mathscr{B}_j$ removed from its top. Then show

$$\begin{aligned}\pi_1(\mathsf{R}^3 - \mathscr{N}) &= \pi_1(\mathscr{A} \cup \mathscr{C}_1 \cup \cdots \cup \mathscr{C}_n)\\ &= \pi_1(\mathscr{A} \cup \mathscr{C}_1 \cup \cdots \cup \mathscr{C}_{i-1} \cup \mathscr{C}_{i+1} \cup \cdots \cup \mathscr{C}_n).)\end{aligned}$$

EXERCISE 4.2.4.2. Generalize the Wirtinger method to compute $\pi_1(\mathsf{R}^3 - \mathscr{G})$, where $\mathscr{G}$ is any graph embedded in R^3. Show in particular that the relation at a vertex of degree n has the form

$$a_{i_1}a_{i_2}\cdots a_{i_n} = 1$$

when $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_n}$ is the clockwise sequence of edges into the vertex.

4.2.5 The Simplest Knot and Link

We now compute the groups for the trefoil knot and the two-crossing link.

(i) The two-crossing link (Figure 162). At X we read off the relation

$$a_2 a_1 a_2^{-1} a_1^{-1} = 1 \quad \text{or} \quad a_1 a_2 = a_2 a_1$$

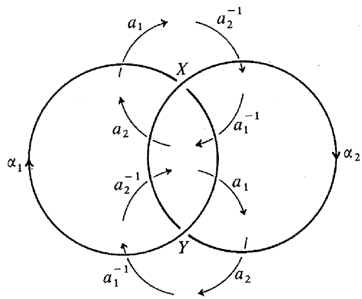


Figure 162

and at Y we find the same relation. Thus the group of the two-crossing link is

$$\langle a_1, a_2; a_1 a_2 = a_2 a_1 \rangle$$

or the free abelian group of rank 2. The group of (R^3-two unlinked circles) is the free group of rank 2 (why?) and hence we have a proof of the non-triviality of the link.

(ii) The trefoil knot (Figure 163). At X we read off

$$a_1 a_3^{-1} a_2^{-1} a_3 = 1. \tag{1}$$

At Y

$$a_3 a_2^{-1} a_1^{-1} a_2 = 1. \tag{2}$$

At Z

$$a_2 a_1^{-1} a_3^{-1} a_1 = 1. \tag{3}$$

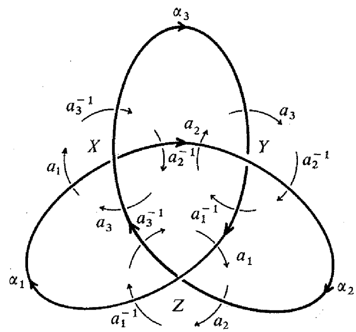


Figure 163

Solving (1) for a_1 and substituting the result in (2) and (3) we find that each yields the equation

$$a_3 a_2 a_3 = a_2 a_3 a_2. \tag{4}$$

It follows that

$$(a_2 a_3 a_2)^2 = (a_3 a_2 a_3)(a_2 a_3 a_2) = (a_3 a_2)^3$$

which we write

$$a^2 = b^3 \tag{5}$$

by setting $a = a_2 a_3 a_2$, $b = a_3 a_2$. But $a_2 = ab^{-1}$, $a_3 = b^2 a^{-1}$ so a, b are in fact generators and (4) is a consequence of (5). Thus we have the presentation

$$G_{2,3} = \langle a, b; a^2 = b^3 \rangle$$

for the group of the trefoil knot.

We now investigate whether $G_{2,3}$ is infinite cyclic. Notice that the group S_3 of permutations on three symbols is also a model of the relation $a^2 = b^3$, namely, take

$$a = (1\ 2), \qquad b = (1\ 2\ 3).$$

Hence any relation in $G_{2,3}$ is also valid in S_3 under this interpretation of a, b. It follows that $ab = ba$ is not a relation of $G_{2,3}$, since $ab \neq ba$ in S_3, and therefore $G_{2,3}$ is not infinite cyclic because all elements commute in the infinite cyclic group. (The representation of $G_{2,3}$ by S_3 is due to Wirtinger, who used it to construct a covering of $\mathbf{S}^3$ branched over the trefoil knot, see 1.1.4. The same covering had already been considered by Heegaard 1898, who made the surprising discovery that the covering manifold is also $\mathbf{S}^3$.)

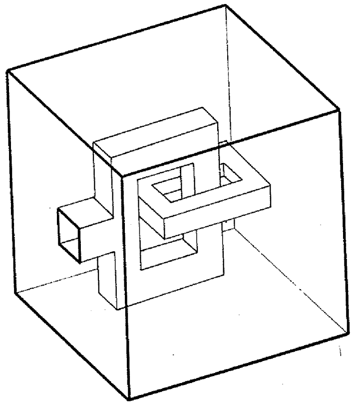


Figure 164

EXERCISE 4.2.5.1. Show that the cube with holes shown in Figure 164 has a deformation retraction onto the torus, and hence gives an alternative derivation of the group of the two-crossing link.

4.2.6 The Fox–Artin Wild Arc

A simple polygonal arc $\mathscr{A}$ in R^3 has the property that $\pi_1(\mathsf{R}^3 - \mathscr{A}) = \{1\}$. Fox and Artin 1948 call a simple arc $\mathscr{A}$ in R^3 *wild* if there is no homeomorphism of R^3 which maps $\mathscr{A}$ onto a polygon, in particular if $\pi_1(\mathsf{R}^3 - \mathscr{A}) \neq \{1\}$. Figure 165 shows an example of a wild arc (the limit points P and Q are included in the arc).

The generators we shall use for $\pi_1(\mathsf{R}^3 - \mathscr{A})$ are loops a_n, b_n, c_n for all integers n, placed as shown in Figure 166. $\mathsf{R}^3 - \mathscr{A}$ is the union of sets $\mathscr{C}_n$ obtained by removing cubes centred on P, Q at the positions shown in

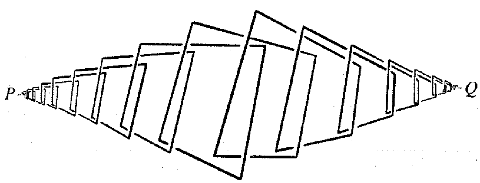


Figure 165

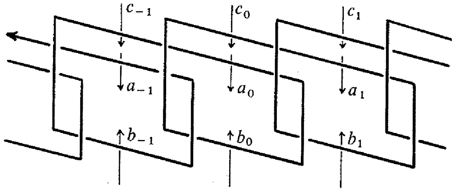


Figure 166

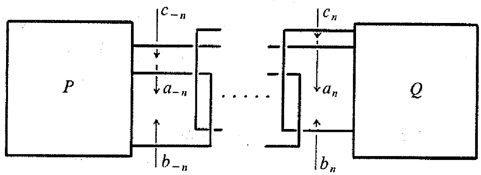


Figure 167

Figure 167. The generators of $\pi_1(\mathscr{C}_n)$ are a_m, b_m, c_m for $-n \leq m \leq n$ and the relations are

$$\left.\begin{aligned} a_{m+1} &= c_{m+1}^{-1} c_m c_{m+1} \\ b_m &= c_{m+1}^{-1} a_m c_{m+1} \\ c_{m+1} &= b_m^{-1} b_{m+1} b_m \end{aligned}\right\} \quad \begin{aligned} &\text{Wirtinger relations at the crossings;} \\ \\ &\qquad -n \leq m \leq n \end{aligned}$$

together with relations

$$c_{-n} a_{-n} = b_{-n}, \qquad c_n a_n = b_n$$

at the ends (shrinking the cubes to points and using Exercise 4.2.4.2).

By removing a small tunnel neighbourhood of the arc (of diameter which tends to 0 as $n \to \infty$) we can replace $\mathscr{C}_n$ by a finite simplicial complex, so it follows from 4.1.7 that the generators of $\pi_1(\mathrm{R}^3 - \mathscr{A})$ are a_n, b_n, c_n as claimed, and the relations are (for all integers n)

$$c_n a_n = b_n \tag{1}$$

$$a_{n+1} = c_{n+1}^{-1} c_n c_{n+1} \tag{2}$$

$$b_n = c_{n+1}^{-1} a_n c_{n+1} \tag{3}$$

$$c_{n+1} = b_n^{-1} b_{n+1} b_n. \tag{4}$$

Substituting (2) in (1) and (3) gives

$$c_n c_n^{-1} c_{n-1} c_n = b_n, \quad \text{that is,} \quad b_n = c_{n-1} c_n \tag{5}$$

and

$$b_n = c_{n+1}^{-1} c_n^{-1} c_{n-1} c_n c_{n+1}. \tag{6}$$

Substituting (5) in (4) gives

$$c_{n+1} = c_n^{-1} c_{n-1}^{-1} c_n c_{n+1} c_{n-1} c_n$$

or

$$c_{n-1} c_n c_{n+1} = c_n c_{n+1} c_{n-1} c_n \tag{7}$$

which is the same as the result of eliminating b_n between (5) and (6). Thus we can use the c's as generators, with the defining relations (7).

It can then be verified that these relations hold in the nontrivial group generated by the permutations (1 2 3 4 5) and (1 4 2 3 5) when c_n is interpreted as (1 2 3 4 5) for n odd and (1 4 2 3 5) for n even, hence $\pi_1(\mathsf{R}^3 - \mathscr{A}) \neq \{1\}$ □

In constructing the wild arc $\mathscr{A}$ we have also constructed a *wild ball* (the tunnel neighbourhood of $\mathscr{A}$) and *wild sphere* (the boundary of the wild ball). The first examples of such objects were given by Antoine 1921, based on an even more paradoxical object, a *wild Cantor set* in R^3. The ordinary Cantor set, obtained by the "middle-third" construction on the unit interval, has a simply connected complement in R^3. However, a wide variety of descending sequence constructions lead to homeomorphic images of the Cantor set; the one used by Antoine iterates the construction of linked solid tori inside a solid torus (Figure 168). Four linked tori are constructed again within each inner torus, and so on. The intersection of all these tori is a Cantor set in R^3 called *Antoine's necklace*. Antoine showed geometrically that its complement is not simply connected, and this was confirmed by calculation of the fundamental group by Blankenship and Fox 1950.

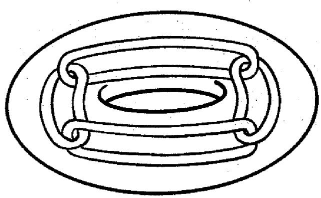


Figure 168

Fox and Artin 1948 also showed the wildness of the arc $\mathscr{A}'$ obtained by altering $\mathscr{A}$ so that the crossings are alternately over and under. $\mathscr{A}'$ is in fact the "chain stitch" of knitting, infinitely extended in both directions. Its group is calculated similarly, but turns out to be slightly more complicated than that of $\mathscr{A}$. It is interesting to note that the infinite chain stitch was pictured in the first ever paper on knot theory, Vandermonde 1771.

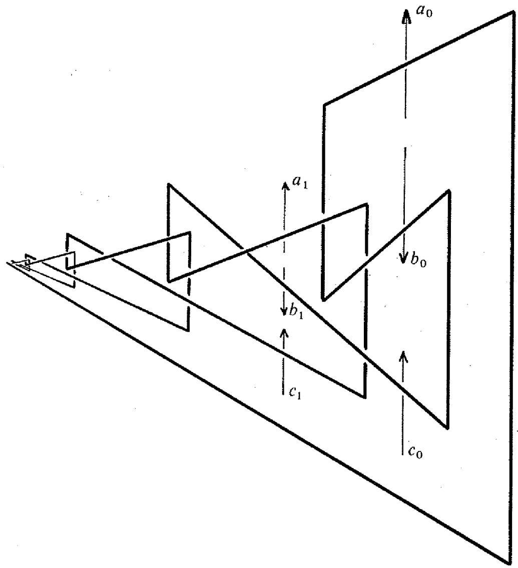


Figure 169

EXERCISE 4.2.6.1 (Fox 1949). Show that the group of the simple closed curve in Figure 169 is generated by $b_0, b_1, b_2, \ldots$ subject to the relations

$$b_1 b_0 b_1^{-1} = b_2 b_1 b_2^{-1} = b_3 b_2 b_3^{-1} = \cdots$$

and find a permutation representation which shows it is nonabelian.

4.2.7 Torus Knots

Consider a solid cylinder $\mathscr{C}$ with m line segments on its curved face, equally spaced and parallel to the axis. If the ends of $\mathscr{C}$ are identified after a twist of $2\pi(n/m)$, where n is an integer relatively prime to m, we obtain a single curve $\mathscr{K}_{m,n}$ on the surface of a solid torus $\mathscr{T}$ (Figure 170). Assuming that $\mathscr{T}$ lies

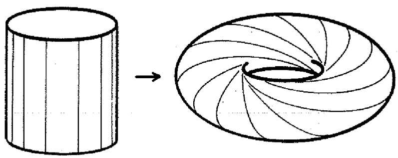


Figure 170

in R^3 in the "standard" way, which means among other things that $\pi_1(\mathsf{R}^3 - \mathcal{T})$ is infinite cyclic, the curve $\mathcal{K}_{m,n}$ is called the (m, n) *torus knot*. We now compute $\pi_1(\mathsf{R}^3 - \mathcal{K}_{m,n})$, following the method of Seifert and Threlfall 1934.

If we drill out a thin tubular neighbourhood $\mathcal{N}$ of $\mathcal{K}_{m,n}$ from R^3, the effect on $\mathcal{T}$ is to gouge out a narrow channel from its surface, and similarly on the surface of $\mathsf{R}^3 - \mathcal{T}$. $\mathcal{T} - \mathcal{N}$ and $(\mathsf{R}^3 - \mathcal{T}) - \mathcal{N}$ then meet along an annulus $\mathcal{L}_{m,n}$ which, like $\mathcal{K}_{m,n}$, results from m parallel strips on the cylinder being joined up after a twist of $2\pi(n/m)$ (Figure 171). $\pi_1(\mathcal{L}_{m,n})$ is infinite cyclic and generated by the centre line $l_{m,n}$ of $\mathcal{L}_{m,n}$. $\pi_1(\mathcal{T} - \mathcal{N})$ is also infinite cyclic and generated by the axis a of $\mathcal{T}$. Since $l_{m,n}$ results from m circuits of $\mathcal{T}$ we have

$$l_{m,n} = a^m$$

in $\pi_1(\mathcal{T} - \mathcal{N})$. Similarly, $\pi_1((\mathsf{R}^3 - \mathcal{T}) - \mathcal{N})$ is infinite cyclic, generated by a loop b through the "hole" in $\mathcal{T}$, and

$$l_{m,n} = b^n$$

in $\pi_1((\mathsf{R}^3 - \mathcal{T}) - \mathcal{N})$.

Then if we expand $\mathcal{T} - \mathcal{N}$ and $(\mathsf{R}^3 - \mathcal{T}) - \mathcal{N}$ slightly across $\mathcal{L}_{m,n}$ to open sets $\mathcal{A}$, $\mathcal{B}$ which intersect in a neighbourhood of $\mathcal{L}_{m,n}$, the Seifert–Van Kampen theorem becomes applicable, and we obtain the presentation

$$G_{m,n} = \langle a, b; a^m = b^n \rangle$$

for the group of the (m, n) torus knot. □

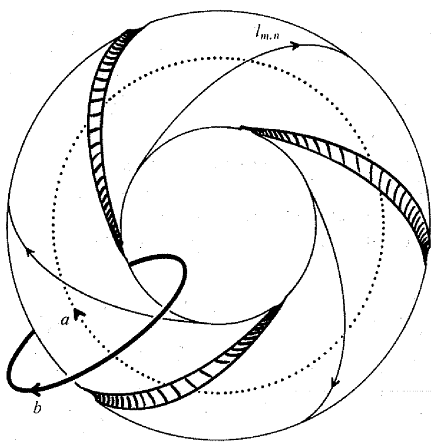


Figure 171

EXERCISE 4.2.7.1. Show that the (m, n) torus knot is the same as the (n, m) torus knot.

EXERCISE 4.2.7.2. Let $\mathscr{H}_n$ be the solid body (*handlebody*) bounded by an orientable surface of genus n which is standardly embedded in R^3, so that $\pi_1(\mathsf{R}^3 - \mathscr{H}_n)$ is the free group of rank n. Show that if $\mathscr{K}$ is a simple curve on the surface of $\mathscr{H}_n$ then $\pi_1(\mathsf{R}^3 - \mathscr{K})$ has a presentation with $n + 1$ generators and n relations. (*Hint*: Attach a thin handle $\mathscr{H}$ to $\mathscr{H}_n$ which follows $\mathscr{K}$ just above the surface of $\mathscr{H}_n$ and has its ends at neighbouring points on $\mathscr{K}$. Show that $\mathscr{H}_n \cup \mathscr{H}$ is a standardly embedded handlebody $\mathscr{H}_{n+1}$, then cut $\mathscr{H}_{n+1}$ so as to obtain a body whose complement is homeomorphic to $\mathsf{R}^3 - \mathscr{K}$.)

4.2.8 Lens Spaces

The (m, n) lens space is a 3-dimensional manifold introduced by Tietze 1908, by means of the following construction. On the surface of a solid ball B^3 one draws an equatorial circle and m equally spaced meridians, dividing the upper hemisphere into triangles $\Delta_1, \ldots, \Delta_m$ and the lower hemisphere into triangles $\Delta'_1, \ldots, \Delta'_m$, where Δ'_i is below Δ_i (Figure 172). The upper hemisphere is then identified with the lower after twisting it through $2\pi(n/m)$, that is, a point P with latitude and longitude (θ, ϕ) is identified with the point P' with latitude and longitude $(-\theta, \phi + 2\pi(n/m))$, where $\phi + 2\pi(n/m)$ is reduced mod 2π. Thus Δ_i is identified with Δ'_{i+n} after inversion ($i + n$ reduced mod m).

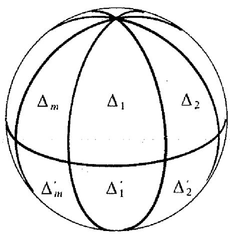


Figure 172

It is evident that if m, n have a common divisor d then the result is expressible more simply as the $(m/d, n/d)$ lens space, so we may as well assume that m, n are relatively prime. Likewise, there is no point in taking $n \geq m$.

Many properties of the (m, n) lens space will come to light in Chapter 8, in particular the fact that it is a manifold and the reason for the name "lens space." For the moment we wish only to compute its fundamental group.

By virtue of 4.1.5, we can forget about the interior of B^3, and just compute π_1 of the surface complex which results from identification of the two hemispheres. Since each point below the equator is identified with a point

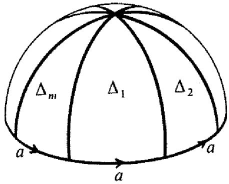


Figure 173

above, it suffices in turn to find out what becomes of the upper hemisphere when the identifications on its boundary, the equator, are carried out.

Since m, n are relatively prime, the numbers

$$1, \ 1 + n, \ 1 + 2n, \ \ldots$$

run through all values $1, 2, \ldots, m$ when reduced mod m. This means that corresponding points on the bases of any two triangles are identified (Figure 173) or that the equator is wrapped m times round a circle corresponding to the base of a triangle. Thus our surface complex is a disc with boundary identified with the path a^m round a circle a. Its fundamental group is therefore $\langle a; a^m \rangle$, the cyclic group of order m, by 3.4.4. □

4.3 Surface Complexes and Subgroup Theorems

4.3.1 Surface Complexes and Groups

A *surface complex* $\mathscr{F}$ consists of three sets $\{P_i\}$, $\{e_j\}$, and $\{\Delta_k\}$ of elements called *vertices*, *edges*, and *faces* respectively, subject to certain incidence relations. The vertices and edges constitute a graph $\mathscr{G}$ (see 2.1.2) called the *1-skeleton* of $\mathscr{F}$, and each face Δ_k is incident with a certain closed path b_k in $\mathscr{G}$ called its *boundary path*.

There is no harm in thinking of $\mathscr{F}$ being realized by actual points, line segments, and discs embedded in some euclidean space, where Δ_k is a disc with its boundary identified with a closed path b_k and different edges and discs are disjoint except where identifications force boundary points into coincidence. Comparison with 4.1.3 will then show that the group we are about to define combinatorially is the familiar fundamental group of $\mathscr{F}$. However, the purely combinatorial approach will suffice for the results we wish to derive, and no appeal will be made to general continuity considerations. Thus the situation is comparable with Chapter 2; the geometric language could in principle be dispensed with, but it seems to convey the most natural explanation of certain group-theoretic results. Indeed, it could be said that those results follow from viewing groups themselves as surface complexes.

The (combinatorial) *fundamental group of* $\mathscr{F}$, $\pi_1(\mathscr{F})$ is the group defined by extending the path product operation to equivalence classes of closed edge paths from some vertex P. Paths p, p' are equivalent if one can be converted to the other by a finite sequence of operations of the following types:

(i) insertion or removal of spurs;
(ii) insertion or removal of boundary paths of faces.

Paths which are equivalent by operations (i) above will be called *freely* equivalent. The equivalence class of p will be denoted $[p]$.

We know from 2.1.7 that generators for all edge paths based at P can be found by constructing a spanning tree $\mathscr{T}$ of $\mathscr{G}$, and for each edge $e_i = P_m P_n$ taking the closed path

$$a_i = w_m e_i w_n^{-1},$$

where w_r denotes the unique reduced path in $\mathscr{T}$ from P to P_r. In fact, any closed path $p(e_i)$ from P is freely equivalent to the corresponding product of the a_i's, $p(a_i)$.

For each boundary path

$$b_k(e_i) = e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n} \quad (\text{where } e_j = \pm 1)$$

of a face Δ_k we have a relation

$$b_k(a_i) = a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_n}^{\varepsilon_n} = 1 \qquad (k)$$

because $b_k(a_i)$ is freely equivalent to the path $wb_k(e_i)w^{-1}$, where w is the unique reduced path from P to the initial point of $b_k(e_i)$. Conversely, any relation in $\pi_1(\mathscr{F})$ is a consequence of the relations (k), because the result of insertion (deletion) of $b_k(e_i)$ in a closed path $p(e_i)$ from P is freely equivalent to the result of insertion (deletion) of $b_k(a_i)$ in the path $p(a_i)$.

It follows immediately that any group $\mathscr{G}$ can be realized as the combinatorial fundamental group of a surface complex $\mathscr{F}$, by taking a bouquet of circles $a_1, a_2, \ldots$ and attaching a face Δ_k with boundary b_k for each relation $b_k(a_i) = 1$ of G. Furthermore, some of the useful topological properties of the (topological) fundamental group have combinatorial counterparts. We now establish some for use in later sections.

(a) $\pi_1(\mathscr{F})$ *does not change under elementary subdivisions of* $\mathscr{F}$ *or their inverses* (1.3.8).

Subdivision of an edge means replacing some e_i by $e_i' e_i''$ and a_i by $a_i' a_i''$ accordingly. But if we extend the spanning tree $\mathscr{T}$ to reach the new vertex it must include exactly one of e_i', e_i'', say e_i''. Then $a_i'' = 1$ and all we have done to $\pi_1(\mathscr{F})$ is to change the presentation by replacing a_i by a_i'.

When a face is subdivided we can assume that the new edge e_l begins at the initial vertex of the boundary path r_k (since a defining relator can be replaced by a cyclic permutation of itself). Then if $r_k = r_k' r_k''$ is the subdivision effected by e_l the boundary paths of the new faces are $r'_k e_l^{-1}$ and $e_l r''_k$, so that the original relation

$$r_k(a_i) = 1$$

is replaced by two relations

$$r'_k(a_i)a_l^{-1} = 1, \qquad a_l r''_k(a_i) = 1.$$

But the latter are equivalent to the former in conjunction with the definition $a_l = r'_k(a_i)$ of the new generator, so we have the same group. □

This is the essence of the theorem of Tietze 1908 that the combinatorial fundamental group is invariant under combinatorial homeomorphisms. It implies in particular that the fundamental group of a finite surface is equal to that of its standard form (1.3.7), a fact we shall use in 4.3.7.

(b) *An elementary collapse across a free edge e_j corresponds to elimination of the generator a_j (so collapsing preserves the fundamental group).*

Since the edge e_j is free it occurs in only one face Δ_k, and only once in the boundary r_k of Δ_k. By cyclic permutation, if necessary, we can express the relation corresponding to r_k as

$$a_j p(a_i) = 1,$$

where $p(a_i)$ does not involve a_j. Hence we can eliminate a_j, since it equals $(p(a_i))^{-1}$, and with it the relation $r_k(a_i) = 1$. But this is precisely what happens when we collapse the face Δ_k across the edge e_j. □

4.3.2 Coverings of Surface Complexes

A surface complex $\tilde{\mathscr{F}}$ is said to *cover* a surface complex $\mathscr{F}$ if there is a map $\phi: \tilde{\mathscr{F}} \to \mathscr{F}$ from the vertices, oriented edges, and faces of $\tilde{\mathscr{F}}$ onto the vertices, oriented edges, and faces of $\mathscr{F}$ respectively, with the following properties:

(1) The restriction of ϕ to the 1-skeleton $\tilde{\mathscr{G}}$ of $\tilde{\mathscr{F}}$ is a covering $\phi|\tilde{\mathscr{G}}: \tilde{\mathscr{G}} \to \mathscr{G}$ of the 1-skeleton $\mathscr{G}$ of $\mathscr{F}$ (cf. 2.2.1).
(2) ϕ preserves boundaries, that is, if $\tilde{b}_k$ is the boundary path of $\tilde{\Delta}_k$ in $\tilde{\mathscr{F}}$ then $\phi(\tilde{b}_k)$ is the boundary path of $\phi(\tilde{\Delta}_k)$.
(3) For each distinct pair $\langle \Delta, \tilde{b} \rangle$, where Δ is a face of $\mathscr{F}$ and $\tilde{b}$ covers the boundary path b of Δ there is a distinct face $\tilde{\Delta}$ with boundary $\tilde{b}$ in $\tilde{\mathscr{F}}$, covering Δ, and all faces of $\tilde{\mathscr{F}}$ arise in this way.

Condition (3) is stronger than we need to prove the subgroup property (4.3.5), which requires only (3′): boundary paths lift to boundary paths. It is designed to secure a local homeomorphism property for topological applications. The motivating example is the covering of the projective plane

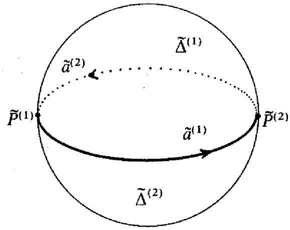


Figure 174

by the sphere: suppose we realize the projective plane $\mathscr{F}$ by a vertex P, edge a, and face Δ with boundary path a^2. Then if we cover the 1-skeleton of $\mathscr{F}$ by the graph with vertices $\tilde{P}^{(1)}$, $\tilde{P}^{(2)}$ and edges $\tilde{a}^{(1)}$, $\tilde{a}^{(2)}$, as shown in Figure 174, there are two distinct paths covering a^2, namely $\tilde{a}^{(1)}\tilde{a}^{(2)}$ and $\tilde{a}^{(2)}\tilde{a}^{(1)}$. We therefore need two faces $\tilde{\Delta}^{(1)}$ and $\tilde{\Delta}^{(2)}$ in $\tilde{\mathscr{F}}$ with $\tilde{a}^{(1)}\tilde{a}^{(2)}$ and $\tilde{a}^{(2)}\tilde{a}^{(1)}$ as their respective boundary paths. Thus $\tilde{\mathscr{F}}$ is the sphere.

The above definition of covering is essentially that given by Reidemeister 1932 in the first systematic treatment of covering complexes. Reidemeister does not explicitly say that faces of $\tilde{\mathscr{F}}$ arise *only* as required by (3), however he uses this assumption in proving that a covering of a closed surface is itself a closed surface (cf. 4.3.4).

We continue with the notational convention of using $\tilde{X}$ to denote any element of $\tilde{\mathscr{F}}$ which covers a particular element X in $\mathscr{F}$, adding superscripts if different instances $\tilde{X}^{(1)}, \tilde{X}^{(2)}, \ldots$ have to be distinguished.

4.3.3 Neighbourhoods

To extend our geometric language to a concept of neighbourhood we imagine the face Δ with boundary

$$b = e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n} \qquad (\varepsilon_j = \pm 1)$$

divided into *sectors* $\Delta(e_{i_1}^{\varepsilon_1})$, $\Delta(e_{i_2}^{\varepsilon_2}), \ldots, \Delta(e_{i_n}^{\varepsilon_n})$ associated with the successive oriented edges in b. The same edge e_i (possibly with different exponents) may of course have several different sectors associated with it. The sectors, from all faces, incident with e_i constitute the *neighbourhood* of e_i (also the neighbourhood of e_i^{-1}) and they may be visualized as a "book" with n triangular "leaves," where n is the number of times e_i occurs, positively or negatively, in boundary paths of faces (counting an occurrence in a boundary path b as many times as there are faces bounded by b). See Figure 175.

We define the neighbourhood of a vertex P by considering all pairs $\{e_i, e_j\}$ such that

$$P = \text{final point of } e_i = \text{initial point of } e_j.$$

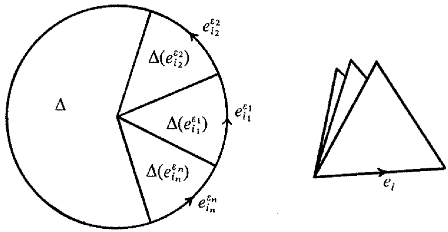


Figure 175

These may include $\{e_i\}$ if e_i is a loop at P, in which case we take the neighbourhood of $\{e_i\}$ to be that of e_i, just defined. The neighbourhood of a proper pair $\{e_i, e_j\}$, which is also the neighbourhood of $\{e_j^{-1}, e_i^{-1}\}$, is the set of sector pairs corresponding to occurrences of $e_i e_j$ in boundary paths of faces (Figure 176), where boundary paths are written cyclically and, as above, an occurrence in a given boundary path b is counted as many times as there are faces bounded by b.

Because we assume that distinct faces do not meet except along common edges, it is reasonable to say that the neighbourhoods of all edge pairs incident with P determine the *neighbourhood of P*.

We then define neighbourhoods of edges to be *homeomorphic*, in the combinatorial sense, when they have the same number of leaves, and define neighbourhoods of vertices to be homeomorphic if there is a one-to-one correspondence between their incident edges such that corresponding pairs have neighbourhoods of the same number of leaves.

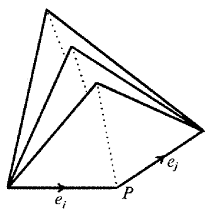


Figure 176

4.3.4 The Local Homeomorphism Property

It is now possible to prove that

If $\phi: \tilde{\mathscr{F}} \to \mathscr{F}$ is a covering, then the neighbourhood of an edge $\tilde{e}_i$ covering e_i is homeomorphic to the neighbourhood of e_i, and the neighbourhood of a vertex $\tilde{P}$ covering P is homeomorphic to the neighbourhood of P.

Suppose an edge e_i has a neighbourhood of n leaves, so that there are n occurrences of e_i in boundary paths of faces in $\mathscr{F}$. We want to match these occurrences with the occurrences of a covering edge $\tilde{e}_i$ in the boundary paths of faces of $\tilde{\mathscr{F}}$.

First consider a boundary path b in $\mathscr{F}$ which bounds only a single face Δ. Let $\tilde{b}^{(1)}, \tilde{b}^{(2)}, \ldots$ be the paths in $\tilde{\mathscr{F}}$ which cover b and contain the edge $\tilde{e}_i$. By (3), these are exactly the boundary paths of faces which cover Δ and are incident with $\tilde{e}_i$. We characterize each occurrence of e_i in b by its *position*, that is, by the number of letters in the initial segment of the word for b ending at the occurrence in question; then the corresponding positions in $\tilde{b}^{(1)}$, $\tilde{b}^{(2)}, \ldots$ have the following properties:

(i) $\tilde{e}_i$ does not fill the same position in two different paths $\tilde{b}^{(j)}$.
(ii) Each position filled by e_i in b is filled by $\tilde{e}_i$ in some $\tilde{b}^{(j)}$, and conversely.

Both of these are immediate consequences of the unique lifting of paths in graph coverings (see (1) of 4.3.2 and 2.2.1). (i) holds because if $\tilde{e}_i$ is in the same position in $\tilde{b}^{(j)}, \tilde{b}^{(k)}$, then $\tilde{b}^{(j)} = \tilde{b}^{(k)}$ since both cover b, and (ii) is obtained by lifting the cyclic permutation of b which starts at the position in question, with $\tilde{e}_i$ being chosen as the initial edge. Then applying the inverse permutation to the covering path yields a $\tilde{b}^{(j)}$ with $\tilde{e}_i$ in the required position. Projection by ϕ gives the converse.

Since $\tilde{b}^{(1)}, \tilde{b}^{(2)}, \ldots$ are exactly the boundary paths of faces $\tilde{\Delta}$ which cover Δ, (i) and (ii) give a one-to-one correspondence between the occurrences of e_i in the boundary path of Δ and the occurrences of its cover $\tilde{e}_i$ in the boundary paths of faces which cover Δ.

If b bounds several faces in $\mathscr{F}$, then each multiple occurrence of e_i in b is matched by an occurrence of $\tilde{e}_i$ of equal multiplicity in some $\tilde{b}^{(j)}$. Finally, adding the contributions from different boundary paths $b_1, b_2, \ldots$ we reach the same number of occurrences for both e_i and $\tilde{e}_i$, hence these edges have homeomorphic neighbourhoods.

The argument is completely analogous when one considers neighbourhoods of a pair $\{e_i, e_j\}$, hence we likewise find that each vertex of $\tilde{\mathscr{F}}$ has a neighbourhood homeomorphic to that of the vertex it covers (bearing in mind that there is already a one-to-one correspondence between the incident edge sets by (1) and 2.2.1). □

The most important consequence of this result is that a covering of a closed surface is itself a closed surface, since a closed surface is simply a surface complex in which each edge has a two-leaved neighbourhood and each vertex has an umbrella neighbourhood (1.3.1). It is easily verified that a neighbourhood homeomorphic to an umbrella is itself an umbrella.

Another consequence is that the *sheet number*, which can be defined for the covering of the 1-skeleton by 2.2.1, is also equal to the number of faces in $\tilde{\mathscr{F}}$ which cover a given face in $\mathscr{F}$.

EXERCISE 4.3.4.1. Show that a cover of a bounded surface is itself a bounded surface.

EXERCISE 4.3.4.2. Show that for each positive integer i the closed orientable surface of genus $n \geq 1$ has an i-sheeted cover, and that any such cover is a closed orientable surface of genus $i(n - 1) + 1$.

EXERCISE 4.3.4.3. Show that a closed finite nonorientable surface has a unique 2-sheeted covering surface.

4.3.5 The Subgroup Property

A covering $\phi: \tilde{\mathscr{F}} \to \mathscr{F}$ induces a monomorphism $\phi_*: \pi_1(\tilde{\mathscr{F}}) \to \pi_1(\mathscr{F})$. Furthermore, the classification of closed path classes based at P in $\mathscr{F}$ according to the final points of the covering paths from a fixed $\tilde{P}$ in $\tilde{\mathscr{F}}$ is exactly the right coset decomposition of $\pi_1(\mathscr{F})$ modulo $\pi_1(\tilde{\mathscr{F}})$, so that the sheet number of the covering equals the index of $\pi_1(\tilde{\mathscr{F}})$ in $\pi_1(\mathscr{F})$.

Equivalent closed paths p, p' based at P lift to equivalent paths $\tilde{p}, \tilde{p}'$ based at $\tilde{P}$, since spurs lift to spurs by (1) and 2.2.2, and boundary paths of faces lift to boundary paths of faces by (3). Conversely, spurs and boundary paths in $\tilde{\mathscr{F}}$ project to spurs and boundary paths in $\mathscr{F}$, by (1) and (2), so there is indeed a one-to-one correspondence ϕ_* induced by ϕ between the closed path classes based at $\tilde{P}$ and certain closed path classes based at P.

Since ϕ maps products to products and inverses to inverses by (1), ϕ_* is a homomorphism, and hence a monomorphism

$$\phi_*: \pi_1(\tilde{\mathscr{F}}) \to \pi_1(\mathscr{F}).$$

The coset decomposition now follows exactly as in the case of graph coverings, 2.2.2. □

4.3.6 Realization of Subgroups

Given a group

$$G = \langle e_1, e_2, \ldots; r_1, r_2, \ldots \rangle$$

we realize it as $\pi_1(\mathscr{F})$, where $\mathscr{F}$ is a surface complex consisting of a single vertex P, edges $e_1, e_2, \ldots$ and faces $\Delta_1, \Delta_2, \ldots$ bounded by the closed paths $r_1, r_2, \ldots$. Then any subgroup H of G can be realized as $\pi_1(\tilde{\mathscr{F}})$, where $\tilde{\mathscr{F}}$ is a surface complex covering $\mathscr{F}$.

The 1-skeleton of $\tilde{\mathscr{F}}$ is a graph $\tilde{\mathscr{G}}$ covering the bouquet of circles $e_1, e_2, \ldots$ which is the 1-skeleton of $\mathscr{F}$. We take a vertex $\tilde{P}^{(j)}$ covering P for each right coset of G modulo H then, as in 2.2.3, $\tilde{\mathscr{G}}$ is uniquely determined by the condition that the outgoing edge labelled e_i from the vertex corresponding to $H[p]$ must end at the vertex corresponding to $H[pe_i]$. $\tilde{\mathscr{F}}$ is completed by adding a face $\tilde{\Delta}_k$ bounded by $\tilde{r}_k$ for each path $\tilde{r}_k$ which covers an r_k in $\mathscr{F}$. (Observe that an r_k always lifts to a closed path, because $H[p]$ and $H[pr_k]$ are the same coset for any p.)

The closed paths $\tilde{p}$ based at some fixed $\tilde{P}$ in $\tilde{\mathscr{F}}$ are exactly those which cover elements of H, since $H[p] = H$ just in case $[p] \in H$. Thus the canonical monomorphism

$$\phi_*: \pi_1(\tilde{\mathscr{F}}) \to \pi_1(\mathscr{F}) \text{ is onto } H. \qquad \square$$

The above construction is readily generalized to the case where $\mathscr{F}$ is an arbitrary surface complex realizing G if one first proves the corresponding generalization for graph coverings (Exercise 2.2.3.1). Then one has that the coverings of an arbitrary surface complex $\mathscr{F}$ realize exactly the subgroups of $\pi_1(\mathscr{F})$, which gives the full force of the parallel between coverings and subgroups discovered by Reidemeister 1928. Reidemeister first proved the result in terms of coverings of manifolds (three or more dimensions), then rephrased it in terms of surface complexes for his book Reidemeister 1932.

It is also possible to show this parallel in a theory of covering spaces based on general continuous maps rather than the combinatorial maps ϕ used above. The local homeomorphism property is then part of the definition of covering, and effort must instead be applied to prove that paths lift uniquely and that homotopic paths lift to homotopic paths. This is not difficult to do using the methods of chapter 3 and may be seen in many texts, for example Massey 1967. It is certainly the best approach when covering spaces are being studied for their own sake, however one does not want to make inessential use of continuity in deriving purely combinatorial results, especially when there is no appreciable gain in simplicity.

In this book we deal only with the combinatorial consequences of covering space theory, notably the results in the three sections which follow.

To illustrate the construction of $\tilde{\mathscr{F}}$ in the theorem we realize the commutator subgroup H of the modular group $G = \langle a, b; a^2, b^3 \rangle$. H consists of all elements for which the exponent sum of $a = 0$ mod 2 and the exponent sum of $b = 0$ mod 3.

There are six right cosets of G modulo H, represented by the elements $1, b, b^2, a, ab, ab^2$. The 1-skeleton of $\tilde{\mathscr{F}}$ is then easily seen to be the graph (1) in Figure 177, which is in fact the Cayley diagram of G/H. We have labelled the vertices by the corresponding coset representatives.

To show how the faces of $\tilde{\mathscr{F}}$ fit on we view a 3-dimensional form of the graph (2). The vertical faces cover the face bounded by a^2 in $\mathscr{F}$, and are double, while the horizontal faces cover the face bounded by b^3 in $\mathscr{F}$ and are triple. It will not affect $\pi_1(\tilde{\mathscr{F}})$ if the multiple faces are replaced by single ones, then a collapse of the faces to lines yields the graph (3), whose fundamental group is the free group of rank 2. Hence by 4.3.1(b)

$$H = \pi_1(\tilde{\mathscr{F}}) = \text{free group of rank 2}. \qquad \square$$

This result was first proved algebraically, by Magnus 1931.

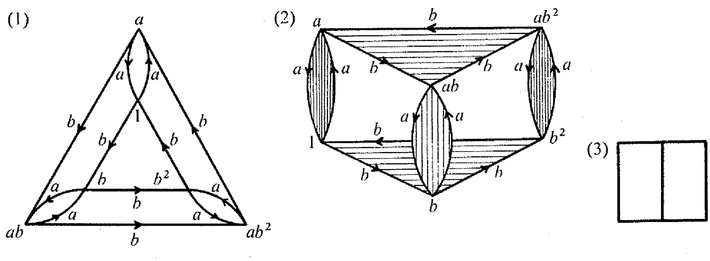


Figure 177

Exercise 4.3.6.1. Show that a subgroup H of index i in a group

$$G = \langle e_1, \ldots, e_m; r_1, \ldots, r_n \rangle$$

has a presentation with $im - i + 1$ generators and in relations.

Exercise 4.3.6.2. Define a suitable notion of rigid motion for surface complexes and show that if $\tilde{\mathscr{F}}$ realizes a normal subgroup H of G then the covering motion group is G/H (cf. 2.2.7). What is the 1-skeleton of $\tilde{\mathscr{F}}$ in this case?

Exercise 4.3.6.3. The covering $\tilde{\mathscr{F}}$ of $\mathscr{F}$ corresponding to the subgroup $\{1\}$ of $\pi_1(\mathscr{F})$ is called the *universal cover* of $\mathscr{F}$. Show that it covers any other cover of $\mathscr{F}$.

4.3.7 Subgroups of Surface Groups

We say that G is a *surface group* if $G = \pi_1(\mathscr{F})$, where $\mathscr{F}$ is a closed surface. Thus the surface groups are those whose presentations are given in 4.2.1 and the free groups on 1, 2, 3, ..., or a countable infinity of generators (by exercise 4.2.2.1). Then we have the result:

Every subgroup of a surface group is itself a surface group.

If $G = \pi_1(\mathscr{F})$ and $\mathscr{F}$ is a finite surface, then $\mathscr{F}$ is in fact the surface complex realization of the standard presentation of G (4.2.1). But any covering of a closed surface is itself a closed surface (4.3.4); in particular the $\tilde{\mathscr{F}}$ constructed in 4.3.6 to realize an arbitrary subgroup of G is a closed surface.

If $\mathscr{F}$ is not finite, then $G = \pi_1(\mathscr{F})$ is a free group of at most countable rank by 4.2.2. Then so is any subgroup H of G by the proof of the Nielsen–Schreier theorem (2.2.4), which means H is also a surface group. □

More exhaustive results detailing the possible subgroups of individual surface groups, and their relations with genus and orientability are easily derived by simple arguments about covering surfaces (for example, in Reidemeister 1928). The remarkable fact is that this geometrically transparent theorem, known to Fricke and Klein 1897, was not given a purely algebraic proof until 1971 (Hoare, Karrass, and Solitar 1971), and then only by means of the Reidemeister–Schreier process which, as the next section will show, is merely the covering complex construction stripped of its geometric garb.

4.3.8 The Reidemeister–Schreier Process

Exercise 4.3.6.1 shows that a subgroup H of finite index in a finitely presented group G is also finitely presented. If H is "given" in the sense that the corresponding covering is obtainable, then generators and relations for H can be computed by the method of 4.3.1. Like its special case, the Schreier method for finding free generators for a subgroup of a free group (2.2.6), the method can be given a purely algebraic formulation.

We realize G as $\pi_1(\mathscr{F})$, where $\mathscr{F}$ is a surface complex with a single vertex P and H as $\pi_1(\tilde{\mathscr{F}})$, where $\tilde{\mathscr{F}}$ is a covering of $\mathscr{F}$, with basepoint $\tilde{P}^{(0)}$ covering P. We can use the same letter p to denote a path in $\tilde{\mathscr{F}}$ from $\tilde{P}^{(0)}$ and its projection in $\mathscr{F}$, since the covering path is uniquely determined by p and $\tilde{P}^{(0)}$. Finally, we let $\bar{p}$ denote the unique reduced path from $\tilde{P}^{(0)}$ to the final point of p in some fixed spanning tree $\tilde{\mathscr{T}}$ of $\tilde{\mathscr{F}}$, and let w_j be the path so determined from $\tilde{P}^{(0)}$ to a given vertex $\tilde{P}^{(j)}$.

Then for each edge from $\tilde{P}^{(j)}$ labelled e_i we have a Schreier generator $w_j e_i (\overline{w_j e_i})^{-1}$, and the rule for expressing a closed path

$$p = e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_n}^{\varepsilon_n}$$

in terms of Schreier generators is to replace each $e_{i_l}^{\varepsilon_l}$ by the corresponding Schreier generator, namely

$$\overline{e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_{l-1}}^{\varepsilon_{l-1}}}\, e_{i_l}^{\varepsilon_l} (\overline{e_{i_1}^{\varepsilon_1} e_{i_2}^{\varepsilon_2} \cdots e_{i_{l-1}}^{\varepsilon_{l-1}} e_{i_l}^{\varepsilon_l}})^{-1}.$$

In particular, the relations of $\pi_1(\tilde{\mathscr{F}})$ result from writing the boundary $\tilde{b}_k$ of each face of $\tilde{\mathscr{F}}$ in this way. But the resulting path is freely equivalent to $w_j r_k w_j^{-1}$, where r_k is the boundary path in $\mathscr{F}$ (that is, the relator of $\pi_1(\mathscr{F})$) which lifts to $\tilde{b}_k$, and w_j is the approach path to the initial vertex of $\tilde{b}_k$. Thus we obtain the relations of $\pi_1(\tilde{\mathscr{F}})$ from the expressions

$$w_j r_k w_j^{-1}$$

as r_k runs through the relators of $\pi_1(\mathscr{F})$ and w_j runs through the reduced paths from $\tilde{P}^{(0)}$ in $\tilde{\mathscr{T}}$, by rewriting them in terms of the Schreier generators.

Recalling that the w_j can be interpreted as Schreier coset representatives (2.2.6), and $^{-}$ as the function which sends an element of G to its coset representative, the above process for obtaining generators and relations of H becomes purely algebraic. □

The first process for computing presentations of subgroups was given by Reidemeister 1927, where it was used to compute invariants which distinguish certain knots. His process was essentially that described above except that the coset representatives were more arbitrary. Schreier's special coset representatives appeared in Schreier 1927, where the spanning tree interpretation was also given. The following year, as was pointed out in 4.3.6, Reidemeister became aware of the full parallel between covering spaces and subgroups, and the corresponding interpretation of the Reidemeister–Schreier process. It was then apparent that what had first appeared to be a contribution of group theory to topology was equally a contribution of topology to group theory. (Ironically, when Reidemeister wrote his 1927 paper he was unaware that the same results on knots had already been obtained by Alexander in 1920, by direct consideration of covering spaces. See also 5.3.4 and 7.2.)

The method is clearly effective when H is given as the kernel of a homomorphism of G onto a finite group, because if one knows the images of the generators of G the coset of any element can be computed immediately, and coset representatives selected. This in fact was Reidemeister's assumption. The more challenging case where H is known to be of finite index, but is given by generators only, can be handled by the Todd–Coxeter coset enumeration algorithm (0.5.9).

4.3.9 The Kurosh Subgroup Theorem

If the reader reviews the brief introduction to free products in 3.4.5 it will be evident that subgroups H_i of groups G_i yield a subgroup $*_i H_i$ of the free product $*_i G_i$. We wish to know to what extent the converse holds. The commutator subgroup of the modular group (4.3.6) shows that free subgroups can arise even though none of the factors G_i in the free product contain an infinite cyclic group. It is also evident that a subgroup H_i of G_i can appear in $*_i G_i$ conjugated by an arbitrary element g of $*_i G_i$, that is, as gH_ig^{-1}. However, this is as far as it goes, because the Kurosh theorem states:

If $G = *_i G_i$ and H is a subgroup of G, then H is the free product of a free group (possibly trivial) with a free product of conjugates of subgroups of the G_i.

Let $\mathscr{A}_i$ be a surface complex with a single vertex P_i which realizes G_i. Then $*_i G_i$ is realized by the complex $\mathscr{F}$ obtained by joining each P_i to a new vertex P by an edge e_i (Figure 178). Let $\tilde{\mathscr{F}}$ be a covering of $\mathscr{F}$ which realizes the subgroup H. $\tilde{\mathscr{F}}$ consists of disjoint pieces $\tilde{\mathscr{A}}_i^{(j)}$ which cover the $\mathscr{A}_i$, together with edges $\tilde{e}_i$ covering the e_i. We take some $\tilde{P}^{(0)}$ covering P as the base point for $\pi_1(\tilde{\mathscr{F}})$.

We construct a spanning tree $\mathscr{T}$ for $\tilde{\mathscr{F}}$ by first taking a spanning tree $\mathscr{T}_i^{(j)}$ in each $\tilde{\mathscr{A}}_i^{(j)}$, then adding just enough of the edges $\tilde{e}_i$ to connect these

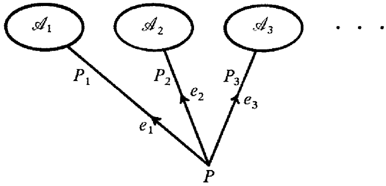


Figure 178

trees into one. The point in $\tilde{\mathscr{A}}_i^{(j)}$ which is then connected to $\tilde{P}^{(0)}$ by the shortest reduced path in $\mathscr{T}$ will be called $\tilde{P}_i^{(j)}$. Note that the reduced path from $\tilde{P}^{(0)}$ to any other vertex in $\tilde{\mathscr{A}}_i^{(j)}$ passes through $\tilde{P}_i^{(j)}$, otherwise $\mathscr{T}$ would not be a tree.

The Schreier generators of $\pi_1(\tilde{\mathscr{F}})$ divide into sets $A_i^{(j)}$ corresponding to edges in $\tilde{\mathscr{A}}_i^{(j)}$ and a set E corresponding to the edges $\tilde{e}_i$. Since none of the generators in E corresponds to an edge in the boundary of a face, they generate a free group.

The edges corresponding to generators in $A_i^{(j)}$ bound faces in $\tilde{\mathscr{A}}_i^{(j)}$ only and the relations involving these generators are therefore

$$wr_k w^{-1} = 1,$$

where w is an approach path to a face and r_k the boundary. We can factor w into $w_i^{(j)}w'$, where $w_i^{(j)}$ is the reduced path in $\mathscr{T}$ from $\tilde{P}^{(0)}$ to $\tilde{P}_i^{(j)}$ and w' is the approach path in $\tilde{\mathscr{A}}_i^{(j)}$ itself. The generators in $A_i^{(j)}$ can similarly be written as Schreier generators within $\tilde{\mathscr{A}}_i^{(j)}$, conjugated by $w_i^{(j)}$.

Thus if we take out the conjugating factor $w_i^{(j)}$ we have precisely the generators and relations derived from the covering $\tilde{\mathscr{A}}_i^{(j)}$ of $\mathscr{A}_i$, in other words, a subgroup of G_i. The subgroup actually determined by the generators in $A_i^{(j)}$ is therefore conjugate to a subgroup of G_i, with $w_i^{(j)}$ as the conjugating factor.

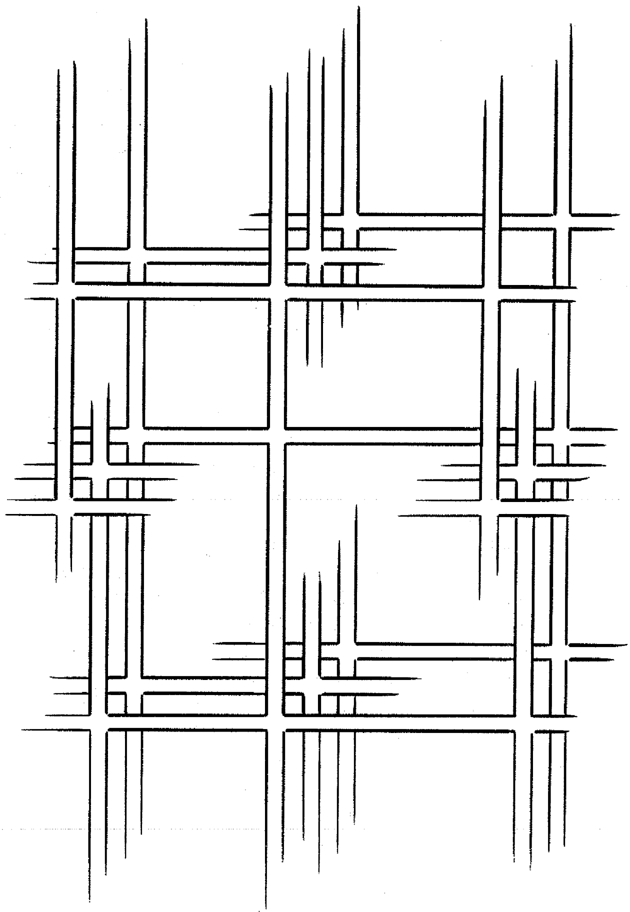
Since the generating sets E and $A_i^{(j)}$ are disjoint and satisfy only relations among themselves, the group they generate collectively is equal to the free product of the groups they generate individually, by 3.4.5. □

The first proof of this theorem, by Kurosh 1934, used a cancellation argument. The proof by covering a surface complex is due to Baer and Levi 1936.

Exercise 4.3.9.1. Use a suitable form of the Axiom of choice to justify the construction of the spanning tree $\mathscr{T}$ of $\tilde{\mathscr{F}}$ by first taking a spanning tree $\mathscr{T}_i^{(j)}$ in each $\tilde{\mathscr{A}}_i^{(j)}$, then adding just enough of the edges $\tilde{e}_i$ to connect these trees into one.

CHAPTER 5

Homology Theory and Abelianization



5.1 Homology Theory

5.1.1 Introduction

With hindsight, one can say that homology theory began with the Descartes–Euler polyhedron formula (1.3.8). It took a further step with Riemann's definition of the connectivity of a surface, and the generalization to higher-dimensional connectivities by Betti 1871. All these results have to do with the computation of numerical invariants of a manifold by means of decomposition into "cells"; the computations involve only the numbers of cells and the incidence relations between them, and it is shown that certain numbers are independent of the particular cellular subdivision chosen.

However, these results remained isolated until they were forged into a *theory of homology* by Poincaré 1895. Poincaré felt that many branches of mathematics were clamouring for such a theory, but his immediate objectives were to generalize a duality relation observed by Betti, and to give a completely general version of the Euler formula. The *Betti numbers*, as they have been called since Poincaré introduced the term, generalize the notion of connectivity number (genus) for an orientable surface. If n is the maximum number of closed cuts which can be made in a surface without separating it, then such a maximal system of curves $a_1, \ldots, a_n$ constitutes a *basis* in the sense that for any other curve p some "sum" of $a_1, \ldots, a_n$ bounds a piece of surface in combination with p. The latter property of n, made precise, is meaningful for manifolds $\mathscr{M}$ of possibly higher dimension than 2, and it serves to define the *one-dimensional Betti number* B_1. One immediately generalizes to the k-dimensional Betti number B_k when curves are replaced by suitable k-dimensional submanifolds of $\mathscr{M}$. Betti had observed that $B_1 = B_2$ when the dimension m of $\mathscr{M}$ was 3.

The appropriate notions of "sum" and "boundary," and the correct choice of k-dimensional manifolds admissible as basis elements, were found only after considerable trial and error. "Appropriate" initially meant satisfying the relation $B_k = B_{m-k}$, since this was the relation Poincaré tried to prove in his 1895 paper. Heegaard 1898 showed this work to be in error by constructing a counterexample. Poincaré then changed the definition and proved the theorem again in Poincaré 1899, inventing the tool of simplicial decomposition for the purpose. He also made a thorough analysis of his error, uncovering the important concept of *torsion* in Poincaré 1900, and exposing the breakdown of his earlier proof as failure to observe torsion.

Torsion is present when an element a does not form a boundary taken once, but does when taken more than once. An example is the curve a in the projective plane $\mathscr{P}$ which generates $\pi_1(\mathscr{P})$. Then a^2 is the boundary of a disc, though a itself does not separate $\mathscr{P}$. Poincaré justified the term "torsion" by showing that $(m-1)$-dimensional torsion is present only in an m-manifold which is nonorientable, and hence twisted onto itself in some sense.

In his first topology paper, Poincaré 1892 showed that the Betti numbers alone did not determine a manifold up to homeomorphism. By 1900 he was hoping that torsion numbers would supply the missing information, and his paper of that year contains a decomposition of the homology information in each dimension k into the Betti number B_k and a finite set of numbers called *k-dimensional torsion coefficients.* Since Noether 1926 it has been customary to encode this information in an abelian group H_k called the *k-dimensional homology group,* and Poincaré's construction can in fact be seen as the decomposition of a finitely generated abelian group into cyclic factors (see the structure theorem 5.2). The word "torsion," which appears so inexplicably in most algebra texts, entered the theory of abelian groups as a result of the derivation of the one-dimensional torsion coefficients by abelianization of the fundamental group in Tietze 1908 (see 5.1.3. and 5.3).

The insufficiency of homology theory to solve the main problems of topology became evident when Poincaré 1904, in the climax to his brilliant series of papers on topology, showed that the Betti and torsion numbers do not suffice to determine even the 3-sphere. He then conjectured that the 3-sphere is characterized, among finite 3-manifolds, by having a trivial fundamental group. This question, now known as the *Poincaré conjecture,* is still open. It must be admitted that the fundamental group is also an inadequate invariant (even for 3-manifolds, as was proved by Alexander 1919a), however, it is far more discriminating than H_1, and for manifolds of dimension ≤ 3 it contains essentially all the information available from homology.

This fact is one reason for the short shrift given to homology theory in this book. The other reason is that, as history shows, homology theory is loaded with subtleties, and an inordinate amount of preparation is required for correct definitions and the desired theorems. This preparation is largely wasted in the low dimensions which are our main concern, since most of the results can be derived rigorously from the fundamental group with far less preparation. (We have the classification of surfaces particularly in mind, see 5.3.3.)

For further information on the history of homology theory the reader is referred to Bollinger 1972. Among the texts which deal with homology theory, a reasonably concrete one is Cairns 1961, and Giblin 1977 may be recommended to readers who wish to encounter homology theory in the familiar context of surfaces.

5.1.2 Foundational Questions

One of the difficulties of homology theory above dimension 1 lies in its intimate relationship with foundational questions, namely, the very nature of *dimension, boundary,* and *separation.* We have already noted the difficulty involved in proving that a simple closed curve separates the plane into two regions (the Jordan curve theorem). Another theorem we omit with regret is the result that nonorientable closed surfaces do not embed in R^3.

It is not difficult to see that a surface which bounds a polyhedral subregion $\mathscr{R}$ of R^3 receives a coherent orientation from one of $\mathscr{R}$, and $\mathscr{R}$ is certainly orientable as a triangulated subset of R^3. Thus a polyhedral surface which separates R^3 must be orientable. But proving that *any* closed surface in R^3 separates it is a generalization of the Jordan curve theorem (due to Brouwer 1912b), difficult to prove even in the polyhedral case.

Such results are best proved in a comprehensive treatment which deals with foundations and homology theory simultaneously (for example, the book of Cairns 1961). The separation and nonembedding theorems then fall out of a result known as the Alexander duality theorem (Alexander 1923a), while the topological invariance of dimension (Brouwer 1911) and boundary also occur naturally in the development.

5.1.3 The First Homology Group

We shall define the first homology group of a complex $\mathscr{C}$, $H_1(\mathscr{C})$, as the abelianization of $\pi_1(\mathscr{C})$, that is, the result of adding relations $a_i a_j = a_j a_i$ for all generators a_i, a_j of $\pi_1(\mathscr{C})$. We postpone until 5.3 the proof that the abelianization is independent of the presentation of $\pi_1(\mathscr{C})$. In the normal development of homology theory

$$H_1(\mathscr{C}) = \text{abelianization of } \pi_1(\mathscr{C}) \tag{*}$$

is a theorem (first proved by Poincaré 1895) rather than a definition, but since we are using the fundamental group as our foundation, we shall merely sketch the intuitive connection between homology and homotopy to show that the definition is reasonable. With a little refinement, the argument which follows will serve as a proof of (*) when $H_1(\mathscr{C})$ is defined independently.

A 1-*chain* c is a sum $e_{i_1} + e_{i_2} + \cdots e_{i_m}$ of oriented edges in $\mathscr{C}$, where $+$ is a purely formal commutative operation. Since we are now using additive notation the two orientations of e_i will be denoted $+e_i$ and $-e_i$. The boundary $\partial(+e_i)$ of the oriented edge $+e_i$ from P_j to P_k is the formal sum $P_k - P_j$, $\partial(-e_i) = -\partial(+e_i)$, and

$$\partial(e_{i_1} + e_{i_2} + \cdots + e_{i_m}) = \partial(e_{i_1}) + \partial(e_{i_2}) + \cdots + \partial(e_{i_m}).$$

A 1-chain c is called *closed*, or a 1-cycle, if $\partial c = 0$. Thus $\partial c = 0$ if c is the sum of oriented edges in a closed path, and it is not hard to see in general that a 1-cycle c can be decomposed into

$$c_1 + c_2 + \cdots + c_l,$$

where each c_i is the sum of edges in a closed path p_i. The p_i do not necessarily all emanate from the same vertex P, however we can replace each p_i by $w_i p_i w_i^{-1}$, where w_i is an approach path from P to p_i, without changing the formal sum of edges, so each 1-cycle c does in fact correspond to a closed path p. Because the formal $+$ is commutative, all paths p' which result from p by allowing generators of $\pi_1(\mathscr{C})$ to commute yield the same 1-cycle c.

A 1-cycle c is called *null-homologous* or *bounding* if there is a 2-chain Γ whose boundary is c. A 2-chain Γ is a formal sum $\Delta_{j_1} + \Delta_{j_2} + \cdots + \Delta_{j_n}$ of faces, the boundary $\partial\Delta$ of a Δ is the formal sum of oriented edges in its boundary path, and

$$\partial(\Delta_{j_1} + \Delta_{j_2} + \cdots + \Delta_{j_n}) = \partial\Delta_{j_1} + \partial\Delta_{j_2} + \cdots + \partial\Delta_{j_n}.$$

Thus if c is null homologous it can be decomposed into

$$c_1 + c_2 + \cdots + c_n,$$

where each $c_i = \partial\Delta_{j_i}$, so the product of the corresponding closed paths $w_i p_i w_i^{-1}$ based at P is null-homotopic, since each p_i is the boundary path of a face. It follows that a closed path p based at P corresponds to a null-homologous 1-cycle just in case it can be converted to a null-homotopic path by allowing the generators of $\pi_1(\mathscr{C})$ to commute.

The first homology group is the quotient of the free abelian group of 1-cycles by the subgroup of null-homologous 1-cycles. From what we have just said, the same group is obtained from the free group on the generators of $\pi_1(\mathscr{C})$ by allowing these generators to commute and equating null-homotopic elements to the identity—that is, by abelianizing the fundamental group.

5.1.4 Geometric Interpretation of Null-homologous Paths

If p is a null-homologous path in a complex $\mathscr{C}$, then p is the boundary of a singular perforated orientable surface in $\mathscr{C}$. That is, there is an orientable surface $\mathscr{F}$, with a single perforation bounded by a curve c, and a continuous map $f\colon \mathscr{F} \to \mathscr{C}$ such that $f(c) = p$.

To avoid cumbersome notation we suppress the distinction between a path p and its homotopy class $[p]$. If p is null-homologous then by definition p is in the commutator subgroup of $\pi_1(\mathscr{C})$. This is the normal subgroup generated by the commutators $aba^{-1}b^{-1}$, where $a, b \in \pi_1(\mathscr{C})$, and hence by 0.5.4 p is expressible as

$$\prod_{i=1}^{n} w_i a_i b_i a_i^{-1} b_i^{-1} w_i^{-1}.$$

Then the closed path $p^{-1} \prod w_i a_i b_i a_i^{-1} b_i^{-1} w_i^{-1}$ is null-homotopic and hence the boundary of a singular disc in $\mathscr{C}$ (by Exercise 3.1.5.1.).

More precisely, there is a disc $\mathscr{D}$ with boundary divided into successive segments labelled

$$p^{-1}, w_1, a_1, b_1, a_1^{-1}, b_1^{-1}, w_1^{-1}, \ldots, w_n, a_n, b_n, a_n^{-1}, b_n^{-1}, w_n^{-1}$$

and a continuous map $h: \mathscr{D} \to \mathscr{C}$ such that the like-labelled segments become identified by h with observation of orientation. (Of course h may have to make more identifications than these since, for example, different a_i's and b_i's may actually stand for the same curve in $\mathscr{C}$.)

But there is already a continuous map g of $\mathscr{D}$ onto a perforated orientable surface $\mathscr{F}$ whose boundary is the image of the segment labelled p, namely the map which identifies like-labelled segments with observation of orientation. For we have seen in 1.3 that a schema in which each edge label occurs twice, with opposite exponents, is always an orientable surface. Hence it will suffice to factor $h: \mathscr{D} \to \mathscr{C}$ into $g: \mathscr{D} \to \mathscr{F}$ and a continuous map $f: \mathscr{F} \to \mathscr{C}$.

If x is a point on $\mathscr{F}$ we let $g^{-1}(x)$ denote the set of its preimage points in $\mathscr{D}$. It is then meaningful to define

$$f(x) = h(g^{-1}(x))$$

because h identifies all points identified by g. A sufficiently small δ-disc neighbourhood of x is the g-image of a δ-disc neighbourhood of $g^{-1}(x)$ when $g^{-1}(x)$ is a point in the interior of $\mathscr{D}$ (necessarily single), and the g-image of a (finite) union of δ-semidisc neighbourhoods when $g^{-1}(x)$ consists of points on the boundary of $\mathscr{D}$. Continuity of h implies that h maps these neighbourhoods into a given ε-disc neighbourhood of $h(g^{-1}(x))$ in $\mathscr{C}$ for δ sufficiently small, which means that $f(x) = h(g^{-1}(x))$ defines a continuous function f.

(Readers who have not done Exercise 3.1.5.1 should note that it can be done by similarly factoring the continuous map h: rectangle $\to \mathscr{C}$ into g: rectangle $\to$ disc, where g identifies three sides of the rectangle to a point, and f: disc $\to \mathscr{C}$, where $f(x) = h(g^{-1}(x))$.) □

This theorem appeared in Seifert and Threlfall 1934. Their proof makes use of the fact that any element in the commutator subgroup is a product of commutators (see Exercise 5.1.4.1).

An example of a null-homologous path which is not null-homotopic is the boundary path p of the "handle" $\mathscr{H}$ (Figure 179). In terms of the generators of $\pi_1(\mathscr{H})$,

$$p = aba^{-1}b^{-1}$$

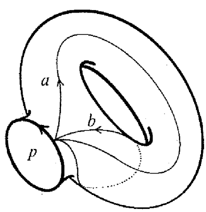


Figure 179

so p is certainly null-homologous, however, this element does not equal 1 in $\pi_1(\mathscr{H})$, since $\pi_1(\mathscr{H})$ is the free group generated by a, b.

EXERCISE 5.1.4.1. Show that $waba^{-1}b^{-1}w^{-1}$ is freely equivalent to a product of commutators, so that the commutator subgroup of a group G is in fact generated by the commutators of G.

Give an example to show that the commutators of generators of G do not in general suffice to generate its commutator subgroup.

EXERCISE 5.1.4.2. If p is the boundary of a singular perforated orientable surface in $\mathscr{C}$ prove that $[p]$ is in the commutator subgroup of $\pi_1(\mathscr{C})$.

5.2 The Structure Theorem for Finitely Generated Abelian Groups

5.2.1 Introduction

The fundamental theorem for finite abelian groups appears in Kronecker 1870. In this paper, Kronecker gives what we would recognize as the abstract definition of a finite abelian group—a finite set closed under a commutative, associative binary operation f, with the property that $a' \neq a''$ implies $f(a, a') \neq f(a, a'')$—then proves that such a group is a direct product of cyclic groups. Kronecker's proof is so brief and lucid we shall reproduce it almost verbatim below.

A different proof, using matrices, was discovered by Poincaré 1900. Poincaré's method is actually intended to compute the Betti number and torsion coefficients (of given dimension) of a complex, but this is tantamount to decomposing a finitely generated abelian group into certain cyclic factors, the number of infinite cyclic factors being the Betti number, and the orders of the finite factors being the torsion coefficients. His result is therefore a generalization of Kronecker's—what we now know as the structure theorem for finitely generated abelian groups—however, we shall see how Kronecker's proof can be augmented to deal with elements of infinite order. (This seems to have first been done by Noether 1926.)

Kronecker's proof begins with the following remarks.

(1) The exponents k of all powers a^k equal to 1 for a fixed element a are integer multiples of some positive integer n called the *period* of a.
(2) If n is a period, so is any divisor of n.
(3) If a', a'' have periods n', n'' which are relatively prime, then $a'a''$ has period $n'n''$.
(4) If n_1 is the lowest common multiple of the periods of elements in the group, then there is in fact an element of period n_1. For if

$$n_1 = p^\alpha q^\beta r^\gamma \cdots$$

is the prime factorization of n_1, there must be periods n containing $p^\alpha, q^\beta, r^\gamma, \ldots$ as factors, and hence by (2), elements $a', a'', a''', \ldots$ of periods $p^\alpha, q^\beta, r^\gamma, \ldots$ respectively. Then by (3) the element $a'a''a''' \cdots$ has period $p^\alpha q^\beta r^\gamma \cdots = n_1$.

It will be seen from the proof which follows that Kronecker is implicitly using coset decompositions and coset representatives, however, the directness of his argument is more obvious if these terms are not mentioned.

5.2.2 Kronecker's Theorem

If A is a finite abelian group, then $A = A_1 \times A_2 \times \cdots \times A_s$, where $A_1, A_2, \ldots$ are cyclic groups of orders $n_1, n_2, \ldots$ and each n_{i+1} is a divisor of n_i.

Let n_1 denote, as in (4), the maximal period among elements of A. Then n_1 is a multiple of the period of each element a, and we have

$$a^{n_1} = 1$$

for an arbitrary $a \in A$.

If a_1 is an element with period n_1, we shall call elements a', a'' *equivalent relative* to a_1 if

$$a'a_1^k = a'' \quad \text{for some } k.$$

This is indeed an equivalence relation, and the equivalence classes form a finite abelian group under the obvious multiplication (it is, of course, the quotient of A by the cyclic subgroup generated by a_1). The properties (1)–(4) relativize to corresponding properties of equivalence. In particular, there is an equivalence class of maximal period n_2, which means that for any representative a^* of the class, $(a^*)^{n_2}$ is the least of its powers equivalent to 1. Since $(a^*)^{n_1}$ equals 1 and is *a fortiori* equivalent to it, the relativized version of (1) says that n_2 is a divisor of n_1.

Now if $(a^*)^{n_2} = a_1^k$ and one raises both sides to the power n_1/n_2 then

$$1 = (a^*)^{n_1} = a_1^{kn_1/n_2}$$

so when k/n_2 is set equal to m we have

$$a_1^{mn_1} = 1$$

from which it follows, since n_1 is the period of a_1, that m is an integer.

The equation

$$a_2 a_1^m = a^*$$

then defines an element a_2 equivalent to a^* whose n_2th power is not merely equivalent to 1, but equal to it.

We now call elements a', a'' *equivalent relative to* a_1, a_2 if

$$a'a_1^h a_2^k = a'' \quad \text{for some } h, k$$

and similarly obtain a group of equivalence classes whose maximal period, n_3, divides n_2, and a representative a_3 of the class of maximal period such that $a_3^{n_3} = 1$.

The procedure terminates when we have a set of elements $a_1, a_2, \ldots, a_s$ such that any a is equivalent to 1 relative to $a_1, a_2, \ldots, a_s$, that is, when any a is expressible as

$$a = a_1^{h_1} a_2^{h_2} \cdots a_s^{h_s} \qquad (0 \leq h_i < n_i).$$

It also follows that the expression is unique, for the equivalence classes relative to $a_1, \ldots, a_{s-1}$ must constitute a cyclic group with a_s as a representative generator. An element a is therefore uniquely determined by the integers $h_1, \ldots, h_{s-1}$, which determine it relative to an equivalence class representative, and the integer h_s which determines the equivalence class representative itself, $a_s^{h_s}$.

Thus A is the direct product $A_1 \times A_2 \times \cdots \times A_s$, where A_i is the cyclic group generated by a_i, and the order n_i of A_i is such that n_{i+1} divides n_i.

□

Note that the orders $n_1, n_2, \ldots, n_s$ of factors in a representation of A as a direct product of cyclic groups are uniquely determined by the requirement that n_{i+1} divides n_i. For in a group $A_1 \times A_2 \times \cdots \times A_s$ in which the orders n_i of A_i have this property, n_1 is indeed the maximal period of an element, n_2 the maximal period in the quotient modulo the subgroup generated by an element of order n_1, etc.

One can actually factor further into cyclic subgroups whose orders are prime powers—for example, the cyclic group of order 6 is the direct product of cyclic groups of orders 2 and 3—however, the numbers $n_1, n_2, \ldots, n_s$ are those most suitable to describe the "torsion" in the group. If A is the first homology group of a complex, n_1 represents the maximum number of times a nonbounding curve has to be traversed before it becomes bounding, n_2 is the maximum when curves are considered "relative to a curve of period n_1," and so on.

5.2.3 A Factorization Theorem

If A is an abelian group and B a subgroup such that A/B is free abelian, then

$$A = B \times \frac{A}{B}$$

(Note: in what follows we understand "free generators," "nontrivial relation," and so on in the context of abelian groups. For example, $a_1 a_2 a_1^{-1} a_2^{-1} = 1$ is now a *trivial* relation.)

Let $x_1, x_2, \ldots$ be free generators of A/B and for each i choose a $c_i \in A$ such that $\phi(c_i) = x_i$, where $\phi: A \to A/B$ is the canonical homomorphism. Then the c_i freely generate a subgroup C of A, isomorphic to A/B, since any nontrivial relation between the c_i's would yield the corresponding relation between the x_i's under the map ϕ.

It follows that any $a \in A$ has a unique factorization $a = bc$, where $b \in B$, $c \in C$. For c must satisfy $\phi(c) = \phi(a)$, and there is exactly one such c by the construction and freeness of C; and b is then uniquely determined as ac^{-1}. The latter is indeed an element of B, since $\phi(ac^{-1}) = \phi(a)(\phi(c))^{-1}$.

Now if $a_1 = b_1c_1$, $a_2 = b_2c_2$ we have $a_1a_2 = (b_1b_2)(c_1c_2)$, so multiplication takes place componentwise on the B and C factors. In other words, $A = B \times C$ or $B \times A/B$, since C is isomorphic to A/B. □

5.2.4 Free Abelian Groups of Finite Rank

If $\mathbf{Z}^n$ denotes the abelian group freely generated by $a_1, \ldots, a_n$, then any set of free generators for $\mathbf{Z}^n$ has n elements. Also, any subgroup of $\mathbf{Z}^n$ is free abelian, with $\leq n$ generators.

The typical element $a = a_1^{k_1}a_2^{k_2} \cdots a_n^{k_n}$ of $\mathbf{Z}^n$ can be represented by the integer vector $\boldsymbol{a} = (k_1, k_2, \ldots, k_n)$ and the product operation in $\mathbf{Z}^n$ then corresponds to vector sum.

By elementary linear algebra, a set of $>n$ such vectors is linearly dependent with rational, and hence in fact integer, coefficients (multiplying through by a common denominator). This means there is a nontrivial relation between any set of $>n$ members of $\mathbf{Z}^n$; so any set of free generators has $\leq n$ members. Conversely, if $b_1, \ldots, b_m$ generate $\mathbf{Z}^n$, then the elements $a_1, \ldots, a_n$ in particular must be products of them. In other words, the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are linear combinations of the $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$. Since $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are linearly independent, $m \geq n$ by the same argument. Hence $m = n$.

The number n is the number of factors in the decomposition of $\mathbf{Z}^n$ into the direct product of infinite cyclic groups, so we have now shown that this decomposition is unique.

To show the second part of the theorem, suppose that C is a subgroup of $\mathbf{Z}^n$. We observe that a subgroup of $\mathbf{Z}^1$ is certainly free, on $\leq$ one generator, and continue by induction on n as follows.

The projection $\pi: \mathbf{Z}^n \to \mathbf{Z}^1$ which sends each a_i to a_1 maps C onto a free subgroup C_1 of $\mathbf{Z}^1$. Then by the factorization theorem

$$C = B_1 \times C_1,$$

where B_1 is the kernel of the projection $C \to C_1$, that is, a subgroup of the free abelian group generated by $a_2, \ldots, a_n$. By induction we can assume that B_1 is free abelian on $\leq n - 1$ generators, that is, the direct product of $\leq n - 1$

infinite cyclic groups, and then C is the direct product of $\leq n$ infinite cyclic groups. □

EXERCISE 5.2.4.1. Prove that finitely generated abelian groups are finitely presented. (Of course, this result will follow from the structure theorem, but it is of interest to see what it really depends on.)

5.2.5 Torsion-free Abelian Groups

An abelian group A is called torsion-free if it has no elements of finite order. A finitely-generated torsion-free abelian group is free.

Let $a_1, \ldots, a_n$ be a maximal subset of the generators of A which generate freely. Then for each $i > n$ the generator a_i enters a nontrivial relation with $a_1, \ldots, a_n$, which we may assume to be

$$w(a_1, \ldots, a_n) = a_i^{k_i}.$$

Thus if B denotes the free abelian group generated by $a_1, \ldots, a_n$ we have $a_i^{k_i} \in B$ for each $i > n$. Let k be a common multiple of the k_i's and consider the homomorphism $\phi: A \to B$ which sends each a_i to a_i^k. Since no a_i has finite order the kernel is trivial and hence we have a monomorphism. The image subgroup of B is free by 5.2.4, so A itself is free. □

We mention in passing that an infinitely-generated torsion-free abelian group may not be free—an interesting example is the group D of rationals of the form $p/2^q$ (p, q integers) under addition. Exercise 5.2.5.1 develops some of the properties of this group, which actually occurs in topology (see Rolfsen 1976, p. 186).

For the theory of infinitely-generated abelian groups, which is quite well developed, see Fuchs 1960.

EXERCISE 5.2.5.1. (1) Show that any finite set of elements $p_1/2^{q_1}, \ldots, p_n/2^{q_n}$ of D with $q_1 \leq \cdots \leq q_n$ generate an infinite cyclic subgroup containing no element $< 1/2^{q_n}$. Deduce that D is not finitely generated and that any finite set of ≥ 2 elements satisfy a nontrivial relation.

(2) Show that D has a presentation

$$\langle a_1, a_2, a_3, \ldots; a_1 = a_2^2, a_2 = a_3^2, a_3 = a_4^2, \ldots \rangle.$$

(3) Show that every proper subgroup $\neq \{1\}$ and containing the element a_1 of D is infinite cyclic.

(4) Show that D/Z, the result of adding the relation $a_1 = 1$ to D, is an infinite group whose proper subgroups are all finite cyclic.

EXERCISE 5.2.5.2. Show that the positive rationals under multiplication constitute an infinitely generated free abelian group.

5.2.6 The Torsion Subgroup

Suppose A is any finitely-generated abelian group, and let T be the subgroup of elements of finite order. T is called the torsion subgroup. *Then T is a finite abelian group and $A = F \times T$ where F is a free abelian group.*

Let $\mathbf{Z}^k$ be the free abelian group on the generators of A and let B be the subgroup of $\mathbf{Z}^k$ which maps onto T under the canonical homomorphism $\phi: \mathbf{Z}^k \to A$. By 5.2.4, B is finitely generated, so the images of its generators give a finite set of generators for T. But a finitely generated abelian group in which every element has finite order is obviously finite, hence T is a finite abelian group.

Now consider the coset decomposition of A modulo T. If any coset is of order m this means $x^m \in T$ for any of its representatives x. But then x^m is of finite order, hence so is x and the coset in question can only be T itself. Thus A/T is torsion-free and hence free by 5.2.5.

It then follows by the factorization theorem that $A = F \times T$ where F is the free abelian group A/T. □

The proof of the structure theorem is now complete. We have decomposed the given finitely generated abelian group A into the direct product of a free abelian group A/T and a finite abelian group T. This decomposition is unique because in any abelian group $F \times T$, where F is free and T is finite the torsion subgroup is obviously T. The free abelian group A/T decomposes uniquely into the direct product of n infinite cyclic groups by 5.2.4, while T decomposes uniquely into cyclic groups of orders $n_1, \ldots, n_s$, where n_{i+1} divides n_i, by Kronecker's theorem and the remark following it in 5.2.2. A is therefore uniquely determined by the number n (Betti number) and the numbers $n_1, \ldots, n_s$ (torsion coefficients).

5.2.7 Computability of the Betti Number and Torsion Coefficients

The above proof of the structure theorem does not make clear how to actually compute the decomposition of a given finitely-generated abelian group A into cyclic factors. The proof using matrices is quite explicit in this respect (see for example Cairns 1961), however, we can also obtain an algorithm for computing the decomposition from its mere existence by the following cheap trick:

Given a presentation of A, systematically apply all possible Tietze transformations until an abelian presentation of the form

$$\langle a_1, \ldots, a_n, b_1, \ldots, b_s; b_1^{n_1}, \ldots, b_s^{n_s} \rangle,$$

where each n_{i+1} divides n_i, is obtained. Then n is the Betti number of A and $n_1, \ldots, n_s$ are its torsion coefficients. The structure theorem implies the existence of such a presentation, so we must be able to reach it in a finite sequence of Tietze transformations by the Tietze theorem (0.5.8). The uniqueness of $n, n_1, \ldots, n_s$ guarantees that the first presentation found to have the required form will give the correct numbers.

EXERCISE 5.2.7.1. Use the connected sum (Exercise 4.1.6.1) to construct a 3-dimensional manifold $\mathscr{M}$ with

$$\pi_1(\mathscr{M}) = \langle a_1, \ldots, a_n, b_1, \ldots, b_s; b_1^{n_1}, \ldots, b_s^{n_s} \rangle$$

and hence with arbitrary finitely generated first homology group. ($\pi_1(\mathscr{M})$ is *not* required to be abelian.)

5.3 Abelianization

5.3.1 Presentation Invariance

The group which results from

$$G = \langle a_1, \ldots, a_m; r_1, \ldots, r_n \rangle$$

by addition of the relations $a_i a_j = a_j a_i$*—the* abelianization *of* G*—is independent of the presentation of* G.

The resulting group will be the same if we add relations $gh = hg$ for *all* elements $g, h \in G$, since all elements commute if the generators do. But then 0.5.6 tells us we are factoring G by the normal subgroup $[G, G]$ generated by all commutators $ghg^{-1}h^{-1}$ in G (the commutator subgroup), and the latter definition is independent of the presentation. □

In future we shall use $G/[G, G]$ to denote the abelianization of G. The notation $[G, G]$ for the commutator subgroup is an extrapolation of the notation $[g, h]$ often used for the commutator $ghg^{-1}h^{-1}$.

EXERCISE 5.3.1.1. Use Tietze transformations to give an alternative proof that the result of adding relations $a_i a_j = a_j a_i$ is independent of the presentation. (Tietze 1908 actually introduced his transformations for this purpose, though his version of the proof was to show that Betti numbers and torsion coefficients were invariant.)

5.3.2 Rank of a Free Group

The rank of a finitely generated free group is independent of the choice of free generators.

If $G = \langle a_1, \ldots, a_m; — \rangle$, then $G/[G, G]$ is the free abelian group of rank m. But rank is presentation invariant for free abelian groups (5.2.4), hence it is invariant for G also. □

This is the proof of invariance of rank given by Schreier 1927.* Comparison with the proof of Nielsen 1921 (Exercises 2.2.4.1–2.2.4.3) shows the power of elementary linear algebra in this context. Abelianization erases just enough of the group's structure to simplify the problem, while preserving the information necessary for its solution. A similar example is given in the following exercise.

EXERCISE 5.3.2.1. Show that if G has a presentation with m generators and $<m$ relations, then G is infinite.

5.3.3 Surface Groups

If $G = \pi_1(\mathscr{F})$, where $\mathscr{F}$ is a closed finite surface, then we know from 4.2.1 that $\pi_1(\mathscr{F})$ has one of the following forms, according as $\mathscr{F}$ is orientable or nonorientable of genus $n \geq 1$.

(i) $G = \langle a_1, b_1, \ldots, a_n, b_n; a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1} \rangle$,
(ii) $G = \langle a_1, \ldots, a_n; a_1^2 a_2^2 \cdots a_n^2 \rangle$.

It follows that the respective abelianizations are

(i)′ $G/[G, G]$ = free abelian of rank $2n$,
(ii)′ $G/[G, G]$ = (free abelian of rank $n - 1$) × (cyclic of order 2),

For (i)′ we observe that $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_n b_n a_n^{-1} b_n^{-1} = 1$ is true in any abelian group, hence $G/[G, G]$ is free abelian on the $2n$ free generators $a_1, b_1, \ldots, a_n, b_n$.

For (ii)′ we first make the following Tietze transformations to obtain a more convenient presentation:

$$\begin{aligned} G/[G, G] &= \langle a_1, \ldots, a_n; a_1^2 \cdots a_n^2 = 1, a_i a_j = a_j a_i \rangle \\ &= \langle a_1, \ldots, a_n; (a_1 \cdots a_n)^2 = 1, a_i a_j = a_j a_i \rangle \\ &= \langle a_1, \ldots, a_n, a_{n+1}; a_{n+1}^2 = 1, a_1 \cdots a_n = a_{n+1}, a_i a_j = a_j a_i \rangle \\ &= \langle a_2, \ldots, a_n, a_{n+1}; a_{n+1}^2 = 1, a_i a_j = a_j a_i \rangle. \end{aligned}$$

Any element can now be represented uniquely in the form

$$a_2^{x_2} \cdots a_n^{x_n} a_{n+1}^{x_{n+1}}$$

These forms multiply componentwise, with the $a_2, \ldots, a_n$ components being free, while a_{n+1} is of order, 2, hence $G/[G, G]$ is the direct product of

* On the other hand, the rank of an infinitely-generated free group is invariant on purely set-theoretic grounds, since the cardinality of any set of free generators is equal to the cardinality of the set of all words. This implies that the number of edges omitted from a spanning tree $\mathscr{T}$ in a graph of infinite connectivity is independent of $\mathscr{T}$ (cf. 2.1.7, 2.1.8)—a result which does not seem to have a direct graph-theoretic proof.

the free abelian group of rank $n - 1$ generated by $a_2, \ldots, a_n$ with the cyclic group of order 2 generated by a_{n+1}. □

Thus the orientable surfaces are topologically distinguished from each other by first homology groups of different rank (Betti number), while the nonorientable surfaces are likewise distinguished from each other by different Betti numbers, and from the orientable surfaces by the presence of a torsion coefficient 2.

This completes the topological classification of surfaces which was begun in 1.3. We note that the one-dimensional homology invariants suffice to distinguish the different normal forms obtained in 1.3, and that the 2-sphere is characterized as the only closed finite surface with Betti number 0 and no torsion.

We also have a proof of the *Hauptvermutung* for finite triangulated 2-manifolds, since the classification theorem of 1.3.7 reduces these to normal forms by elementary subdivisions and amalgamations, and we now know that the normal forms are topologically distinct. Thus if two finite triangulated surfaces are homeomorphic, they have the same normal form, and hence a common simplicial refinement. It also follows that the Euler characteristic (1.3.8) and orientability character are topological invariants of these manifolds, since we already know that they are combinatorial invariants.

Exercise 5.3.3.1. What is the relationship between Betti number and Euler characteristic?

Exercise 5.3.3.2. Prove that the surface groups (i) and (ii) are not free.

5.3.4 Knot Groups

Since abelianization destroys some of the structure of the group (unless it is already abelian) we should not expect it to always be a source of information. Its most conspicuous failure occurs in the case of knot groups, where abelianization always collapses the group to an infinite cyclic group.

We know from the Wirtinger presentation (4.2.3, 4.2.4) that any knot group can be given by generators $a_1, \ldots, a_m$ and relations of the form either

$$a_j a_i = a_{i+1} a_j \qquad \text{for } i = 1, \ldots, m \text{ and indices reduced mod } m$$

or

$$a_i a_j = a_j a_{i+1}$$

Under abelianization, these relations simply say that

$$a_i = a_{i+1}$$

so all the generators become one, with no relations, and we have an infinite cyclic group. □

It was this difficulty which stimulated Reidemeister 1927 to find a procedure for finding presentations of subgroups (4.3.8) which, as he pointed out, marked the first step beyond abelianization as a means of extracting information from group presentations. When abelianization is applied to the subgroups obtained, the result is not always trivial and it can in fact be used to distinguish a large number of knots. Abelianization of a subgroup of course corresponds to the one-dimensional homology of a covering space, and Alexander's anticipation of Reidemeister's results was obtained by this direct route (see 7.2).

The groups of infinite knots such as the Fox–Artin wild arc (4.2.6) also collapse under abelianization. If $\mathscr{A}$ denotes the Fox–Artin arc the relations of $\pi_1(\mathsf{R}^3 - \mathscr{A})$ are

$$c_{n-1}c_nc_{n+1} = c_nc_{n+1}c_{n-1}c_n$$

so we simply get $c_n = 1$ by abelianization, and hence

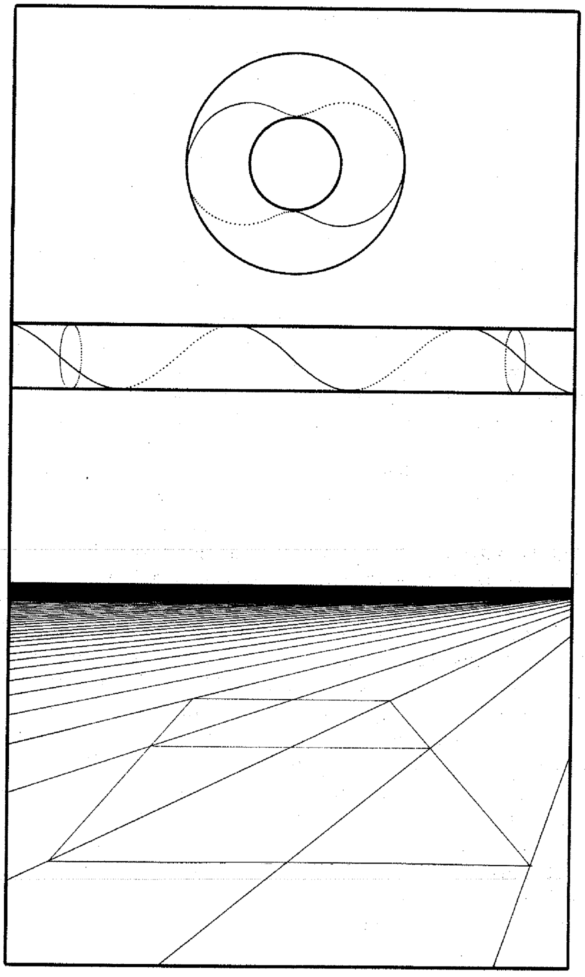
$$H_1(\mathsf{R}^3 - \mathscr{A}) = \{1\}.$$

It is actually a consequence of the Alexander duality theorem that the complement of a simple arc has trivial homology, so homology groups are not sensitive enough to detect wildness.

EXERCISE 5.3.4.1. If G is the group of a link of m curves, show that $G/[G, G]$ is free abelian of rank m.

CHAPTER 6

Curves on Surfaces



6.1 Dehn's Algorithm

6.1.1 Introduction

The fundamental problem in the topological classification of curves on surfaces is to decide whether a given closed curve contracts to a point. We shall call this the *contractibility problem.* Jordan 1866b recognized that the problem could be expressed in algebraic terms, but his work contained errors. He showed that each curve could be deformed into a product of certain canonical curves—essentially the generators of the fundamental group—and realized that the canonical curves satisfied certain relations. However, he seemed not to notice that he actually had a group (surprisingly, in view of the subsequent appearance of his pioneering treatise on group theory, Jordan 1870), and failed to get the right relations.

The problem was solved, at least from a geometric point of view, with the introduction of the universal covering surface in the 1880s. If a curve p on the surface $\mathscr{F}$ is lifted to a curve $\tilde{p}$ in the universal covering surface $\tilde{\mathscr{F}}$ then p contracts to a point on $\mathscr{F}$ if and only if $\tilde{p}$ is closed. The reason is simple: if p can be contracted to a point it can be contracted with its basepoint fixed; this contraction lifts to a contraction of $\tilde{p}$ to a point with both its endpoints fixed, which is possible only if the endpoints coincide. Conversely, if $\tilde{p}$ is closed, it contracts to a point, since $\tilde{\mathscr{F}}$ is the plane or sphere. By performing the contraction in one polygon of $\tilde{\mathscr{F}}$ at a time, it can be projected to a contraction of p to a point on $\mathscr{F}$.

If p is "given" in an effective way, say as a sequence of edges in the canonical polygon for $\mathscr{F}$, then one can effectively construct the lift $\tilde{p}$ of p by building a large enough portion of $\tilde{\mathscr{F}}$, and then see whether $\tilde{p}$ is closed. Bearing in mind that the polygons which tessellate $\tilde{\mathscr{F}}$ do not have to be congruent, this algorithm is perfectly concrete and combinatorial. It is admittedly not very convenient because of the very dense packing of polygons required even for genus 2 (see Figure 111). Nevertheless, around the turn of the century the algorithm was considered sufficiently obvious not to require more than a passing mention (for example, in Poincaré 1904 and Dehn 1910).

It was Dehn who first appreciated the algebraic significance of the problem and found a practical solution. He observed that the labelled net of polygons on the universal covering surface was in fact the Cayley diagram of $\pi_1(\mathscr{F})$, and that the contractibility problem was therefore the same as the *word problem* for $\pi_1(\mathscr{F})$. Furthermore, geometric properties entailed an algebraic process for solving the word problem without actual construction of the net. This process is now known as *Dehn's algorithm.* His first proof depended on the metric in the hyperbolic plane (Dehn 1912a), but he then saw how the algorithm could be justified by purely topological properties of the net (Dehn 1912b).

The topological argument of Dehn 1912b is actually designed to solve the more difficult *conjugacy problem* (deciding whether two elements are conjugate) and a far less intricate argument suffices for the word problem. We present this argument in 6.1.3, 6.1.4 after dealing with the trivially solvable cases of the word problem in 6.1.2.

6.1.2 Some Special Cases

To free ourselves from distractions in the proof of Dehn's algorithm, we first deal with the "small" surfaces to which it does not apply.

(1) The sphere $\mathbf{S}^2$. $\pi_1(\mathbf{S}^2) = \{1\}$, so every word $= 1$.
(2) The projective plane $\mathscr{P}$. $\pi_1(\mathscr{P}) = \langle a; a^2 \rangle$ so the word a^m is 1 just in case m is even.
(3) The torus $\mathscr{T}$. $\pi_1(\mathscr{T}) = \langle a, b; ab = ba \rangle$. It is immediate from the Cayley diagram (Figure 180) that each word has a normal form $a^m b^n$, and that $a^m b^n = 1$ just in case $m = n = 0$.
(4) The Klein bottle $\mathscr{B}$. $\pi_1(\mathscr{B}) = \langle a, b; abab^{-1} \rangle$. The Cayley diagram is Figure 181 (obtained by constructing the universal covering surface from copies of the rectangle in Figure 182 so that the neighbourhood of

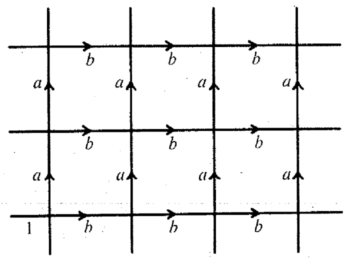


Figure 180

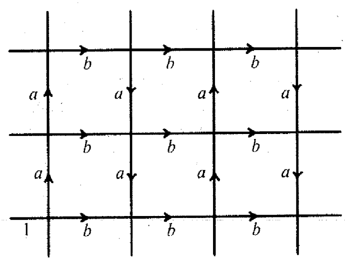


Figure 181

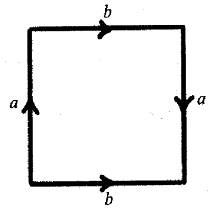


Figure 182

each vertex has a, b both incoming and outgoing.) Then the most convenient way to solve the word problem is to actually trace the path in the Cayley diagram, though an algebraic algorithm is not hard to find.

EXERCISE 6.1.2.1. Formulate an algebraic algorithm which solves the word problem in the Klein bottle group.

EXERCISE 6.1.2.2. Show that there are no nontrivial elements of finite order in the Klein bottle group.

6.1.3 The Subpath Property

We can now assume that our surface $\mathscr{F}$ has a canonical polygon with $2n$ sides, where $n \geq 3$. The Cayley diagram of $\pi_1(\mathscr{F})$ is therefore a planar net of $2n$-gons which meet $2n$ at each vertex (cf. the construction of the universal covering surface in 1.4.2).

Then if p is any closed path in the net, p contains either a spur or a subpath consisting of more than half the edges in the boundary of a polygon in succession.

We construct the net in the euclidean plane as in 1.4.2, using concentric circles of increasing radius $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \ldots$. Each circle will be subdivided into arcs called *circumferential edges* of the net, and all other edges in the net will be line segments between $\mathscr{C}_i$ and $\mathscr{C}_{i+1}$, called *radial edges*.

Thus $\mathscr{C}_1$ will be subdivided into $2n$ arcs to form the first canonical polygon. Each vertex P on $\mathscr{C}_1$ will emit $2n - 2$ radial edges to $\mathscr{C}_2$, so as to arrange that $2n$ polygons meet at P.

These radial edges divide $\mathscr{C}_2$ into $2n(2n - 2)$ arcs, each of which is further subdivided by new vertices so as to convert each region between $\mathscr{C}_1$ and $\mathscr{C}_2$ into a $2n$-gon. (A region with one vertex on $\mathscr{C}_1$ therefore requires $2n - 3$ new vertices, a region with two vertices on $\mathscr{C}_1$ requires $2n - 4$.)

Each vertex on $\mathscr{C}_2$ then emits radial edges to $\mathscr{C}_3$; $2n - 3$ if it is the endpoint of a radial edge from $\mathscr{C}_1$, $2n - 2$ otherwise, and so on. See Figure 183.

Now consider a closed edge path p in the net starting at a vertex on $\mathscr{C}_1$. Assume all spurs have been removed from p and let c be its outermost circumferential portion, on $\mathscr{C}_k$ say. This portion must then be a "turning

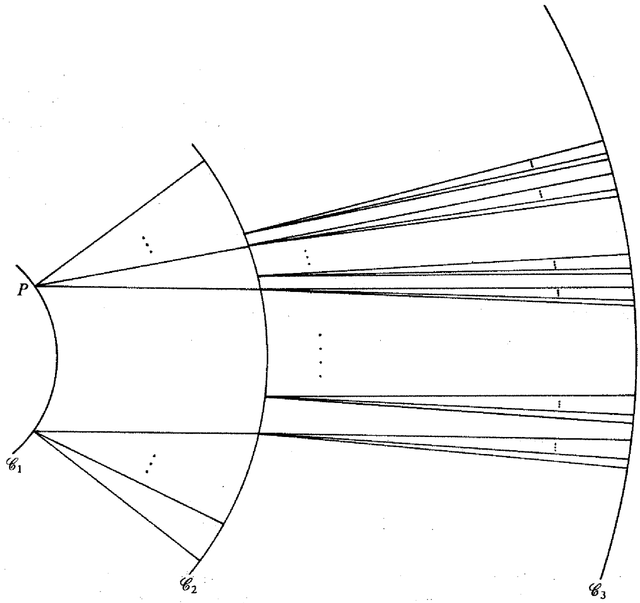


Figure 183

point" of p, where p reaches $\mathscr{C}_k$ on a radial edge e from $\mathscr{C}_{k-1}$, runs along $\mathscr{C}_k$ for some distance, then returns to $\mathscr{C}_k$ on another radial edge as in Figure 184 (leaving aside the trivial case $k = 1$). But then some initial segment of ec traverses the perimeter of some polygon $\mathscr{P}$ with the exception of at most two edges, since $\mathscr{P}$ can have at most two vertices on $\mathscr{C}_{k-1}$. Since we assume that $\mathscr{P}$ has at least six edges, p has therefore traversed more than half the edges of $\mathscr{P}$ in succession. □

Despite appearances, the above proof is purely topological, because we can define the ordering of the discs bounded by $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \ldots$ by set inclusion, rather than in terms of the lengths of their radii.

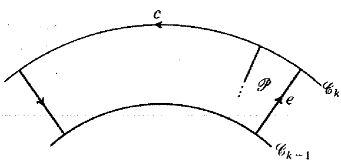


Figure 184

6.1.4 Dehn's Algorithm

It is immediate from the subpath property that if w is a word in $\pi_1(\mathscr{F})$ which equals 1 then w can be shortened either by cancellation (removal of a spur) or by replacing a subword p_1 which is more than half some cyclic permutation r^* of the defining relator by the complementary word p_2^{-1}, where $p_1p_2 = r^*$. (This corresponds to pulling the subpath p_1 across the polygon $\mathscr{P}$ to the position p_2^{-1} on the other side, cf. 1.4.3.)

Then any w which equals 1 in $\pi_1(\mathscr{F})$ will actually be reduced to 1 in a finite number of steps by this process, while a $w \neq 1$ will have a reduced form $\neq 1$ which cannot be shortened. This is Dehn's algorithm for the word problem in $\pi_1(\mathscr{F})$. As we have seen, it applies to any $\mathscr{F}$ whose canonical polygon has ≥ 6 sides, and hence to all but the small surfaces already dealt with in 6.1.2.

6.2 Simple Curves on Surfaces

6.2.1 Poincaré's Model of the Hyperbolic Plane

We mentioned in 1.4.2 that Poincaré used the hyperbolic plane in order to obtain a tessellation of the universal covering of an orientable surface $\mathscr{F}$ of genus $g > 1$ by congruent canonical polygons. Furthermore, Poincaré 1883 introduced a particular model of the hyperbolic plane which permits a convenient euclidean interpretation of noneuclidean "polygons." The "plane" is the open unit disc $\mathscr{D}$, "lines" are circular (or straight) arcs in $\mathscr{D}$ orthogonal to its boundary, and "points" are points of $\mathscr{D}$. It can be verified that there is exactly one line between any two points, and a distance is definable so that the line is the shortest curve between any two points. It turns out that "angle" coincides with the euclidean angle (between curves) so a polygon with given angles is determined by circular arcs orthogonal to the boundary of $\mathscr{D}$ which meet each other at the given angles.

All the axioms of plane geometry, except of course the parallel axiom, are satisfied in this model. For further details, see for example Magnus 1974.

Regular tessellations in the Poincaré model are well-known in complex function theory, and had in fact been introduced by Schwarz 1872. Schwarz's example was the tessellation by the triangle with angles $\pi/5$, $\pi/4$, $\pi/2$, shown in Figure 185. However, it was Poincaré who first observed that the polygons in such tessellations could be regarded as *congruent* when a suitable metric was introduced.

Exercise 6.2.1.1. Give a straight-edge-and-compass construction of the right-angled pentagon in the centre of the Schwarz tessellation.

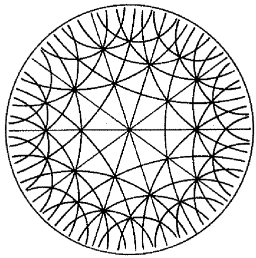


Figure 185

6.2.2 Simple Curves on the Torus

To provide some guidance in the use of the hyperbolic plane as universal covering surface, we first explain what happens when $\mathscr{F}$ is covered by the euclidean plane, that is, when $\mathscr{F}$ is the torus and the covering tessellation consists of rectangles. We then find that the closed simple curves are just the (m, n) torus curves with m, n relatively prime (cf. 4.2.7).

Let a curve p be given as a product of the canonical curves a, b and let $\tilde{p}$ be the covering path from a fixed $\tilde{P}^{(0)}$, ending at $\tilde{P}^{(1)}$ say (Figure 186). Then the straight line $\tilde{P}^{(0)}\tilde{P}^{(1)}$ is homotopic to $\tilde{P}$ and covers a path $\rho(p)$ homotopic to p. In a natural sense $\rho(p)$ is a *geometrically reduced* form of p, and we shall now show that it serves as a test case for the existence of simple curves homotopic to p.

The totality of paths covering $\rho(p)$ is obtained by drawing lines l equal to $\tilde{P}^{(0)}\tilde{P}^{(1)}$ in length and direction out of each vertex in the tessellation. If any of these lines meet at a point $\tilde{Q}$ other than an endpoint, then the underlying point Q on $\mathscr{F}$ is a point where $\rho(p)$ meets itself; thus $\rho(p)$ is simple just

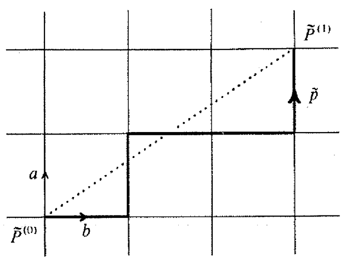


Figure 186

in case this does not happen. Equivalently, the only vertices on $\tilde{P}^{(0)}\tilde{P}^{(1)}$ are $\tilde{P}^{(0)}$, $\tilde{P}^{(1)}$ themselves, which means that if

$$p = a^m b^n,$$

then m, n are relatively prime, and $\rho(p)$ is the (m, n) torus curve of 4.2.7.

Conversely, if another vertex $\tilde{P}^{(2)}$ falls on $\tilde{P}^{(0)}\tilde{P}^{(1)}$ we have to show that any curve p' homotopic to p intersects itself. Assume for simplicity that P is fixed under the homotopy and let $\tilde{p}'^{(0)}_\infty$ be an infinite path in the plane constructed by first lifting p' to a $\tilde{p}'^{(0)}$ beginning at $\tilde{P}^{(0)}$ (and hence ending at $\tilde{P}^{(1)}$) and attaching further lifts successively at each end. Let $\tilde{p}'^{(2)}_\infty$ be the infinite covering path constructed in the same way but beginning at $\tilde{P}^{(2)}$. It will suffice to show that $\tilde{p}'^{(0)}_\infty$ intersects $\tilde{p}'^{(2)}_\infty$.

Consider the infinite straight line $\mathscr{L}$ through $\tilde{P}^{(0)}$, $\tilde{P}^{(1)}$. Translation along this line through distance $\tilde{P}^{(0)}\tilde{P}^{(1)}$ maps $\tilde{p}'^{(0)}_\infty$ onto itself while translation through distance $\tilde{P}^{(0)}\tilde{P}^{(2)}$ maps $\tilde{p}'^{(0)}_\infty$ onto $\tilde{p}'^{(2)}_\infty$. It follows that $\tilde{p}'^{(0)}_\infty$, $\tilde{p}'^{(2)}_\infty$ both have the same maximum distances to left and right of $\mathscr{L}$ (the distances are well-defined since they are determined by the compact sets $\tilde{p}'^{(0)}$ and $\tilde{P}^{(0)}\tilde{P}^{(1)}$) and hence they must intersect. For $\tilde{p}'^{(2)}_\infty$ cannot lie to the left of $\tilde{p}'^{(0)}_\infty$ at a point where $\tilde{p}'^{(0)}_\infty$ achieves its maximum distance to the left of $\mathscr{L}$, nor to the right of $\tilde{p}'^{(0)}_\infty$ at a point where $\tilde{p}'^{(0)}_\infty$ achieves its maximum distance to the right, yet by the Jordan curve theorem $\tilde{p}'^{(2)}_\infty$ must lie on just one side of $\tilde{p}'^{(0)}_\infty$ if they do not meet.

Exactly the same argument applies if the homotopy shifts the basepoint of p, since a shift of the basepoint from P to Q corresponds to a translation of the whole plane by the vector $\tilde{P}^{(0)}\tilde{Q}^{(0)}$. □

An equivalent condition for p to be homotopic to a simple curve is that $p \neq q^s$ for any $s > 1$. Poincaré 1904 showed that this is the condition for the *homology* class of p to contain a simple curve, whatever the genus of $\mathscr{F}$ (see 6.4.7). However, for genus >1, where homology and homotopy no longer coincide, the condition for the homotopy class to contain a simple closed curve is more complicated, as we shall see.

EXERCISE 6.2.2.1. Show that the homology class of a curve p on the perforated torus contains a simple curve iff $p \neq q^s$, $s > 1$.

EXERCISE 6.2.2.2. What is the minimum number of double points of a curve homotopic to $a^m b^n$, when m, n are arbitrary integers?

6.2.3 Poincaré's Algorithm

Given a curve p on an orientable surface $\mathscr{F}$ of genus $g > 1$ we construct the tessellation of the unit disc by $4g$-gons which meet $4g$ at each vertex and are regular and congruent in the Poincaré metric. We interpret the tessellated disc in the natural way as the universal covering surface $\tilde{\mathscr{F}}$, and lift the curve p to a curve $\tilde{p}$ on $\tilde{\mathscr{F}}$ from vertex $\tilde{P}^{(0)}$ to $\tilde{P}^{(1)}$ say. Then replace $\tilde{p}$ by the "straight line" $\tilde{P}^{(0)}\tilde{P}^{(1)}$ (that is, the arc of the unique circle through $\tilde{P}^{(0)}$, $\tilde{P}^{(1)}$ which is orthogonal to the boundary circle.) A similar straight line l is constructed for each lift $\tilde{p}^{(i)}$ of p to $\tilde{\mathscr{F}}$. Then p is homotopic to a simple curve just in case none of these lines meet, except at endpoints (Poincaré 1904).

Since the line $\tilde{P}^{(0)}\tilde{P}^{(1)}$ has the same endpoints as $\tilde{p}$ it is homotopic to $\tilde{p}$ on $\tilde{\mathscr{F}}$ and covers a path $\rho(p)$ homotopic to p on $\mathscr{F}$. The totality of lines l then comprise all the lifts of $\rho(p)$, and any double point of $\rho(p)$ will lift to a point where two of these l's meet (not counting meetings at endpoints). Thus if none of the l's meet, p is homotopic to the simple curve $\rho(p)$.

Conversely, if two of the l's, say l_1 and l_2, meet they must either overlap or cross (Figure 187). In Case (1), where l_1, l_2 overlap, we consider the infinite (in the Poincaré metric) line $\mathscr{L}$ which is their common prolongation. By considering translations along the line $\mathscr{L}$ one shows exactly as for the torus that any curve p' homotopic to p must intersect itself.

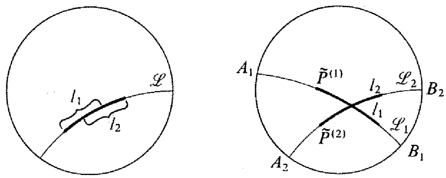


Figure 187

In Case (2) we consider the infinite prolongations $\mathscr{L}_1$, $\mathscr{L}_2$ of l_1, l_2 respectively. Since l_1 and l_2 cross, the limit points A_1, B_1 of $\mathscr{L}_1$ on the boundary circle separate the limit points A_2, B_2 of $\mathscr{L}_2$.

Now let p' be any curve homotopic to p on $\mathscr{F}$, with the initial point P fixed for simplicity. Let $\tilde{p}'^{(1)}_{\infty}$ be the infinite covering path constructed by first lifting p' from the initial point $\tilde{P}^{(1)}$ of l_1, then attaching further lifts successively at each end. Clearly all the endpoints of successive lifts lie on $\mathscr{L}_1$, so $\tilde{p}'^{(1)}_{\infty}$ has the same limit points A_1 and B_1, on the boundary circle as $\mathscr{L}_1$. Let $\tilde{p}'^{(2)}_{\infty}$ be constructed similarly by first lifting p' from the initial point $\tilde{P}^{(2)}$ of l_2; then $\tilde{p}'^{(2)}_{\infty}$ has the same limit points, A_2 and B_2, on the boundary circle as $\mathscr{L}_2$. Since A_1, B_1 separate A_2, B_2 it follows from the Jordan curve theorem that $\tilde{p}'^{(1)}_{\infty}$ intersects $\tilde{p}'^{(2)}_{\infty}$, and since the point of intersection cannot be the endpoint of lifts of p' on both $\mathscr{L}_1$, $\mathscr{L}_2$ (otherwise $\mathscr{L}_1$, $\mathscr{L}_2$ would be identical) it must cover a point where p' intersects itself.

The same argument applies when the homotopy shifts P to another point Q, since this corresponds to a translation of the whole hyperbolic plane.

□

EXERCISE 6.2.3.1. Give an algorithm for deciding whether two curves on $\mathscr{F}$ are homotopic to disjoint curves.

6.2.4 Effectiveness of Poincaré's Algorithm

There are two problems to be resolved before Poincaré's algorithm can be considered fully effective. Firstly, we cannot actually construct infinitely many lines l. However, as Poincaré points out, every polygon $\mathscr{P}$ in the tessellation carries a similar, finite, set of arcs $\rho(p_i)$ (corresponding to segments of the curve $\rho(p)$ cut off by the canonical curves on $\mathscr{F}$), so it suffices to construct them for just one $\mathscr{P}$. Then $\rho(p)$ is simple just in case none of the arcs $\rho(p_i)$ intersect.

The second problem is that our dependence on the metric seems to require an infinitely accurate geometric construction. However, translation into analytic geometry shows that the coordinates of all points of intersection are algebraic numbers (we have in fact a straight-edge and compass construction), so we can use known theorems about the solution of polynomial equations to decide in a finite number of steps whether two arcs actually intersect. Dehn 1912a made this observation in a similar context, and Calugareanu 1966 salvaged Poincaré's algorithm by working out the algebra explicitly. The algorithm can also be extended to nonorientable and bounded surfaces incidentally, as was shown by Reinhart 1962.

Complete removal of the metric from the algorithm, in the spirit of Dehn's solution of the word problem, was first achieved by Zieschang 1965. Zieschang's argument depends on a deep theorem of Whitehead 1936 on automorphisms of free groups, and a simple approach has not yet been found. Poincaré's approach remains the most intuitively transparent and serves to remind us that topology has a lot to gain from nontopological methods.

The methods of Poincaré, in particular the use of the unit disc and the characterization of homotopy classes by limit points on the disc boundary, were perfected by Nielsen in the 1920s and 1930s. His results on mappings of surfaces (see for example Nielsen 1927) remain among the deepest we know in combinatorial topology, and in some cases nonmetric proofs have still not been found.

EXERCISE 6.2.4.1. Show that there are two nontrivial simple closed curves on the Möbius band, up to homotopy.

EXERCISE 6.2.4.2. Show that a, b, ab, b^2 are the only nontrivial simple closed curves on the Klein bottle, up to homotopy (without fixed basepoint).

6.2.5 Baer's Theorem

If p and q are homotopic (without a fixed basepoint) simple curves on an orientable surface $\mathscr{F}$, then p and q are isotopic on $\mathscr{F}$ (Baer 1928).

To avoid *Hauptvermutung*-type problems, we assume p, q are polygonal, that is, edge paths in a simplicial decomposition of $\mathscr{F}$. Any overlapping of p, q can then be removed by small isotopies, so that p, q have only finitely many points of intersection.

The result is obvious if $\mathscr{F}$ is the sphere, otherwise we lift p, q to paths $\tilde{p}^{(0)}, \tilde{p}^{(1)}, \ldots$ and $\tilde{q}^{(0)}, \tilde{q}^{(1)}, \ldots$ in the universal covering plane. The $\tilde{p}^{(i)}$ unite into disjoint simple curves $\tilde{p}_0, \tilde{p}_1, \ldots$ each of which consists of infinitely many $\tilde{p}^{(i)}$'s joined end to end. Similarly, the $\tilde{q}^{(i)}$ unite into disjoint infinite simple curves $\tilde{q}_0, \tilde{q}_1, \ldots$. A $\tilde{p}_j$ may meet a $\tilde{q}_k$ however, namely at points lying over the points of intersection of p and q.

Special case: p and q do not meet, or have only a single point P in common. If the former, deform p isotopically until p, q just touch at a point P. Now let $\tilde{p}_0$, $\tilde{q}_0$ be infinite covering curves which pass through some point $\tilde{P}^{(0)}$ over P, meeting again at other points $\tilde{P}^{(l)}$ over P (since p, q are homotopic, lifts of them beginning at $\tilde{P}^{(l)}$ both end at $\tilde{P}^{(l+1)}$, as shown in Figure 188). No other $\tilde{p}_j$ or $\tilde{q}_j$ enters the shaded region between $\tilde{p}_0$ and $\tilde{q}_0$, as this would imply further points of intersection between p, q on $\mathscr{F}$. Therefore, an isotopic deformation of $\tilde{p}_0$ into $\tilde{q}_0$ which is periodic with respect to the tessellation of the universal covering plane (as can easily be arranged) projects to an isotopy between p and q on $\mathscr{F}$.

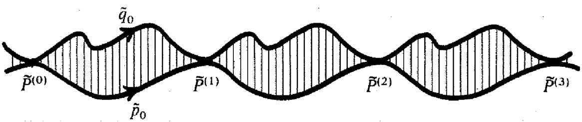


Figure 188

General case: p and q have more than one point of intersection. We again take a $\tilde{p}_0$ and $\tilde{q}_0$ which pass through a $\tilde{P}^{(0)}$ over P, and hence meet at other points $\tilde{P}^{(l)}$ over P. Since p, q meet at points other than P, $\tilde{q}_0$ may have intersections with $\tilde{p}_0$ other than the $\tilde{P}^{(l)}$, and other $\tilde{p}_j$'s may also meet $\tilde{q}_0$. Since we assume p, q have only finitely many points in common, there are only finitely many intersections in the interval between a $\tilde{P}^{(l)}$ and $\tilde{P}^{(l+1)}$, and hence one which is "innermost" in the following sense: either the $\tilde{p}_j$ in question just touches $\tilde{q}_0$ (Figure 189) or else $\tilde{p}_j$ accounts for two successive points of intersection on $\tilde{q}_0$ (Figure 190). In either case, an isotopy of P_j, which projects to an isotopy of p, reduces the number of points of intersection (Figure 191), so after a finite sequence of such isotopies we return to the special case. □

Figure 189

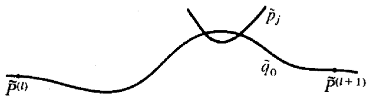


Figure 190

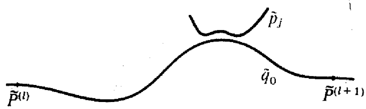


Figure 191

The special case shows that a simple curve which is null-homotopic on $\mathscr{F}$ in fact bounds a (nonsingular) disc—a nice demonstration of the way the universal covering surface can make facts leap to the eye.

The idea of finding a place where the nature of the intersection between two figures can be simplified by an isotopy has also been applied to surfaces in 3-manifolds, and some similar theorems have been obtained. For further information, see Laudenbach 1974.

EXERCISE 6.2.5.1. Extend the proof of Baer's theorem to nonorientable, bounded, and infinite surfaces.

6.3 Simplification of Simple Curves by Homeomorphisms

6.3.1 Introduction

In this section we study simple closed curves on a closed orientable surface $\mathscr{F}$. (Some suggestions on extending the results to nonorientable surfaces are contained in the exercises). If such a curve p is nonseparating then it is a consequence of the classification of bounded surfaces (1.3.9) that there is a homeomorphism $h: \mathscr{F} \to \mathscr{F}$ which maps p onto a canonical nonseparating curve $h(p)$. For example we can take $h(p)$ to be a "meridian" curve m on one of the handles of $\mathscr{F}$ (Figure 192).

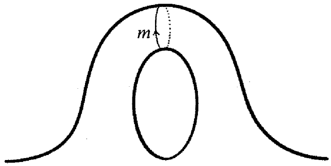


Figure 192

Simply observe that when we cut $\mathscr{F}$ along p the result $\mathscr{F}'$ has the same number of boundary curves (two), Euler characteristic, and orientability as the surface $\mathscr{F}''$ obtained by cutting $\mathscr{F}$ along the meridian m. Since these invariants define a bounded surface up to homeomorphism we have a homeomorphism $f: \mathscr{F}' \to \mathscr{F}''$.

Now if we identify the two edges p^+ and p^- of the cut p by any homeomorphism $g: p^+ \to p^-$, and the edges m^+ and m^- of m by the corresponding map $fgf^{-1}: m^+ \to m^-$, then the identification spaces of $\mathscr{F}'$ and $\mathscr{F}''$ both become $\mathscr{F}$, and the homeomorphism $h: \mathscr{F} \to \mathscr{F}$ induced by $f: \mathscr{F}' \to \mathscr{F}''$ maps p onto m. □

Since we depend on triangulations to define Euler characteristic and orientability, and to prove the classification theorem (which can be viewed as the construction of a homeomorphism between triangulated surfaces with the same invariants), the above proof requires p to be polygonal. The result is in fact true in general, but we need it only for polygonal curves, and these will be our only concern in this section. Thus until further notice a curve will be an edge path in some simplicial decomposition of the surface.

A similar proof shows that if p is a simple, separating curve, then we can map it onto a canonical separating curve $h(p)$ by a homeomorphism, the curve in question being determined by the Euler characteristics of the two components into which p divides $\mathscr{F}$. The obvious choice for canonical separating curves on an orientable $\mathscr{F}$ are those shown in Figure 193. Of course, having proved these theorems we can see that there *is* no preferred decomposition of $\mathscr{F}$ into "sphere" plus "handles," and that any simple curve is just as "canonical" as another. Any set of closed curves $a_1, b_1, \ldots, a_n, b_n$ which produce the canonical polygon for $\mathscr{F}$ (1.3.7) when cut may be called *a* canonical curve system, and the process of cutting along them is called a *handle decomposition* of $\mathscr{F}$.

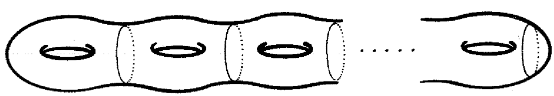


Figure 193

EXERCISE 6.3.1.1. Verify that the separating curves shown above do in fact partition $\mathscr{F}$ in all possible ways, as far as the Euler characteristic is concerned.

EXERCISE 6.3.1.2. Find canonical separating and nonseparating curves on the sphere with n crosscaps.

6.3.2 Twist Homeomorphisms

For applications to 3-manifolds in Chapter 8 it is convenient to be able to map simple curves onto canonical curves by homeomorphisms which affect the surface less radically than the cut- and-paste homeomorphisms used in proving the classification theorem. The *twist homeomorphisms* of Max Dehn are suitable for this purpose.

A twist homeomorphism relative to a simple closed curve c on $\mathscr{F}$ is determined by taking an annular neighbourhood $\mathscr{N}_c$ of c (always possible since $\mathscr{F}$ is orientable), cutting $\mathscr{N}_c$ from $\mathscr{F}$ as in Figure 194(1), then pasting it back after one of its boundaries has been given a full twist relative to the other as in Figure 194(2). As a result, a transverse segment AB of $\mathscr{N}_c$ becomes an arc which makes a full circuit of $\mathscr{N}_c$. The fact that points which are close together before the cutting and pasting are close together after it is easily formalized to show that the operation defines a homeomorphism of $\mathscr{F}$.

Dehn's work on simple curves on surfaces was done in the early 1920s but not published by him at the time. The first appearance of twist homeomorphisms in print seems to be Goeritz 1933. They are often known as Dehn twists, and since their revival by Lickorish 1962 it has been tempting to call them "Lickorish twists."

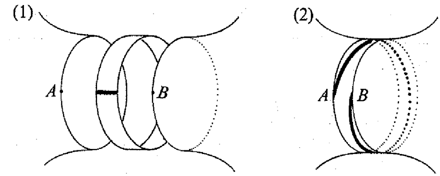


Figure 194

EXERCISE 6.3.2.1. Let t_a, t_b denote the twist homeomorphisms of the torus relative to the canonical curves a, b. Find a combination of t_a, t_b which maps a onto a curve isotopic to b.

6.3.3 Curves Which Meet at a Single Point and Cross

If p, q are simple closed curves on an orientable surface $\mathscr{F}$ with a single common point, where they cross, then p can be mapped onto q by a twist homeomorphism followed by an isotopy.

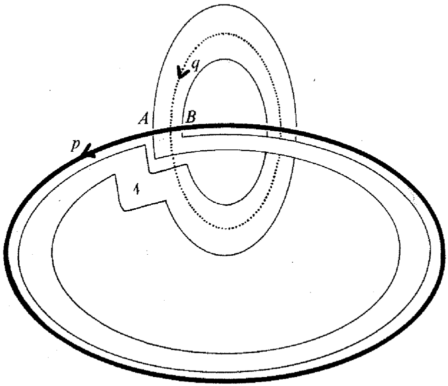


Figure 195

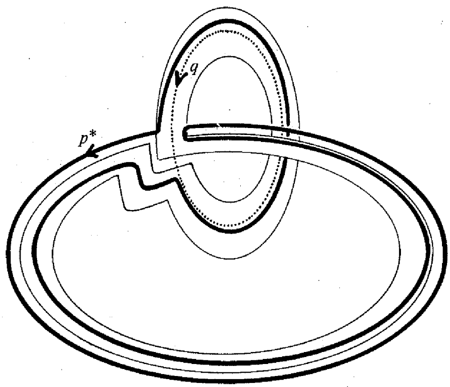


Figure 196

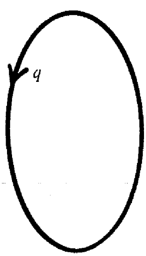


Figure 197

Given the curves p, q construct the strip $\mathcal{N}$ shown in Figure 195 which includes a small segment AB of p, all but a small segment of q, and elsewhere follows close to p. Then a twist of $\mathcal{N}$ maps p into the curve p^*, shown in Figure 196, which is obviously isotopic to q (Figure 197). □

Exercise 6.3.3.1. Show that a curve p satisfies the hypothesis of the theorem if and only if it is a nonseparating curve.

Exercise 6.3.3.2. Let $p_1, p_2, \ldots, p_{2n}$ be simple closed curves on an $\mathcal{F}$ of genus n with a single common point, where they cross each other. Show that the result of cutting $\mathcal{F}$ along $p_1, p_2, \ldots, p_{2n}$ is a disc. Is this necessarily a handle decomposition (6.3.1)?

Exercise 6.3.3.3. Let p, q be simple closed curves on a nonorientable surface which meet at a single point and cross. Under what circumstances is there a homeomorphism mapping p onto q? Find a twist homeomorphism which (together with an isotopy) realizes the mapping in this case.

6.3.4 Removal of Intersections

We shall assume that all points of intersection between two curves p, q are crossings. If not, a small isotopy of p near a point where it touches q removes the contact point altogether (Figure 198). When p, q are oriented, p can cross q from right to left (1), or from left to right (2), as shown in Figure 199. Assigning $+1$ and -1 respectively to these two types of crossing, we obtain the *algebraic intersection* of p with q by summing the values obtained in a circuit around p. We shall now prove:

If p, q are simple closed curves on $\mathcal{F}$, then there is a combination of twist homeomorphisms and isotopies which maps p onto a curve p^* which meets q at most twice, with zero algebraic intersection.

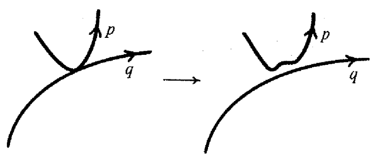


Figure 198

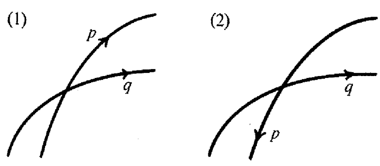


Figure 199

If p, q have only one point of intersection we can use the twist homeomorphism of 6.3.3 to obtain a p^* disjoint from q, as Figure 196 shows.

We can therefore assume there are at least two points where p intersects q with the same algebraic value. Without loss of generality we can assume they occur as successive points of intersection on q (Case 1), or as first and third in a sequence of three successive intersections on q (Case 2).

Case 1. Let A, B be successive points of intersection on q, where p passes from left to right, say (Figure 200). The curve c shown follows p after its exit at X, stays close to p without crossing it, and hence returns on the right-hand side of p at Y (since $\mathscr{F}$ is orientable) after traversing part of p just once. It therefore has fewer intersections with q than p has (since A, B have been replaced by D) and exactly one with p itself. By 6.3.3 we can then map p onto c by a twist homeomorphism and isotopy. (Sliding the portions of c near p back onto p by an isotopy we can in fact get a curve which departs from p only in a narrow neighbourhood of q, but has fewer intersections with q.)

Case 2. Let A, B, C be the successive points of intersection (Figure 201). Assume that the path d which exits at X and follows p returns first at Y, rather than on the middle branch of p. (If not, construct d so that it exits on the bottom branch of p. It must then return on the top branch, and a similar argument applies.) We then perform a twist homeomorphism using a neighbourhood of d, which maps p onto the curve p' shown in Figure 202 which is isotopic to the curve p'' shown in Figure 203 and in turn isotopic to the curve p''' in Figure 204 which has fewer intersections with q. (Notice also that p''' departs from p only in a narrow neighbourhood of q, shown dotted in Figure 204.)

Since the construction of a c or p''' reduces the number of intersections with q, after a finite number of steps we obtain a p^* $(=c$ or $p''')$ for which neither Case 1 nor Case 2 holds, and which therefore satisfies the conditions of the theorem. □

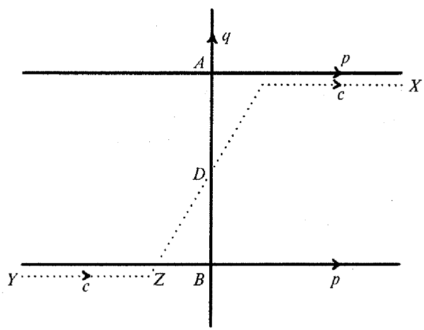


Figure 200

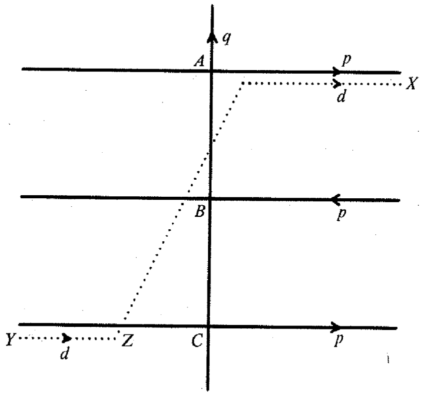


Figure 201

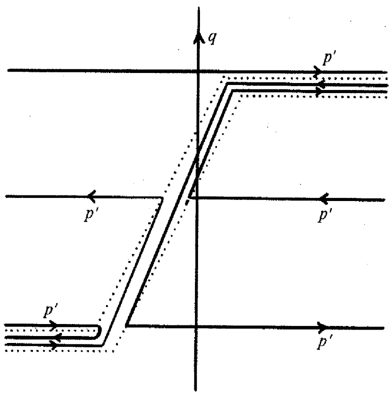


Figure 202

The above proof is due to Lickorish 1962, who notes the corollary that if $q_1, \ldots, q_r$ are disjoint simple curves on $\mathscr{F}$, then any simple curve p can be mapped by twist homeomorphisms and isotopies onto a simple p^* which meets each q_i at most twice, with zero algebraic intersection. For since the above construction produces a p^* which departs from p only in a narrow neighbourhood of q (though possibly omitting some portions of p entirely) we can remove the intersections between p and any q_i without affecting its intersections with other q_j's, eventually obtaining a p^* with the property described.

Lickorish uses this corollary to continue the simplification of p as follows. We continue to denote each successive modification of p by p^*.

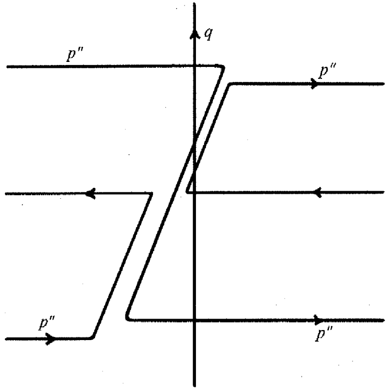


Figure 203

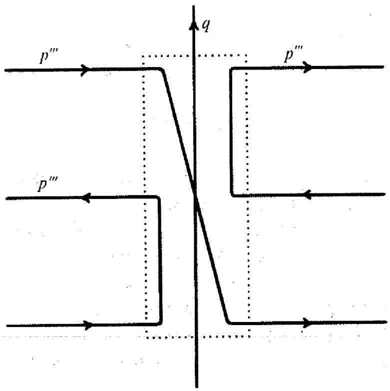


Figure 204

6.3.5 Taking a Curve off the Handles

Given a particular handle decomposition of $\mathscr{F}$, choose a "meridian" q_i and a "base curve" q_j on each handle as in Figure 205. If p^* does not meet q_i we shall say p^* is "off" the handle in question. (It may of course still pass "through" the handle, meeting the base curve on each side.) Using twist homeomorphisms and isotopies, we can obtain a p^* which is off all the handles.

Take the meridian and base curves to be the $q_1, q_2, \ldots, q_r$ mentioned in the corollary at the end of 6.3.4, and let p^* be the curve obtained in its conclusion.

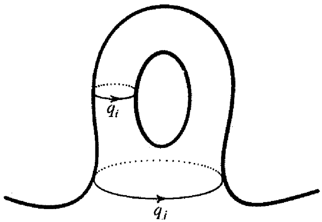


Figure 205

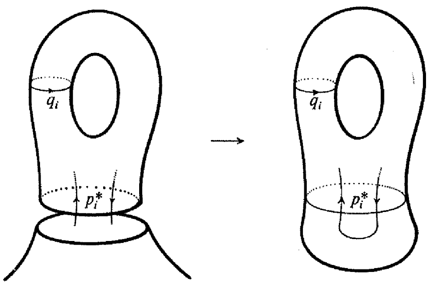


Figure 206

We can view the portion p_i^* of p^* on a given handle as a simple closed curve on the torus by cutting the handle from $\mathscr{F}$ along q_j and closing q_j with a disc, closing p_i^* by an arc on the disc between its intersections with q_j (if they exist), as in Figure 206. We then have a simple closed curve on the torus, and the known types of these curves (6.2.2), plus the fact that p^* meets q_i and q_j at most twice, with zero algebraic intersection, allow us to conclude that p^* is either null-homotopic or homotopic to q_i itself. In either case an isotopy will pull p_i^*, and hence p^*, away from q_i, so that p^* is off the handle. □

Exercise 6.3.5.1. Consider the handlebody $\mathscr{H}$ obtained by taking $\mathscr{F}$ together with its interior when $\mathscr{F}$ is "standardly embedded" in $\mathbb{R}^3$. Show that a simple closed curve which is off the handles of $\mathscr{F}$ bounds a disc in $\mathscr{H}$.

Exercise 6.3.5.2. Give an algorithm for deciding whether a given simple curve on $\mathscr{F}$ bounds a disc in $\mathscr{H}$. (*Hint*: consider the universal covering space of $\mathscr{H}$.)

6.3.6 Mapping onto a Canonical Curve

The construction of 6.3.5 not only takes p^* off the handles, it also arranges that p^* passes at most once "through" each handle, since p^* meets each base curve at most twice, with zero algebraic intersection. We can then map p^* onto a canonical curve by twist homeomorphisms and isotopies.

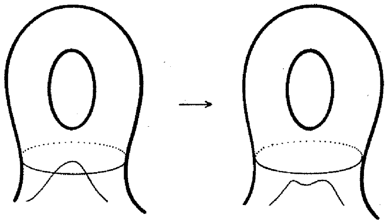


Figure 207

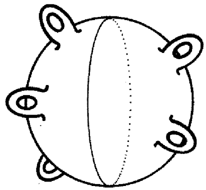


Figure 208

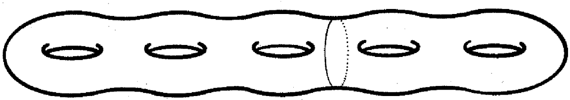


Figure 209

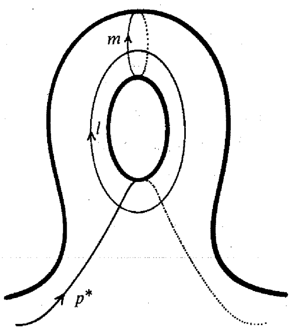


Figure 210

Case 1. p^* passes through no handles. That is, all its intersections with base curves q_j can be removed by isotopies (Figure 207). Then p^* can be viewed as a curve on the "sphere part" of $\mathscr{F}$ and deformed into a semicircle, for example see Figure 208. This is obviously the same as the canonical separating curve on $\mathscr{F}$ previously pictured as in Figure 209.

Case 2. p^* passes through a handle. This corresponds to the case where the portion p_i^* on the torus constructed in 6.3.5 is not null homotopic. Then p_i^*, and hence p^*, has a single point in common with a suitable latitude curve l, which in turn has a single point in common with a meridian m (Figure 210). Hence by two applications of 6.3.3 we can map p^* onto the canonical curve m. □

Exercise 6.3.6.1. Show that the mapping in Case 2 can also be achieved by an isotopy.

6.4 The Mapping Class Group of the Torus

6.4.1 Introduction

Just as we can construct a discrete group $\pi_1(\mathscr{C})$ from closed paths in a complex $\mathscr{C}$ by factoring out "topologically uninteresting" paths, we can construct a discrete group $M(\mathscr{C})$ from the self-homeomorphisms of $\mathscr{C}$ by factoring out topologically uninteresting maps. We could take topologically uninteresting maps to be those homotopic to the identity, or isotopic to the identity. In the case where $\mathscr{C}$ is a surface $\mathscr{F}$ these concepts coincide (as Baer's theorem suggests), but we take the isotopy relation for definiteness, and call $M(\mathscr{F})$ the *homeotopy*, or *mapping class*, group of $\mathscr{F}$.

Thus the homeotopy group of $\mathscr{F}$ is the quotient of the group of self-homeomorphisms of $\mathscr{F}$, under composition, by the normal subgroup of isotopies.

The easiest nontrivial case is the mapping class group of the torus $\mathscr{T}$, which was known implicitly in the theory of elliptic functions as far back as Clebsch and Gordan 1866, p. 304. It was not described as a group until later, in Klein and Fricke 1890, and the topological interpretation appeared in Tietze 1908. Tietze's interpretation yielded the important insight that $M(\mathscr{T})$ is the automorphism group of $\pi_1(\mathscr{T})$, a result whose extension to other surfaces $\mathscr{F}$ (Baer 1928) reduces the computation of $M(\mathscr{F})$ to a purely algebraic problem concerning $\pi_1(\mathscr{F})$. The details of this extension may be inferred from Baer's theorem and the argument for the torus we shall give below, however we omit them because they yield little information about $M(\mathscr{F})$ *except* its relation to $\pi_1(\mathscr{F})$.

Exercise 6.4.1.1. Why do the isotopies constitute a normal subgroup among the group of self-homeomorphisms?

6.4.2 Canonical Curve Pairs on the Torus

It follows immediately from the classification of bounded surfaces (cf. 6.3.1) that the canonical polygon for the torus (Figure 211) can be produced by cutting the torus along *any* two simple closed curves a, b which meet, and cross, at a single point. For under these circumstances a, b must both be nonseparating; a cut along a will produce a surface with two boundary curves, a^+ and a^-, homeomorphic to an annulus, and b will be a "crosscut" from a^+ to a^- which reduces the surface to a disc. The four edges of the disc must then be identified as in the canonical polygon in order to recover the torus.

Thus from an intrinsic point of view any pair of simple curves a, b which meet, and cross, at a single point can be considered canonical, and choosing one pair as generators to express all other curves is analogous to an arbitrary choice of coordinate axes.

The curves a, b shown in the standard picture of the torus (Figure 212) which we call a *meridian* and *latitude* curve respectively, can be distinguished from other canonical curve pairs only relative to an embedding of the torus in R^3. Namely, if we take the *solid torus*, or ring, $\mathscr{R}$ consisting of the surface $\mathscr{T}$ and the region "inside" it, then a is determined up to isotopy as the curve which is not null-homotopic on $\mathscr{T}$ but bounds a disc in $\mathscr{R}$.

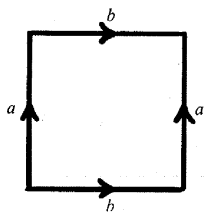


Figure 211

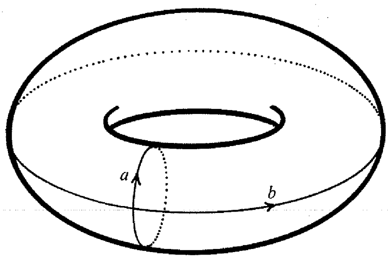


Figure 212

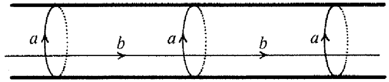


Figure 213

A quick way to see this is to consider the universal covering space $\tilde{\mathcal{R}}$ of $\mathcal{R}$, which is an infinite solid cylinder (Figure 213). By the standard argument, a curve which contracts to a point in $\mathcal{R}$ must lift to a closed curve in $\tilde{\mathcal{R}}$. The only curves which do so are those of the form a^m, and a is the only simple nontrivial curve among these. The latitude curve is not determined up to isotopy, even on the solid torus. One can either choose an arbitrary curve which crosses a once as the latitude, or else impose a further condition, such as requiring it to bound a surface in the region "outside" $\mathcal{T}$, which does determine the latitude curve for a given embedding.

These facts will not be used in the remainder of this chapter, but they are very much to the point in 7.1, when we study knotted embeddings of the solid torus in R^3.

EXERCISE 6.4.2.1. What types of curve cross a at a single point?

EXERCISE 6.4.2.2. If the solid torus $\mathcal{R}$ is embedded in R^3 in a knotted way, no nontrivial curve on $\mathcal{T}$ will bound a disc in $\mathsf{R}^3 - \mathcal{R}$. Why must there be a nontrivial curve on $\mathcal{T}$ which bounds a surface in $\mathsf{R}^3 - \mathcal{R}$?

6.4.3 Classification of Canonical Curve Pairs

An (m_1, n_1) torus curve and an (m_2, n_2) torus curve in standard form (that is, projections of straight line segments through the origin in the universal covering plane) have a single common point if and only if

$$m_1 n_2 - m_2 n_1 = \pm 1.$$

They necessarily cross at this point.

Suppose that the net generated by a, b on the universal covering plane has its vertices at the lattice points (i, j) where i, j are integers. Then the condition

$$m_1 n_2 - m_2 n_1 = \pm 1$$

or

$$\begin{vmatrix} m_1 & n_1 \\ m_2 & n_2 \end{vmatrix} = \pm 1$$

is precisely the condition for the linear transformation

$$T: \begin{cases} x' = m_1 x + n_1 y \\ y' = m_2 x + n_2 y \end{cases}$$

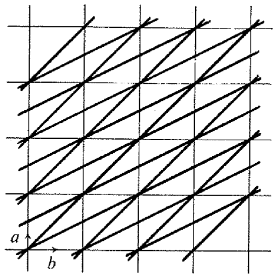


Figure 214

to map the set of lattice points one-to-one onto itself. This transformation maps the net of squares with (vector) sides a, b onto the net of parallelograms with sides $m_1a + n_1b$, $m_2a + n_2b$. For example, Figure 214 shows the net generated by $a + b$, $-a - 2b$. The fact that T is one-to-one means in particular that no vertex falls in the interior of either vector $m_1a + n_1b$, $m_2a + n_2b$ from the origin. Thus the corresponding (m_1, n_1) and (m_2, n_2) curves on the torus have only a single common point, at which they cross.

Conversely, suppose we have (m_1, n_1) and (m_2, n_2) torus curves in standard form, with a single common point. The above argument about the vectors $m_1a + n_1b$ and $m_2a + n_2b$ and the net of parallelograms they generate can then be reversed, showing that the transformation T is one-to-one on the set of lattice points, and hence of determinant ± 1. □

EXERCISE 6.4.3.1 Show that the set of transformations T with determinant ± 1 is isomorphic to the group of automorphisms of the free abelian group of rank 2.

EXERCISE 6.4.3.2. Show that

$$\begin{vmatrix} m_1 & n_1 \\ m_2 & n_2 \end{vmatrix}$$

in general is the algebraic intersection (6.3.4) of the (m_1, n_1) curve and the (m_2, n_2) curve.

6.4.4 Generating Homeomorphisms for Transformations of Canonical Curves

A homeomorphism of the torus maps the canonical pair a, b onto another canonical pair and hence by 6.4.3 onto an (m_1, n_1) curve and an (m_2, n_2) curve, with

$$m_1n_2 - m_2n_1 = \pm 1.$$

Any such transformation of canonical pairs can be realized by a combination of isotopies with the reflection which sends a, b to a^{-1}, b and the twist homeomorphisms t_a, t_b determined by a, b.

We first find generators for the linear transformations

$$T:\begin{cases} x' = m_1 x + n_1 y \\ y' = m_2 x + n_2 y \end{cases} \quad \text{with} \quad \begin{vmatrix} m_1 & n_1 \\ m_2 & n_2 \end{vmatrix} = \pm 1$$

then interpret the generators as homeomorphisms of the torus.

T can be represented by the matrix

$$\begin{pmatrix} m_1 & n_1 \\ m_2 & n_2 \end{pmatrix}$$

with determinant ± 1 and in fact the group $GL(2, \mathbf{Z})$ of such matrices is generated by

$$R = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

This idea is to work backwards from

$$\begin{pmatrix} m_1 & n_1 \\ m_2 & n_2 \end{pmatrix}$$

to $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. First we get a matrix of the form $\left(\begin{smallmatrix} i & j \\ 0 & k \end{smallmatrix}\right)$ using right multiplication by A^{-1}, B^{-1} to perform the euclidean algorithm on the bottom row, then observe that $i, j = \pm 1$ because the determinant remains ± 1 throughout. Now multiply by R and/or $(AB^{-1})^3 = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ to get $\left(\begin{smallmatrix} 1 & j \\ 0 & 1 \end{smallmatrix}\right)$, which equals A^j, and can therefore be multiplied by A^{-j} to get $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

The matrix R obviously represents the reflection which sends a, b to a^{-1}, b. The transformation with matrix A can in fact be realized by the twist homeomorphism t_a. For

$$\begin{pmatrix} m_1 & n_1 \\ m_2 & n_2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} m_1 & m_1 + n_1 \\ m_2 & m_2 + n_2 \end{pmatrix}$$

and recalling the construction of an (m, n) curve by drawing m lines on a cylinder, then joining the ends after a twist of $2\pi(n/m)$, we see that the extra twist of 2π given by t_a changes an (m, n) curve into an $(m, m + n)$ curve. The effect of t_a on the pair represented by the matrix

$$\begin{pmatrix} m_1 & n_1 \\ m_2 & n_2 \end{pmatrix}$$

is therefore to produce the pair with matrix

$$\begin{pmatrix} m_1 & m_1 + n_1 \\ m_2 & m_2 + n_2 \end{pmatrix},$$

as required.

Similarly, t_b realizes the transformation with matrix B. Of course t_a, t_b do not realize *linear* transformations of the plane, however it is easy to deform them into linear transformations by isotopies. □

EXERCISE 6.4.4.1 (Tietze 1908). What are the generating homeomorphisms for the solid torus?

EXERCISE 6.4.4.2. What are the generating homeomorphisms for the Klein bottle? Show that there is only one twist homeomorphism, and that its square is isotopic to the identity.

6.4.5 Homeomorphisms Are Determined up to Isotopy by the Transformation of the Canonical Curves

The proof of this result in the general case requires point set subtleties such as the Jordan-Schoenflies theorem, however the general idea is the following, which can easily be developed into a formal proof in the case of simplicial homeomorphisms.

Suppose $h_1: \mathcal{T} \to \mathcal{T}$ and $h_2: \mathcal{T} \to \mathcal{T}$ are homeomorphisms which map a, b onto pairs of canonical curves which are the same, up to isotopy, as a standard (m_1, n_1), (m_2, n_2) pair. We want to show that h_1 is isotopic to h_2.

It is clear that the covering curves of $h_1(a)$, $h_1(b)$ in the plane can *simultaneously* be isotopically straightened into the straight line segments from $(0, 0)$ to (m_1, n_1) and (m_2, n_2); see Figure 215. By breaking the isotopy down into a series of isotopies which take place within sufficiently small parts of the plane we can induce a series of isotopies on $\mathcal{T}$, the result of which is to reduce $h_1(a)$, $h_1(b)$ to a standard (m_1, n_1), (m_2, n_2) pair. Thus h_1 is isotopic to a map h_1' which maps a, b onto this standard pair, and similarly

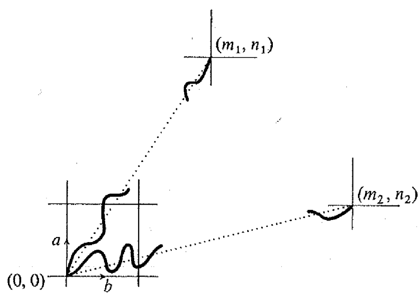


Figure 215

h_2 is isotopic to an h_2' which maps a, b onto the same pair. We then have that

$$f = h_1' h_2'^{-1} : \mathscr{T} \to \mathscr{T}$$

is a homeomorphism which maps a, b onto themselves.

To show h_1, h_2 are isotopic it will then suffice to show f is isotopic to the identity. The required isotopy can be constructed in the following steps (successive improvements of f will still be denoted f).

Step 1. Deform f into the identity on a, b using the fact that any orientation-preserving homeomorphism $h: \mathbf{S}^1 \to \mathbf{S}^1$ is isotopic to the identity.

Step 2. Deform f into the identity on strip neighbourhoods $\mathscr{A}$, $\mathscr{B}$ of a, b (Figure 216).

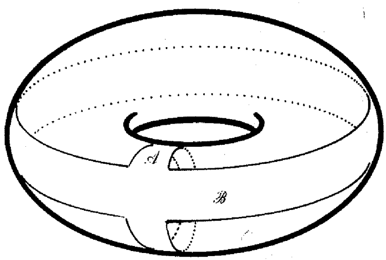


Figure 216

Step 3. Observe that $\mathscr{T} - (\mathscr{A} \cup \mathscr{B})$ is a disc, and f can be deformed into the identity on it because *any* homeomorphism of the disc which is the identity on the boundary is isotopic to the identity on the whole disc. A clever proof of this fact is due to Alexander 1923b: for $0 \le t \le 1$ and x the vector from the centre to an arbitrary point on the disc define

$$f_t(x) = \begin{cases} x & \text{for} \quad |x| \ge t \\ tf\left(\frac{1}{t}x\right) & \text{for} \quad t > |x| > 0 \end{cases}$$

when $t > 0$, and let $f_0(x) = x$. This function f_t is an isotopy between $f_1 = f$ and $f_0 =$ identity. □

A more general result was proved by Baer 1928. Baer's theorem (6.2.5) in fact generalizes to canonical *systems* of curves on any surface $\mathscr{F}$, then a similar argument shows that a homeomorphism $h: \mathscr{F} \to \mathscr{F}$ is determined up to isotopy by the image of the canonical curve system on $\mathscr{F}$.

Exercise 6.4.5.1. Prove that any orientation-preserving homeomorphism $h: \mathbf{S}^1 \to \mathbf{S}^1$ is isotopic to the identity.

EXERCISE 6.4.5.2. Give a rigorous version of the above proof, assuming h_1, h_2 are simplicial maps.

EXERCISE 6.4.5.3. Prove Baer's generalized theorem.

6.4.6 The Mapping Class Group

Now that we know a homeomorphism $h: \mathscr{T} \to \mathscr{T}$ is determined up to isotopy by the images $h(a)$, $h(b)$ of the canonical curves, and thus by the matrix

$$\begin{pmatrix} m_1 & n_1 \\ m_2 & n_2 \end{pmatrix},$$

where $h(a)$, $h(b)$ are curves of type (m_1, n_1), (m_2, n_2) respectively, we can associate the mapping class of h with this matrix, which has determinant ± 1. Since composition of mappings corresponds to right matrix multiplication (6.4.4) we in fact have a homomorphism ϕ of $M(\mathscr{T})$ onto the group $GL(2, \mathbf{Z})$ of 2×2 integer matrices with determinant ± 1. This homomorphism is one-to-one, since a mapping which sends a, b to a pair not homotopic to a, b is *a fortiori* not isotopic to the identity, hence $M(\mathscr{T})$ is in fact isomorphic to $GL(2,\mathbf{Z})$.

This matrix representation gives quite a clear picture of the group (for example one has an immediate solution of the word problem by multiplying out the matrices), however, it is also possible to give a finite presentation (see Nielsen 1924b, which also solves the much harder problem for 3×3 integer matrices with determinant ± 1).

EXERCISE 6.4.6.1. Show that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ also generate $GL(2, \mathbf{Z})$.

EXERCISE 6.4.6.2. What is the mapping class group of the Klein bottle?

6.4.7 Automorphisms of $H_1(\mathscr{F})$ when $\mathscr{F}$ Is Orientable of Genus > 1

The simple construction of the mapping class group of the torus $\mathscr{T}$ can be attributed to the fact that $\pi_1(\mathscr{T}) = H_1(\mathscr{T})$ and the automorphisms of abelian groups are relatively easy to determine. When $\mathscr{F}$ is of genus $g > 1$ and $\pi_1(\mathscr{F}) \neq H_1(\mathscr{F})$ the representation of a mapping by a matrix reflects only the homology class of the image curves, not their isotopy class. Furthermore, not every automorphism of $H_1(\mathscr{F}) = \mathbf{Z}^{2g}$ can actually be realized by a homeomorphism. Those that can constitute a subgroup of the group $GL(2g, \mathbf{Z})$ of $2g \times 2g$ integer matrices with determinant ± 1, called the *symplectic group* over the integers.

Nevertheless, the mere fact that the matrix of an automorphism of $\mathbf{Z}^{2g}$ must be integral enables us to prove that the homology class of a simple curve is not a multiple of any other homology class.

Let $a_1, b_1, \ldots, a_g, b_g$ be the canonical generators of $\pi_1(\mathscr{F})$ and $H_1(\mathscr{F})$. An element

$$p = a_1^{m_1} b_1^{n_1} \cdots a_g^{m_g} b_g^{n_g}$$

of $H_1(\mathscr{F})$ can be represented by the vector $(m_1, n_1, \ldots, m_g, n_g)$. By 6.3.1 there is a homeomorphism $h: \mathscr{F} \to \mathscr{F}$ which maps p onto the curve a_1 represented by the vector $(1, 0, \ldots, 0)$.

This homeomorphism induces an automorphism of $\pi_1(\mathscr{F})$ (3.1.7) and hence an automorphism A of $H_1(\mathscr{F})$. We can interpret A as a $2g \times 2g$ integer matrix such that

$$A(m_1, n_1, \ldots, m_g, n_g)^{\mathrm{T}} = (1, 0, \ldots, 0)^{\mathrm{T}},$$

where T denotes the transpose. But then the greatest common divisor d of $m_1, n_1, \ldots, m_g, n_g$ must be 1, since $A(m_1/d, n_1/d, \ldots, m_g/d, n_g/d)^{\mathrm{T}}$ is an integer vector which equals $(1/d, 0, \ldots, 0)$.

Thus $(m_1, n_1, \ldots, m_g, n_g)$ is not a multiple of any other homology class. □

When combined with its converse, this theorem yields the Poincaré 1904 characterization of homology classes which contain simple curves. A quick proof of the converse using twist homeomorphisms has been given by Meyerson 1976, and it is developed in the following exercise.

Exercise 6.4.7.1. Take generators $\alpha_1, \beta_1, \ldots, \alpha_g, \beta_g$ for $H_1(\mathscr{F})$ on a handle decomposition of $\mathscr{F}$ as shown in Figure 217 and denote the element $p = \alpha_1^{m_1}\beta_1^{n_1} \cdots \alpha_g^{m_g}\beta_g^{n_g}$ of $H_1(\mathscr{F})$ by $(m_1, n_1, \ldots, m_g, n_g)$.

(1) Show that twists about α_i, β_i send $(\ldots, m_i, n_i, \ldots)$ to $(\ldots, m_i \pm n_i, n_i, \ldots)$ and $(\ldots, m_i, n_i, \ldots)$ to $(\ldots, m_i, n_i \pm m_i, \ldots)$ respectively, where $\pm$ depends on the direction of twist.

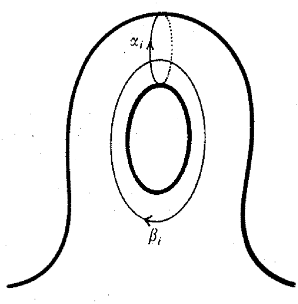


Figure 217

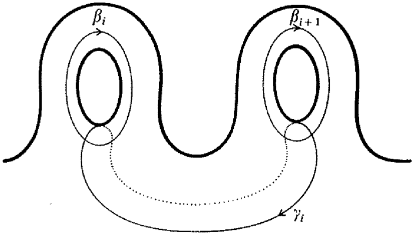


Figure 218

(2) Deduce that these twists can be used to send $(m_1, n_1, \ldots, m_g, n_g)$ to $(d_1, 0, d_2, 0, \ldots, d_g, 0)$, where $d_i = \gcd(m_i, n_i)$.

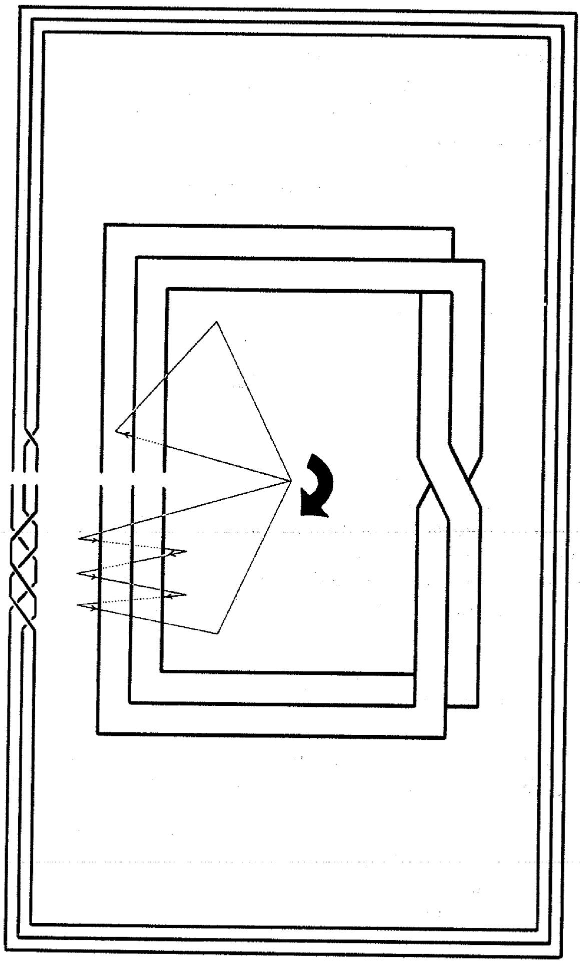
(3) Let γ_i, $1 \leq i < g$, be the curve shown in Figure 218. Show that a twist about γ_i sends $(\ldots, m_i, n_i, m_{i+1}, n_{i+1}, \ldots)$ to $(\ldots, m_i \pm n_i \pm n_{i+1}, n_i, m_{i+1} \pm n_i \pm n_{i+1}, n_{i+1}, \ldots)$.

(4) Deduce that $(d_1, 0, d_2, 0, \ldots, d_g, 0)$ can be sent to $(d, 0, \ldots, 0)$ by α, β, γ twists, where $d = \gcd(d_1, \ldots, d_g)$.

(5) Conclude that if p is not a multiple of any element of $H_1(\mathscr{F})$, then p is homologous to a simple curve.

CHAPTER 7

Knots and Braids



7.1 Dehn and Schreier's Analysis of the Torus Knot Groups

7.1.1 Introduction

We have seen (4.2.7) that the (m, n) torus knot has group

$$G_{m,n} = \langle a, b; a^m = b^n \rangle.$$

It is obvious that $G_{m,n} = G_{n,m}$, which reflects the less obvious fact that the (m, n) torus knot is the same as the (n, m) torus knot. $G_{m,n}$ does *not* reflect the orientation of the knot in R^3, since the knot and its mirror image have homeomorphic complements and hence the same group. Since Listing 1847, at least, it has been presumed that there is no ambient isotopy in R^3 between the two trefoil knots (Figure 219) and the same applies to the general (m, n) knot.

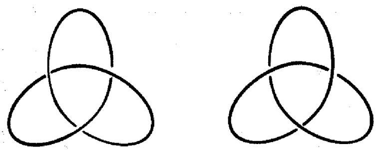


Figure 219

However, apart from this the torus knots are completely classified by their groups (Schreier 1924) and a brilliant argument of Dehn 1914 even allows us to distinguish a knot from its mirror image by deeper study of the group. Dehn did not use the above presentation of $G_{m,n}$ but a more complicated presentation like Wirtinger's, and Schreier's analysis of $G_{m,n}$ was intended, among other things, to simplify Dehn's argument.

7.1.2 The Centre of $G_{m,n}$

By definition, the centre C of $G_{m,n}$ is the subgroup of elements which commute with all other elements of $G_{m,n}$. C is equal to the subgroup generated by a^m.

The element a^m certainly commutes with a, and since it equals b^n it also commutes with b. Then a^m commutes with every element of $G_{m,n}$ and hence belongs to the centre.

Now since a^m commutes with every element of $G_{m,n}$, the elements

$\{a^{\pm km} | k = 0, 1, 2, \ldots\}$ constitute a normal subgroup N, and $G_{m,n}/N$ is obtained by adjoining the relation $a^m = 1$, that is,

$$\begin{aligned} G_{m,n}/N &= \langle a, b; a^m = b^n, a^m = 1 \rangle \\ &= \langle a, b; a^m = 1, b^n = 1 \rangle. \end{aligned}$$

We shall write this group as $\langle \bar{a}, \bar{b}; \bar{a}^m = 1, \bar{b}^n = 1 \rangle$ to emphasize the interpretation of the generators as cosets:

$$\bar{a} = Na, \qquad \bar{b} = Nb.$$

To show that N is the centre C of $G_{m,n}$ it will suffice to show that $G_{m,n}/N$ has no nontrivial elements in its centre. Since $\bar{a}^m = 1$ and $\bar{b}^n = 1$, any element of $G_{m,n}/N$ has a unique normal form

$$\bar{a}^{x_1}\bar{b}^{y_1}\bar{a}^{x_2}\bar{b}^{y_2} \ldots \bar{a}^{x_p}\bar{b}^{y_p},$$

where

$$0 \leq x_1 < m, 0 < x_i < m \quad \text{for } i > 1,$$

$$0 \leq y_p < n, 0 < y_i < n \quad \text{for } i < p,$$

are integers. But such an expression commutes with $\bar{a}$ only if it has $\bar{a}$ at both ends, and with $\bar{b}$ only if it has $\bar{b}$ at both ends, so a nontrivial element cannot commute with both $\bar{a}, \bar{b}$. □

7.1.3 Elements of Finite Order in $G_{m,n}/C$

Elements of finite order in $G_{m,n}/C$ are conjugate to a power of $\bar{a}$ or a power of $\bar{b}$.

Let g be an element of order r and let $\bar{a}^{x_1} \cdots \bar{b}^{y_p}$ be its normal form. Since no shortening of the word

$$(\text{normal form of } g)^r$$

is possible unless the normal form begins and ends with the same letter, we must have one of x_1, y_p equal to 0. In the case $p = 1$ this makes g a power of $\bar{a}$ or $\bar{b}$, so the assertion is *a fortiori* true for $p = 1$.

Now suppose p is arbitrary and the assertion is true for all values $<p$. If, say, $y_p = 0$, then

$$\bar{a}^{x_p} g \bar{a}^{-x_p} = \bar{a}^{x_1 + x_p} \ldots \bar{b}^{y_{p-1}}. \tag{*}$$

But if g is of order r, so is $\bar{a}^{x_p} g \bar{a}^{-x_p}$, and hence so is ω, the right-hand side of (*), which means it is conjugate to a power of $\bar{a}$ or $\bar{b}$, by induction. The same is then true of

$$g = \bar{a}^{-x_p} \omega \bar{a}^{x_p}.$$

(Analogously when $x_1 = 0$.) □

7.1.4 Presentation-Invariant Determination of $\{m, n\}$ from $G_{m,n}$

7.1.2 tells us that we can define the group $\langle \bar{a}, \bar{b}; \bar{a}^m = 1, \bar{b}^n = 1\rangle$ independently of a presentation of $G_{m,n}$, namely as the quotient of $G_{m,n}$ by its centre C. Now let

$$E = \text{abelianization of } \frac{G_{m,n}}{C}.$$

E is obviously a group of order mn, so we have determined the number mn from $G_{m,n}$ in an invariant manner.

Now 7.1.3 tells us that no element of finite order in $G_{m,n}/C$ can have order greater than that of $\bar{a}$ or $\bar{b}$. The number $\max(m, n)$ is therefore invariantly determined as the maximum finite order of an element in $G_{m,n}/C$. The number $\min(m, n)$ is then determined as $mn/\max(m, n)$, so we have determined the pair $\{m, n\}$ independently of the presentation of $G_{m,n}$.

It follows that if $\{m, n\} \neq \{m', n'\}$, then $G_{m,n}$ and $G_{m',n'}$ are nonisomorphic groups.

An immediate consequence of this theorem is that if $\{m, n\} \neq \{m', n'\}$, then the (m, n) and (m', n') torus knots have nonhomeomorphic complements, hence they are distinct knots. In particular, this proves that there are infinitely many knots.

The group $\langle \bar{a}, \bar{b}; \bar{a}^m, \bar{b}^n\rangle$ is of course the free product of cyclic groups of orders m and n, so Schreier has incidentally derived four important properties of such groups:

(1) Normal form of elements (hence a solution of the word problem).
(2) The centre is trivial.
(3) An element of finite order is conjugate to a power of a generator.
(4) The orders $\{m, n\}$ of the factors are uniquely determined by the group.

The particular free product $\langle a, b; a^2, b^3\rangle$ had already been studied by Klein and Fricke 1890 in its realization as the group of transformations generated by

$$a(z) = -\frac{1}{z}, \qquad b(z) = \frac{1}{-z+1}$$

(these are the transformations $f(z) = (pz + q)/(rz + s)$ where p, q, r, s are integers such that $ps - rq = 1$—the modular group). In this case the order of procedure was different since initially only the generators were known. The normal form was derived first and then used to show that the relations $a^2 = b^3 = 1$ sufficed to define the group (since they suffice to obtain the normal form).

Exercise 7.1.4.1. Show that the transformations $a(z)$, $b(z)$ above define the group $\langle a, b; a^2, b^3\rangle$ (For more help, see Magnus, Karrass, and Solitar 1966, p. 44.)

Exercise 7.1.4.2. Construct the Cayley diagram of $\langle a, b; a^2, b^3\rangle$.

7.1.5 Latitude and Meridian Curves on a Knotted Ring

We now adapt Dehn's train of thought on the two trefoil knots with the Schreier analysis of $G_{2,3}$ in mind. The two knots are represented by knotted rings $\mathscr{R}_1, \mathscr{R}_2$ in R^3 and we take generators a_1, b_1 for $\pi_1(\mathsf{R}^3 - \mathscr{R}_1)$ and a_2, b_2 for $\pi_1(\mathsf{R}^3 - \mathscr{R}_2)$ as shown in Figure 220 (a_2, b_2 are the mirror images of a_1, b_1).

On the torus surface $\mathscr{T}_1$ of $\mathscr{R}_1$ we choose latitude and meridian curves c_1, d_1 which serve to determine a right-hand screw in R^3 (Figure 221). We can specify the meridian d_1 up to isotopy and orientation as a simple closed curve which is null-homotopic in $\mathscr{R}_1$ but not on $\mathscr{T}_1$ (6.4.2). A latitude is then a simple closed curve on $\mathscr{T}_1$ which meets d_1 exactly once. Such curves are not unique up to isotopy, and we shall choose the one most simply expressed in terms of the generators of $\pi_1(\mathsf{R}^3 - \mathscr{R}_1)$, namely

$$c_1 = a_1^2.$$

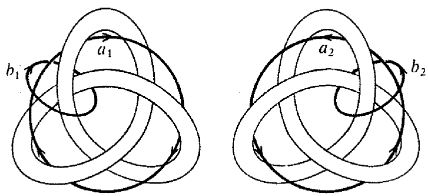


Figure 220

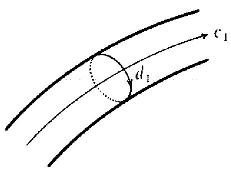


Figure 221

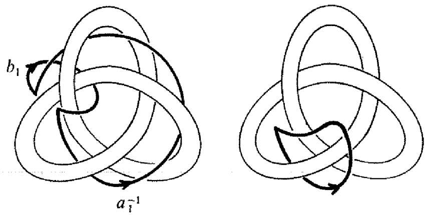


Figure 222

It is clear from the diagram above that a_1^2 can be deformed onto a latitude curve on $\mathscr{T}_1$. Figure 222 shows that

$$d_1 = a_1^{-1}b_1.$$

EXERCISE 7.1.5.1. Draw a picture of the latitude curve corresponding to a_1^2.

7.1.6 Group-Theoretic Consequences of an Ambient Isotopy Between the Two Trefoil Knots

If such an isotopy exists, the generators a_2, b_2 of $\pi_1(\mathrm{R}^3 - \mathscr{R}_2)$ can be regarded as elements $a_2(a_1, b_1)$, $b_2(a_1, b_1)$ of $\pi_1(\mathrm{R}^3 - \mathscr{R}_1)$. In fact, since

$$\langle a_1, b_1; a_1^2 = b_1^3 \rangle = \pi_1(\mathrm{R}^3 - \mathscr{R}_1) = \pi_1(\mathrm{R}^3 - \mathscr{R}_2) = \langle a_2, b_2; a_2^2 = b_2^3 \rangle,$$

the correspondence

$$a_1 \mapsto a_2, \qquad b_1 \mapsto b_2$$

defines an automorphism of the group.

We now see how the latitude and meridian are expressed in terms of a_2, b_2. Since the meridian d_1 is unique up to isotopy and orientation, in terms of a_2, b_2 it can only be

$$d_1 = (a_2^{-1}b_2)^{\pm 1}.$$

By sliding the meridian and making a rotation of $\mathscr{R}_2$, if necessary, we can arrange that $d_1 = d_2 = a_2^{-1}b_2$ (see Figure 223). The latitude c_1 must then have the orientation of a_2^{-2} in order to determine a right-hand screw with d_1, and its most general possible form is therefore

$$\begin{aligned} c_1 &= a_2^{-2}d_2^n \qquad \text{(cf. 6.4.3)} \\ &= a_2^{-2}(a_2^{-1}b_2)^n. \end{aligned}$$

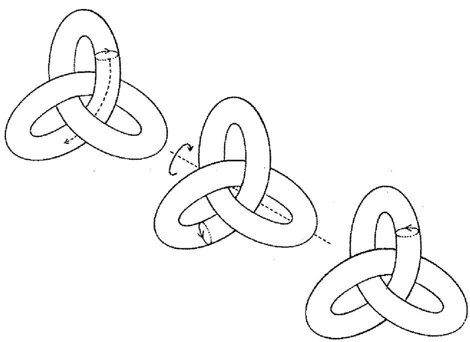


Figure 223

When we substitute $a_2 = a_2(a_1, b_1)$ and $b_2 = b_2(a_1, b_1)$ in these expressions for c_1, d_1 and compare with

$$c_1 = a_1^2, \qquad d_1 = a_1^{-1}b_1$$

from 7.1.5, we must get equations

$$a_1^2 = a_2^{-2}(a_2^{-1}b_2)^n \tag{1}$$

$$a_1^{-1}b_1 = a_2^{-1}b_2 \tag{2}$$

which are valid in $G_{2,3} = \langle a_1, b_1; a_1^2 = b_1^3 \rangle$.

We now have a purely algebraic question: *is there an automorphism* $a_1 \mapsto a_2$, $b_1 \mapsto b_2$ *of* $G_{2,3}$ *which satisfies the Equations* (1), (2)? Dehn was able to answer this question in the negative by explicitly finding all automorphisms of $G_{2,3}$. The task is made much easier by Schreier's analysis, which we now resume.

7.1.7 Automorphisms of $G_{m,n}/C$

When $m \neq n$ *all the automorphisms are given by* $\bar{a} \mapsto \bar{t}^{-1}\bar{a}^r\bar{t}$, $b \mapsto \bar{t}^{-1}\bar{b}^s\bar{t}$, *where* r *is prime to* m *and* s *is prime to* n.

Since $G_{m,n}/C = \langle \bar{a}, \bar{b}; \bar{a}^m, \bar{b}^n \rangle$ the above substitutions certainly determine automorphisms. We have to show there are no others. Each automorphism must send $\bar{a}$ and $\bar{b}$ to elements $\bar{a}'$ and $\bar{b}'$ of orders m and n respectively, and hence conjugate to powers of generators by 7.1.3. They cannot be conjugate to powers of the same generator, because the other generator would then have exponent sum 0 in each product of $\bar{a}'$ and $\bar{b}'$, and hence would not appear in the group they generate. By further consideration of exponent sums we find that the powers r and s of $\bar{a}$ and $\bar{b}$ which occur must be respectively prime to m and n. Then since $m \neq n$ the only way $\bar{a}'$, $\bar{b}'$ can have periods m, n respectively is if

$$\bar{a}' = \rho^{-1}\bar{a}^r\rho, \qquad \bar{b}' = \sigma^{-1}\bar{b}^s\sigma,$$

for some elements ρ, σ of $G_{m,n}/C$.

Now if $\bar{a} \mapsto \rho^{-1}\bar{a}^r\rho$, $\bar{b} \mapsto \sigma^{-1}\bar{b}^s\sigma$ determines an automorphism, we can compose it with the inner automorphism $\bar{a} \mapsto \sigma\bar{a}\sigma^{-1}$, $\bar{b} \mapsto \sigma\bar{b}\sigma^{-1}$ to obtain the automorphism

$$\bar{a} \mapsto \tau^{-1}\bar{a}^r\tau, \qquad \bar{b} \mapsto \bar{b}^s,$$

where $\tau = \rho\sigma^{-1}$, and it will suffice to show that the latter has the required form. Well, since $\bar{a}$ must be generated from $\tau^{-1}\bar{a}^r\tau$ and $\bar{b}^s$, τ can only have the form $\bar{a}^x\bar{b}^y$ (otherwise sufficient cancellation is not possible) and then the automorphism is

$$\bar{a} \mapsto \bar{b}^{-y}\bar{a}^{-x}\bar{a}^r\bar{a}^x\bar{b}^y = \bar{b}^{-y}\bar{a}^r\bar{b}^y$$

$$\bar{b} \mapsto \bar{b}^s = \bar{b}^{-y}\bar{b}^s\bar{b}^y$$

and so we have the required form $\bar{a} \mapsto \bar{t}^{-1}\bar{a}^r\bar{t}$, $\bar{b} \mapsto \bar{t}^{-1}\bar{b}^s\bar{t}$. □

7.1.8 Automorphisms of $G_{m,n}$

When $m \neq n$ all automorphisms are given by $a \mapsto t^{-1}a^{\varepsilon}t$, $b \mapsto t^{-1}b^{\varepsilon}t$, where $\varepsilon = \pm 1$ and t is an arbitrary element of $G_{m,n}$.

Because the centre C is an invariantly defined subgroup of $G_{m,n}$, any automorphism of $G_{m,n}$ maps C onto itself and induces automorphisms of C and $G_{m,n}/C$. The only automorphisms of the infinite cyclic group C are $a^m \mapsto a^{\pm m}$, $b^n \mapsto b^{\pm n}$ so we must have

$$a'^m = a^{\varepsilon m} = b^{\varepsilon n} = b'^n$$

if $a \mapsto a'$, $b \mapsto b'$ is an automorphism of $G_{m,n}$.

By 7.1.7, the automorphism $\bar{a} \mapsto \bar{a}'$, $\bar{b} \mapsto \bar{b}'$ induced in $G_{m,n}/C$ must have the form $\bar{a}' = \bar{t}^{-1}\bar{a}^r\bar{t}$, $\bar{b}' = \bar{t}^{-1}\bar{b}^s\bar{t}$. Now $\bar{a}' = \bar{t}^{-1}\bar{a}^r\bar{t} = \bar{t}^{-1}Ca^r\bar{t} = \{t^{-1}a^{r+hm}t\}$, where $h = 0, \pm 1, \pm 2, \ldots$ and t is a representative of $\bar{t}$, so a' is therefore some particular $t^{-1}a^{r+hm}t$. Similarly, $b' = t^{-1}b^{s+kn}t$ for some integer k. Then it is easy to see, by considering exponent sums, that

$$a'^m = a^{\varepsilon m} = b^{\varepsilon n} = b'^n$$

only if $r + hm = \varepsilon = s + kn$. □

EXERCISE 7.1.8.1 (Schreier 1924). Show that the automorphism group of $G_{m,n}$ for $m \neq n$ has presentation

$$\langle \alpha, \beta, \gamma; \alpha^m, \beta^n, \gamma^2 \rangle,$$

where α, β, and γ are

$$a \mapsto a, \qquad b \mapsto a^{-1}ba,$$
$$a \mapsto b^{-1}ab, \qquad b \mapsto b,$$
$$a \mapsto a^{-1}, \qquad b \mapsto b^{-1},$$

respectively.

7.1.9 Nonexistence of an Ambient Isotopy between the Two Trefoil Knots

Knowing that the automorphism $a_1 \mapsto a_2$, $b_1 \mapsto b_2$ of $\pi_1(\mathsf{R}^3 - \mathscr{R}_1)$ has the form

$$a_2 = t^{-1}a_1^{\varepsilon}t, \qquad b_2 = t^{-1}b_1^{\varepsilon}t$$

we can now write Equations (1), (2) of 7.1.6 as

$$\begin{aligned} a_1^2 &= (t^{-1}a_1^{\varepsilon}t)^{-2}(t^{-1}a_1^{-\varepsilon}tt^{-1}b_1^{\varepsilon}t)^n \\ &= t^{-1}a_1^{-2\varepsilon}(a_1^{-\varepsilon}b_1^{\varepsilon})^n t, \end{aligned} \tag{1}$$

$$\begin{aligned} a_1^{-1}b_1 &= (t^{-1}a_1^{\varepsilon}t)^{-1}(t^{-1}b_1^{\varepsilon}t) \\ &= t^{-1}a_1^{-\varepsilon}b_1^{\varepsilon}t. \end{aligned} \tag{2}$$

Since the defining relation of $\pi_1(\mathsf{R}^3 - \mathscr{R}_1)$ is $1 = a_1^2 b_1^{-3}$, it follows that in any expression equal to 1

$$\text{exponent sum of } a_1 = 2k$$

$$\text{exponent sum of } b_1 = -3k$$

for some integer k. Applying this to (2) we find $\varepsilon = 1$, then applying it to (1) after $\varepsilon = 1$ is substituted, we get $n = -12$, so that (1) becomes

$$a_1^2 = t^{-1} a_1^{-2} (a_1^{-1} b_1)^{-12} t \tag{3}$$

But (3) is *not* true in $\pi_1(\mathsf{R}^3 - \mathscr{R}_1) = G_{2,3}$; it is not even true in $G_{2,3}/C$, in which

$$a_1^2 = t^{-1} a_1^{-2} (a_1^{-1} b_1)^{-12} t$$

iff

$$1 = t^{-1} (a_1^{-1} b_1)^{-12} t \quad \text{because } a_1^2 = 1$$

iff

$$1 = (a_1^{-1} b_1)^{-12}$$

—which is clearly false.

This contradiction completes the proof that the two trefoil knots are not equivalent. In his proof, Dehn drew on a solution for the word problem for $G_{2,3}$ which he had already found in Dehn 1910. The method can be generalized to show that no torus knot is equivalent to its own mirror image. The exercise below shows that some knots are equivalent to their mirror images—such knots are called *amphicheiral* following Listing 1847—and hence are not torus knots.

EXERCISE 7.1.9.1. Show that the knot in Figure 224 is amphicheiral.

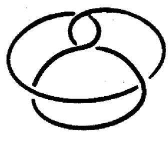


Figure 224

7.2 Cyclic Coverings

7.2.1 Introduction

Knot groups in general have not proved to be nearly as tractable as the torus knot groups. A solution of the word problem for all knot groups was obtained by Waldhausen 1968b, and the isomorphism problem has been solved recently by an extension of the same methods (Waldhausen 1978), but both are very intricate. In trying to distinguish knots the approach has therefore been to use less discriminating, but easily computable invariants.

The most important of these are the homology invariants of cyclic covering spaces. As we mentioned in 1.1.4, there are good reasons for generalizing the notion of Riemann surface to branched covers of $\mathbf{S}^3$, in which case the branch set becomes a set of closed curves which may be knotted or linked. Heegaard 1898 was the first to observe that knots and links give rise to interesting 3-dimensional manifolds in this way, in particular lens spaces (though they had not yet received that name).

Heegaard's examples in fact suggest that *torsion* is a convenient distinguishing feature of the covering spaces. He found that the 2-sheeted cover of $\mathbf{S}^3$ branched over the trefoil is the (3, 1) lens space, so this distinguishes the trefoil knot from the circle, whose branched covers are evidently $\mathbf{S}^3$ itself. However, Heegaard was not interested in distinguishing knots, nor had he quite isolated the concept of torsion (which was left to Poincaré), so the potential of his discovery remained unfulfilled until Alexander revived it about 20 years later.

7.2.2 The 2-sheeted Cover of $\mathbf{S}^3$ Branched over the Trefoil Knot

Heegaard gives a rather complicated method for reducing the branched cover to what we now call its *Heegaard diagram* (see next chapter), then merely quotes the results of applying it in a few special cases. We shall demonstrate his result for the 2-sheeted cover of the trefoil—the (3, 1) lens space—by comparing it with the lens space definition given in 4.2.8.

Rather than use Heegaard's cone construction, we span the trefoil knot $\mathscr{K}$ by a *nonsingular* surface in $\mathbf{S}^3$ (an idea due to Seifert 1934) which serves as a "door" through which we pass to the second sheet. The simplest such surface is a Möbius band $\mathscr{M}$ with three half twists (Figure 225). It is important to remember that when we cut $\mathbf{S}^3$ along $\mathscr{M}$ there are "two sides" to the cut,

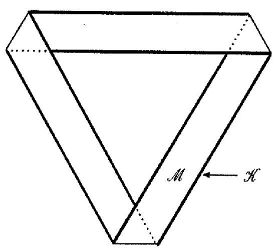


Figure 225

which meet along $\mathscr{K}$. These two sides form a twisted annulus (the double cover of the Möbius band) whose edges are sewn together along $\mathscr{K}$, and the part of $\mathbf{S}^3$ it encloses is a solid ring with $\mathscr{K}$ as a (2, 3) torus curve on its surface (Figure 226). Notice how a latitude curve l on this ring transfers back to the original surface, namely Figure 227. When we press this curve through to the other side of the cut, to see how it looks on the boundary surface of the second sheet, we get Figure 228—a (1, 3) torus curve. Thus we have that each sheet is an $\mathbf{S}^3$ with a solid ring drilled out of it, but the boundary tori are identified in such a way that a (1, 0) latitude curve on one is mapped to a (1, 3) curve on the other.

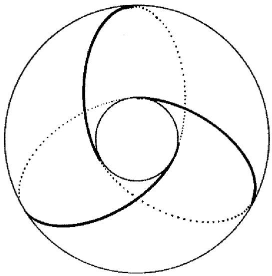


Figure 226

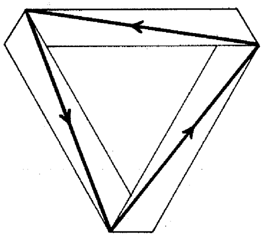


Figure 227

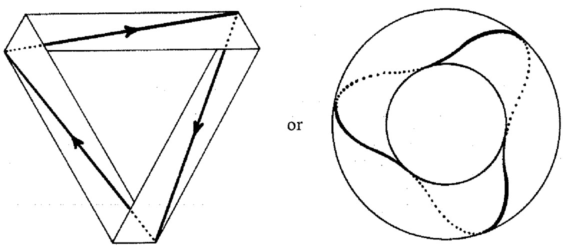


Figure 228

But for any unknotted ring $\mathcal{R}$, $\overline{\mathbf{S}^3 - \mathcal{R}}$ is itself a ring $\mathcal{R}'$, and an (m, n) curve on $\mathcal{R}$ becomes an (n, m) curve on $\mathcal{R}'$ (see Exercise 7.2.2.1 below), so another way to describe the covering space is: the union of two rings $\mathcal{R}'_1$, $\mathcal{R}'_2$ glued together so that a meridian on one is identified with a (3, 1) curve on the other. As we shall see in 8.3.2, this determines the manifold uniquely, but for the present we are content to show that the (3, 1) lens space can also be described in this way.

Take a lens-shaped solid with upper and lower faces divided into three equal sectors, and identify the upper sectors with the lower after a twist of $2\pi/3$ (Figure 229). Now if we take the "core" out of the lens and draw a curve p round its middle (Figure 230) the result is a solid ring (since top and bottom faces are identified) with a meridian curve p. The remaining portion of the lens is split into three wedges by vertical cuts x, y, z through the boundaries of the sectors (Figure 231). The arcs identified with the three pieces of p are also shown. We now identify the 1, 2, 3 regions on the wedges, giving Figure

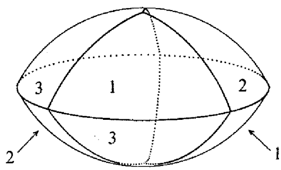


Figure 229

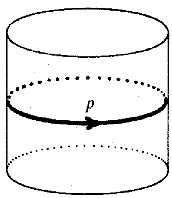


Figure 230

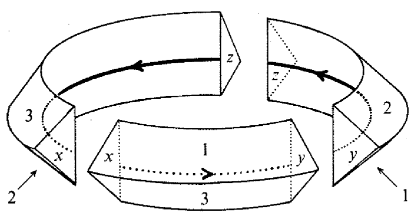


Figure 231

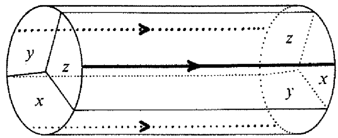


Figure 232

232, which becomes a solid ring when x, y, z are rejoined, and the arcs join up to form a (3, 1) torus curve! □

EXERCISE 7.2.2.1. Viewing $\mathbf{S}^3$ as the (1, 0) lens space, use the core construction to show that it is the union of two solid rings, $\mathscr{R}_1$, $\mathscr{R}_2$ and that an (m, n) torus curve on the boundary of $\mathscr{R}_1$ is an (n, m) torus curve on the boundary of $\mathscr{R}_2$.

EXERCISE 7.2.2.2. What space is obtained as the 2-sheeted branched cover over the $(2, 2n + 1)$ torus knot? Deduce that there are infinitely many knots.

7.2.3 Alexander's Results

Heegaard's result lay dormant (although noted by Tietze 1908) until the publication of the French translation of his thesis in 1916. The translation was checked for mathematical soundness by J. W. Alexander, fresh from his work on homology groups, and we may surmise that the collision of these ideas led to the fruitful discoveries which were to follow. Alexander must also have read Tietze 1908 at this time, because in short order he disposed of two of the most important of Tietze's conjectures: Alexander 1919a shows that there are nonhomeomorphic lens spaces with the same group, while Alexander 1919b proves that any orientable 3-manifold is a branched cover of $\mathbf{S}^3$. Later in 1920 he finally took the cue from Heegaard's example and looked for torsion in cyclic covers of $\mathbf{S}^3$ branched over various knots.

The torsion of a cyclic cover is obviously an invariant of the knot used as branch set so, with the known computibility of the homology groups, one had the first computable knot invariants.

With hindsight, and especially with the simplifying device of Seifert surfaces, it is not hard to see how the twisted position of a knot in $\mathbf{S}^3$ might induce torsion in its cyclic covers. Nevertheless, the extent to which different knots induce different torsion is remarkable. Alexander found that by using just 2- and 3-sheeted covers he was able to distinguish all knots with up to eight crossings. His results were delivered in a paper to the U.S. National Academy of Sciences in 1920 but not published until Alexander and Briggs 1927.

Alexander computed the first homology group H_1 from a cell decomposition of the cover obtained by cutting $\mathbf{S}^3$ along a surface rather like the Heegaard cone. The computation of H_1 from a cell decomposition is just

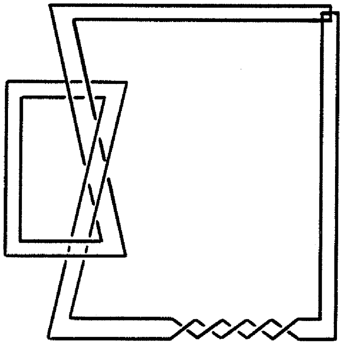


Figure 233

like the computation of π_1 (4.1.6) except that one allows the generators to commute.

It should be mentioned that not all knots have torsion in their cyclic covers, so Alexander's method does not invariably distinguish knots, even from the trivial knot. Examples with no torsion are "doubled knots" such as that shown in Figure 233 (see Rolfsen 1976, p. 157). The doubling back and twisting of the thread cancels the torsion induced by its trefoil-like shape.

7.2.4 Reidemeister's Subgroup Method

It is perhaps fortunate that Alexander put off publication of his results, because of the way Reidemeister 1927 came to rediscover them. Reidemeister's method differs from Alexander's in two small but significant ways:

(i) The knot $\mathscr{K}$ is removed from the space so that one has an *unbranched* cover of the knot complement $\mathbf{S}^3 - \mathscr{K}$.
(ii) The fundamental group π_1 of the cover is computed first, then abelianized to give H_1.

Since (ii) is a quite unnecessary detour it is clear that Reidemeister was primarily interested in handling noncommutative groups, while the fortunate choice of an unbranched cover meant that π_1 of the cover was in fact a *subgroup* of $\pi_1(\mathbf{S}^3 - \mathscr{K})$. Thus the stage was set for Reidemeister's subgroup method, and its covering space interpretation.

Reidemeister gives the method first (essentially the Reidemeister–Schreier process of 4.3.8, without the Schreier condition on coset representatives), then the following geometric interpretation. To form the m-sheeted cover of $\mathbf{S}^3 - \mathscr{K}$ we take m copies of it, say $(\mathbf{S}^3 - \mathscr{K})_1, \ldots, (\mathbf{S}^3 - \mathscr{K})_m$, joined along copies $\mathscr{C}_1, \ldots, \mathscr{C}_m$ of the Heegaard cone $\mathscr{C}$, which serve as doors from

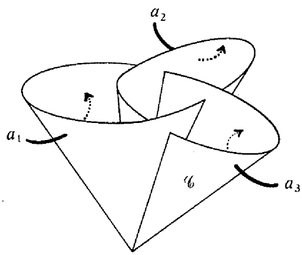


Figure 234

one sheet to the next (Figure 234). The Wirtinger generators $a_1, a_2, \ldots$ can be viewed as the basic routes for passing from the sheet $(\mathbf{S}^3 - \mathscr{K})_i$ to $(\mathbf{S}^3 - \mathscr{K})_{i+1}$, or from $(\mathbf{S}^3 - \mathscr{K})_m$ to $(\mathbf{S}^3 - \mathscr{K})_1$. It follows that the *closed* paths in the covering space $\widetilde{(\mathbf{S}^3 - \mathscr{K})}$ are exactly those which cover a path

$$a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \cdots a_{i_k}^{\varepsilon_k}$$

in $\mathbf{S}^3 - \mathscr{K}$ whose exponent sum $\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_k$ is an integer multiple of m.

It is clear that the null-homotopic paths in $\widetilde{(\mathbf{S}^3 - \mathscr{K})}$ are exactly those which cover null-homotopic paths in $\mathbf{S}^3 - \mathscr{K}$, so $\pi_1\widetilde{(\mathbf{S}^3 - \mathscr{K})}$ is in fact isomorphic to the subgroup of $\pi_1(\mathbf{S}^3 - \mathscr{K})$ generated by the above elements. This subgroup is just the kernel of the homomorphism

$$\phi\colon \pi_1(\mathbf{S}^3 - \mathscr{K}) \to \mathbf{Z}_m,$$

where $\mathbf{Z}_m$ is the cyclic group of order m and ϕ is defined for any element x by taking the exponent sum of x and reducing it mod m. This situation is indeed ripe for the Reidemeister–Schreier process, since coset representatives are readily computed.

Exercise 7.2.4.1. Define the above homomorphism ϕ independently of the presentation of $\pi_1(\mathbf{S}^3 - \mathscr{K})$.

Exercise 7.2.4.2. Show that π_1 of the m-sheeted cyclic cover of $\mathbf{S}^3$ branched over $\mathscr{K}$ is obtained from $\pi_1\widetilde{(\mathbf{S}^3 - \mathscr{K})}$ by adding the relation

$$A_1 = 1,$$

where A_1 is the element of $\pi_1\widetilde{(\mathbf{S}^3 - \mathscr{K})}$ corresponding to a_1^m in $\pi_1(\mathbf{S}^3 - \mathscr{K})$. Show also that this relation implies

$$A_i = 1 \qquad i = 2, 3, \ldots,$$

where A_i is the element corresponding to a_i^m.

7.2.5 The Knot Problem

As mentioned in 0.1.1, the homeomorphism problem for knot complements has been solved only recently, and with great difficulty; a full presentation of the result would be a book in itself. A key ingredient in the algorithm is a method of Haken 1961 for manipulating surfaces inside 3-manifolds. In particular, to recognize whether a knot $\mathscr{K}$ is trivial, Haken spans it by a surface $\mathscr{S}$, then tries to reduce the genus of $\mathscr{S}$ as far as possible. Haken's method guarantees that an $\mathscr{S}^*$ with minimal genus will eventually be found; then $\mathscr{K}$ is trivial if and only if genus $(\mathscr{S}^*) = 0$. A simplification of Haken's 130-page proof was given by Schubert 1961, but it is still difficult.

One of the few others to master the method was Waldhausen, who used it in Waldhausen 1968b to solve the word problem for all knot groups. Waldhausen's algorithm applies uniformly to all knot groups, so one can use it to decide whether a given knot group is abelian; namely, see whether all commutators of generators equal 1. This fact also implies an algorithm for the trivial knot, because of a classical criterion of Dehn 1910: a knot $\mathscr{K}$ is trivial if and only if $\pi_1(S^3 - \mathscr{K})$ is abelian. Dehn's criterion depends on the notorious "Dehn's lemma," a result about manipulating surfaces in 3-manifolds which first revealed how difficult such questions were. One of Dehn's manipulations was incorrect, and the result was not reinstated until finally proved by Papakyriakopoulos 1957.

The step to the general knot problem, strangely enough, depended on further progress in 2-manifold topology. This was finally achieved when Hemion 1979 solved the conjugacy problem for mapping class groups. The method for distinguishing one knot complement from another was then sketched by Waldhausen 1978, using a combination of Hemion's result with the Haken method.

A recently announced result of Thurston offers a more elegant and classical method for recognizing the trivial knot and solving the word problem (without settling the knot problem, however). Thurston claims to have proved the long-standing conjecture that knot groups are *residually finite.* A group G is residually finite if, for each $g \in G, g \neq 1$, there is a homomorphism of G onto a finite group which does not map g to 1. It is easy to show that any finitely presented residually finite group has a solvable word problem, in fact by a uniform procedure, so we also have an algorithm for deciding whether a knot group is abelian (see Exercise 7.2.5.1 below).

These results seem to vindicate Max Dehn's confidence in the knot group as a tool for understanding knots.

EXERCISE 7.2.5.1. (1) Show how to effectively enumerate all homomorphisms of a finitely presented group G onto finite groups, using (say) groups of finite permutations.

(2) Deduce that if G is residually finite one can effectively enumerate all words in the generators of G which are $\neq 1$.

(3) Deduce an algorithm for the word problem for G.

7.3 Braids

7.3.1 The Closed Braid Form of a Knot

The theory of braids was introduced by Artin 1926 as a possible approach to the knot problem. It had been observed by Alexander 1923c that any knot had a certain normal form, called a *closed braid*. A knot $\mathscr{K}$ is in closed braid form if there is a straight line $\mathscr{L}$, called the *axis*, such that the vector from $\mathscr{L}$ to a point P on $\mathscr{K}$ rotates in a fixed sense as P traverses $\mathscr{K}$ in a fixed sense. There is always a corresponding projection of $\mathscr{K}$, with $\mathscr{L}$ represented by a single point O, such that the radius vector OP rotates in a fixed sense. For example, the usual projection Figure 235, of the trefoil knot has this property. The usual projection, Figure 236(1), of the figure-eight knot is not in braid form, but it has the braid form (2). Alexander proved that any knot (or link) has a closed braid form. Kneser (as reported in Artin's paper) observed that a proof of this fact also followed from the result of Brunn 1897 that every knot has a projection with just one, necessarily multiple, crossing.

The following proof of Alexander's theorem includes a proof of Brunn's theorem, hence it may be what Kneser had in mind.

Step 1. Represent the knot or link $\mathscr{K}$ as a finite set of simple arcs in the plane connected by small "bridges" at the crossings. For example, one such representation of the trefoil knot is shown in Figure 237. Now make a "rubber-sheet" deformation of the plane which pulls the endpoints of the bridges onto a straight line $\mathscr{L}$ we shall call the axis (Figure 238). $\mathscr{L}$ divides $\mathscr{K}$ into a number of simple arcs $\mathscr{A}_1, \ldots, \mathscr{A}_m$ (nine in our example).

Step 2. By slightly rotating the half-plane which contains $\mathscr{A}_i$ about $\mathscr{L}$, we can move each $\mathscr{A}_i$ into a different half-plane without changing the knot

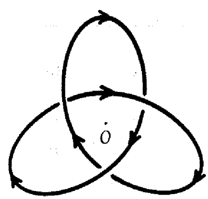


Figure 235

(1)

(2)

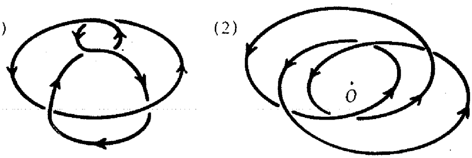


Figure 236

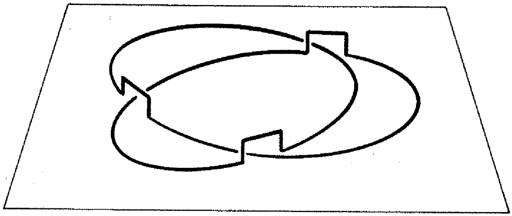


Figure 237

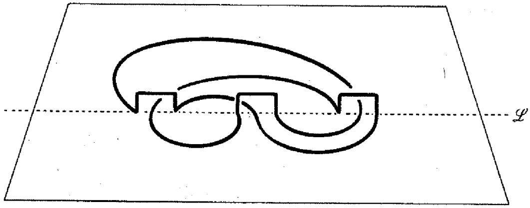


Figure 238

or link $\mathscr{K}$. Now that $\mathscr{A}_i$ has a half-plane to itself, it can be isotopically deformed into an arc $\mathscr{A}'_i$ consisting of three sides of a square. Looking along $\mathscr{L}$, we now see m line segments radiating from a single point.

Step 3. We orient $\mathscr{K}$ and slightly twist the half-plane containing $\mathscr{A}'_i$ so that the outer edge of $\mathscr{A}'_i$ goes out of parallel with $\mathscr{L}$ (Figure 239). The modified arcs $\mathscr{A}'_i$, which we call $\mathscr{A}''_i$, then present only a single multiple point when viewed along $\mathscr{L}$, so we have proved Brunn's theorem. The direction of twist is chosen so that each $\mathscr{A}''_i$ has a clockwise orientation when viewed from a point P on $\mathscr{L}$ to one side of $\mathscr{K}$.

Step 4. Take a cylinder with axis $\mathscr{L}$ and radius small enough not to include any $\mathscr{A}''_i$ and replace each V-shaped segment of $\mathscr{K}$ inside it by a circular arc on the surface of $\mathscr{L}$. Choose from the two possible arcs the one which presents a clockwise orientation when viewed from P. Thus if $\mathscr{A}''_i$ is followed by $\mathscr{A}''_j$, the two cases are Figure 240 and Figure 241.

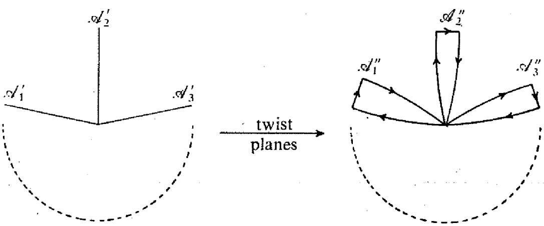


Figure 239

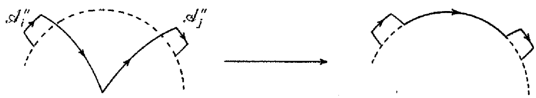


Figure 240

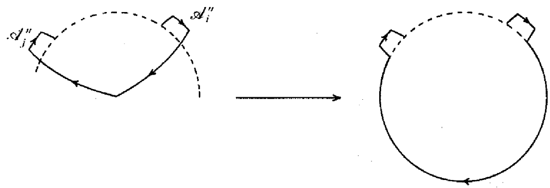


Figure 241

Since the cross-section of the cylinder containing this V-shaped segment contains no other part of $\mathcal{K}$, the operation can be realized by an isotopy of $\mathcal{K}$ and hence does not change its knot type. When viewed from P, which is a finite distance away, the arcs on the cylinder will not be superimposed, and we shall see $\mathcal{R}$ as a closed braid. □

EXERCISE 7.3.1.1. Show that the closed braid form of a knot $\mathcal{K}$ can be further specialized to consist of a spiral AB together with a single wandering strand of the braid, BA, which subtends an angle $<2\pi$ at the axis, for example Figure 242. (Suggestion: Find a suitable way to "comb" the braid.) This result was stated by Alexander 1932. He also noted that $\mathcal{K}$ is trivial if BA crosses each strand of the spiral only once, and conjectured that if BA could not be reduced to this form by "obvious transformations," then $\mathcal{K}$ was nontrivial.

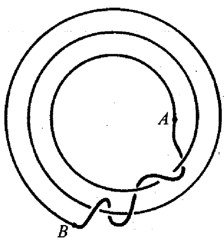


Figure 242

7.3.2 Markov Operations

It is clear that any knot $\mathcal{K}$ has infinitely many different closed braid forms, in particular the operation in Figure 243 changes the number of loops around the axis without altering the knot type. This operation and its inverse are

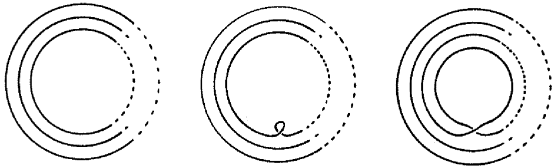


Figure 243

called *Markov operations*. Markov operations differ from proper braid operations (to be discussed in 7.3.4) in that the braid form is not maintained relative to a fixed axis throughout the operation.

Markov 1936 showed that if $\mathscr{B}_1, \mathscr{B}_2$ are two braid forms of the same knot $\mathscr{K}$, then $\mathscr{B}_1$ is convertible to $\mathscr{B}_2$ by a finite sequence of Markov operations and proper braid operations. The proof is quite lengthy and the only published version is in Birman 1975.

With Markov's theorem, the problem of deciding whether knots $\mathscr{K}_1, \mathscr{K}_2$ are equivalent is the same as deciding whether their braid forms $\mathscr{B}_1, \mathscr{B}_2$ (obtained, say, by the method of 7.3.1) are equivalent under Markov operations and proper braid operations. The problem of equivalence under proper braid operations was posed by Artin 1926, and an algorithm for it was found by Garside 1969, however the effect of adding Markov operations has so far been too difficult to handle. Artin confined himself mainly to *open* braids, which are what we usually mean by the term "braid," and found the elegant group-theoretic interpretation of them which follows.

7.3.3 The Braid Group B_n

To define a braid on n threads we take a rectangle with sets of points $A_1, \ldots, A_n$ and $A'_1, \ldots, A'_n$ on its two horizontal sides at corresponding positions, as in Figure 244. Each A_i is connected to A'_{j_i} by a simple polygonal arc in such a way that

(i) No two arcs meet (in particular, each goes to different endpoints).
(ii) Each arc meets a given horizontal plane in at most one point.

The arcs are called the *threads* of the braid. By placing one rectangle on top of another and then erasing the common edge we obtain a natural *product* of braids, which is obviously associative.

Two braids are considered to be the same if one can be deformed into the other by an isotopy in $\mathbb{R}^3$. There is no loss of generality in assuming the figure remains a braid throughout the deformation. Such deformations

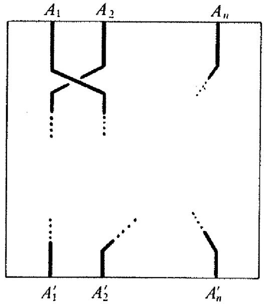


Figure 244

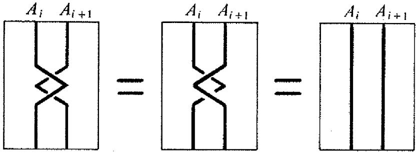


Figure 245

enable each braid to be brought into a standard form whose projection has the following properties

(iii) No multiple point of the projection is more than double (a "crossing").
(iv) There is at most one crossing on each horizontal line.

Then if we denote a braid with a crossing of the ith thread over the $(i + 1)$th by σ_i, and one with a crossing of the ith thread *under* the $(i + 1)$th by σ_i^{-1}, any braid in standard form can be described as a product of σ_i's and σ_i^{-1}'s by reading its crossings from the top down.

We observe that $\sigma_i\sigma_i^{-1} = \sigma_i^{-1}\sigma_i =$ a braid with no crossings (Figure 245). Hence any braid β can be cancelled by its formal inverse β^{-1} which results from writing β in reverse order and reversing all exponents. This yields the following proposition:

The braids on n threads form a group B_n under the product operation. The identity element 1 is the braid with no crossings and the generators are the braids σ_i in which the only crossing is one at which the ith thread passes over the $(i + 1)$th. The inverse σ_i^{-1} is the braid whose only crossing has the ith thread passing under the $(i + 1)$th.

EXERCISE 7.3.3.1. Describe the usual hair braids (Figure 246) as elements of B_3.

Figure 246

7.3.4 Defining Relations of B_n

Any deformation of a braid can be realized by a series of small deformations of the following two types (called *braid operations*).

(i) Pulling one thread across its neighbour (Figure 247): This corresponds to insertion of a term $\sigma_i\sigma_i^{-1}$ or $\sigma_i^{-1}\sigma_i$ in the product describing the braid, hence it has no consequences additional to those we already know from the group property of braids.

(ii) Sliding a crossing up or down (Figures 248–250): The simplest effect is to exchange two consecutive terms in the product (Figure 248). This is possible just in case $k \neq i - 1, i + 1$. When the crossings have successive indices the possible situations are shown in Figures 249, 250. In both cases one obtains the relation

$$\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}.$$

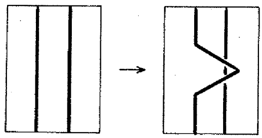
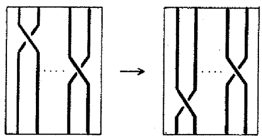


Figure 247



$\sigma_i\sigma_k = \sigma_k\sigma_i$

Figure 248

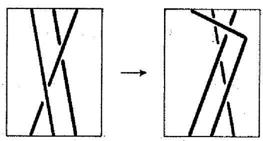


Figure 249

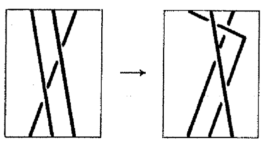


Figure 250

Since the braid operations suffice to produce all equivalents of a given braid, it follows that B_n is generated by $\sigma_1, \ldots, \sigma_{n-1}$ with the defining relations

$$\sigma_i \sigma_k = \sigma_k \sigma_i \qquad k \neq i-1, i+1$$

$$\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}.$$

EXERCISE 7.3.4.1. Show that B_3 is the same as the trefoil knot group.

EXERCISE 7.3.4.2. What group results from B_n when we add the relations $\sigma_i^2 = 1$ for $i = 1, \ldots, n-1$?

7.3.5 Artin's Solution of the Word Problem for B_n

In geometric terms the problem of deciding whether a word w equals 1 in B_n is just the problem of deciding whether a given braid on n threads is trivial. This geometric problem has an evident geometric solution, namely, see if each thread of the braid can be straightened by "combing." More precisely, take a loop t_k round the kth thread and pull it through the braid from top to bottom. If, for each k, t_k emerges at the bottom as a loop around the kth thread only, then the braid is trivial, and conversely.

An elegant algebraic formulation of this process was given by Artin 1926. He actually closes the braid by connecting top to bottom by circular arcs, and throws the loops $t_1, \ldots, t_n$ around its threads at the top. However, his subsequent interpretation of $t_1, \ldots, t_n$ as generators of a free group is easier to defend if instead the threads of the braid are prolonged straight to infinity in both directions, which we shall therefore do (Figure 251). It is then easy to show that $t_1, \ldots, t_n$ are indeed free generators for $\pi_1(\mathsf{R}^3 - \mathscr{B})$, where $\mathscr{B}$ denotes the infinitely extended braid. Furthermore, free generators $t'_1, \ldots, t'_n$ are obtained by throwing loops around the threads at any level. Passing

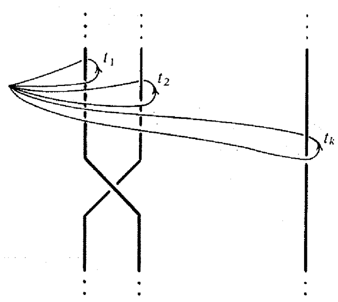


Figure 251

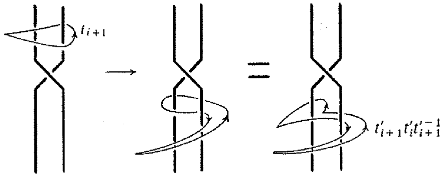


Figure 252

from the t_j generators to expressions for them in the t'_j generators, then dropping primes, defines an automorphism of F_n, the free group of rank n.

The automorphisms which occur are generated by automorphisms σ_i^* induced by pulling $t_1, \ldots, t_n$ past the level of a crossing σ_i. It is clear that

$$\sigma_i^*(t_j) = t_j \quad \text{if } j \neq i, i+1,$$

$$\sigma_i^*(t_i) = t_{i+1},$$

and Figure 252 shows that

$$\sigma_i^*(t_{i+1}) = t_{i+1}t_it_{i+1}^{-1}.$$

The automorphism corresponding to a given braid

$$\sigma_{i_1}^{\varepsilon_1} \cdots \sigma_{i_m}^{\varepsilon_m} = \beta$$

is therefore

$$(\sigma_{i_1}^{\varepsilon_1})^* \cdots (\sigma_{i_m}^{\varepsilon_m})^* = \beta^*$$

and the braid is trivial just in case this automorphism sends each t_k to t_k.

To decide whether this is the case, one computes $\beta^*(t_k)$ for each k and then uses free reduction to see whether the result equals t_k. In effect, Artin has reduced the word problem for B_n to the easy word problem for F_n.

Exercise 7.3.5.1. Show that the problem of deciding whether two closed braids α, β of n threads are equivalent is the same as the conjugacy problem for B_n, that is, the problem of deciding whether there is an element γ in B_n such that

$$\alpha = \gamma\beta\gamma^{-1}.$$

Exercise 7.3.5.2 (Artin 1926). Using the closed braid representation of a knot, and a construction of the closed braid by identification of the ends of an open braid β, deduce that any knot group has a presentation of the form

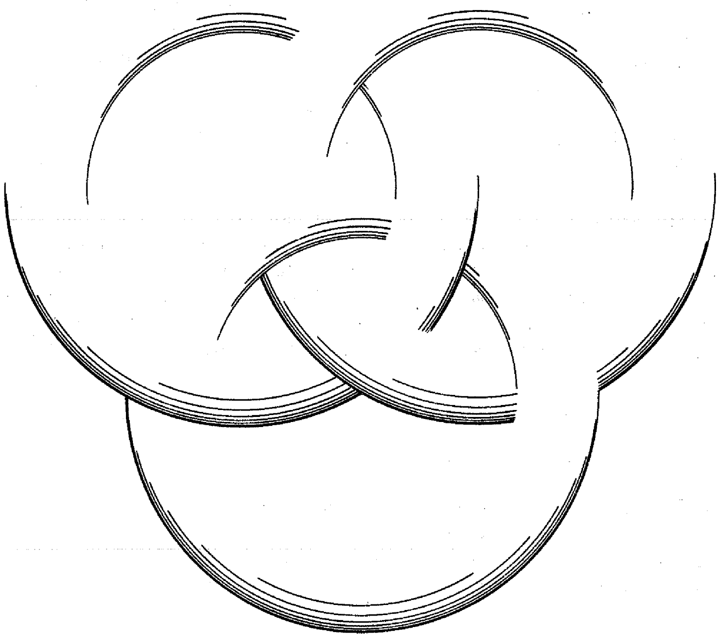
$$t_i = \beta^*(t_i) \qquad i = 1, \ldots, n,$$

where n is the number of threads in β.

Exercise 7.3.5.3. Compute presentations of the groups of the 2-crossing link and the trefoil knot by the method of 7.3.5.2.

CHAPTER 8

Three-Dimensional Manifolds



8.1 Open Problems in Three-Dimensional Topology

8.1.1 Review of 2-manifolds

In Chapters 1 and 6 we have seen how to solve a wide variety of problems concerning 2-manifolds. The success of 2-manifold topology is quite surprising when compared with dimension 3, where almost no algorithms are known. To see the contrast more clearly, let us first review what we know about 2-manifolds.

The homeomorphism problem was essentially solved by the nineteenth-century mathematicians who thought of handles, crosscaps, and the relation

$$\text{handle} + \text{crosscap} = 3 \text{ crosscaps}.$$

The classification theorem (1.3.7) confirms that the number of handles (or crosscaps, respectively) and the number of boundary curves determine the homeomorphism type, and also give an algorithm for computing it. A convenient alternative is to compute the Euler characteristic and orientability character, and in particular the 2-sphere is recognizable as the only surface with Euler characteristic 2.

Various equivalence relations between curves on a surface $\mathscr{F}$—homology, homotopy, and isotopy—can be decided by solving the corresponding algebraic problems in $H_1(\mathscr{F})$ and $\pi_1(\mathscr{F})$ (the word problem and the conjugacy problem). It can also be decided whether a given homology or homotopy class contains a simple curve, in other words, an embedded submanifold $\mathbf{S}^1$ of $\mathscr{F}$. Equivalence of two embeddings of $\mathbf{S}^1$ up to homeomorphism of $\mathscr{F}$ is decided by the Euler characteristics and orientability character of the pieces into which they divide $\mathscr{F}$, and up to isotopy by the solution of the conjugacy problem and Baer's theorem.

The algorithms used to solve these problems apply whether the 2-manifold is given as a Riemann surface, polygon schema, or simplicial complex.

It is worth mentioning that the homeomorphism problem for 2-complexes can also be solved (Papakyriakopoulos 1943. A more accessible proof is Whittlesey 1958). This is despite the fact that any finitely presented group G can be realized as π_1 of a 2-complex, so there are 2-complexes with unsolvable contractibility problem.

8.1.2 Methods of Constructing 3-manifolds

To begin seeing 3-manifolds from a combinatorial point of view it is probably most convenient to consider simplicial decompositions. This is completely general by virtue of the Moise triangulation theorem (0.2.5). A 3-manifold is then a union of solid tetrahedra with disjoint interiors, with at most two tetrahedra meeting at each face (*exactly* two for a closed manifold), and finitely many at each edge and vertex. Furthermore, the neighbourhood surface of each vertex must be a 2-sphere, so that each point in the complex has a neighbourhood homeomorphic to the 3-ball (see also 8.2).

If the complex is finite, then all the above properties are decidable, so we have an algorithm for deciding whether a finite 3-complex is a manifold (and of course, for distinguishing between closed and bounded manifolds).

As in dimension 2, we can amalgamate simplexes until we have a single polyhedron, homeomorphic to a ball, with faces identified in pairs. This *cell decomposition* method of constructing 3-manifolds is therefore completely general and could serve as an alternate definition provided we include the statement that the neighbourhood surface of each vertex is a sphere.

The remaining methods of construction we shall mention yield only the orientable 3-manifolds, however these greatly overshadow the nonorientable 3-manifolds in importance. (Note that the lens spaces, including the (2, 1) lens space or *projective space*, are orientable.)

The first such method, the Heegaard splitting, decomposes the manifold into two pieces homeomorphic to subsets of R^3, namely handlebodies (Heegaard 1898, see also Dyck 1884). This is faintly analogous to the Clifford decomposition of an orientable surface into two pieces of R^2 (1.1.3). However, the perforated discs of a Clifford decomposition produce only one orientable 2-manifold, no matter how they are joined together, whereas two handlebodies can produce infinitely many different manifolds via different homeomorphisms between their boundary surfaces. It is fairly easy to show that every finite orientable 3-manifold has a Heegaard splitting (8.3.1), by starting with a simplicial decomposition.

The second construction of orientable 3-manifolds is by *surgery*, the first example of which was given by Dehn 1910. One removes some solid tori $\mathscr{T}_1, \ldots, \mathscr{T}_n$ from an S^3 and "sews them back differently," that is, identifies the boundary of the hole left by $\mathscr{T}_i$ with the boundary of another solid torus $\mathscr{T}'_i$ via a homeomorphism different from the one defined by the inclusion of $\mathscr{T}_i$ in S^3. To obtain all orientable 3-manifolds the $\mathscr{T}_i$ have to be knotted or linked in most cases. In 8.4 we shall follow Lickorish 1962 in deriving the surgery construction from a Heegaard splitting.

The third construction is by branched coverings of S^3, which has already been discussed in 1.1.4 as the 3-dimensional analogue of Riemann surfaces. Like surgery, it is generally based on knots or links, and can in fact be derived from the surgery construction (Lickorish 1973, see 8.5).

Each of the above methods yields a finite description of the 3-manifold which can be effectively translated into a simplicial decomposition. In fact the different forms of description are intertranslatable, which is often useful for showing that two manifolds are the same. The problem is that the same manifold has infinitely many descriptions, so we cannot always be sure whether different descriptions actually represent different manifolds.

EXERCISE 8.1.2.1. Give an algorithm for effectively enumerating all finite 3-manifolds (with repetitions).

8.1.3 The Homeomorphism Problem

Since we do not yet know how to recognize the simplest 3-manifold $\mathbf{S}^3$ (see 8.1.4) it must be admitted that the homeomorphism problem is far from being solved. The principal obstacles seem to be

(1) absence of plausible normal forms,
(2) lack of bounds on the length of constructions which convert one description of a manifold to another.

If we had plausible normal forms, comparable to handle and crosscap forms for 2-manifolds, a way to solve the homeomorphism problem might become evident. However, it should also be remembered that we had to discover the relation

$$\text{handle} + \text{crosscap} = 3 \text{ crosscaps}$$

before obtaining normal forms of 2-manifolds, and even if 3-manifolds possess a comparable set of building blocks, the relations between them may be unmanageable.

This is precisely the difficulty with the methods of construction described above. We know for example a set of "simple" relations between Heegaard splittings which enable us to enumerate all splittings in a given homeomorphism class (Singer 1933). Such results show that, like the word problem for groups, the homeomorphism problem for 3-manifolds is a recursively enumerable problem (0.4) but they do not help to solve it. To do that, one needs to know when to stop looking for a description D_2 among the infinitely many descriptions equivalent to a given D_1—it is a question of bounding the length of the search. For example, it may be possible to define a "complexity" $|D|$ of a simplicial decomposition D for which one can compute a bound $b(|D_1|, |D_2|)$ on the complexity of a common refinement of D_1, D_2. Then one could enumerate all simplicial decompositions of complexity $\leq b(|D_1|, |D_2|)$ and see if they included a common refinement of D_1, D_2. If not, then D_1, D_2 would represent nonhomeomorphic manifolds, by the *Hauptvermutung.*

If the homeomorphism problem is solvable, then some such computable bound must exist, and it will provide a solution whether or not normal forms are available.

The homeomorphism problem has been solved for the subcase of lens spaces (Reidemeister 1935). We shall see in 8.3 that the lens spaces are the 3-manifolds of *Heegaard genus* 1, that is, they possess Heegaard splittings into solid tori. In general, the Heegaard genus of a 3-manifold $\mathscr{M}$ is the minimum n for which $\mathscr{M}$ splits into handlebodies of genus n. We do not know how to compute it, and little is known about the $\mathscr{M}$ with Heegaard genus > 1.

8.1.4 Recognizing the 3-sphere

The problem of recognizing the 3-sphere is particularly tantalizing in view of the ease with which we can recognize the 2-sphere. However, all the methods which apply to the 2-sphere break down. Homology invariants are inadequate (Poincaré 1904), and so too are various methods which might be expected to simplify the description of a manifold until its spherical nature, or lack of it, became obvious.

One such method is called *shelling a simplicial decomposition*: one removes any simplex, then tries to remove a sequence of *free* simplexes until only one simplex is left. A free n-simplex is one which meets the boundary in a piece homeomorphic to an $(n - 1)$-ball, so its removal does not change the homeomorphism type, and hence a complete shelling guarantees that the initial manifold was a sphere. It is well known (and often used in proofs of the Euler polyhedron formula) that any triangulation of $\mathbf{S}^2$ can be shelled, so one way to decide whether a given 2-manifold is $\mathbf{S}^2$ is to systematically try to shell it in all possible ways.

This does not work for 3-manifolds because there are unshellable triangulations of $\mathbf{S}^3$ (W. B. R. Lickorish, private communication). The proof is a little long to include here, but one can get the idea of it from the construction of an unshellable 3-ball, which is easier (8.1.5).

An algorithm for the 3-sphere needs to be at least strong enough to recognize the trivial knot. The reduction of the latter problem to the former is obtained by taking two copies of

$$(\mathbf{S}^3 - \text{tubular neighbourhood } \mathcal{N} \text{ of the given knot})$$

and pasting them together so that meridian and latitude on one $\partial\mathcal{N}$ are identified with latitude and meridian on the other. It can be shown (using Dehn's lemma for example) that the resulting manifold is $\mathbf{S}^3$ if and only if the knot is trivial.

Exercise 8.1.4.1. Prove that any triangulation of the 2-sphere, or the disc, can be shelled.

8.1.5 An Unshellable Triangulation of the 3-ball

The first such example was discovered by Frankl 1931, and a simpler example has been given by Bing 1964. Bing's example shows knots once again causing trouble in 3-manifold topology.

Take a cube $\mathscr{C}$ with a knotted hole and plug the top end of the hole with a small cube $\mathscr{P}$, so that $\mathscr{C} \cup \mathscr{P}$ is a topological ball (Figure 253). We triangulate $\mathscr{C} \cup \mathscr{P}$ as follows. Divide $\mathscr{C}$ into small cubes the same size as $\mathscr{P}$ (assume the size and shape of the hole are chosen so as to make this possible), then triangulate each cube by dividing it into two triangular prisms and dividing each prism into three tetrahedra.

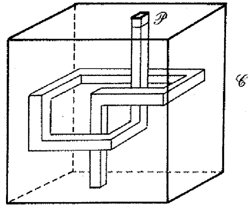


Figure 253

Now $\mathscr{C} \cup \mathscr{P}$ contains a knotted curve which lies in its boundary except for one interior edge, namely, a curve which pierces the plug but elsewhere travels on the surface of the solid. A curve of this description remains after the removal of a free simplex, and hence at each stage of shelling. But no such curve lies on the single simplex which is the final stage of shelling, so no shelling can be completed. □

EXERCISE 8.1.5.1. Close the Bing cube to a triangulation of $\mathbf{S}^3$ containing a *knotted triangle*. (The triangle is of course not the boundary of a simplex in the $\mathbf{S}^3$, but its edges are edges of the triangulation.)

8.1.6 The Poincaré Conjecture

Certainly the most famous problem in 3-manifold topology, the Poincaré conjecture, is not in itself an algorithmic problem. However, it has important implications for such problems and may be dependent on some of them for its solution. Poincaré asked whether every finite simply connected 3-manifold is an $\mathbf{S}^3$. An affirmative answer would reduce the problem of recognizing the 3-sphere to the problem of recognizing whether a 3-manifold group is trivial, and vice versa.

On the other hand, in searching for a counterexample to the Poincaré conjecture—an $\mathscr{M}$ such that $\pi_1(\mathscr{M}) = \{1\}$ and not homeomorphic to $\mathbf{S}^3$—the obstacle is precisely the lack of an algorithm for recognizing the 3-sphere. We can in principle enumerate all $\mathscr{M}$ such that $\pi_1(\mathscr{M}) = \{1\}$ (by enumerating, say, the simplicial decompositions which yield manifolds, computing their groups by 4.1.6, and seeing which presentations reduce to 1 by Tietze transformations), but we do not know how to enumerate all the $\mathscr{M}$ not homeomorphic to $\mathbf{S}^3$.

Interplay between the Poincaré conjecture and the problem of recognizing the 3-sphere also appears when we look at 2-spheres in a 3-manifold $\mathscr{M}$. Laudenbach 1974 shows that, assuming the Poincaré conjecture, an $\mathbf{S}^2$ which is homotopic to a point in $\mathscr{M}$ is also isotopic to a point. It then follows that the $\mathbf{S}^2$ will bound a 3-ball in $\mathscr{M}$ (Sanderson 1957), which can be decided if we have an algorithm for the 3-sphere.

8.1.7 The Word Problem for 3-manifold Groups

Many groups are known not to be π_1 of a 3-manifold, for example $\mathbf{Z} \times \mathbf{Z} \times \mathbf{Z} \times \mathbf{Z}$ (Stallings 1962), so 3-manifold groups may not be sufficiently general to include groups with unsolvable word problem. As mentioned in 7.2.1, Waldhausen solved the word problem for knot groups, and his solution in fact covers a large class of 3-manifold groups. The recent results of Thurston 1977 also solve the word problem in many 3-manifold groups. Both these authors use geometric methods, and Thurston's are reminiscent of classical surface topology.

Of course, to solve the word problem in $\pi_1(\mathscr{M})$ we only need to construct the universal covering space of $\mathscr{M}$. The trouble is that we do not know what the potential covering spaces might be in general. They include $\mathbf{S}^3$ and $\mathbf{R}^3$, but also more peculiar manifolds. Alexander 1932 gives the example of the manifold $\mathscr{M}$ obtained by identifying the inner and outer surfaces of a spherical shell. Its universal covering space $\tilde{\mathscr{M}}$ is homeomorphic to the open region of $\mathbf{R}^3$ between two concentric spheres. It is not known whether all universal covers are homeomorphic to submanifolds of $\mathbf{S}^3$.

EXERCISE 8.1.7.1. (1) Show that $\mathscr{M} = \mathbf{S}^1 \times \mathbf{S}^2$. (2) Why is $\tilde{\mathscr{M}}$ not homeomorphic to $\mathbf{S}^3$ or $\mathbf{R}^3$?

8.1.8 Above Dimension 3

It is not yet known whether all 4-manifolds can be triangulated. In any case, the problem of deciding whether a given 4-complex is a manifold awaits an algorithm for recognizing the 3-sphere, since the hard condition to check is whether the neighbourhood complex of each vertex is a 3-sphere.

We do know that any finitely presented group can be realized as π_1 of a 4-manifold (Dehn 1910), so there are 4-manifolds with unsolvable contractibility problem. A remarkable refinement of this result, due to Markov 1958, shows that *the homeomorphism problem for triangulated 4-manifolds is unsolvable* (see Chapter 9). Unsolvability follows in all dimensions > 4, so dimension 3 is our last chance for a positive result.

Finally, two negative results about dimensions ≥ 5 may be mentioned. The *Hauptvermutung* is false (Milnor 1961) and there is no algorithm for recognizing $\mathbf{S}^5$ (S. P. Novikov, a sketch of the proof is given in Volodin, Kuznetsov, and Fomenko 1974).

EXERCISE 8.1.8.1 (Seifert and Threlfall 1934). (1) Using a 4-dimensional form of the connected sum construction (Exercise 4.1.6.1) construct a 4-manifold $\mathscr{M}_n$ such that $\pi_1(\mathscr{M}_n) = F_n$ (free group of rank n).

(2) If p is a curve in $\mathscr{M}_n$, show that $\pi_1(\mathscr{M}_n) = \pi_1(\mathscr{M}_n - p)$ and likewise $= \pi_1(\mathscr{M}_n - \mathscr{N})$, where $\mathscr{N}$ is a tubular neighbourhood of p, that is, the set swept out by a small 3-ball moving with just its centre point on p.

(3) Identify the boundary $S^1 \times S^2$ of $\mathcal{N}$ with the boundary of a $B^2 \times S^2$. Show that $p = 1$ in π_1 of the resulting manifold.

(4) Deduce the construction of a 4-manifold to realize an arbitrary finitely presented group.

(5) Why does a similar 3-dimensional construction break down?

8.2 Polyhedral Schemata

8.2.1 Manifolds and Pseudomanifolds

Poincaré 1895 introduces the construction of 3-manifolds by identifying faces of simply-connected polyhedra, like the polygon schemata for 2-manifolds. The strict analogy with a polygon schema would be

(1) a finite set of polyhedra (topological 3-balls) called *cells*, with disjoint interiors,
(2) faces of cells identified in pairs, with vertices corresponding to vertices,
(3) resulting in a connected complex.

However, Poincaré observes that these conditions do not guarantee that the outcome will be a manifold. They *do* guarantee that the cells incident with a given edge form a closed cycle (analogous to the "umbrella" round a vertex in a polygon schema) so interior points of edges have 3-ball neighbourhoods, as of course do interior points of cells. But the neighbourhood surface of a vertex need not be a sphere—as we shall see in 8.2.3—and in this case the vertex will not have a 3-ball neighbourhood.

A complex satisfying (1), (2), (3) is called a (finite) 3-dimensional *pseudomanifold*, and it is a manifold if and only if it satisfies the additional condition

(4) The neighbourhood surface of each vertex is a 2-sphere.

Poincaré determines whether the neighbourhood surface is a sphere by computing its Euler characteristic. Visualizing the neighbourhood of points distant $\leq \varepsilon$ from a given vertex P, he observes that vertices on the boundary surface $\mathscr{F}_P$ of this neighbourhood correspond to edges containing P, edges on $\mathscr{F}_P$ correspond to faces containing P, and faces on $\mathscr{F}_P$ correspond to "corners" at P, that is, the vertices of the polyhedron schema which are identified in P.

It is then a matter of counting vertices, edges, and faces correctly by observing identifications in the schema. We shall not explain this in further detail, since it turns out to be even more convenient to compute the Euler characteristic of the pseudo-manifold itself.

EXERCISE 8.2.1.1. Show that the result of removing an ε-neighbourhood of each vertex in a pseudomanifold is a manifold with boundary.

8.2.2 The Euler Characteristic of a Pseudomanifold

Let V, E, F as usual denote the numbers of vertices, edges, and faces in a cell decomposition of a 3-dimensional pseudomanifold $\mathscr{M}$, and in addition let C denote the number of cells. Then $V - E + F - C$ is called the *Euler characteristic* $\chi(\mathscr{M})$ of $\mathscr{M}$, and $\chi(\mathscr{M}) = 0$ if and only if $\mathscr{M}$ is a manifold.

It is clear that elementary subdivision (0.2.4) preserves the Euler characteristic in three dimensions just as it does in two (1.3.8), hence we can assume that the subdivision of $\mathscr{M}$ is simplicial. The collection of tetrahedral cells, with their face identifications, will be called the *schema* of $\mathscr{M}$.

Let V', E', F' be the numbers of vertices, edges, and faces in the schema, *ignoring* identifications, and let v, e, f be the total numbers of vertices, edges, and faces in the neighbourhood surfaces of vertices of $\mathscr{M}$ (when identifications are taken into account).

$$v = 2E \tag{1}$$

because each edge of $\mathscr{M}$ accounts for two vertices on neighbourhood surfaces (near its two ends).

$$e = E' \tag{2}$$

because if we look at a vertex of a tetrahedron of the schema we see equal numbers of edges from the schema (drawn heavily) and in the neighbourhood surface—three each (Figure 254). If we survey all vertices in the schema, then all schema edges will be counted twice (since in a simplicial decomposition each edge has two endpoints), but so will all edges in the neighbourhood surfaces, since they are identified in pairs. Hence (2).

$$f = V' \tag{3}$$

because each face on a neighbourhood surface corresponds to the corner of a tetrahedron in the schema. Finally

$$F' = 2F \tag{4}$$

because faces of the schema are identified in pairs.

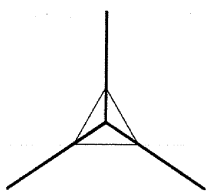


Figure 254

Now since the Euler characteristic of a closed surface is ≤ 2 and we have V neighbourhood surfaces,

$$v - e + f \leq 2V \tag{5}$$

and since the boundary of each of the C cells is a sphere,

$$V' - E' + F' = 2C. \tag{6}$$

Substituting v, e, f, F' from (1), (2), (3), (4) in (5) and (6) gives

$$2E - E' + V' \leq 2V \tag{7}$$

and

$$V' - E' + 2F = 2C. \tag{8}$$

Then subtracting (8) from (7) gives

$$2E - 2F \leq 2V - 2C.$$

That is,

$$V - E + F - C \geq 0$$

and $=$ holds just in case it holds in (5), namely if and only if all neighbourhood surfaces are spheres, which means $\mathscr{M}$ is a manifold. □

The above proof is adapted from Blackett 1967. The half which says 3-manifolds have Euler characteristic 0 can be derived more elegantly by Poincaré's method of *dual cell decomposition* (Poincaré 1899), designed to prove the more general property of Betti numbers mentioned in 5.1.1, and which shows in particular that any manifold of odd dimension has Euler characteristic 0.

The following exercise explores Poincaré's construction in dimension 3.

EXERCISE 8.2.2.1. Let $\mathscr{M}$ be a 3-manifold with a simplicial decomposition of V vertices, E edges, F faces, and C cells.

(1) Construct a new cell decomposition of $\mathscr{M}$ which has a new vertex in the interior of each old cell, a new edge through each old face, a new face pierced by each old edge, and a new cell enclosing each old vertex.

(2) Show that the new and old cell decompositions possess a common refinement, obtainable from either of them by elementary subdivisions.

(3) If V^*, E^*, F^*, C^* denote the numbers of vertices, edges, faces, and cells in the new decomposition, then

$$V^* = C, \qquad E^* = F, \qquad F^* = E, \qquad C^* = V.$$

(4) Deduce that

$$V - E + F - C = V^* - E^* + F^* - C^* = 0.$$

8.2.3 An Example

Consider the pseudomanifold $\mathscr{C}$ defined by a cigar-shaped cell with four 2-gonal faces, the top being identified with the bottom and the back with the front (Figure 255). $\mathscr{C}$ has two vertices, one edge, two faces, and one cell so $\chi(\mathscr{C}) = 2$ and it is not a manifold. The neighbourhood surface of each vertex is a torus, as is clear from the edge identifications when we cut the corners off at A and B (Figure 256). The bounded 3-manifold which results from deleting the neighbourhoods of the vertices is easily glued together in ordinary space and is homeomorphic to the space bounded by two coaxial torus surfaces.

It is also obtained by removing a tubular neighbourhood of the two-crossing link in $\mathbf{S}^3$. The easiest way to see this is to view $\mathbf{S}^3$ as a lens-shaped cell with top and bottom faces identified. A groove of semicircular cross-section is gouged out of the lens rim, and a vertical hole is drilled through its centre (Figure 257). These close up to linked circular tunnels when the top and bottom faces are identified. Then cutting as in Figure 258 yields the schema above.

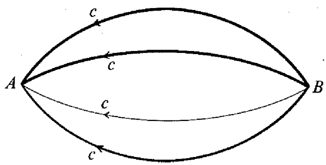


Figure 255

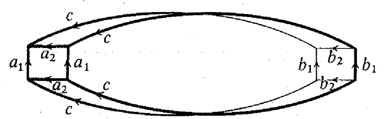


Figure 256

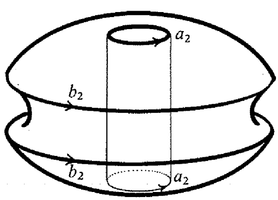


Figure 257

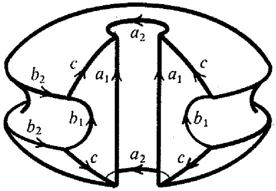


Figure 258

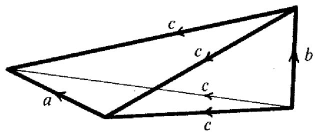


Figure 259

Exercise 8.2.3.1. Stretch the vertices A, B of $\mathscr{C}$ into edges a, b as in Figure 259, still identifying the front face with the back and the top with the bottom. Show that the resulting complex is an $\mathbf{S}^3$ in which a, b constitute the two-crossing link.

8.2.4 Remarks

Polyhedral schemata have never been used in a systematic way for the construction of 3-manifolds, even though some interesting manifolds originally arose in this way (for example lens spaces, cf. 4.2.8. See also Threlfall and Seifert 1930, 1932 and Weber and Seifert 1933 for manifolds obtained from the Platonic solids). Apparently polyhedral schemata do not admit anything like the reductions applicable to polygon schemata, but it is not clear that anyone has worked very hard on the problem. Only recently, Thurston 1977 has found polyhedral forms of many 3-manifolds which can be used to tesselate hyperbolic 3-space, yielding a theory like the classical theory of 2-manifolds.

8.3 Heegaard Splittings

8.3.1 Existence

Given a triangulation of a finite 3-manifold $\mathscr{M}$, we decompose it into a tubular neighbourhood of the 1-skeleton, and the complement of this neighbourhood. A glance at the two pieces which result from a given tetrahedral cell (Figure

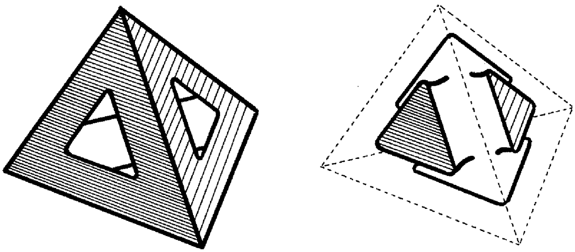


Figure 260

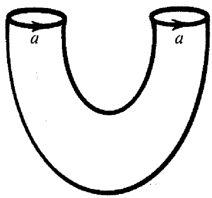


Figure 261

260) clearly suggests how the "edge pieces" and the "interior pieces" respectively unite into handlebodies $\mathscr{H}_1$, $\mathscr{H}_2$. Each body is certainly the tubular neighbourhood of a graph $\mathscr{G}_i$ $(i = 1, 2)$, and the standard ball with handles can be obtained by taking a spanning tree $\mathscr{T}_i$ of $\mathscr{G}_i$ and letting the tubular neighbourhood of $\mathscr{T}_i$ be the "ball part" of $\mathscr{H}_i$.

This argument contains a hidden assumption of orientability because we are assuming that the tubular neighbourhood of a closed curve in $\mathscr{M}$ is a solid torus, and not the solid Klein bottle obtained by identifying the ends of a solid cylinder as in Figure 261. It is clear that any 3-manifold containing a solid Klein bottle is nonorientable, and the converse follows by the same argument used to show that a nonorientable 2-manifold contains a Möbius band (1.2.1).

Exercise 8.3.1.1 Show that $\mathbf{S}^1 \times \mathbf{S}^1 \times \mathbf{S}^1$ splits into handlebodies of genus 3.

8.3.2 Heegaard Diagrams

A manifold $\mathscr{M}$ which splits into handlebodies $\mathscr{H}_1$, $\mathscr{H}_2$ is determined up to homeomorphism by the map $h: \partial\mathscr{H}_1 \to \partial\mathscr{H}_2$ which identifies the handlebody boundaries. In fact, $\mathscr{M}$ is determined by the images $h(m_1), \ldots, h(m_n)$ on $\partial\mathscr{H}_2$ (its *Heegaard diagram*) of the canonical meridians $m_1, \ldots, m_n$ shown on $\partial\mathscr{H}_1$ in Figure 262.

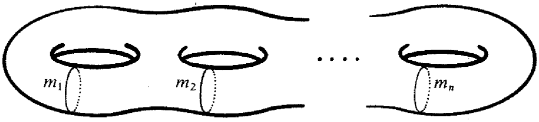


Figure 262

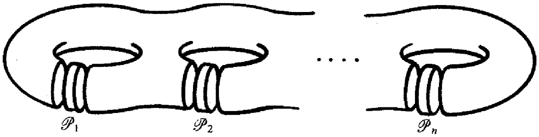


Figure 263

We write $\mathscr{M} = \mathscr{H}_1 \cup_h \mathscr{H}_2$ to indicate that $\mathscr{M}$ is obtained from $\mathscr{H}_1$, $\mathscr{H}_2$ by identifying $\partial\mathscr{H}_1$ with $\partial\mathscr{H}_2$ via h. It will suffice to show that $\mathscr{M}$ can be reconstructed, up to homeomorphism, from knowledge of the $h(m_i)$ on $\partial\mathscr{H}_2$.

We span the meridian m_i on $\partial\mathscr{H}_1$ by a disc $\mathscr{D}_i \subset \mathscr{H}_1$ and then detach a thin closed "plate" neighbourhood $\mathscr{P}_i$ of $\mathscr{D}_i$ (Figure 263). $\mathscr{P}_i$ is attached to $\mathscr{H}_2$ by identifying its rim $\mathscr{P}_i \cap \partial\mathscr{H}_1$ with a thin closed annular neighbourhood of $h(m_i)$. Assuming that the annular neighbourhoods are thin enough not to meet each other, the resulting bounded manifold $\mathscr{M}'$ is unique up to homeomorphism. But the piece of $\mathscr{H}_1$ which remains is a topological ball, so $\mathscr{M}' = \overline{\mathscr{M}\text{-ball}}$, and this determines $\mathscr{M}$ up to homeomorphism. □

This construction is essentially that of Heegaard 1898. As stated in 8.1.3, it is not easy to determine when two Heegaard diagrams determine the same manifold. Waldhausen 1968a has shown that any Heegaard diagram of $\mathbf{S}^3$ is isotopic to the diagram in Figure 264; however, even isotopy of Heegaard diagrams is not easy to recognize. The catch is that an isotopy within the handlebody itself cannot necessarily be achieved in ordinary space. For example, Figure 265(1) is isotopic to Figure 265(2) but not in ordinary space. (Why?) The isotopy can be discovered if the embedding of Figure 265(2) in R^3 is changed to that shown in Figure 266.

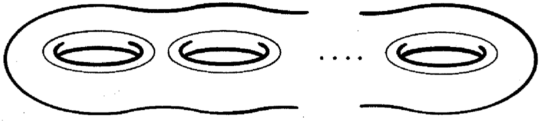


Figure 264

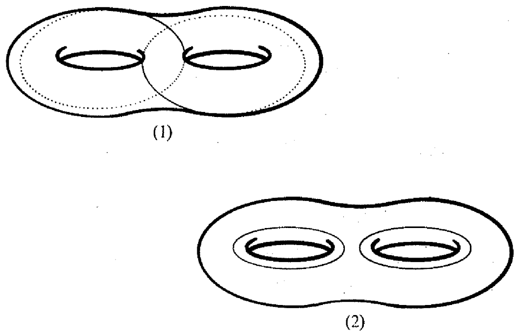


Figure 265

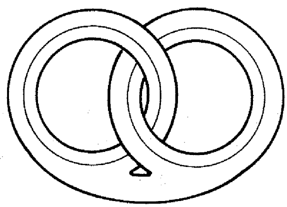


Figure 266

EXERCISE 8.3.2.1. Find the isotopy just claimed.

EXERCISE 8.3.2.2. Give an algorithm which decides when a set of n disjoint simple closed curves on a handlebody $\mathscr{H}$ of genus n is a Heegaard diagram.

8.3.3 Reading a Presentation of the Fundamental Group

The fundamental group of a handlebody $\mathscr{H}$ of genus n is the free group F_n of rank n, generated by loops a_i through the handles. If p_i denotes the element of $\pi_1(\mathscr{H})$ corresponding to the curve $h(m_i)$, then attaching a plate to an annular neighbourhood of $h(m_i)$ introduces the relation $p_i = 1$. These facts are immediate from suitable deformation retractions and the Seifert–Van Kampen theorem. Since attachment of the ball which completes $\mathscr{M}$ does not change π_1 (4.1.5), it follows that

$$\pi_1(\mathscr{M}) = \langle a_1, \ldots, a_n; p_1, \ldots, p_n \rangle.$$

In particular, any 3-manifold group has a *balanced presentation*—one with equal numbers of generators and relations. It also follows that the minimum number of generators required to present $\pi_1(\mathscr{M})$ is a lower bound on the Heegaard genus (8.1.3) of $\mathscr{M}$. For example, we can construct an $\mathscr{M}_n$ with $\pi_1(\mathscr{M}_n) = F_n$ by taking the $h(m_i)$ to be curves which are already trivial in $\pi_1(\mathscr{H}_2)$, such as standard meridians. Then

$$\text{Heegaard genus of } \mathscr{M}_n \leq n$$

but since there is no presentation of F_n with $< n$ generators (Exercise 2.2.4.3, or 5.3.2) we have

$$\text{Heegaard genus of } \mathscr{M}_n = n.$$

(Papakyriakopoulos 1957 proved, assuming the Poincaré conjecture, that $\mathscr{M}_n$ is the unique finite 3-manifold whose $\pi_1 = F_n$.)

EXERCISE 8.3.3.1. Show that $\mathscr{M}_1 = \mathbf{S}^1 \times \mathbf{S}^2$.

EXERCISE 8.3.3.2. Show that $\mathbf{S}^1 \times \mathbf{S}^1 \times \mathbf{S}^1$ has Heegaard genus 3.

8.3.4 Lens Spaces

To find the Heegaard diagram of the (m, n) lens space we use the core dissection of 7.2.2. The diagram turns out to be a solid torus with an (m, n) curve.

Realize the (m, n) lens space as a lens-shaped cell with its faces divided into m equal sectors, the top face being identified with the bottom after a twist of $2\pi(n/m)$ (Figure 267). The identifications are indicated by the numbers (which are assumed to be reduced mod m) and it is convenient to imagine that the top and bottom face meet at angle $2\pi/m$ when $m > 1$.

We now remove a vertical core from the lens and draw m equally spaced vertical lines on it, connecting the ends of the sectors (Figure 268). After the twist of $2\pi(n/m)$ required to identify the top and bottom faces, these vertical lines join up to form an (m, n) torus curve p.

The remainder of the lens is divided into wedges of angle $2\pi/m$ by vertical cuts through the boundaries of the sectors (Figure 269). When the regions numbered $1, 2, \ldots, m$ on the wedges are identified the result is a cylinder with

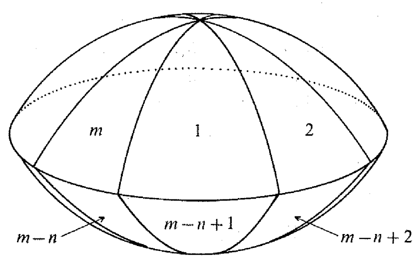


Figure 267

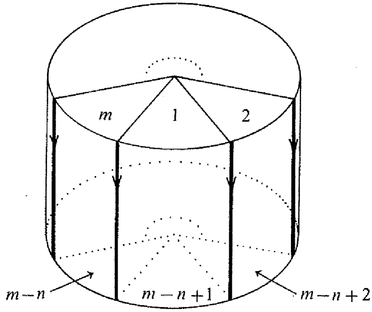


Figure 268

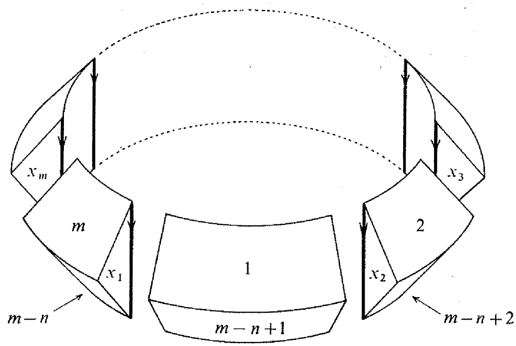


Figure 269

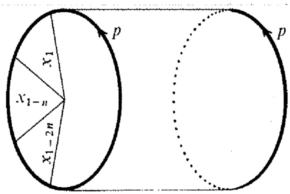


Figure 270

the curve p as the edge at each end (Figure 270). The cut marks follow the same cycle $x_1, x_{1-n}, x_{1-2n}, \ldots$ (indices mod m) on both ends, so when the cuts are rejoined the cylinder ends close to form a solid torus, with p as a meridian curve.

The impression of p on the core torus, an (m, n) curve, is therefore the Heegaard diagram of the (m, n) lens space. □

If we associate a "(0, 1) lens space," to which the lens construction obviously does not apply, with the (0, 1) Heegaard diagram (p = meridian on both tori) then the lens spaces exhaust all Heegaard diagrams on the torus. Thus the lens spaces are precisely the manifolds of Heegaard genus 1.

EXERCISE 8.3.4.1. What is the (0, 1) lens space?

EXERCISE 8.3.4.2. Show that the ends of the cylinder formed by the wedges in the above construction have to be joined after a twist of $2\pi(n'/m)$, where

$$n'n \equiv -1 \bmod m.$$

Deduce that the (m, n) lens space and the (m, n') lens space are homeomorphic when

$$n'n \equiv \pm 1 \bmod m.$$

EXERCISE 8.3.4.3 (Hotelling 1925). Consider the 3-manifold $\mathscr{M}$ consisting of all unit tangents to the sphere $x^2 + y^2 + z^2 = 1$ in $\mathbb{R}^3$ (Figure 271). Show that the submanifold of tangents t to points P in a given hemisphere is homeomorphic to a solid torus, and deduce that $\mathscr{M}$ is the (2, 1) lens space.

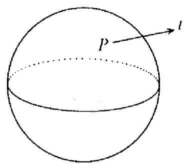


Figure 271

8.3.5 Alexander's Proof that the (5, 1) Lens Space and the (5, 2) Lens Space Are not Homeomorphic

We have seen (4.2.8) that the (5, 1) and (5, 2) lens spaces have the same fundamental group. Tietze 1908 conjectured that they are not homeomorphic. This conjecture was proved by Alexander 1919a as follows.

Suppose on the contrary that the (5, 1) space and the (5, 2) space are homeomorphic, so their Heegaard diagrams can be viewed as different decompositions of one and the same space $\mathscr{M}$.

The diagram of the (5, 1) space corresponds to a decomposition into solid tori $\mathscr{H}_1$, $\mathscr{H}_2$ with a meridian on $\mathscr{H}_1$ identified with a (5, 1) curve $p = ab^5$ on $\mathscr{H}_2$ (Figure 272). When p is traversed in the direction shown its points of intersection 1, 2, 3, 4, 5 with the meridian a are encountered successively.

On the other hand, for the (5, 2) space the decomposition is into solid tori $\mathscr{H}'_1$, $\mathscr{H}'_2$ with a meridian on $\mathscr{H}'_1$ identified with a (5, 2) torus curve p' on

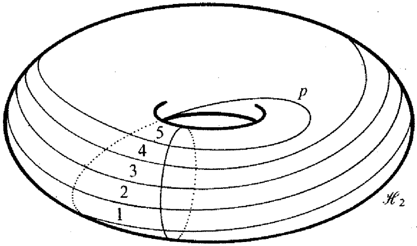


Figure 272

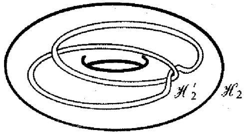


Figure 273

$\mathscr{H}'_2$. The corresponding intersections of p' with a meridian are encountered in the order 1, 3, 5, 2, 4 as p' is traversed.

Now $\mathscr{H}_1$ can be shrunk in cross-section, and $\mathscr{H}_2$ subjected to a complementary expansion, while retaining the same Heegaard diagram. Furthermore, the new $\mathscr{H}_1$ can be an arbitrarily thin neighbourhood of any curve inside the old $\mathscr{H}_1$ and isotopic to its axis. Similarly, we can shrink down $\mathscr{H}'_2$ and deform it enough to miss the shrunken $\mathscr{H}_1$. Thus there is no loss of generality in assuming that $\mathscr{H}'_2$ is disjoint from $\mathscr{H}_1$, that is, inside $\mathscr{H}_2$.

$\mathscr{H}'_2$ may lie inside $\mathscr{H}_2$ in a complicated way, for example as in Figure 273, the important point is that the curve p' on it must be nullhomotopic in $\mathscr{M} - \mathscr{H}'_2$, since it bounds a disc in $\mathscr{H}'_1$, and hence homotopic to p, which bounds a disc in $\mathscr{H}_1$. *A fortiori* then, p and p' will be null-homologous in $\mathscr{M} - \mathscr{H}'_2$, and we shall now compute the consequences of this relation by describing both curves in terms of a meridian a' on $\mathscr{H}'_2$ and the latitude b on $\mathscr{H}_2$.

$H_1(\mathscr{H}_2)$ is freely generated by b hence $H_1(\mathscr{H}_2 - \mathscr{H}'_2)$ is freely generated by a', b (cf. 5.3.4) and consequently $H_1(\mathscr{M} - \mathscr{H}'_2)$ is generated by a', b with the relation $p = 1$, where p is described in terms of a', b. The last follows by an "abelianized" Seifert–Van Kampen theorem, since $\mathscr{M} - \mathscr{H}'_2$ is obtained by glueing a plate to $\mathscr{H}_2 - \mathscr{H}'_2$ along the curve p, then attaching a ball.

Now let θ be the number of times $\mathscr{H}'_2$ winds around $\mathscr{H}_2$ in the b direction (= the algebraic sum of the signed intersections of $\mathscr{H}'_2$ with a meridian disc in $\mathscr{H}_2$. For the example shown, $\theta = 2$). Then

$$p = ab^5 = a'^{\theta}b^5 \qquad (1)$$

in $H_1(\mathscr{H}_2 - \mathscr{H}'_2)$, because a typical meridian disc of $\mathscr{H}_2$, bounded by a, contains θ meridian sections of $\mathscr{H}'_2$, hence $a = a'^{\theta}$ in $H_1(\mathscr{H}_2 - \mathscr{H}'_2)$. The defining relation $p = 1$ of $H_1(\mathscr{M} - \mathscr{H}'_2)$ is therefore

$$a'^{\theta}b^5 = 1. \tag{2}$$

On the other hand

$$p' = a'^{5k+2}b^{\pm 5\theta} \tag{3}$$

for some k, because when we describe p' in the positive sense we traverse five arcs which meet a' at the points 1, 3, 5, 2, 4 in turn, so each arc winds around $\mathscr{H}'_2$ "$k + \frac{2}{5}$" turns in the a' direction, for some fixed integer k. At the same time, we make $\pm 5\theta$ circuits inside $\mathscr{H}_2$ in the b direction.

It follows from (2) and (3) that if $p' = p = 1$ in $H_1(\mathscr{M} - \mathscr{H}'_2)$, we must have

$$a'^{5k+2}b^{\pm 5\theta} = (a'^{\theta}b^5)^n \quad \text{for some } n$$

and equating exponents we get

$$5k + 2 = n\theta \tag{4}$$

$$\pm 5\theta = 5n \tag{5}$$

Hence

$$\theta = \pm n$$

and if we substitute this in (4), we get

$$5k + 2 = \pm n^2.$$

But this is impossible, since all squares are of the form $5k$ or $5k \pm 1$. □

EXERCISE 8.3.5.1. Prove the number-theoretic result used in the last line of the proof.

EXERCISE 8.3.5.2. Show that $\mathscr{H}_2 - \mathscr{H}'_2$ is the complement of a two-component link in $\mathbf{S}^3$, hence the claim that $H_1(\mathscr{H}_2 - \mathscr{H}'_2)$ is free abelian is a consequence of Exercise 5.3.4.1.

8.3.6 Heegaard Diagrams of Bounded 3-manifolds

Every set of disjoint simple closed curves on a handlebody determines a bounded 3-manifold $\mathscr{M}$, namely, the $\mathscr{M}$ obtained by glueing plates to annular neighbourhoods of the curves.

Conversely, every finite orientable bounded 3-manifold $\mathscr{M}$ is obtained in this way.

We begin by splitting $\mathscr{M}$ into two handlebodies $\mathscr{H}_1$, $\mathscr{H}_2$ where $\mathscr{H}_1$ is a tubular neighbourhood of the 1-skeleton of $\mathscr{M}$ and $\mathscr{H}_2 = \overline{\mathscr{M} - \mathscr{H}_1}$.

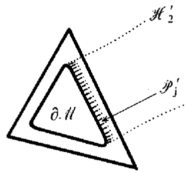


Figure 274

$\mathscr{H}_1$ and $\mathscr{H}_2$ are not identified along their whole boundaries, indeed the points where $\partial\mathscr{H}_1, \partial\mathscr{H}_2$ are not identified constitute $\partial\mathscr{M}$.

We glue $\mathscr{H}_2$ onto $\mathscr{H}_1$ in the following steps.

Step 1. Render $\mathscr{H}_2$ simply connected by removing suitable meridian plates $\mathscr{P}_i$. Since the pieces of $\partial\mathscr{M}$ on $\partial\mathscr{H}_2$ are topological discs (from the middle of faces in the triangulation of $\mathscr{M}$) we can place the $\mathscr{P}_i$ so that they do not meet $\partial\mathscr{M}$, by pushing their rims away from these discs where necessary. The $\mathscr{P}_i$ are then attached to $\mathscr{H}_1$ along their rims according to the identification map between $\partial\mathscr{H}_1, \partial\mathscr{H}_2$.

Step 2. $\mathscr{H}'_2 = \overline{\mathscr{H}_2 - \cup_i \mathscr{P}_i}$ is a topological ball which meets $\partial\mathscr{M}$ in certain discs $\mathscr{D}_j$. For all but one of these discs, cut a plate $\mathscr{P}'_j$ from $\mathscr{H}'_2$ which has $\mathscr{D}_j$ as its top face, and glue $\mathscr{P}'_j$ to $\mathscr{H}_1$ along its rim, which is an annulus common to $\partial\mathscr{H}_1, \partial\mathscr{H}_2$ (Figure 274).

Step 3. $\mathscr{H}''_2 = \overline{\mathscr{H}'_2 - \cup_j \mathscr{P}'_j}$ is a ball which meets $\partial\mathscr{M}$ in a single disc. $\mathscr{M} - \mathscr{H}''_2$ is therefore homeomorphic to $\mathscr{M}$ and we can throw $\mathscr{H}''_2$ away. □

EXERCISE 8.3.6.1. Construct a Heegaard diagram of the solid torus with the toroidal hole shown in Figure 275.

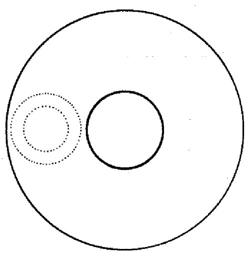


Figure 275

8.3.7 Fundamental Groups of Bounded 3-manifolds

As with ordinary diagrams, one can read off a presentation of $\pi_1(\mathscr{M})$ directly from the Heegaard diagram of a bounded 3-manifold $\mathscr{M}$, with the number of generators equal to the genus of the diagram.

Moreover, since *any* set of disjoint simple closed curves on a handlebody determines a bounded 3-manifold, we can easily state a geometric criterion for a finitely presented group G to be π_1 of a bounded orientable 3-manifold. Namely, G must have a presentation

$$\langle a_1, \ldots, a_n; p_1, \ldots, p_r \rangle$$

such that the $p_1, \ldots, p_r$, when interpreted as products of canonical generators $a_1, \ldots, a_n$ for π_1 (handlebody of genus n), can be realized by disjoint simple closed curves.

8.3.8 Heegaard Diagrams of Knot and Link Complements

The general construction of 8.3.6 is not very economical with respect to Heegaard genus, and a more efficient method in the case of knot (or link) complements is to find a handlebody $\mathscr{H}_n$, standardly embedded in $\mathbf{S}^3$, with the knot as a curve $\mathscr{K}$ on $\partial\mathscr{H}_n$. We replace this curve by a thin handle $\mathscr{H}$ just outside $\mathscr{H}_n$ except at two neighbouring discs where it meets $\partial\mathscr{H}_n$, so that cutting $\mathscr{H}_n$ by meridian discs produces a solid torus in the form of the given knot. On the other hand, we can unravel $\mathscr{H}$ from $\mathscr{H}_n$ by dragging one of its ends round the curve $\mathscr{K}$, producing a standard handlebody $\mathscr{H}_{n+1}$. In the process, the meridian curves on $\mathscr{H}_n$ are dragged into positions on $\mathscr{H}_{n+1}$ which bound discs cutting $\mathscr{H}_{n+1}$ into the solid torus of knot $\mathscr{K}$. (This was in Exercise 4.2.7.2.)

The complementary handlebody in $\mathbf{S}^3$ can be viewed as a ball with holes, one of which (corresponding to $\mathscr{H}$) is knotted. Instead of cutting, we glue on plates to seal off unknotted holes, so it is a matter of seeing what happens to the plate rims p when the knotted hole is unravelled.

Some examples clarify this process.

Example 1. The two-crossing link (Figure 276). On a standard handlebody of genus 2, p looks like Figure 277, which we can read as $p = a_1 a_2 a_1^{-1} a_2^{-1}$, thus obtaining the standard presentation of the group.

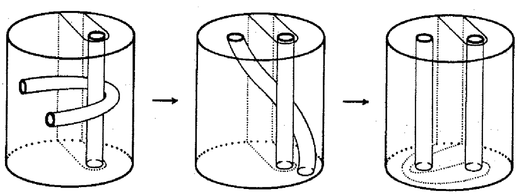


Figure 276

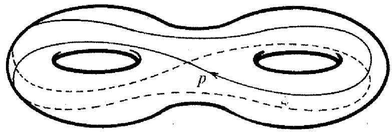


Figure 277

EXAMPLE 2. The trefoil knot (Figure 278). In terms of generators a_1, a_2 (clockwise round the right and left holes respectively),

$$p = a_1 a_2 a_1 a_2^{-1} a_1^{-1} a_2^{-1}$$

which gives a standard presentation of the trefoil knot group (cf. 4.2.5), $\langle a_1, a_2 ; a_1 a_2 a_1 = a_2 a_1 a_2 \rangle$.

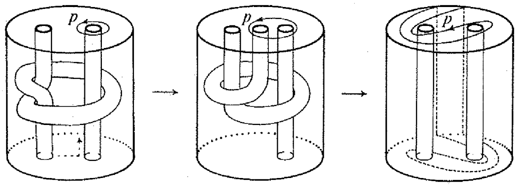


Figure 278

EXERCISE 8.3.8.1. Show that the figure-eight knot complement, Figure 279(1) has a Heegaard diagram Figure 279(2) and deduce that its group has a presentation

$$\langle a_1, a_2 ; a_1 a_2 a_1 a_2^{-1} a_1^{-1} a_2 a_1 a_2 a_1^{-1} a_2^{-1} \rangle.$$

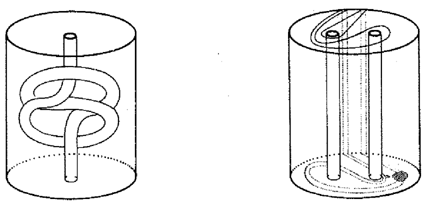


Figure 279

8.4 Surgery

8.4.1 Dehn's Construction of the Poincaré Homology Sphere

Poincaré 1904 constructed a 3-manifold with trivial homology but non-trivial fundamental group, the so-called *Poincaré homology sphere.* His construction was a Heegaard diagram of genus 2. Different constructions were later given by Dehn 1910 (surgery on a trefoil knot, see below), Threlfall and Seifert 1930 (identifying opposite faces of a dodecahedron with a $\frac{1}{5}$ twist) and Weber and Seifert 1933 (2-fold cover of $\mathbf{S}^3$ branched over a (3, 5) torus knot), though it was not at first realized that these manifolds were identical. A quick pictorial proof that the Poincaré and Dehn constructions are equivalent may be found in Rolfsen 1976, along with many other constructions of the Poincaré homology sphere.

Dehn's construction is noteworthy as the first example of *surgery* on 3-manifolds, and it is the most directly convincing as far as the trivial homology is concerned. Surgery in general is the process of removing a solid torus $\mathscr{R}$ from a manifold $\mathscr{M}$, and identifying the boundary of the hole with the boundary of another solid torus $\mathscr{R}'$, via a homeomorphism different from the one defined by the inclusion of $\mathscr{R}$ in $\mathscr{M}$ (often called "sewing $\mathscr{R}$ back differently"). Dehn takes $\mathscr{R}$ in $\mathbf{S}^3$, knotted in the form of a trefoil knot. Then we know from 5.3.4 that $H_1(\mathbf{S}^3 - \mathscr{R})$ is the infinite cyclic group generated by a_1 (Figure 280); on the other hand, since $\mathscr{R}$ is knotted there are many curves in $\mathbf{S}^3 - \mathscr{R}$ which are homologous to a_1 but not homotopic to it. It we take such a curve p embedded in $\partial\mathscr{R}$, then remove $\mathscr{R}$ and glue back another solid torus $\mathscr{R}'$ with its meridian identified with p, we obtain a homology sphere $\mathscr{M}$.

For $\mathscr{R}'$ can be attached by first glueing a plate along its rim to p, then attaching a ball. The abelianized Seifert–Van Kampen theorem says that

$$H_1(\mathscr{M}) = \langle a_1; a_1 = p = 1\rangle = \{1\}$$

while the ordinary Seifert–Van Kampen theorem says that $\pi_1(\mathscr{M})$ is obtained by adding the relation $p = 1$ to $\pi_1(\mathbf{S}^3 - \mathscr{R})$, where p is expressed as an element of $\pi_1(\mathbf{S}^3 - \mathscr{R})$.

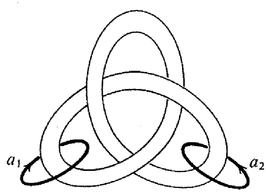


Figure 280

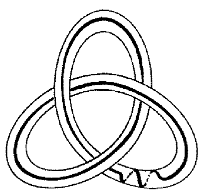


Figure 281

We can ensure that $\pi_1(\mathscr{M}) \neq \{1\}$ by suitable choice of p. It is necessary, and in fact sufficient, for p not to bound a disc in $\mathscr{R}$, though we shall verify this fact only for the simplest such p, which is shown in Figure 281.

8.4.2 The Fundamental Group of Poincaré's Homology Sphere

We use the Wirtinger generators a_1, a_2 for $\pi_1(\mathbf{S}^3 - \mathscr{R})$ so that

$$\pi_1(\mathbf{S}^3 - \mathscr{R}) = \langle a_1, a_2; a_1 a_2 a_1 = a_2 a_1 a_2 \rangle.$$

The expression for p is $a_1 a_2^2 a_1 a_2^{-3}$, so it follows by the Seifert–Van Kampen theorem that

$$\pi_1(\mathscr{M}) = \langle a_1, a_2; a_1 a_2 a_1 = a_2 a_1 a_2, a_1 a_2^2 a_1 a_2^{-3} \rangle,$$

which the substitution $b = a_1 a_2$ reduces to

$$\pi_1(\mathscr{M}) = \langle a_2, b; (a_2 b)^2 = b^3 = a_2^5 \rangle$$

—a group which is nontrivial because it has a nontrivial subgroup of the icosahedral group as homomorphic image. The details follow.

To find the expression for p we deform it as shown in Figure 282 and read the sequence of undercrossings in terms of the Wirtinger generators:

$$\begin{aligned} p &= a_1 a_2 (a_2 a_1 a_2^{-1}) a_2^{-2} \\ &= a_1 a_2^2 a_1 a_2^{-3}. \end{aligned}$$

(Notice incidentally that p is homologous to a_1, since a_1 and a_2 are homologous.) Thus we have

$$\pi_1(\mathscr{M}) = \langle a_1, a_2; a_1 a_2 a_1 = a_2 a_1 a_2, a_1 a_2^2 a_1 a_2^{-3} \rangle.$$

Now substituting $b = a_1 a_2$ so that $a_1 = b a_2^{-1}$ we get

$$\begin{aligned} \pi_1(\mathscr{M}) &= \langle a_2, b; b^2 a_2^{-1} = a_2 b, b a_2 b a_2^{-4} \rangle \\ &= \langle a_2, b; b^2 a_2^{-1} = a_2 b, b(b^2 a_2^{-1}) a_2^{-4} \rangle \\ &= \langle a_2, b; b^2 a_2^{-1} = a_2 b, b^3 = a_2^5 \rangle \\ &= \langle a_2, b; b^3 = (a_2 b)^2, b^3 = a_2^5 \rangle \\ &= \langle a_2, b; (a_2 b)^2 = b^3 = a_2^5 \rangle. \end{aligned}$$

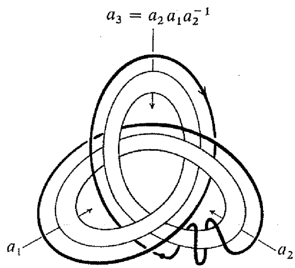


Figure 282

These relations can be modelled by the group of rigid motions of the icosahedron generated by a_2 = rotation of $2\pi/5$ about a diagonal through a vertex P and b = rotation of $2\pi/3$ about a diagonal through a face with vertex P. Since the latter group is nontrivial, so is $\pi_1(\mathscr{M})$. □

Poincaré 1904 found the homomorphism into the icosahedral group, and Dehn 1910 found the complete Cayley diagram of $\pi_1(\mathscr{M})$, showing that it had order 120. To this day, this is the only known homology sphere $\neq \mathbf{S}^3$ with finite fundamental group. Dehn did not point out that he had the same group as Poincaré, let alone the same manifold. The latter was proved around 1930 by a combination of results by Kneser (showing that the Dehn homology sphere equals dodecahedral space), Threlfall and Seifert. See Threlfall and Seifert 1930, 1932 and Weber and Seifert 1933.

Dehn observed that the homology spheres obtained by other surgeries on the trefoil knot (that is, other choices of the curve p) had infinite fundamental groups. A proof that these groups are all distinct is given in de Rham 1969.

8.4.3 Construction of Finite Orientable 3-manifolds by Surgery

Every finite orientable 3-manifold $\mathscr{M}$ is obtainable by removing a finite number of disjoint solid tori from $\mathbf{S}^3$ and sewing them back differently.

The proof is by induction on the Heegaard genus n of $\mathscr{M}$. If $n = 0$, then $\mathscr{M}$ is already $\mathbf{S}^3$ and no surgery is necessary. Suppose then that $n > 0$ and consider a splitting of $\mathscr{M}$ into handlebodies $\mathscr{H}_1, \mathscr{H}_2$ of genus n which identifies the meridians $m_1, \ldots, m_n$ on $\partial\mathscr{H}_1$ with curves $h(m_1), \ldots, h(m_n)$ on $\partial\mathscr{H}_2$. The latter are disjoint simple curves so $h(m_1)$ in particular may be mapped onto a canonical latitude curve on $\partial\mathscr{H}_2$ (Figure 283) by a homeomorphism t of $\partial\mathscr{H}_2$ which is a composite of twist homeomorphisms and isotopies (6.3.6).

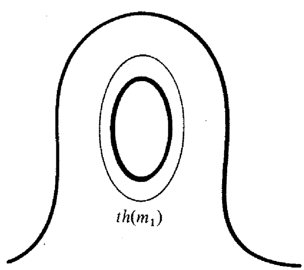


Figure 283

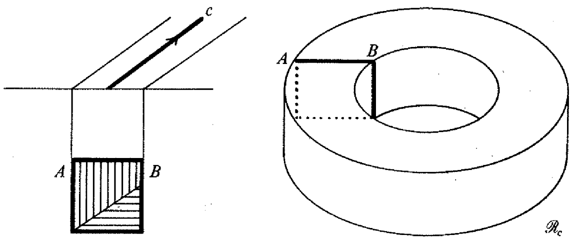


Figure 284

The homeomorphism t will not in general extend to $\mathcal{H}_2$ itself (although the isotopies will), but we *can* extend it to ($\mathcal{H}_2$ – a finite set of disjoint solid tori) as follows. Given a twist homeomorphism t_c about a curve c we excavate a "tunnel" below c by removing a thin solid torus $\mathcal{R}_c$ (Figure 284). The twist homeomorphism t_c can then be extended to $\overline{\mathcal{H}_2 - \mathcal{R}_c}$ by extending the twist of the annular neighbourhood of c inwards as far as the roof of the tunnel. The twist in the roof of the tunnel leaves its boundary looking as in Figure 285. The meridian through AB has been transformed into a (1, 1) torus curve, while the latitude retains its form. Sewing back $\mathcal{R}_c$ so that its meridian and latitude are identified with these two curves on $\partial(\overline{\mathcal{H}_2 - \mathcal{R}_c})$ recovers $\mathcal{H}_2$, but with $\partial\mathcal{H}_2$ transformed by the twist homeomorphism t_c.

To extend the whole homeomorphism t to ($\mathcal{H}_2$ – some solid tori) we excavate a tunnel under each twist curve in turn, halving the depth and cross-section of each successive tunnel so that it just misses its predecessors. When the necessary isotopies are combined with the twists the result is an extension of t to

$$(\mathcal{H}_2 - \text{finite set of disjoint solid tori}) = \mathcal{H}'_2$$

and $th(m_1)$ is a canonical latitude curve on $\mathcal{H}'_2$.

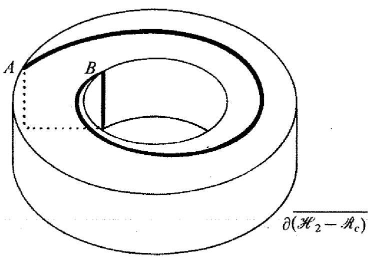


Figure 285

If $\mathscr{H}''_2$ denotes the result of sewing back the tori $\mathscr{R}_c$ into $\mathscr{H}'_2$ in the *ordinary* way, then $\mathscr{H}''_2$ is a handlebody which yields $\mathscr{H}_2$ by surgery. If we perform this surgery on the manifold $\mathscr{M}''$ obtained by attaching $\mathscr{H}_1$ to $\mathscr{H}''_2$ via *th*, the result is therefore $\mathscr{M}$.

But the manifold $\mathscr{M}''$ has Heegaard genus $\leq n - 1$, because we can simultaneously reduce $\mathscr{H}_1$, $\mathscr{H}''_2$ to handlebodies of genus $n - 1$ by transferring a meridian plate corresponding to m_1 in $\mathscr{H}_1$ to a neighbourhood of the latitude $th(m_1)$ on $\mathscr{H}''_2$. It follows by induction that $\mathscr{M}''$ is obtainable by surgery on $\mathbf{S}^3$, and hence $\mathscr{M}$ is also, since the result of two successive surgeries is itself a surgery. (The tori can be made arbitrarily thin and deformed so as to miss each other, cf. 8.3.5). □

The above theorem was first proved using differential topology by Wallace 1960. The above proof is essentially that of Lickorish 1962, with some simplifications due to Hempel 1962. Hempel also notes that any knotting in the tori used for surgery can be removed by further surgery. It suffices to observe that any polygonal curve can be unknotted by changing finitely many crossings from over to under, and then to prove the following lemma, which we leave as an exercise.

Exercise 8.4.3.1. Let $\mathscr{K}$ be a knot and $\mathscr{K}'$ the knot which results from reversing one crossing of $\mathscr{K}$. Show that the manifold $\mathbf{S}^3 - \mathscr{K}$ may be changed to $\mathbf{S}^3 - \mathscr{K}'$ by surgery using a single unknotted torus. (*Hint*: consider Figure 286.)

Exercise 8.4.3.2. Show that any bounded orientable 3-manifold is obtainable by surgery on a submanifold of $\mathbf{S}^3$.

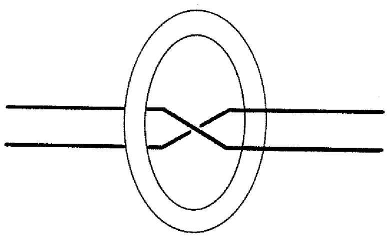


Figure 286

8.4.4 The Role of Knots and Links

All the surgeries used in the Lickorish proof are of the same type—a solid torus $\mathscr{R}$ is removed, then sewn back with its meridian identified with a (1, 1) curve on the tunnel boundary and its latitude preserved. This also applies to the surgeries used for unknotting the tori in the main proof; in this case one removes a solid torus around a crossing, then cuts the cylindrical "neck"

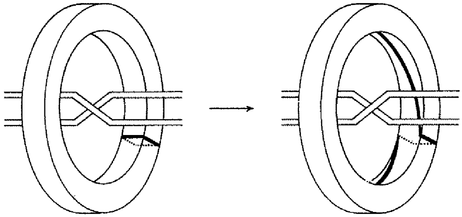


Figure 287

containing the crossing and rejoins it after a full twist, again converting a meridian to a (1, 1) curve (Figure 287). We may call these *Lickorish surgeries*.

Since we assume the tori are unknotted, any *single* Lickorish surgery yields $\mathbf{S}^3$, since $\mathbf{S}^3$ is the (1, 1) lens space. The point is that surgery on one solid torus changes the embedding of the tori linked with it, so that although they remain in $\mathbf{S}^3$ they may become knotted. It is easy to construct examples by reversing crossings as above. Lickorish surgery on a *knotted* solid torus can definitely produce a nonspherical manifold. In fact, Dehn's construction of the Poincaré homology sphere can be described in precisely this way. For if one removes the trefoil-knotted solid torus with the curve p shown in Figure 281, unknots it and spreads it out, p is seen to be a (1, 1) torus curve.

The crossing-reversing trick can be used to produce the trefoil knot in many different ways, yielding numerous surgery constructions of the Poincaré homology sphere. For example, one can use the link in Figure 288, known

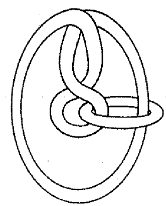


Figure 288

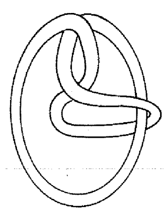


Figure 289

as Whitehead's link: Lickorish surgery on the horizontal torus, with suitable orientation, changes the embedding of the other torus to that shown in Figure 289, which is easily seen to be a trefoil knot. Therefore Lickorish surgery on both components will produce the Poincaré homology sphere.

EXERCISE 8.4.4.1. (Rolfsen 1976). Show that the Poincaré homology sphere may be constructed by Lickorish surgery on the "Borromean rings" (Figure 290).

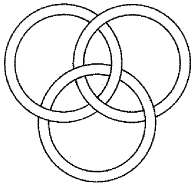


Figure 290

8.5 Branched Coverings

8.5.1 Wirtinger's Construction of the Lens Spaces as Branched Covers of S^3

Tietze 1908 reports the following unpublished result of Wirtinger:

The (m, n) *lens space is an m-fold cover of* S^3 *branched over the two-crossing link.*

We define the (m, n) lens space as usual by identifying top and bottom faces of a lens-shaped solid after a twist of $2\pi(n/m)$. Top and bottom faces are divided into m equal sectors by meridians, and the lens can be decomposed into m tetrahedra by vertical cuts through these meridians (Figure 291). The picture shows the typical tetrahedron $T_i = A_iB_iC_iD_i$. The construction of the (m, n) lens space requires us to identify $A_iB_iC_i$ with $A_{i+n}B_{i+n}D_{i+n}$ (vertex by vertex), and $B_iC_iD_i$ with $A_{i+1}C_{i+1}D_{i+1}$ (vertex by vertex), where indices are taken mod m.

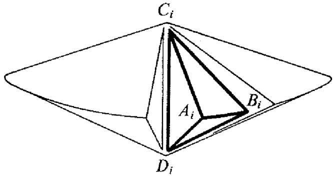


Figure 291

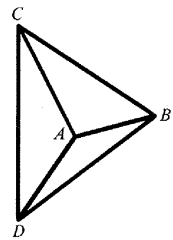


Figure 292

This schema gives an m-fold covering of the manifold determined by the single tetrahedron $ABCD$ (Figure 292) with ABC identified with ABD and BCD with ACD, branched over the curves corresponding to the edges AB and CD. Since we pass from T_i to T_{i+1} in a circuit around CD, the associated permutation of the sheets is the cyclic permutation $(1\ 2 \ldots m)$, and since we pass from T_i to T_{i+n} in a circuit around AB, its associated permutation is $(1 + n\ 1 + 2n \ldots)$.

The manifold determined by $ABCD$ is easily seen to be $\mathbf{S}^3$ (cf. Exercise 8.2.3.1), with the branch curves AB and CD forming the two-crossing link. □

This result is a prototype of a general representation of orientable 3-manifolds as branched covers of $\mathbf{S}^3$ found by Lickorish 1973. The first such general representation was found by Alexander 1919b, but with rather arbitrary branch curves. In the Lickorish representation the branch curves are linked, but unknotted, circles which lie inside the solid tori of the surgery representation (8.4.3).

EXERCISE 8.5.1.1. (Tietze 1908). Show that $\mathbf{S}^3$ is an m-sheeted *unbranched* cover of the (m, n) lens space (*Hint*: glue m copies of the lens together) and deduce that $\mathbf{S}^3$ can be viewed as an m^2-sheeted cover of itself, branched over the two-crossing link.

8.5.2 Constructing a Branched Cover from a Surgery Representation

The surgery theorem of 8.4.3 says that any orientable 3-manifold $\mathscr{M}$ contains certain disjoint solid tori $\mathscr{Q}_1, \ldots, \mathscr{Q}_n$ such that $\overline{\mathscr{M} - (\mathscr{Q}_1 \cup \cdots \cup \mathscr{Q}_n)}$ is homeomorphic to $\overline{\mathbf{S}^3 - (\mathscr{R}_1 \cup \cdots \cup \mathscr{R}_n)}$ where $\mathscr{R}_1, \ldots, \mathscr{R}_n$ are solid tori in $\mathbf{S}^3$ which are unknotted but may be linked. Suppose the homeomorphism is

$$h: \overline{\mathscr{M} - (\mathscr{Q}_1 \cup \cdots \cup \mathscr{Q}_n)} \to \overline{\mathbf{S}^3 - (\mathscr{R}_1 \cup \cdots \cup \mathscr{R}_n)}.$$

Then a branched covering $f\colon \mathscr{M} \to \mathbf{S}^3$ is constructed by dividing $\mathscr{M}$ into the following pieces:

(1) $\overline{\mathscr{M} - (\mathscr{Q}_1 \cup \cdots \cup \mathscr{Q}_n)}$,
(2) for each $i = 1, 2, \ldots, n$, a shell neighbourhood $\mathscr{N}_i$ of $\partial\mathscr{Q}_i$ lying inside $\mathscr{Q}_i$,
(3) the solid torus $\mathscr{P}_i = \overline{\mathscr{Q}_i - \mathscr{N}_i}$.

Then we map

(1) $\overline{\mathscr{M} - (\mathscr{Q}_1 \cup \cdots \cup \mathscr{Q}_n)}$ onto $\overline{\mathbf{S}^3 - (\mathscr{R}_1 \cup \cdots \cup \mathscr{R}_n)}$ by h,
(2) $\mathscr{N}_i$ onto $\mathscr{R}_i$ as a branched double cover,
(3) $\mathscr{P}_i$ homeomorphically onto the solid torus $\overline{\mathbf{S}^3 - \mathscr{R}_i}$.

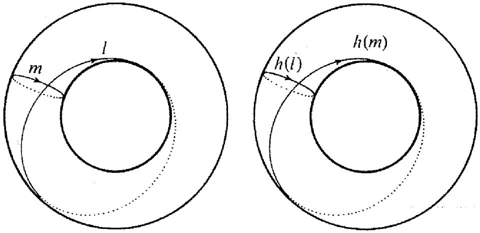
If we are able to find the required double cover, and make the three pieces of the map match on the boundaries, this will give an $(n + 1)$-fold cover of $\mathbf{S}^3$ by $\mathscr{M}$, branched over curves in the $\mathscr{R}_i$.

The double cover is not hard to find (8.5.4), but matching the boundary maps requires a careful choice of canonical curves on the tori, which we now discuss.

8.5.3 Choice of Canonical Curves

Since $\mathscr{R}_i$ is unknotted in $\mathbf{S}^3$, its meridian and latitude are determined up to isotopy and sign as curves which bound discs in $\mathscr{R}_i$ and $\overline{\mathbf{S}^3 - \mathscr{R}_i}$ respectively. For our purposes this gives sufficient determination of the (1, 1) curve on $\partial\mathscr{R}_i$ which is the h-image of the meridian m on $\mathscr{Q}_i$ according to the surgery theorem. If we choose the curve l shown on $\mathscr{Q}_i$ as the other canonical curve, we find that the latitudinal twist of surgery transforms it into the meridian of $\mathscr{R}_i$ (Figure 293). We now extend the map h across $\mathscr{N}_i$ in the natural way, so that $h(\mathscr{N}_i)$ is a shell neighbourhood of $\partial\mathscr{R}_i$ in $\mathscr{R}_i$ and if m', l' are the natural projections of m, l on the "inner" boundary of $\mathscr{N}_i$, $h(m')$ and $h(l')$ are the projections of $h(m)$, $h(l)$ on the inner boundary of $h(\mathscr{N}_i)$.

To obtain the required branched cover $f\colon \mathscr{N}_i \to \mathscr{R}_i$ we compose $h|\mathscr{N}_i$ with a branched cover $g\colon h(\mathscr{N}_i) \to \mathscr{R}_i$ which is the identity on $\partial\mathscr{R}_i$ and maps



$\mathscr{Q}_i$ $\mathscr{R}_i$

Figure 293

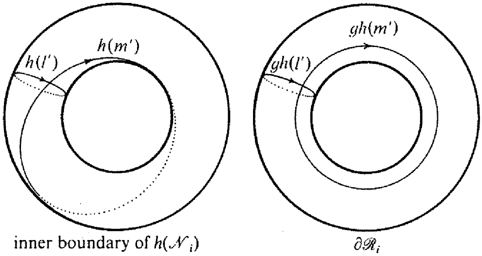


Figure 294

the inner boundary of $h(\mathcal{N}_i)$ onto $\partial\mathcal{R}_i$ in such a way that it can be extended to a homeomorphism $\mathcal{P}_i \to \overline{\mathbf{S}^3 - \mathcal{R}_i}$. For the latter it is necessary and sufficient that the meridian m' on $\mathcal{P}_i$ map to a meridian $gh(m')$ on $\overline{\mathbf{S}^3 - \mathcal{R}_i}$, that is, a latitude on $\mathcal{R}_i$. We shall in fact construct g to behave as shown in Figure 294. Thus g gives the inner boundary of $h(\mathcal{N}_i)$ a meridian twist relative to the outer boundary, $\partial\mathcal{R}$.

EXERCISE 8.5.3.1. Verify the claimed effect of the latitudinal twist on l.

8.5.4 A Branched Cover of (Disc $\times$ $\mathbf{S}^1$) by (Annulus $\times$ $\mathbf{S}^1$)

Let $\mathcal{D}$ denote the disc and $\mathcal{A}$ the annulus $\overline{\mathcal{D} - \mathcal{D}'}$, where $\mathcal{D}'$ is a disc in the interior of $\mathcal{D}$. Then there is a branched covering map $g: \mathcal{A} \times \mathbf{S}^1 \to \mathcal{D} \times \mathbf{S}^1$ with the following properties:

(1) *There is a single branch curve, a circle which winds twice around the interior of $\mathcal{D} \times \mathbf{S}^1$ and is unknotted when $\mathcal{D} \times \mathbf{S}^1$ is standardly embedded in $\mathbf{S}^3$.*
(2) *Assuming meridians on the two torus boundaries $\partial\mathcal{D} \times \mathbf{S}^1$ and $\partial\mathcal{D}' \times \mathbf{S}^1$ of $\mathcal{A} \times \mathbf{S}^1$ are chosen to lie in a cross-section $\mathcal{A} \times \theta$, then g is the identity on $\partial\mathcal{D} \times \mathbf{S}^1$ and performs a meridian twist on $\partial\mathcal{D}' \times \mathbf{S}^1$.*

First note the branched double cover of $\mathcal{D}$ by $\mathcal{A}$ obtained by cutting and cross-joining two copies of $\mathcal{D}$ (Figure 295). If we do this for each cross-section $\mathcal{D} \times \theta$ in two copies of $\mathcal{D} \times \mathbf{S}^1$ we obtain a branched double cover

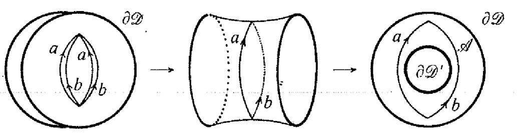


Figure 295

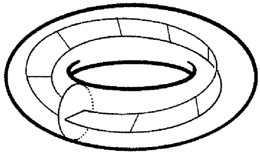


Figure 296

of $\mathscr{D} \times \mathbf{S}^1$ by $\mathscr{A} \times \mathbf{S}^1$. Furthermore, we can rotate the cut as θ runs from 0 to 2π so that it sweeps out a Möbius band in the interior of $\mathscr{D} \times \mathbf{S}^1$, the single edge of which is the branch curve of the covering (Figure 296). The twisting of the cut through π imparts a meridian twist of 2π in one boundary of $\mathscr{A} \times \mathbf{S}^1$ relative to the other, because a turn of the cut through ϕ occurs in the two covering discs, and hence turns one boundary of $\mathscr{A}$ through 2ϕ relative to the other.

In particular, if we keep $\partial\mathscr{D}$ fixed in each cross-section of the covering, the covering map

$$g: \mathscr{A} \times \mathbf{S}^1 \to \mathscr{D} \times \mathbf{S}^1$$

will be the identity on $\partial\mathscr{D} \times \mathbf{S}^1$ and will perform a meridian twist on $\partial\mathscr{D}' \times \mathbf{S}^1$. □

This completes the proof that orientable 3-manifolds can be represented as branched covers of $\mathbf{S}^3$, and makes a fitting conclusion to the journey we began in Chapter 1 with the representation of orientable 2-manifolds as Riemann surfaces.

EXERCISE 8.5.4.1. Describe the permutations of the $n + 1$ sheets about the branch curves in the above construction.

CHAPTER 9

Unsolvable Problems

$$a_i b_j = b_j a_i$$

$$w \in \langle w_1, \dots, w_n \rangle \,?$$

9.1 Computation

9.1.1 Turing Machines

The concept of a Turing machine was introduced in 0.4.1, as a machine that controls a read/write head moving on an infinite tape. We shall now explain the concept in more detail and give a few examples that illustrate how Turing machines compute functions and solve problems.

A *Turing machine* consists of a finite *alphabet* set $\mathscr{A} = \{\text{blank}, S_1, \ldots, S_m\}$, a finite *state* set $\mathscr{Q} = \{q_1, \ldots, q_n\}$, and a *response* function $\mathscr{R}: \mathscr{A} \times \mathscr{Q} \to \mathscr{A} \times \{\text{left, right}\} \times (\mathscr{Q} \cup \{\text{halt}\})$. Since the response function is also a finite set it can be described completely by listing its members. Each member is traditionally written, without commas and brackets, as a *quintuple*

$$q_iS_jS_{j'}Lq_{i'}, \quad \text{or} \quad q_iS_jS_{j'}Rq_{i'},$$

which is understood to mean: "in state q_i, if the scanned symbol is S_j, replace it by $S_{j'}$, move one square to the left (or right, respectively), and go into state $q_{i'}$ (where $q_{i'}$ = halt is allowed)."

Thus we can simply define a Turing machine M to be a finite list of quintuples with the property that no two quintuples start with the same two symbols q_iS_j. The latter condition expresses the single-valuedness of the response function, also called *determinacy of M*—the uniqueness of response to a given symbol when in a given state. It is usual to save space by omitting quintuples commencing with a pair q_iS_j for which no response is required. One can assume some standard response, for example,

$$q_iS_jS_jL \text{ halt},$$

for all such pairs.

Since nothing is said about a state q_i other than the responses it makes to symbols S_j, it is possible (and useful) to regard q_i as merely a label for a set of commands to the read/write head. The formalism of Turing machines is then nothing but a *language for programming a read/write head* on an infinite tape divided into squares. It is probably the simplest programming language, though unfortunately not the most readable. Apart from the read/write and movement commands, the language includes nothing but the two things most detested by modern programmers—line numbers (the subscripts i) and GOTO commands (the q_i).

One appreciates their detestation of these features as soon as one tries to read a Turing machine program written by someone else. It helps to write comments next to the quintuples, but only slightly. Readers are encouraged to study the following examples and then to try some of their own.

Example 1. *Adding a one to a block of consecutive ones.*

It is assumed that the tape is initially blank except for finitely many ones on consecutive squares. The machine starts on the leftmost of these squares in state q_1.

$q_1\ 1\ 1\ \mathrm{R}\ q_1$	Move right as long as the scanned symbol is 1
$q_1\ \square\ 1\ \mathrm{R}\ q_2$	Write 1 in the first blank square

Note the obvious notation $\square$ for the blank square. The machine halts after writing the extra 1, since no action is specified for state q_2.

This machine M illustrates a standard input/output convention for Turing machines. A positive integer n is input as a block of n ones on an otherwise blank tape, and M starts on the leftmost one in state q_1. The output is the number of ones on the tape when M halts, at which time the tape is blank except for a block of ones. M may then be considered to compute the function $f(\text{input}) = \text{output}$, in this case $f(n) = n + 1$ for positive integers n.

Example 2. *Doubling a block of ones.*

Under the same input/output convention as in Example 1, this machine computes $f(n) = 2n$. It does so by "copying" the input block one symbol at a time, "marking" the copied symbols as it does so. To follow the program it is important to realise that q_1 is not just the initial state. It is a state that recurs each time the head returns to the input block after copying a symbol. The program is written with this general situation in mind.

$q_1\ 1\ 1'\ \mathrm{L}\ q_2$	Mark the leftmost 1 in the input block with $'$.
$q_2\ 1'\ 1'\ \mathrm{L}\ q_2$	Move left across any $1'$ (copied part of input block),
$q_2\ 1\ 1\ \mathrm{L}\ q_2$	and 1 (these are in the copy).
$q_2\ \square\ 1\ \mathrm{R}\ q_3$	Write 1 in the first blank square.
$q_3\ 1\ 1\ \mathrm{L}\ q_3$	Move right across any 1,
$q_3\ 1'\ 1'\ \mathrm{R}\ q_1$	and $1'$ (at which point the initial state is resumed).
$q_1\ 1'\ 1'\ \mathrm{R}\ q_1$	
$q_1\ \square\ \square\ \mathrm{L}\ q_4$	If each symbol in the input block is now marked,
$q_4\ 1'\ 1\ \mathrm{L}\ q_4$	move left across the input block, erasing marks.

Example 3. *Deciding whether a number is even.*

This example shows how machine states play a role like mental states. The machine has a state q_{YES} that "remembers" when an even number of ones have been scanned and a state q_{NO} that "remembers" when an odd number of ones have been scanned. The initial state is q_{YES} and, as usual, the machine starts on the leftmost one of the block.

$$q_{YES}\ 1\ \square\ \mathrm{R}\ q_{NO}$$

$$q_{NO}\ 1\ \square\ \mathrm{R}\ q_{YES}$$

The machine halts on the first blank encountered, in state q_{YES} if the number of ones in the block is even and in state q_{NO} otherwise.

Thus the final state answers the question implicit in the input n: is n even? If we take a *problem*, as in 0.4.1, to be a set of questions, then this machine solves the problem of deciding evenness. Many other conventions for answering questions are possible, and they are equivalent in the sense that the same sets of questions are answerable. In the case of questions with a YES/NO answer we shall compute the *characteristic function* of the problem using the output convention of Examples 1 and 2. If the problem consists of questions Q_i then its characteristic function is

$$f(Q_i) = \begin{cases} 1 & \text{if the answer to } Q_i \text{ is YES} \\ 0 & \text{if the answer to } Q_i \text{ is NO} \end{cases}$$

In particular, the characteristic function for the evenness problem is computed when we add the following quintuple to the preceding machine

$$q_{YES}\ \square\ 1\ \mathrm{R}\ q_3.$$

(This is why the original machine was made to erase all ones on the tape. A blank tape signifies output 0.)

EXERCISE 9.1.1.1. Construct a Turing machine that copies an arbitrary block of *a*s and *b*s.

EXERCISE 9.1.1.2. Explain how to convert an arbitrary Turing machine M to a machine M', which produces tape expressions the same as those produced by M except for "end markers" [and] on the nonblank portion of tape.

EXERCISE 9.1.1.3. Construct a Turing machine to erase the tape between [and].

9.1.2 Church's Thesis

The simple examples of computation given in 9.1.1 fall far short of justifying the claim, made in 0.4.1, that the concept of Turing machine is a general definition of computer. This claim is known as *Church's thesis* after Alonzo Church, who first proposed a mathematical definition of computability in 1933. Church's definition, called λ-definability, was not at first very compelling, but it gained credibility with the work of Turing 1936. The Turing machine, with its read/write head and internal states, distilled the essence of computation as experienced by a human being—the eye that scans and recognizes, the hand that writes, and the mental states that direct the actions of eye and hand. It was therefore reasonable to expect that it could realize any proposed model of computation. Turing 1937 showed immediately that this was true of λ-definability, and it has proved to be true of all subsequent definitions of computability. In particular, all known programming languages can be written in the language of Turing machines.

All this is extraordinarily impressive—Gödel 1946 called it a "kind of miracle"—but it does not give us a theorem that computability = Turing machine computability. It is better to regard Church's thesis as an *axiom*, not a "self-evident" fact. Human computational processes are surely never going to be as narrow, simple, and laborious as Turing machine processes, hence Church's thesis will probably stand in need of eternal verification. However, we can make verification easier by devising programming languages that better reflect broader notions of computation and contain single commands for the most common routines. This is what has happened in the development of "higher-level" languages, which may be seen as a movement toward easier verification of Church's thesis. The ultimate programming language, if it exists, will be one which makes Church's thesis self-evident.

Despite the provisional nature of Church's thesis, we cannot avoid assuming it in proving unsolvability of algorithmic problems. An algorithmic problem is a set of questions and it is *solvable* if the answers are a computable function of the questions. Thus to prove such a problem *unsolvable* we have to prove that the answer function is not computable—and hence, by Church's thesis, not computable by Turing machine. This is the only way the statement to be proved becomes mathematically definite.

On the other hand, as long as we have to use Church's thesis we may as well enjoy it. It is a wonderful shortcut in proving the existence of Turing machines—if something is computable there is a Turing machine that computes it—so we shall use it as a shortcut whenever possible.

9.1.3 The Halting Problem

The simplest way to devise a problem unsolvable by a Turing machine is to consider questions about Turing machines themselves. Assuming only that each machine M has a description $\ulcorner M \urcorner$ that can be given as input and any fixed convention for answering YES and NO, then no machine can correctly answer all the following questions.

$$Q_M\text{: Does } M\text{, given input } \ulcorner M \urcorner\text{, eventually answer NO?}$$

A Turing machine S may be deemed to have received question Q_M when it is given input $\ulcorner M \urcorner$, since $\ulcorner M \urcorner$ contains all the necessary information. But S cannot correctly answer Q_S! If S eventually answers NO, then the correct answer to Q_S is YES, and if S eventually answers YES then the correct answer to Q_S is NO.

If we use the YES/NO convention of 9.1.1, according to which YES is signaled by halting on a tape with 1 on it and NO by halting on a blank tape, then we call the problem the *self-halting problem* because it consists of the questions

Q_M: Does M, given input ⌜M⌝, eventually halt on a blank tape?

All that is now required to make the halting problem mathematically definite is to delimit the class of Turing machines M and define their descriptions ⌜M⌝.

Without loss of generality we can assume that each machine alphabet is a finite subset of $\{\square, 1, 1', 1'', \ldots\}$ and that each state set is a finite subset of $\{q, q', q'', \ldots\}$. Each quintuple can therefore be written in the finite alphabet $\{q, \square, 1, ', \mathrm{R}, \mathrm{L}\}$ by viewing $'$ as a whole, rather than part, symbol. The quintuples of any machine can be concatenated into a single word without ambiguity if q_i is translated as $q^{(i)}$ (i primes). For example, Example 1 of 9.1.1 has quintuples

$$q' \; 1 \; 1 \; \mathrm{R} \; q'$$

$$q' \; \square \; 1 \; \mathrm{R} \; q''$$

and the quintuples can be recovered from the single word $q' \, 1 \, 1 \, \mathrm{R} \, q' \, q' \, \square \, 1 \, \mathrm{R} \, q''$. Thus if we define a *standard* Turing machine M to be one whose alphabet is a subset of $\{\square, 1, 1', 1'', \ldots\}$ and whose state set is a subset of $\{q, q', q'', \ldots\}$ then we can describe M with a single word on the alphabet $\{q, \square, 1, ', \mathrm{R}, \mathrm{L}\}$. Finally we can rewrite this word in the standard alphabet by replacing

$\square$ with $\square$,

1 with 1,

$'$ with $1'$,

q with $1''$,

L with $1'''$, and

R with $1''''$.

The resulting word is what we call the *standard description* ⌜M⌝ of M.

With this definition of ⌜M⌝, the questions Q_M comprising the self-halting problem have a mathematically precise meaning. Thus *the self-halting problem is a mathematical problem, not solvable by a Turing machine and hence, by Church's thesis, not solvable by any algorithm.* The mathematics of the self-halting problem is admittedly remote from group theory and topology, nevertheless they can be connected. We shall start building a bridge between them in 9.1.5. In the meantime, we shall look at some variant halting problems. These are useful for technical reasons and as an illustration of the idea of *reducing* one problem to another.

The first variant is actually more natural than the self-halting problem. It is the *general halting problem*, consisting of the questions

$Q_{M,I}$: Does M, given input I, eventually halt on a blank tape?

Of course, now that we know that the self-halting problem is unsolvable, it is obvious that the general halting problem also is. The questions $Q_{M,I}$ *include* the questions Q_M (for $I = \ulcorner M \urcorner$), hence any machine for answering the $Q_{M,I}$ will answer the Q_M, and we have shown that no such machine exists. We say that the self-halting problem *reduces to* the general halting problem. In general we say that *problem* P_1 *reduces to problem* P_2 when a solution of P_1 can be computed from a solution of P_2. It follows that *if* P_1 *is reducible to* P_2, *and* P_1 *is unsolvable, then* P_2 *is unsolvable.*

Reduction of one unsolvable problem to another is the most common method of proving unsolvability. In fact, the self-halting problem is the only problem we shall need whose unsolvability has to be proved directly. Unsolvability extends from the halting problem to the homeomorphism problem by a series of reductions, as we shall see in the remainder of this chapter.

EXERCISE 9.1.3.1. Use reduction (and Church's thesis) to show that there is no algorithm that correctly answers all the questions

$Q'_{M,I}$: Does M, given input I, eventually halt?

9.1.4 Universal Turing machines

The halting problems discussed in 9.1.3 involved all Turing machines, and it is less clear that there is a *single* machine with an unsolvable halting problem. Such a machine comes to light when we reflect on the process of reconstructing a machine M from its standard description $\ulcorner M \urcorner$.

First we separate $\ulcorner M \urcorner$ into its quintuples and find those beginning with the initial state $q_1 = q'$. Then, given a standard description $\ulcorner I \urcorner$ of M's input I, we can "simulate" the computation of M on I by reading and marking the leftmost symbol of $\ulcorner I \urcorner$, marking the quintuple of $\ulcorner M \urcorner$ that deals with this symbol, and thereafter moving from marked quintuple to marked symbol, carrying out the required action, and updating the marks accordingly. It is clear that a human computer can simulate the action of M on I in this way and hence, by Church's thesis, so can a Turing machine. We call such a machine a *universal Turing machine.* The existence of universal machines was first pointed out by Turing 1936.

Since the coding of I into $\ulcorner I \urcorner$ replaces □ with □ (see 9.1.3), if M eventually converts I into a blank tape, then the universal machine will convert $\ulcorner M \urcorner \ulcorner I \urcorner$ to $\ulcorner M \urcorner$ (assuming marks are removed after the simulation is completed). Therefore, if we construct a special universal machine T that erases $\ulcorner M \urcorner$ only after it finishes simulating a computation of M that halts on a blank tape then

T, given input $\ulcorner M \urcorner \ulcorner I \urcorner$, eventually halts on a blank tape

$\Leftrightarrow$ M, given input I, eventually halts on a blank tape.

Thus the unsolvability of the general halting problem (9.1.3) implies the unsolvability of the halting problem for T, even on the special inputs ⌜M⌝ ⌜I⌝. This in turn implies the unsolvability of the *general halting problem for T*, consisting of the questions

Q_E: Does T, given input E, eventually halt on blank tape?

We do not actually need this strong unsolvability result for the main results of this chapter, the unsolvability of the isomorphism problem and the homeomorphism problem. However, an explicit universal machine is needed to produce an explicit group with unsolvable word problem, so some readers may wish to do the following lengthy exercise.

EXERCISE 9.1.4.1. Construct a universal Turing machine.

9.1.5 Z^2-Machines

The difficulty in proving the homeomorphism problem unsolvable lies in reducing the halting problem to the word problem for groups and, to a lesser extent, reducing the word problem to the homeomorphism problem. In this section we shall pave the way for the reduction to the word problem by viewing computation as the composition of functions on subsets of Z^2.

At any instant in the computation of a machine M, the future actions of M are determined by a word

$$S_{k_u} \dots S_{k_2} S_{k_1} q_i S_{j_1} S_{j_2} \dots S_{j_v}$$

we shall call the *complete state* of M. Here $S_{k_u} \dots S_{k_2} S_{k_1} S_{j_1} S_{j_2} \dots S_{j_v}$ is the expression on the marked portion of tape; q_i is the current (internal) state of M; and S_{j_1} is the scanned symbol. (Thus the position of the q symbol in the complete state gives the position of the read/write head.) The *computation* of M can be identified with the sequence of complete states, and the transformation of one complete state to its successor (resulting from the response of M to $q_i S_{j_1}$) is called a *step* of computation.

We now replace the complete state with the pair

$$(S_{k_u} \dots S_{k_2} S_{k_1} q_i,\ S_{j_v} \dots S_{j_2} S_{j_1})$$

obtained by splitting the complete state at the scanned symbol and writing the right-hand portion backward. This pair, which obviously carries the same information as the complete state, is called the *complete state pair*. Its advantage over the complete state is that a step of computation changes only the right-hand ends of the elements of the pair, and these changes are easy to express arithmetically when we interpret the symbols $\square$, $S_1, \dots, S_m, q_1, \dots, q_n$ as the digits for $0, 1, \dots, m+n+1$ in base $b = m+n+2$ numerals.

Before looking at changes in the complete state pair, note the unobtrusive way in which extra blank tape is conjured up by the numerical interpretation. Since $\square$ is interpreted as 0, the blank tape to the left of S_{k_u} and to the right of S_{j_v} is represented by the unwritten zeros to the left of the numerals $S_{k_u} \dots S_{k_1} q_1$ and $S_{j_v} \dots S_{j_2} S_{j_1}$.

Now let us see what steps of computation look like when complete state pairs are interpreted as pairs of numerals. A "left-moving" quintuple $q_i S_j S_{j'} L q_{i'}$ transforms the complete state

$$S_{k_u} \dots S_{k_2} S_{k_1} q_i S_j S_{j_2} \dots S_{j_v}$$

into

$$S_{k_u} \dots S_{k_2} q_{i'} S_{k_1} S_{j'} S_{j_2} \dots S_{j_v}$$

and hence transforms the complete state pair

$$(S_{k_u} \dots S_{k_2} S_{k_1} q_i,\ S_{j_v} \dots S_{j_2} S_j)$$

into

$$(S_{k_u} \dots S_{k_2} q_{i'},\ S_{j_v} \dots S_{j_2} S_{j'} S_{k_1}).$$

More concisely, the quintuple $q_i S_j S_{j'} L q_{i'}$ induces the transformation of complete state pairs $(q_i S_j)$: $(U S_{k_1} q_i,\ V S_j) \mapsto (U q_{i'},\ V S_{j'} S_{k_1})$ for any digit S_{k_1} and any "q-free" numerals $U = S_{k_u} \dots S_{k_2}$ and $V = S_{j_v} \dots S_{j_2}$.

These transformations can be expressed in terms of multiplication by b and b^2 and addition of one- and two-digit numbers as follows:

$$(q_i S_j): (b^2 U + S_{k_1} q_i,\ bV + S_j) \mapsto (bU + q_{i'},\ b^2 V + S_{j'} S_{k_1})$$

where $U, V \in \mathbf{Z}$ are positive integers whose base b numerals have the special form. In fact, there is no harm in allowing arbitrary $U, V \in \mathbf{Z}$. A pair with a negative U (respectively, V) component transforms into another pair with a negative U (respectively, V) component, and a positive pair with more than one q-symbol transforms into another positive pair with more than one q-symbol. Hence, a complete state pair cannot arise from a noncomplete state pair.

This leads us to associate $q_i S_j S_{j'} L q_{i'}$ with a set of transformations of pairs of integers we call *l-transformations*. They are defined for all $U, V \in \mathbf{Z}$ and are of the form

$$(l) \qquad (b^2 U + A_l,\ bV + B_l) \mapsto (bU + C_l,\ b^2 V + D_l)$$

where $A_l = S_{k_1} q_i$, $B_l = S_j$, $C_l = q_{i'}$, $D_l = S_{j'} S_{k_1}$. Since S_{k_1} can take $m + 1$ different values there are $m + 1$ different transformations associated with a given "left-moving" quintuple $q_i S_j S_{j'} L q_{i'}$. Similarly, we associate each "right-moving" quintuple with a set of *r-transformations* of the form

$$(r) \qquad (bU + A_r,\ b^2 V + B_r) \mapsto (b^2 U + C_r,\ bV + D_r)$$

where $U, V \in \mathbf{Z}$ are arbitrary and $\{l\}$ and $\{r\}$ are disjoint sets of indices.

Corresponding to each Turing machine M we now have a finite set Z_M of l- and r-transformations that "simulate" M in the following sense. The pairs

(X, Y) of integers whose base b numerals form the successive complete state pairs of M are produced by successive application of l- or r-transformations, and exactly one such transformation applies to each (X, Y) in the sequence except the last—corresponding to halting of M—to which no transformation applies. We call Z_M a Z^2-*machine*. (It is a slight variation of the "modular machine" of Cohen and Aanderaa 1980.)

EXERCISE 9.1.5.1. What are the values of A_r, B_r, C_r, D_r in the transformations (r) corresponding to the quintuple $q_i S_j S_{j'} \mathrm{R} q_{i'}$?

9.1.6 The Halting Problem for Z^2-Machines

The simulation of Turing machines by Z^2-machines naturally reduces the general halting problem, or the general halting problem for the universal machine T, to a halting problem for Z^2-machines. A direct reduction, however, does not lead to the most simply stated problem, since the final complete state of Z_T involves the final state q_f of T (which could be any one of several states). Namely,

$$\begin{aligned} &T \text{ halts on a blank tape in state } q_f \\ &\quad\Leftrightarrow \text{ the final pair produced by } Z_T \text{ is } (q_f, 0). \end{aligned}$$

In order to avoid mention of q_f we modify T and Z_T as follows.

Introduce a new state q_0 and add the quintuple $q_f \square \square \mathrm{R} q_0$ to T for each state q_f in which T can halt on a $\square$. This creates a universal machine T^*, which halts on a blank tape *in state* q_0 if and only if T halts on a blank tape. Then

$$\begin{aligned} &T \text{ halts on a blank tape} \\ &\quad\Leftrightarrow \text{ the final pair produced by } Z_{T^*} \text{ is } (q_0, 0). \end{aligned}$$

Now add extra l-transformations to Z_{T^*}, which convert $(q_0, 0)$ to $(0, 0)$ and which apply only after $(q_0, 0)$ has been created, namely,

$$(*) \qquad (b^2 U + S_{k_1} q_0, bV) \mapsto (bU + S_{k_1}, b^2 V).$$

If we now allow Z^2-machines to include transformations $(*)$ and denote the thus expanded Z_{T^*} by $Z^*_{T^*}$, then

$$\begin{aligned} &T \text{ halts on a blank tape} \\ &\quad\Leftrightarrow \text{ the final pair produced by } Z^*_{T^*} \text{ is } (0, 0). \end{aligned}$$

The general halting problem for T is thereby reduced to the following problem about integers, which we call the *halting problem for* Z^2-*machines*: *for each* $(X, Y) \in \mathsf{Z}^2$ *decide whether* (X, Y) *is convertible to* $(0, 0)$ *by the set* $Z^*_{T^*}$ *of l- and r-transformations*. The latter problem is therefore unsolvable.

EXERCISE 9.1.6.1. Show unsolvability of the following problem without appealing to the existence of a universal machine. Given a set S of l- and r-transformations, and $(X, Y) \in \mathbb{Z}^2$, decide whether S converts (X, Y) to $(0, 0)$.

9.2 HNN Extensions

9.2.1 Representation of Computation in Groups

In 9.1.5 we showed that computation could be realized by transformations of subsets of $\mathbb{Z}^2$ (l- and r-transformations). The next step is to encode pairs $(X, Y) \in \mathbb{Z}^2$ by elements $p(X, Y)$ of a certain group K, and l- and r-transformations by certain isomorphisms of subgroups of K. This seemingly roundabout realization of computation, due to Cohen and Aanderaa 1980, supersedes earlier more direct approaches. In the direct approach, complete states of a Turing machine are represented by words, and quintuples by relations, in a relatively obvious way. The trouble is that symbols in a group are necessarily subject to relations with no computational counterpart, such as $q_i q_i^{-1} = 1$, and enormous ingenuity is required to prevent these relations creating "fake computations." The most important outcome of the direct approach was an appreciation of the value of the HNN construction, due to Higman, Neumann, and Neumann 1949. This construction was found to be the key to the success of Novikov 1955 and other early proofs of the unsolvability of the word problem. (For more historical information, see Stillwell 1982.) It is also crucial in the work of Cohen and Aanderaa, which leads to the unsolvability of the word problem much more easily.

Cohen and Aanderaa begin with the group

$$K = \langle x, y, z; xy = yx \rangle$$

and they encode the pair $(X, Y) \in \mathbb{Z}^2$ by the element $p(X, Y) = y^{-Y}x^{-X}zx^{X}y^{Y} \in K$. This is not the simplest encoding of pairs one can imagine, but simpler ones lack certain properties that will be essential later. For the moment we only wish to observe how an l-transformation

$$(l) \qquad (b^2U + A_l, bV + B_l) \mapsto (bU + C_l, b^2V + D_l)$$

can be reflected in K.

The corresponding transformation of group elements,

$$p(b^2U + A_l, bV + B_l) \mapsto p(bU + C_l, b^2V + D_l),$$

is induced by

$$x^{b^2} \mapsto x^b, \qquad y^b \mapsto y^{b^2}, \qquad p(A_l, B_l) \mapsto p(C_l, D_l),$$

which defines a map ϕ_l of the subgroup of K generated by x^{b^2}, y^b, $p(A_l, B_l)$

onto the subgroup generated by x^b, y^{b^2}, $p(C_l, D_l)$. We shall in general denote the group generated by $g_1, g_2, \ldots, g_m$ by $\langle g_1, g_2, \ldots, g_m \rangle$, hence in this notation

$$\phi_l\colon \langle x^{b^2}, y^b, p(A_l, B_l)\rangle \to \langle x^b, y^{b^2}, p(C_l, D_l)\rangle.$$

The map ϕ_l is in fact an isomorphism (see 9.3.2). There is a similar isomorphism $\phi_r\colon \langle x^b, y^{b^2}, p(A_r, B_r)\rangle \to \langle x^{b^2}, y^b, p(C_r, D_r)\rangle$ for each r-transformation.

This is where the HNN construction makes itself useful. Given any group G and pairs of elements $b_i, c_i \in G$ such that $b_i \mapsto c_i$ defines an isomorphism between the subgroups $\langle\{b_i\}\rangle$ and $\langle\{c_i\}\rangle$, respectively, *we can embed G in a group H in which the isomorphism $b_i \mapsto c_i$ is induced by conjugation by an element $t \in H - G$, that is, $t^{-1}b_it = c_i$.* In fact, H has a presentation consisting of the presentation of G, plus a generator t, plus the relations $t^{-1}b_it = c_i$. We abbreviate this definition of H by

$$H = G \cup \langle t; \{t^{-1}b_it = c_i\}\rangle$$

and call H an HNN *extension* of G with *stable letter t*.

According to the preceding result, which will be proved in 9.2.3, K can be embedded in a group in which the isomorphisms ϕ_l and ϕ_r are induced by conjugation by letters t_l and t_r. That is, if ϕ_l sends $p(X, Y)$ to $p(X', Y')$ then $t_l^{-1}p(X, Y)t_l = p(X', Y')$, and similarly for ϕ_r. Since ϕ_l sends $p(X, Y)$ to $p(X', Y')$ only if (X, Y) goes to (X', Y') by transformation (l), we have finally arrived at a group in which steps of computation are reflected by equations between words.

The crucial properties of the HNN construction are the embedding property already mentioned and a property of the stable letters called Britton's lemma. The embedding property says that, after any number of HNN extensions starting from G, elements of G are equal only if they were already equal in G. In particular, an element $p(X, Y) \in K$ will remain unequal to $p(X', Y')$ for $(X, Y) \neq (X', Y')$, and continue to be a faithful representative of the complete state pair (X, Y). Britton's lemma says, roughly, that words are only equal for "obvious" reasons, and hence that only the "intended" equations reflecting computations are true. To be specific, an equation expressing halting on a blank tape is true only if halting on a blank tape really occurs. This enables us to reduce the halting problem to the word problem and hence conclude that the word problem is unsolvable.

In 9.2.2–9.2.4 we shall prove these properties of HNN extensions. We shall then be in a position to prove the basic unsolvability results about groups in 9.3.

Exercise 9.2.1.1. Show that K is an HNN extension of the free group $\langle x, z; -\rangle$.

Exercise 9.2.1.2. Show that the surface group (cf. 4.2.1)

$$\langle a_1, b_1, \ldots, a_n, b_n; a_1b_1a_1^{-1}b_1^{-1}\ldots a_nb_na_n^{-1}b_n^{-1}\rangle$$

is an HNN extension of the free group $\langle b_1, a_2, b_2, \ldots, a_n, b_n; -\rangle$.

EXERCISE 9.2.1.3. Check that ϕ_l sends $p(b^2U + A_l, bV + B_l)$ to $p(bU + C_l, b^2V + D_l)$ for any $U, V \in \mathbb{Z}$.

9.2.2 Normal Forms

Suppose G is a group and $\phi: B \to C$ is an isomorphism of subgroups. Corresponding to ϕ we have the following HNN extension of G with stable letter t:

$$H = G \cup \langle t; \{t^{-1}bt = \phi(b) | b \in B\}\rangle.$$

If B has a finite generating set $\{b_i\}$, as will be the case in all our work, then H is defined by adding the finitely many relations $t^{-1}b_it = \phi(b_i)$ to G. We now study the elements of H, following the exposition in Stillwell 1982.

A typical word in H looks like

$$g_0 t^{\varepsilon_1} g_1 t^{\varepsilon_2} \ldots t^{\varepsilon_k} g_k$$

where each $\varepsilon_i = \pm 1$ and each g_i is a word in the generators of G (possibly 1). Each g_i can be factored into an element of B or C and a "residue," that is, a coset representative of g_i modulo B or C. Writing the relation $t^{-1}bt = \phi(t)$ as $t\phi(b) = bt$ or $t^{-1}b = \phi(b)t^{-1}$, we see that an element $\phi(b) \in C$ can always pass to the left across t, becoming b on the other side, while $b \in B$ can always pass to the left across t^{-1}, becoming $\phi(b)$ on the other side. This suggests normalizing the word by draining off elements of B or C to the left, leaving residues stuck between the ts.

To make this process precise (though not necessarily computable), we chose specific coset representatives; g^B of the coset Bg, g^C of the coset Cg, with 1 as the representative of both B and C. Then we work from right to left as follows.

If $\varepsilon_k = -1$ we factorize g_k into $B_k g_k^B$, where $B_k \in B$, so that

$$t^{\varepsilon_k} g_k = t^{-1} B_k g_k^B = \phi(B_k) t^{-1} g_k^B = \phi(B_k) t^{\varepsilon_k} g_k^B.$$

Similarly, if $\varepsilon_k = +1$ we factorize g_k into $C_k g_k^C$, where $C_k \in C$ and

$$t^{\varepsilon_k} g_k = t C_k g_k^C = \phi^{-1}(C_k) t g_k^C = \phi^{-1}(C_k) t^{\varepsilon_k} g_k^C.$$

(Since ϕ is an isomorphism, ϕ^{-1} is well defined.) We now have either $g_{k-1}\phi(B_k)$ or $g_{k-1}\phi^{-1}(C_k)$ between $t^{\varepsilon_{k-1}}$ and t^{ε_k}. We factorize it similarly, according to the sign of ε_{k-1}, and continue passing elements of B or C to the left, leaving coset representatives behind. If at any stage t and t^{-1} appear with only 1 between them, they are canceled. The final result is a word of the form

$$g_0' t^{\delta_1} g_1' t^{\delta_2} \ldots t^{\delta_e} g_e', \qquad \delta_i = \pm 1$$

where g_0' is an arbitrary element of G, $\delta_i = -1 \Rightarrow g_i'$ is a coset representative of G mod B, $\delta_i = +1 \Rightarrow g_i'$ is a coset representative of G mod C, and t, t^{-1} do not occur as consecutive letters. This word is still not unique, because g_i' can be any word from the equivalence class $[g_i']$ in G. Let us denote the class of words that results from replacing g_i' with other representatives of $[g_i']$ by

$$[g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\dots t^{\delta_e}[g'_e]$$

and call this class a *normal form* of the element of H it represents.

EXERCISE 9.2.2.1. Viewing $\langle a_1, b_1, a_2, b_2; a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\rangle$ as an HNN extension, as in Exercise 9.2.1.2, check that the word $a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}$ has normal form [1].

9.2.3 Uniqueness of Normal Forms

The normal form of an element of H is unique.

We shall faithfully represent H as a group of permutations of the set N of normal forms. For each $h \in H$ we shall define a mapping $\Phi_h: N \to N$ with the properties

(i) Φ_1 is the identity,
(ii) $\Phi_{h_1}\Phi_{h_2} = \Phi_{h_1h_2}$, and
(iii) $\Phi_{\text{normal form } h}(1) =$ normal form h.

From (i) and (ii) it follows that $h \mapsto \Phi_h$ is a homomorphism, in particular $\Phi_h\Phi_{h^{-1}} =$ identity, so Φ_h is invertible, hence a permutation. From (iii) it follows that normal forms are unique, because Φ_{h_1}, Φ_{h_2} for different normal forms h_1, h_2 send 1 to different places, so that Φ_{h_1}, Φ_{h_2} are different permutations and hence represent different elements h_1, h_2.

The mapping Φ_h is defined to be "multiply on the left by h and reduce to normal form, one letter of h at a time." Then (i)–(iii) are clear, but since the word for an element $h \in H$ is not unique, the problem is to show that Φ_h is well defined. This requires showing that equivalent words determine the same permutation, in other words, that each defining relator of H determines the identity permutation. For a relator g of G this is clear, because

$$\Phi_g([g'_0]t^{\delta_1}\dots t^{\delta_e}[g'_e]) = [gg'_0]t^{\delta_1}\dots t^{\delta_e}[g'_e] = [g'_0]t^{\delta_1}\dots t^{\delta_e}[g'_e]$$

since $g = 1$ in G. (Although we are supposed to apply g one letter at a time, it is clear that these letters just accumulate to the left of g'_0, since no interaction with t is possible. We shall similarly present just the end results of letter-by-letter accumulations later.)

The relators involving t are tt^{-1}, $t^{-1}t$ and $t^{-1}bt\phi(b)^{-1}$. We check just the last of these, since the first two are similar, but easier.

$$\begin{aligned}
&\Phi_{t^{-1}bt\phi(b)^{-1}}([g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\dots)\\
&= \Phi_{t^{-1}}\Phi_b\Phi_t\Phi_{\phi(b)^{-1}}([g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\dots)\\
&= \Phi_{t^{-1}}\Phi_b\Phi_t([\phi(b)^{-1}g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\dots)\\
&= \Phi_{t^{-1}}\Phi_b(\text{normal form of } t[\phi(b)^{-1}g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\dots).
\end{aligned}$$

There are now three cases to consider:

(a) $[\phi(b)^{-1}g'_0] = 1$, i.e., $\phi(b) = g'_0$, and $\delta_1 = -1$;
(b) $[\phi(b)^{-1}g'_0] = 1$, i.e., $\phi(b) = g'_0$, and $\delta_1 = +1$; and
(c) $[\phi(b)^{-1}g'_0] \neq 1$.

In case (a), t and t^{δ_1} cancel and we continue with

$$\begin{aligned}
&\Phi_{t^{-1}}\Phi_b([g'_1]t^{\delta_2}\ldots)\\
&\quad = \Phi_{t^{-1}}([bg'_1]t^{\delta_2}\ldots)\\
&\quad = \text{normal form of } t^{-1}[bg'_1]t^{\delta_2}\ldots\\
&\quad = [\phi(b)]t^{-1}[g'_1]t^{\delta_1}\ldots \quad \text{since } g'_1 \text{ is a coset representative}\\
&\quad = [g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\ldots \quad \text{since } \phi(b) = g'_0,\ \delta_1 = -1 \quad \text{by hypothesis.}
\end{aligned}$$

In case (b) we continue with

$$\begin{aligned}
&\Phi_{t^{-1}}\Phi_b(t[1]t[g'_1]t^{\delta_2}\ldots) = \Phi_{t^{-1}}([b]t[1]t[g'_1]t^{\delta_2}\ldots)\\
&\quad = \text{normal form of } t^{-1}[b]t[1]t[g'_1]t^{\delta_2}\ldots\\
&\quad = \text{normal form of } [\phi(b)]t^{-1}t[1]t[g'_1]t^{\delta_2}\ldots\\
&\quad = [\phi(b)]t[g'_1]t^{\delta_2}\ldots, \quad \text{canceling } t^{-1}t\\
&\quad = [g'_0]t^{\delta_1}[g'_1]t^{\delta_2}\ldots \quad \text{since } \phi(b) = g'_0,\ \delta_1 = 1 \quad \text{by hypothesis.}
\end{aligned}$$

In case (c) we let $\phi(b)^{-1}g'_0 = \phi(\bar{b}_0)^{-1}g_0^C$, where $\phi(\bar{b}_0)^{-1} \in C$ and g_0^C is the coset representative, and continue with

$$\begin{aligned}
&\Phi_{t^{-1}}\Phi_b(\text{normal form of } t[\phi(\bar{b}_0)^{-1}g_0^C]t^{\delta_1}\ldots)\\
&\quad = \Phi_{t^{-1}}\Phi_b([\bar{b}_0^{-1}]t[g_0^C]t^{\delta_1}\ldots)\\
&\quad = \Phi_{t^{-1}}([b\bar{b}_0^{-1}]t[g_0^C]t^{\delta_1}\ldots)\\
&\quad = \text{normal form of } t^{-1}[b\bar{b}_0^{-1}]t[g_0^C]t^{\delta_1}\ldots\\
&\quad = \text{normal form of } [\phi(b\bar{b}_0^{-1})]t^{-1}t[g_0^C]t^{\delta_1}\ldots \quad \text{since } b\bar{b}_0^{-1} \in B\\
&\quad = [\phi(b\bar{b}_0^{-1})g_0^C]t^{\delta_1}\ldots, \quad \text{canceling } t^{-1}t\\
&\quad = [\phi(b)\phi(\bar{b}_0)^{-1}g_0^C]t^{\delta_1}\ldots \quad \text{since } \phi \text{ is a homomorphism}\\
&\quad = [g'_0]t^{\delta_1}\ldots \quad \text{since } \phi(\bar{b}_0)^{-1}g_0^C = \phi(b)^{-1}g'_0 \quad \text{by definition.}
\end{aligned}$$

Thus $\Phi_{t^{-1}bt\phi(b)^{-1}}$ is indeed the identity. □

Since each $[g] \in G$ is identical with its normal form, the proof shows in particular that distinct $[g_1], [g_2] \in G$ are also distinct in H. That is, *G embeds in H* (Higman, Neumann, and Neumann 1949).

9.2.4 Britton's Lemma

If w is a word involving t that equals 1 in H, then w contains either a subword $t^{-1}bt$, where $b \in B$, or a subword tct^{-1}, where $c \in C = \phi(B)$ (Britton 1963).

The normal form of w is 1, hence ts must be canceled in the normalization process. Suppose, for example, that $t^{-1}bt$ is a subword of w whose ts get canceled in the normalization process. Normalization inserts a word $\bar{b} \in B$ between b and t, and the ts cancel only if $b\bar{b} \in B$, because only then can $b\bar{b}$ be moved to the left, leaving no residue. But $b\bar{b} \in B \Leftrightarrow b \in B$.

Similarly, if tct^{-1} is a subword of w whose ts are canceled during normalization, then $c \in C$. □

This result is Britton's lemma for an HNN extension H of G with a single stable letter t. It easily generalizes to any group H_n that is nth in a sequence

$$H_1 = G \cup \langle t_1; \{t_1^{-1}\, dt_1 = \phi_1(d) | d \in D_1\}\rangle$$
$$\vdots$$
$$H_n = H_{n-1} \cup \langle t_n; \{t_n^{-1}\, dt_n = \phi_n(d) | d \in D_n\}\rangle$$

of HNN extensions with stable letters $t_1, \ldots, t_n$. Britton's lemma for such an extension H_n (which we call an HNN *extension of G with stable letters* $t_1, \ldots, t_n$) reads:

If w is a word involving t_is that equals 1 in H_n, then w contains a subword $t_i^{-1}d_it_i$ where $d_i \in D_i$, or a subword $t_ie_it_i^{-1}$, where $e_i \in E_i = \phi_i(D_i)$.

If we consider the t_i of highest index in w, and view w as an element of the HNN extension H_i of H_{i-1}, then this result is immediate from Britton's lemma for extensions with the single stable letter t_i. □

EXERCISE 9.2.4.1. Derive an algorithm for the word problem for the group $\langle a_1, b_1, \ldots, a_n, b_n; a_1b_1a_1^{-1}b_1^{-1} \ldots a_nb_na_n^{-1}b_n^{-1}\rangle$ by viewing it as an HNN extension and applying Britton's lemma.

9.3 Unsolvable Problems in Group Theory

9.3.1 Faithful Representation of Complete State Pairs

The Cohen and Aanderaa representation of computation by groups was outlined in 9.2.1. We shall now use properties of HNN extensions to show that this representation is "faithful" in a certain sense, and hence gives a reduction of the halting problem to the word problem. The first step is to show that $p(X, Y) = y^{-Y}x^{-X}zx^Xy^Y$ faithfully represents the pair (X, Y).

The group $K = \langle x, y, z; xy = yx \rangle$ containing the elements $p(X, Y)$ is the HNN extension

$$\langle x, z; \text{—} \rangle \cup \langle y; y^{-1}xy = x \rangle$$

of the free group $\langle x, z; \text{—} \rangle$, with stable letter y inducing the identity isomorphism of the subgroup $\langle x \rangle$. This yields the following result, which shows how faithfully (X, Y) is represented by $p(X, Y)$.

The elements $p(X, Y) = y^{-Y}x^{-X}zx^{X}Y^{Y}$ *for* $X, Y \in \mathbb{Z}$ *are free generators of a free subgroup of* K.

Since $p(X, Y)^j = y^{-Y}x^{-X}z^jx^Xy^Y$, any reduced word on generators $p(X_i, Y_i)$ can be written in the form

$$w = y^{-Y_1}x^{-X_1}z^{j_1}x^{X_1}y^{Y_1} \cdot y^{-Y_2}x^{-X_2}z^{j_2}x^{X_2}y^{Y_2} \ldots y^{-Y_k}x^{-X_k}z^{j_k}x^{X_k}y^{Y_k},$$

which simplifies to

$$w = y^{-Y_1}x^{-X_1}z^{j_1}x^{X_1-X_2}y^{Y_1-Y_2}z^{j_2} \ldots x^{X_{k-1}-X_k}y^{Y_{k-1}-Y_k}z^{j_k}x^{X_k}y^{Y_k}$$

by commuting x, y and canceling. The fact that w is a reduced word on the $p(X_i, Y_i)$ means in particular that no consecutive powers of the same $p(X_i, Y_i)$ occur, and hence each term $x^{X_{i-1}-X_i}y^{Y_{i-1}-Y_i}$ between powers of z has either $X_{i-1} - X_i \neq 0$ or $Y_{i-1} - Y_i \neq 0$. Thus the simplified w is freely reduced as a word on x, y, z and hence, by Britton's lemma, $w = 1$ only if it contains a subword $y^{-1}x^jy$ or yx^jy^{-1}. This is not the case, since z appears between any consecutive occurrences of $y^{\pm 1}$. Hence $w \neq 1$ and therefore the elements $p(X_i, Y_i)$ are free generators. □

This result means in particular that $p(X, Y) = p(X', Y') \Leftrightarrow (X, Y) = (X', Y')$, not only in K, but also in any HNN extension of K (by the embedding property, 9.2.3). More generally, $p(X, Y)$ *is in the subgroup generated by a set* $\{p(X_i, Y_i)\} \Leftrightarrow p(X, Y) = p(X_i, Y_i)$ *for some* i, since the $p(X_i, Y_i)$ are free generators. We shall call this property *faithful representation of* (X, Y) *by* $p(X, Y)$ *in* K *and its HNN extensions.*

The following exercise shows that certain simpler methods of encoding pairs in groups are not faithful in the same sense.

Exercise 9.3.1.1. If $p(X, Y) = x^Xy^Y$ in the free group $\langle x, y; \text{—} \rangle$, show that $p(X, Y) = p(X + 1, Y)p(i + 1, Y)^{-1}p(i, Y)$ for any $i \in \mathbb{Z}$. Show that the same relation holds if $p(X, Y) = x^Xzy^Y$ in the free group $\langle x, y, z; \text{—} \rangle$. (An attempted simplification of the Cohen and Aanderaa construction in Stillwell 1982 fails because of this relation.)

9.3.2 Representation of l- and r-Transformations

Cohen and Aanderaa realize an l-transformation

$$(b^2U + A_l, bV + B_l) \mapsto (bU + C_l, b^2V + D_l)$$

by the map ϕ_l:

$$x^{b^2} \mapsto x^b, \qquad y^b \mapsto y^{b^2}, \qquad p(A_l, B_l) \mapsto p(C_l, D_l)$$

of the subgroup $\langle x^{b^2}, y^b, p(A_l, B_l)\rangle$ onto the subgroup $\langle x^b, y^{b^2}, p(C_l, D_l)\rangle$. We now prove the result claimed in 9.2.1:

ϕ_l is an isomorphism.

It suffices to show that $x^{b^2} \mapsto x$, $y^b \mapsto y$, $p(A_l, B_l) \mapsto z$ and $x^b \mapsto x$, $y^{b^2} \mapsto y$, $p(C_l, D_l) \mapsto z$ are both isomorphisms onto $\langle x, y, z; xy = yx\rangle$. (Compose the first with the inverse of the second to get ϕ_l.)

Since $p(A_l, B_l) = y^{-B_l}x^{-A_l}zx^{A_l}y^{B_l}$, conjugation by $x^{-A_l}y^{-B_l}$ is an isomorphism $x^{b^2} \mapsto x^{b^2}$, $y^b \mapsto y^b$, $p(A_l, B_l) \mapsto z$ of $\langle x^{b^2}, y^b, p(A_l, B_l)\rangle$ onto $\langle x^{b^2}, y^b, z\rangle$. And since $\langle x^{b^2}, y^b\rangle$ is the free abelian group generated by x^{b^2}, y^b, we also have the isomorphism $x^{b^2} \mapsto x$, $y^b \mapsto y$, $z \mapsto z$ of $\langle x^{b^2}, y^b, z\rangle = \langle x^{b^2}, y^b\rangle * \langle z\rangle$ onto $\langle x, y, z\rangle = \langle x, y\rangle * \langle z\rangle$.

The composite of these two isomorphisms is $x^{b^2} \mapsto x$, $y^b \mapsto y$, $p(A_l, B_l) \mapsto z$.

We similarly show that $x^b \mapsto x$, $y^{b^2} \mapsto y$, $p(C_l, D_l) \mapsto z$ is an isomorphism, as required. □

There is a similar proof that the map ϕ_r:

$$x^b \mapsto x^{b^2}, \qquad y^{b^2} \mapsto y^b, \qquad p(A_r, B_r) \mapsto p(C_r, D_r)$$

realizing an r-transformation is an isomorphism of $\langle x^b, y^{b^2}, p(A_r, B_r)\rangle$ onto $\langle x^{b^2}, y^b, p(C_r, D_r)\rangle$.

It follows that *we can make an* HNN *extension K_Z of K by adding*

$$\begin{aligned} \textit{generator } t_l, \textit{ relations} \quad t_l^{-1}x^{b^2}t_l &= x^b, \\ t_l^{-1}y^bt_l &= y^{b^2}, \\ t_l^{-1}p(A_l, B_l)t_l &= p(C_l, D_l), \\ \text{generator } t_r, \textit{ relations} \quad t_r^{-1}x^bt_r &= x^{b^2}, \\ t_r^{-1}y^{b^2}t_r &= y^b, \\ t_r^{-1}p(A_r, B_r)t_r &= p(C_r, D_r), \end{aligned}$$

for each transformation (l) *or* (r) *making up a* $\mathbb{Z}^2$*-machine* Z.

By construction, K_Z has the property that if a complete state pair (X', Y') results from a complete state pair (X, Y) by transformation (l) then

$$t_l^{-1}p(X, Y)t_l = p(X', Y').$$

Because if transformation (l) applies to (X, Y) we must have

$$(X, Y) = (b^2U + A_l, bV + B_l) \qquad \text{for some } U, V \in \mathbb{Z},$$

in which case

$$\begin{aligned} t_l^{-1}p(X, Y)t_l &= t_l^{-1}p(b^2U + A_l, bV + B_l)t_l \\ &= t_l^{-1}y^{-bV}x^{-b^2U}p(A_l, B_l)x^{b^2U}y^{bV}t_l \\ &= y^{-b^2V}x^{-bU}p(C_l, D_l)x^{bU}y^{b^2V} \\ &= p(bU + C_l, b^2V + D_l) \\ &= p(X', Y'). \end{aligned}$$

Similarly for a transformation (r). Thus steps of computation of Z are represented by equations in K_Z. In 9.3.3 we shall show that equations of this form hold *only* if the corresponding computational steps exist, and hence that the representation is faithful.

9.3.3 Faithful Representation of Computational Steps

If $t_l^{-1}p(X, Y)t_l = p(X', Y')$ *in* K_Z *then there are* $U, V \in \mathbb{Z}$ *such that*

$$X = b^2U + A_l, \qquad Y = bV + B_l,$$

$$X' = bU + C_l, \qquad Y' = b^2V + D_l.$$

That is, the equation reflects an actual computational step of Z.

Rewriting the equation as

$$t_l^{-1}p(X, Y)t_lp(X', Y')^{-1} = 1,$$

we see from Britton's lemma in K_Z that $p(X, Y)$ belongs to the subgroup $\langle x^{b^2}, y^b, p(A_l, B_l)\rangle$ of K. It is clear that this subgroup contains all elements of the form $p(b^2U + A_l, bV + B_l)$ for $U, V \in \mathbb{Z}$. Conversely, the general element of $\langle x^{b^2}, y^b, p(A_l, B_l)\rangle$ has the form

$$y^{-n_1b}x^{-m_1b^2}p(A_l, B_l)^{l_1}x^{m_2b^2}y^{n_2b}p(A_l, B_l)^{l_2}\dots p(A_l, B_l)^{l_k}x^{m_{k+1}b^2}y^{n_{k+1}b}.$$

By rewriting the x, y term between $p(A_l, B_l)^{l_1}$ and $p(A_l, B_l)^{l_2}$ as $x^{m_1b^2}y^{n_1b}.x^{-(m_1-m_2)b^2}y^{-(n_1-n_2)b}$ we create the term

$$y^{-n_1b}x^{-m_1b^2}p(A_l, B_l)^{l_1}x^{m_1b^2}y^{n_1b} = p(b^2m_1 + A_l, bn_1 + B_l)^{l_1},$$

and by successively rewriting the x, y terms from left to right in a similar way the general element takes the form

$$\left(\prod_i p(b^2U_i + A_l, bV_i + B_l)\right)x^{mb^2}y^{nb}.$$

Since this equals $p(X, Y)$ by hypothesis, it follows by Britton's lemma in K (cf. argument in 9.3.1) that $m = n = 0$ and the product of terms $p(b^2U_i + A_l, bV_i + B_l)$ is freely equal to the single term $p(X, Y)$. Thus

$$p(X, Y) = p(b^2U + A_l, bV + B_l) \qquad \text{for some } U, V,$$

and hence

$$X = b^2U + A_l, \qquad Y = bV + B_l$$

by the faithful representation of pairs (9.3.1).

It now follows from the defining relations of K_Z (9.3.2) that

$$\begin{aligned} t_l^{-1}p(X, Y)t_l &= t_l^{-1}p(b^2U + A_l, bV + B_l)t_l \\ &= p(bU + C_l, b^2V + D_l) \end{aligned}$$

and hence

$$p(X', Y') = p(bU + C_l, b^2V + D_l),$$

whence

$$X' = bU + C_l, \qquad Y' = b^2V + D_l,$$

by the faithful representation of pairs again. □

We similarly prove the corresponding result for a t_r:
If $t_r^{-1}p(X, Y)t_r = p(X', Y')$ in K_Z then there are $U, V \in \mathbf{Z}$ such that

$$\begin{aligned} X &= bU + A_r, & Y &= b^2V + B_r, \\ X' &= b^2U + C_r, & Y' &= bV + D_r. \end{aligned}$$

Thus conjugation by a t_l or t_r faithfully reflects a transformation (l) or (r). However, it is equally true that conjugation by t_l^{-1} or t_r^{-1} does not reflect a transformation (l) or (r) but, rather, its inverse:

$$t_l p(bU + C_l, b^2V + D_l)t_l^{-1} = p(b^2U + A_l, bV + B_l),$$

or

$$t_r p(b^2U + C_r, bV + D_r)t_r^{-1} = p(bU + A_r, b^2V + B_r),$$

This corresponds to a fake backward computational step:

$$(bU + C_l, b^2V + D_l) \mapsto (b^2U + A_l, bV + B_l)$$

or

$$(b^2U + C_r, bV + D_r) \mapsto (bU + A_r, b^2V + B_r).$$

In 9.3.4 we shall show that backward steps cannot create a path from (X, Y) to $(0, 0)$ unless there is a path from (X, Y) to $(0, 0)$ by forward steps alone. In this sense K_Z contains a faithful representation of halting computations.

9.3.4 Faithful Representation of Halting Computations

A halting computation of a $\mathbf{Z}^2$-machine Z (cf. 9.1.6) with initial complete state pair (X, Y) is a series of steps $s_1, \ldots, s_{k-1}$:

$$(X, Y) = (X_1, Y_1) \xrightarrow{s_1} (X_2, Y_2) \xrightarrow{s_2} \ldots \xrightarrow{s_{k-1}} (X_k, Y_k) = (0, 0),$$

where each s_i is a transformation (l) or (r). Now suppose we admit *backward steps* from (X_i, Y_i) to (X_{i+1}, Y_{i+1}), writing $(X_i, Y_i) \overset{s}{\leftarrow} (X_{i+1}, Y_{i+1})$ to indicate that (X_{i+1}, Y_{i+1}) is a pair (not necessarily unique) such that $(X_{i+1}, Y_{i+1}) \overset{s}{\rightarrow} (X_i, Y_i)$. Then we have the following result.

If (X, Y) *is a complete state pair convertible to* $(0, 0)$ *by forward and backward steps, then* (X, Y) *is convertible to* $(0, 0)$ *by forward steps alone.*

If (X, Y) is convertible to $(0, 0)$ then we can obtain a sequence $(X, Y) = (X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k) = (0, 0)$ *without repetitions* such that each (X_{i+1}, Y_{i+1}) results from (X_i, Y_i) by a forward or backward step. It suffices to omit any segments of the original sequence between repeated terms. I claim that all steps in a nonrepeating sequence are forward.

This is because any pairs accessible from a complete state pair are themselves complete state pairs, except $(0, 0)$, which occurs only as the last pair in the sequence by hypothesis, and which only results from the forward step $(q_0, 0) \rightarrow (0, 0)$ by definition of Z^2-machines (9.1.6). Thus there is at least one forward step in the sequence, namely, $(X_{k-1}, Y_{k-1}) \rightarrow (X_k, Y_k) = (0, 0)$, and all the terms $(X_1, Y_1), \ldots, (X_{k-1}, Y_{k-1})$ are complete state pairs. It follows that any backward step occurs in a context $(X_{i-1}, Y_{i-1}) \leftarrow (X_i, Y_i) \rightarrow (X_{i+1}, Y_{i+1})$ where (X_i, Y_i) is a complete state pair, and hence that $(X_{i-1}, Y_{i-1}) = (X_{i+1}, Y_{i+1})$ by the determinacy of machines. This contradicts the absence of repetitions. □

Combining this result with 9.3.3 we get:

A Z^2*-machine* Z *converts a complete state pair* (X, Y) *to* $(0, 0) \Leftrightarrow p(X, Y) = w^{-1}p(0, 0)w$ *in* K_Z, *where* w *is a word on the letters* $t_l^{\pm 1}$ *and* $t_r^{\pm 1}$.

Thus we have reduced the halting problem for Z^2-machines (9.1.6) to a problem about special equations in K_Z. With the help of Britton's lemma, the latter is easily reduced to the generalized word problem for K_Z.

9.3.5 The Generalized Word Problem

The halting problem for a Z^2*-machine* Z *is reducible to the generalized word problem for* K_Z.

In fact we shall show that

$$Z \text{ converts a complete state pair } (X, Y) \text{ to } (0, 0)$$

$$\Leftrightarrow \quad p(X, Y) \in \langle p(0, 0), \{t_l\}, \{t_r\}\rangle.$$

The direction ($\Rightarrow$) is immediate from the ($\Rightarrow$) direction of the result at the end of 9.3.4 (which is essentially the fact that conjugations by t_l, t_r reflect computational steps of Z).

To prove ($\Leftarrow$), suppose $p(X, Y) \in \langle p(0, 0), \{t_l\}, \{t_r\}\rangle$. This means $p(X, Y) = T_1 p(0, 0)^{j_1} T_2 \ldots T_k p(0, 0)^{j_k} T_{k+1}$, where each T_i is a word on the t_l and t_r, and hence

(*) $$T_1 p(0, 0)^{j_1} T_2 \dots T_k p(0, 0)^{j_k} T_{k+1} p(X, Y)^{-1} = 1.$$

By Britton's lemma for K_Z the left-hand side of (*) must contain a subword of one of the following forms

(i) $t_l^{-1} w t_l$, where $w \in \langle x^{b^2}, y^b, p(A_l, B_l)\rangle$,
(ii) $t_l w t_l^{-1}$, where $w \in \langle x^b, y^{b^2}, p(C_l, D_l)\rangle$,
(iii) $t_r^{-1} w t_r$, where $w \in \langle x^b, y^{b^2}, p(A_r, D_r)\rangle$,
(iv) $t_r w t_r^{-1}$, where $w \in \langle x^{b^2}, y^b, p(C_r, D_r)\rangle$.

Since the terms T_i in (*) consist entirely of letters $t_l^{\pm 1}$ and $t_r^{\pm 1}$, w must in fact be $p(0, 0)^{j_i}$, in which case (i) and (iii) simplify to the form $p(u, v)^{j_i}$, where $(0, 0) \rightarrow (u, v)$ in Z, and (ii) and (iv) simplify to $p(u', v')^{l_i}$, where $(0, 0) \leftarrow (u', v')$ in Z.

Repeating this argument until all letters $t_l^{\pm 1}$ and $t_r^{\pm 1}$ are eliminated, (*) simplifies to the form

(**) $$\left(\prod_i p(u_i, v_i)^{j_i}\right) p(X, Y)^{-1} = 1,$$

where each (u_i, v_i) is obtainable from (0, 0) by a series of backward or forward computational steps by Z. It then follows, since the elements $p(u_i, v_i)$ are free generators (9.3.1), that $\prod_i p(u_i, v_i)^{j_i}$ freely reduces to the single term $p(X, Y)$, and hence (X, Y) itself is obtained from (0, 0) by backward or forward computational steps. In other words, (0, 0) is obtainable from (X, Y) by forward or backward steps and hence, by 9.3.4, by forward steps alone.

This shows that the halting problem for Z is reducible to the generalized word problem for K_Z. That is, the problem of deciding, given words $w_1, \dots, w_p$ and w, whether w belongs to the subgroup $\langle w_1, \dots, w_p\rangle$ of K generated by $w_1, \dots, w_p$. Thus if we choose a machine Z for which the halting problem is unsolvable, such as the machine $Z^*_{T^*}$ of 9.1.6, we find:

There is a finitely presented group K_Z with unsolvable generalized word problem.

Exercise 9.3.5.1. Show, without assuming the assistance of a universal machine, the unsolvability of the *general generalized word problem*: given a finitely presented group G, and words $w_1, \dots, w_p, w$, decide whether $w \in \langle w_1, \dots, w_p\rangle$ in G.

9.3.6 The Word Problem

A group K_Z with an unsolvable generalized word problem is easily extended to a finitely presented group G_Z with unsolvable word problem by a trick of Boone 1959. Namely, let

$$G_Z = K_Z \cup \langle k; k^{-1} p(0, 0) k = p(0, 0), \{k^{-1} t_l k = t_l\}, \{k^{-1} t_r k = t_r\}\rangle.$$

This is an HNN extension of K_Z, with stable letter k inducing the identity isomorphism of $\langle p(0, 0), \{t_l\}, \{t_r\}\rangle$.

The generalized word problem for K_Z is reducible to the word problem for G_Z (hence the latter is unsolvable when Z has an unsolvable halting problem).

Since k commutes with $p(0, 0)$, $\{t_l\}$ and $\{t_r\}$, a word $p(X, Y) \in \langle p(0, 0), \{t_l\}, \{t_r\}\rangle$ satisfies

$$kp(X, Y) = p(X, Y)k.$$

Conversely, if this equation holds we have

$$kp(X, Y)k^{-1}p(X, Y)^{-1} = 1.$$

Since the only occurrences of k, k^{-1} on the left-hand side are those explicitly shown, Britton's lemma for G_Z tells us that $p(X, Y) \in \langle p(0, 0), \{t_l\}, \{t_r\}\rangle$. Hence

$$p(X, Y) \in \langle p(0, 0), \{t_l\}, \{t_r\}\rangle \quad \Leftrightarrow \quad kp(X, Y)k^{-1}p(X, Y) = 1.$$

This equivalence reduces the generalized word problem for K_Z (or, rather, the special case of it equivalent to the halting problem by 9.3.5) to the word problem for G_Z. □

A presentation for a G_Z with unsolvable word problem can be constructed, in principle, from a universal machine Z. It would obviously be enormously complicated. In fact, no really simple or natural group with unsolvable word problem is yet known. However, one can use the existence of G_Z to show that the *generalized* word problem for $F_2 \times F_2$ (direct product of free groups of rank 2) is unsolvable. This surprising result was discovered by Mikhailova 1958, and her argument is outlined in the following exercises.

EXERCISE 9.3.6.1. Let $G = \langle a_1, \ldots, a_p; r_1, \ldots, r_q\rangle$ and let $F_p = \langle a_1, \ldots, a_p; —\rangle$. Recalling from 0.5.6 that

$$w = 1 \text{ in } G \quad \Leftrightarrow \quad w = \prod_k g_k r_{j_k}^{-\varepsilon_k} g_k^{-1} \text{ in } F_p, \text{ show that}$$

$$w = 1 \text{ in } G \quad \Leftrightarrow \quad (w, 1) \in \langle (r_1, 1), \ldots, (r_q, 1), (a_1, a_1), \ldots, (a_p, a_p)\rangle \text{ in } F_p \times F_p.$$

EXERCISE 9.3.6.2. Deduce that the word problem for G is reducible to the generalized word problem for $F_p \times F_p$.

EXERCISE 9.3.6.3. Using 2.2.7.3 or otherwise, reduce the generalized word problem for $F_p \times F_p$ to the generalized word problem for $F_2 \times F_2$.

EXERCISE 9.3.6.4. Give a finite presentation of $F_2 \times F_2$.

9.3.7 The Isomorphism Problem

To obtain our final unsolvable problem in group theory, we shall reduce the word problem for G_Z (9.3.6) to the isomorphism problem. The reduction depends only on the following fact: *each $w \neq 1$ in G_Z is of infinite order*. This is certainly true of the group $K = \langle x, y; xy = yx \rangle * \langle z \rangle$ we started with. Elements of K retain their orders under successive HNN extensions, by the embedding property of 9.2.3. And each element involving a stable letter is easily seen to be of infinite order by applying Britton's lemma to the powers of its normal form.

Now for each word w in the generators of G_Z, which we rename $a_1, \ldots, a_p$, we construct the group (which is *not* necessarily an HNN extension)

$$G_Z(w) = G_Z \cup \langle \{k_i\}; \{k_i^{-1} w k_i = a_i\} \rangle.$$

Then the reduction of the word problem for G_Z to the isomorphism problem is immediate from the following result.

$$w = 1 \textit{ in } G \quad \Leftrightarrow \quad G_Z(w) \cong \langle k_1, \ldots, k_p; - \rangle \cong F_p.$$

The direction ($\Rightarrow$) holds because

$$w = 1 \Rightarrow a_i = k_i^{-1} k_i = 1 \Rightarrow G_Z(w) = \langle k_1, \ldots, k_p; - \rangle.$$

Conversely, if $w \neq 1$ in G_Z then w is of infinite order, as is a_i. Hence $w \mapsto a_i$ is an isomorphism and $G_Z(w)$ is an HNN extension of G_Z. In particular, G_Z embeds in $G_Z(w)$ by 9.2.3, and hence $G_Z(w)$ has an unsolvable word problem. Then $G_Z(w)$ cannot be a free group F_p, because the word problem for F_p is solvable for any set of generators. (If $G_Z(w) \cong F_p$ with free generators $x_1, \ldots, x_p$, let $a_i(x_j)$, $k_i(x_j)$ be words expressing the generators a_i, k_i of $G_Z(w)$ in terms of $x_1, \ldots, x_p$. Replace the letters a_i, k_i of a given word W by $a_i(x_j)$, $k_i(x_j)$, and hence decide whether $W = 1$ by free reduction of the x_js, as in 2.1.4). □

Since G_Z was chosen to have an unsolvable word problem, this reduction brings us to our goal: *the isomorphism problem for finitely presented groups is unsolvable.*

EXERCISE 9.3.7.1. Without assuming the existence of a universal machine Z, prove the unsolvability of the *general word problem*: given a finitely presented group G and a word w, decide whether $w = 1$ in G.

EXERCISE 9.3.7.2. Deduce the unsolvability of the isomorphism problem—given finitely presented groups F and G, decide whether $F \cong G$—without assuming a universal machine.

9.4 The Homeomorphism Problem

9.4.1 Manifolds with Given Fundamental Group

A natural way to prove unsolvability of the homeomorphism problem is to show that the isomorphism problem is reducible to it. A reduction would be immediate if we had a construction of complexes $C(\mathscr{P})$ from presentations $\mathscr{P}$

with the following properties

(i) $\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P}') \Rightarrow C(\mathscr{P})$ homeomorphic to $C(\mathscr{P}')$
(ii) $\pi_1(C(\mathscr{P})) = \text{Group}(\mathscr{P})$,

where $\text{Group}(\mathscr{P})$ denotes the group with presentation $\mathscr{P}$. This is because complexes with different fundamental groups are not homeomorphic, hence (i) and (ii) give the equivalence

$$\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P}') \quad \Leftrightarrow \quad C(\mathscr{P}) \text{ homeomorphic to } C(\mathscr{P}'),$$

which yields the required reduction.

Unfortunately, the known methods of constructing complexes from presentations can give nonhomeomorphic $C(\mathscr{P})$, $C(\mathscr{P}')$ for different presentations $\mathscr{P}$, $\mathscr{P}'$ of the same group. In particular, the surface complex constructed in 3.4.4 to realize the group $\langle a_1, \ldots, a_m; r_1, \ldots, r_n \rangle$ as fundamental group actually depends on the generators $a_1, \ldots, a_m$ and relators $r_1, \ldots, r_n$. This is why we are now taking special care to distinguish between a presentation $\mathscr{P}$ and the group it defines, $\text{Group}(\mathscr{P})$. We shall continue to use the notation $\langle a_1, \ldots, a_m; r_1, \ldots, r_n \rangle$ for the group, but write the presentation simply as $a_1, \ldots, a_m; r_1, \ldots, r_n$. It is then appropriate to describe the surface complex of 3.4.4 as $S(a_1, \ldots, a_m; r_1, \ldots, r_n)$.

Now, to see that the complex $S(\mathscr{P})$ depends on the presentation $\mathscr{P}$, consider the two presentations $a_1; a_1$ and $a_1, a_2; a_1, a_2$ of the trivial group. $S(a_1; a_1)$ is a disc, while $S(a_1, a_2; a_1, a_2)$ is the union of a pair of discs with a single common point (Figure 297). These are nonhomeomorphic because, for example, $S(a_1, a_2; a_1, a_2)$ is disconnected by the removal of a single point and $S(a_1; a_1)$ is not.

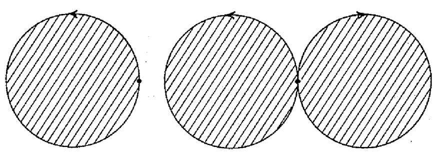


Figure 297

This difficulty was first overcome by Markov 1958, using the Seifert and Threlfall construction of a 4-manifold with given finitely presented fundamental group (Exercise 8.1.8.1). A second approach, based on an idea of Dehn 1912a for construction of a 4-manifold with given fundamental group, was developed by Boone, Haken, and Poénaru 1968. Both these approaches yield the unsolvability of the homeomorphism problem restricted to closed 4-manifolds. We shall be content to get unsolvability for open 5-manifolds, in order to avoid detailed high-dimensional constructions, but we shall follow Boone, Haken, and Poénaru part of the way. Readers familiar with manifolds of dimension >3 should find it easy to strengthen the proof so as to obtain unsolvability for closed 4-manifolds. (The results of Markov and Boone, Haken, and Poénaru are also stronger than ours in a few more technical aspects. However, our proof is substantially simpler than theirs.)

The idea of Dehn 1912a is to embed the surface complex $S(\mathscr{P})$, for a finite presentation $\mathscr{P}$ in R^5. Since $S(\mathscr{P})$ can obviously be triangulated, this is possible by 0.2.3. Then the ε-neighbourhood

$$M_\varepsilon(\mathscr{P}) = \{P \in \mathsf{R}^5 | \text{distance}(P, S(\mathscr{P})) < \varepsilon\}$$

is an open set, and hence a 5-manifold, and for ε sufficiently small it has a deformation retraction to $S(\mathscr{P})$. Assuming ε is chosen sufficiently small we therefore have

$$\pi_1(M_\varepsilon(\mathscr{P})) = \pi_1(S(\mathscr{P})) = \text{Group}(\mathscr{P}).$$

It is unfortunately still not true that $M_\varepsilon(\mathscr{P})$ is homeomorphic to $M_\varepsilon(\mathscr{P}')$ whenever $\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P}')$ (see 9.4.2), but the situation is better than with $S(\mathscr{P})$ and $S(\mathscr{P}')$. The counterexample for $S(\mathscr{P})$ where $\mathscr{P}$ is $a_1; a_1$ and $\mathscr{P}'$ is $a_1, a_2; a_1, a_2$ is *not* a counterexample for $M_\varepsilon(\mathscr{P})$, since the ε-neighbourhoods of $S(a_1; a_1)$ and $S(a_1, a_2; a_1, a_2)$ are homeomorphic (each to an open 5-ball in fact). Figure 298 indicates these neighbourhoods in R^5 schematically. One

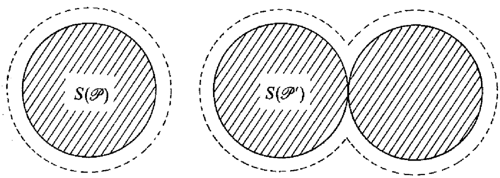


Figure 298

way to recognize the homeomorphism is to view $M_\varepsilon(\mathscr{P}')$ as a neighbourhood (of points at distances between $\varepsilon/2$ and ε) of the topological disc D shown in Figure 299, which is $\varepsilon/2$ larger than $S(\mathscr{P}')$. A homeomorphism $D \to S(\mathscr{P})$ can then be extended to a homeomorphism of their neighbourhoods.

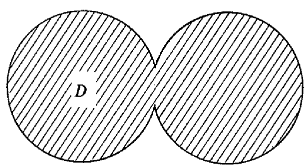


Figure 299

It is hoped that readers will be content with sketchy descriptions of homeomorphisms, like the example just given. We need only three, intuitively simple, homeomorphisms (see 9.4.5), so it seems a waste to develop a detailed theory of high-dimensional manipulation for their sake. Nevertheless, readers who are aware of such details will be in a better position to do the following exercise, which is the key to obtaining the unsolvability result for closed 4-manifolds. See also Boone, Haken, and Poénaru 1968.

EXERCISE 9.4.1.1. Show that the frontier of $M_\varepsilon(\mathscr{P})$ is a closed 4-manifold with the same fundamental group as $M_\varepsilon(\mathscr{P})$.

9.4.2 Reducing Isomorphism to Homeomorphism

If $\mathscr{P}$ and $\mathscr{P}'$ are different presentations of the same group, then we know from 0.5.8 that there is a sequence of Tietze transformations, $T_1^{\pm 1}$ or $T_2^{\pm 1}$, converting $\mathscr{P}$ to $\mathscr{P}'$. Thus the dream of 9.4.1, of finding complexes $C(\mathscr{P})$ such that

$$\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P}') \Rightarrow C(\mathscr{P}) \text{ homeomorphic to } C(\mathscr{P}'),$$

could come true if the $C(\mathscr{P})$ were "invariant under Tietze transformations." That is, if $C(\mathscr{P})$ were homeomorphic to $C(\mathscr{P}')$ whenever $\mathscr{P}'$ resulted from $\mathscr{P}$ by a Tietze transformation.

Since the known constructions of complexes from presentations all fail to give homeomorphic $C(\mathscr{P})$, $C(\mathscr{P}')$ for certain $\mathscr{P}$, $\mathscr{P}'$ defining the same group, they necessarily fail to be invariant under certain Tietze transformations. The examples of $S(a_1; a_2)$ and $S(a_1, a_2; a_1, a_2)$ show that the surface complex $S(\mathscr{P})$ fails to be invariant under T_2 (adding a generator). $M_\varepsilon(\mathscr{P})$ in fact *is* invariant under T_2 (see 9.4.5), but not under T_1 (adding a consequence relator).

Consider, for example, the presentations $a_1; a_1$ and $a_1; a_1, a_1$ of the trivial group. $M_\varepsilon(a_1; a_1)$ is the ε-neighbourhood of a disc in R^5, hence a 5-ball, while $M_\varepsilon(a_1; a_1, a_1)$ is the ε-neighbourhood of a sphere. These are not homeomorphic because $M_\varepsilon(a_1; a_1, a_1)$ contains a noncontractible sphere and $M_\varepsilon(a_1; a_1)$ does not.

Further study of the $M_\varepsilon(\mathscr{P})$ reveals that addition of redundant relators, as in this case, is the whole trouble. Since it may be necessary to add redundant relators to convert $\mathscr{P}$ to a $\mathscr{P}'$ which has more relators than $\mathscr{P}$, this trouble cannot be entirely eliminated. However, it turns out that we can preempt it by adding redundant relators to $\mathscr{P}$ *in advance*. Given a $\mathscr{P}'$ that may define the same group as $\mathscr{P}$, we can find, from the size of $\mathscr{P}$ and $\mathscr{P}'$, a bound t on the number of relators that must be added in the course of converting $\mathscr{P}$ to $\mathscr{P}'$. If we then add t trivial relators 1 to $\mathscr{P}$ at the beginning, it turns out that we can convert the enlarged $\mathscr{P}$ to $\mathscr{P}'$ (plus some trivial relators 1) by *special* Tietze transformations *under which* $M_\varepsilon(\mathscr{P})$ *is invariant.*

Following Markov 1958 and Boone, Haken, and Poénaru 1968, we denote the result of adding t trivial relators to $\mathscr{P}$ by $\mathscr{P} * t$. Then the reduction of the isomorphism problem to the homeomorphism problem may be described as follows.

Given presentations $\mathscr{P}$ and $\mathscr{P}'$, compute t and t' such that $\mathscr{P} * t$ is convertible to $\mathscr{P}' * t'$ by special Tietze transformations if Group($\mathscr{P}$) = Group($\mathscr{P}'$). Then construct the manifolds $M_\varepsilon(\mathscr{P} * t)$ and $M_\varepsilon(\mathscr{P}' * t')$. Since

$$\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P} * t) = \pi_1(M_\varepsilon(\mathscr{P} * t)),$$

and similarly for $\mathscr{P}'$, t', and since $M_\varepsilon(\mathscr{P} * t)$ is homeomorphic to $M_\varepsilon(\mathscr{P}' * t')$ whenever $\mathscr{P} * t$ is convertible to $\mathscr{P}' * t'$ by special Tietze transformations, we have

$$\begin{aligned}\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P}') \quad &\Leftrightarrow \quad \text{Group}(\mathscr{P} * t) = \text{Group}(\mathscr{P}' * t')\\ &\Leftrightarrow \quad M_\varepsilon(\mathscr{P} * t) \text{ is homeomorphic to } M_\varepsilon(\mathscr{P}' * t').\end{aligned}$$

Since $M_\varepsilon(\mathscr{P} * t)$, $M_\varepsilon(\mathscr{P}' * t')$ are computable from $\mathscr{P}$, $\mathscr{P}'$, this equivalence gives the desired reduction.

The details of t, t' and the special Tietze transformations will be worked out in 9.4.3 and 9.4.4. The invariance of $M_\varepsilon(\mathscr{P})$ under the special Tietze transformations will be shown in 9.4.5, to complete the proof.

The only other point worth stressing right now is that the manifolds $M_\varepsilon(\mathscr{P} * t)$, $M_\varepsilon(\mathscr{P}' * t')$ have natural finite descriptions, so it is valid to assume that a (hypothetical) algorithm for the homeomorphism problem is applicable to them. $M_\varepsilon(\mathscr{P})$ can be described by a list of vertices, edges, and faces of the triangulated complex $S(\mathscr{P})$ in R^5, plus the number ε. Since the coordinates of the vertices can be taken to be rational, without loss of generality, as can ε, this is a finite description.

9.4.3 Special Tietze Transformations

The special Tietze transformations we shall use are T_2 and four "small" transformations T_{11}, T_{12}, T_{13}, T_{14}, which replace T_1. The latter are chosen to be "small enough" that their effect on $M_\varepsilon(P)$ is apparent (see 9.4.5), yet comprehensive enough to reconstruct any relator that could be added by T_1. Unlike T_1, however, T_{11}, T_{12}, T_{13}, T_{14} can add consequence relations only when given sufficient "room," in an enlargement of $\mathscr{P}$ to the equivalent presentation $\mathscr{P} * t$ with t copies of the trivial relator 1.

We shall begin by defining T_{11}, T_{12}, T_{13}, T_{14} for an arbitrary finite presentation $a_1, \ldots, a_m; r_1, \ldots, r_n$, then work out the amount of "room" they require to add a single consequence of $r_1, \ldots, r_n$ to the presentation.

T_{11}: Replace r_i by $a_j a_j^{-1} r_i$ or $a_j^{-1} a_j r_i$,

T_{12}: Replace $r_i = uvw$ by a cyclic permutation vwu,

T_{13}: Replace r_i by r_i^{-1}

T_{14}: Replace r_i by $r_i r_j$ with $j \neq i$.

These transformations differ from T_1 not only in being "small," but also in *replacing* a relator rather than adding one. T_{11}, T_{12}, T_{13} obviously give a presentation P' defining the same group as $\mathscr{P}$. T_{14} does also, thanks to the restriction $j \neq i$, which allows r_i to be recovered from the new relator $r_i r_j$ and the relator r_j still present. The fact that T_{11}, T_{12}, T_{13}, T_{14} replace relations rather than adding them is the secret of their success in keeping $M_\varepsilon(\mathscr{P}')$ homeomorphic to $M_\varepsilon(\mathscr{P})$, as we shall see in 9.4.5. At the same time, it is the reason they need an enlargement $\mathscr{P} * t$ of $\mathscr{P}$ to work on.

If $\mathscr{P}$ is the presentation $a_1, \ldots, a_m; r_1, \ldots, r_n$ and $\mathscr{P}'$ is the presentation $a_1, \ldots, a_m; r_1, \ldots, r_n, s$, where s is a consequence of $r_1, \ldots, r_n$, then $\mathscr{P} * 2$ may be converted to $\mathscr{P}' * 1$ by T_{11}, T_{12}, T_{13}, T_{14} and their inverses.

By 0.5.6, any consequence s of $r_1, \ldots, r_n$ has the form

$$s = g_1 r_{j_1}^{\pm 1} g_1^{-1} \cdot g_2 r_{j_2}^{\pm 1} g_2^{-1} \cdot \ldots \cdot g_k r_{j_k}^{\pm 1} g_k^{-1},$$

where $g_1, \ldots, g_k$ are words on $a_1, \ldots, a_m$. The idea is to construct the factors $g_1 r_{j_1}^{\pm 1} g_1^{-1}$, $g_2 r_{j_2}^{\pm 1} g_2^{-1}, \ldots$ successively in the place of the last 1 in $\mathscr{P} * 2$, and to build up s in the place of the first 1 by successive applications of S_{14}. In more detail, the manipulation of the sequence of relators goes as follows.

$r_1, \ldots, r_n, 1, 1$

$\rightarrow r_1, \ldots, r_n, 1, r_{j_1}$ by T_{14}

$\rightarrow r_1, \ldots, r_n, 1, r_{j_1}^{\pm 1}$ by T_{13} (if necessary)

$\rightarrow r_1, \ldots, r_n, 1, a_{i_1}^{-1} a_{i_1} r_{j_1}^{\pm 1}$ by T_{11} (where a_{i_1} = last letter of g_1)

$\rightarrow r_1, \ldots, r_n, 1, a_{i_1} r_{j_1}^{\pm 1} a_{i_1}^{-1}$ by T_{12}

$\rightarrow r_1, \ldots, r_n, 1, a_{i_2} a_{i_1} r_{j_1}^{\pm 1} a_{i_1}^{-1} a_{i_2}^{-1}$ by T_{11}, T_{12}

$\vdots$ (where a_{i_2} = second last letter of g_1)

$\rightarrow r_1, \ldots, r_n, 1, g_1 r_{j_1}^{\pm 1} g_1^{-1}$ by T_{11}, T_{12} repeatedly

$\rightarrow r_1, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, g_1 r_{j_1}^{\pm 1} g_1^{-1}$ by T_{14}

$\rightarrow r_1, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, r_{j_1}^{\pm 1}$ by T_{11}^{-1}, T_{12} repeatedly (reversing earlier steps)

$\rightarrow r_1, \ldots, r_{j_1}^{\mp 1}, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, r_{j_1}^{\pm 1}$ by T_{13} (if necessary)

$\rightarrow r_1, \ldots, r_{j_1}^{\mp 1}, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, r_{j_1}^{\mp 1} r_{j_1}^{\pm 1}$ by T_{14}

$\rightarrow r_1, \ldots, r_{j_1}^{\mp 1}, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, 1$ by T_{12}, T_{11}^{-1} repeatedly

$\rightarrow r_1, \ldots, r_{j_1}, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, 1$ by T_{13} (if necessary)

$\rightarrow r_1, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1}, g_2 r_{j_2}^{\pm 1} g_2^{-1}$ by T_{11}, T_{12}, T_{13}, T_{14} (as used to construct $g_1 r_{j_1}^{\pm 1} g_1^{-1}$)

$\rightarrow r_1, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1} \cdot g_2 r_{j_2}^{\pm 1} g_2^{-1}, g_2 r_{j_2}^{\pm 1} g_2^{-1}$ by T_{14}

$\rightarrow r_1, \ldots, r_n, g_1 r_{j_1}^{\pm 1} g_1^{-1} \cdot g_2 r_{j_2}^{\pm 1} g_2^{-1}, 1$ by T_{11}^{-1}, T_{12}, T_{13}, T_{14}

$\vdots$ (as used to restore $g_1 r_{j_1}^{\pm 1} g_1$ to 1)

$\rightarrow r_1, \ldots, r_n, s, 1.$ □

9.4.4 Presentation of the Same Group

If $\mathscr{P}$(m generators, n relators) and $\mathscr{P}'$(m′ generators, n′ relators) are presentations of the same group, then $\mathscr{P} * (n + n' + 1)$ may be converted to $\mathscr{P}' * (n' + n + 1)$ by T_{11}, T_{12}, T_{13}, T_{14}, T_2 and their inverses.

We imitate the proof of Tietze's theorem (0.5.8), using the construction of 9.4.3 to add consequence relations. Because of the existence of inverse transformations it suffices, as in 0.5.8, to convert both $\mathscr{P} * (n + n' + 1)$ and $\mathscr{P}' * (n' + n + 1)$ to the same presentation $\mathscr{P}''$. As in 0.5.8 we let α_i' be a word on $a_1, \ldots, a_m$ representing a_i', and let α_i be a word on $a_1', \ldots, a_{m'}'$ representing a_i. We also abbreviate the list $a_1, \ldots, a_m$ by a_i, the list $a_1', \ldots, a_{m'}'$ by a_i', etc.. Thus $a_i; r_j * (n + n' + 1)$, for example, stands for the presentation $a_1, \ldots, a_m, r_1, \ldots, r_n, 1, 1, \ldots, 1$ with $n + n' + 1$ relators 1.

$a_i; r_j(a_i) * (n + n' + 1)$

$\to a_i; r_j(a_i), r_j'(\alpha_i') * (n + 1)$ by 9.4.3, n' times

$\to a_i, a_i'; r_j(a_i), r_j'(\alpha_i'), a_i' = \alpha_i' * (n + 1)$ by T_2, m' times

$\to a_i, a_i'; r_j(a_i), r_j'(a_i'), a_i' = \alpha_i' * (n + 1)$ by 9.4.3, n' times, and 9.4.3 inverse n times (alternately adding the consequence $r_j'(a_i')$ of $r_j'(\alpha_i')$ and $a_i' = \alpha_i'$, and subtracting the relator $r_j'(\alpha_i')$)

$\to a_i, a_i'; r_j(a_i), r_j'(a_i'), a_i' = \alpha_i', a_i = \alpha_i * 1$ by 9.4.3, n times.

This presentation $\mathscr{P}''$ is symmetric with respect to primed and unprimed symbols, hence $\mathscr{P}' * (n' + n + 1)$ is also convertible to $\mathscr{P}''$, as required. □

9.4.5 Invariance Under Special Tietze Transformations

If $\mathscr{P}'$ results from $\mathscr{P}$ by T_{11}, T_{12}, T_{13}, T_{14} or T_2 then $M_\varepsilon(\mathscr{P})$ is homeomorphic to $M_\varepsilon(\mathscr{P}')$.

If the transformation is T_{12} (cyclic permutation of a relator) or T_{13} (inversion of a relator), then $M_\varepsilon(\mathscr{P})$ is certainly homeomorphic to $M_\varepsilon(\mathscr{P}')$. In fact the surface complexes $S(\mathscr{P})$ and $S(\mathscr{P}')$ are identical in this case.

In the remaining three cases $S(\mathscr{P})$ and $S(\mathscr{P}')$ are *not* homeomorphic, but their ε-neighbourhoods $M_\varepsilon(\mathscr{P})$ and $M_\varepsilon(\mathscr{P}')$ are. The extra thickness of $M_\varepsilon(\mathscr{P})$ allows us to "slide" and "squash" part of it to form $M_\varepsilon(\mathscr{P}')$ as follows.

T_{11}: Replace r_i with $a_j a_j^{-1} r_i$ ($a_j^{-1} a_j r_i$ is similar).

Let $\mathscr{P}^-$ be $\mathscr{P}$ minus the relator r_i (which is also $\mathscr{P}'$ minus the relator $a_j a_j^{-1} r_i$). Then $S(\mathscr{P})$ is the result of attaching a disc D to the curve r_i in $S(\mathscr{P}^-)$, and $S(\mathscr{P}')$ is the result of attaching a disc D' to the curve $a_j a_j^{-1} r_i$. Then D meets the frontier of $M_\varepsilon(\mathscr{P}^-)$ in a simple closed curve d (since D, by construction of $S(\mathscr{P})$, has identified points only on its boundary) and D' meets the frontier of $M_\varepsilon(\mathscr{P}^-)$ in

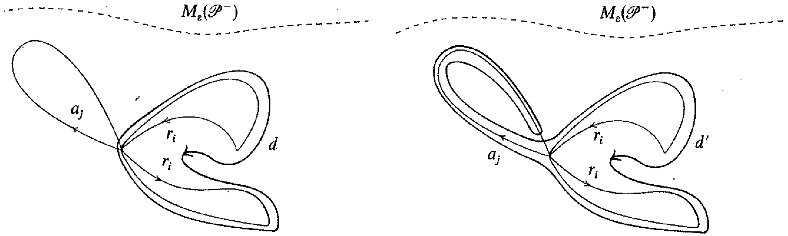


Figure 300

a simple closed curve d', which is close to d except for a "tongue" at distance $\leq \varepsilon$ from the path $a_j a_j^{-1}$ (Figure 300, a schematic view of the situation in R^5). Now $M_\varepsilon(\mathscr{P})$ is the union of $M_\varepsilon(\mathscr{P}^-)$ with the ε-neighbourhood D_ε of a closed disc with boundary d, while $M_\varepsilon(\mathscr{P}')$ is the union of $M_\varepsilon(\mathscr{P}^-)$ with the ε-neighbourhood D'_ε of a closed disc with boundary d' (the discs being subdiscs of D and D', respectively). We obtain a homeomorphism between $M_\varepsilon(\mathscr{P})$ and $M_\varepsilon(\mathscr{P}')$ by sliding d to position d'—extending the "tongue" in the ε-neighbourhood of a_j.

T_{14}: Replace r_i with $r_i r_j$, $\quad j \neq i$.

In this case $M_\varepsilon(\mathscr{P})$ is the union of $M_\varepsilon(\mathscr{P}^-)$ with the ε-neighbourhood D_ε of a closed disc with boundary d, a simple closed curve at distance $\leq \varepsilon$ from r_i. And $M_\varepsilon(\mathscr{P}')$ is the union of $M_\varepsilon(\mathscr{P}^-)$ with the ε-neighbourhood D'_ε of a closed disc with boundary d', a simple closed curve at distance $\leq \varepsilon$ from $r_i r_j$ (Figure 301, which incidentally illustrates how d is a simple curve, even when r_i is not). Since $j \neq i$, the curve r_j is in $\mathscr{P}^-$, and since r_j is a relator the curve is spanned by a disc. We can slide d across the ε-neighbourhood of this disc in $M_\varepsilon(\mathscr{P}^-)$, thus obtaining a homeomorphism between $M_\varepsilon(\mathscr{P})$ and $M_\varepsilon(\mathscr{P}')$.

T_2: Add generator a_{m+1} and relator $a_{m+1}w$, where w is a word in $a_1, \ldots, a_m$.

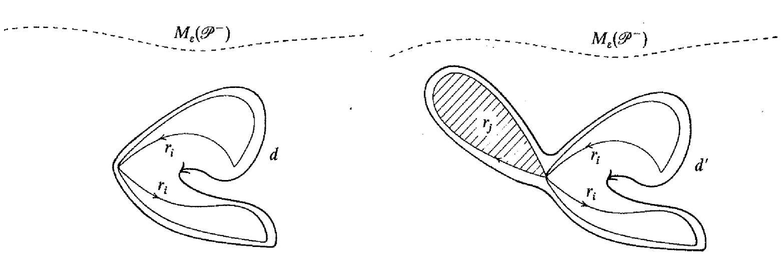


Figure 301

The new generator a_{m+1} adds a "handle" to $M_\varepsilon(\mathscr{P})$—the ε-neighbourhood of the arc of a_{m+1} outside $M_\varepsilon(\mathscr{P})$. $S(\mathscr{P}')$ is formed by attaching a disc to $S(\mathscr{P})$

along $a_{m+1}w$. Suppose this disc meets the frontier of $M_\varepsilon(\mathscr{P})$ plus handle in the simple closed curve $d_{m+1}d$, where d is the intersection with $M_\varepsilon(\mathscr{P})$ and d_{m+1} the intersection with the handle (Figure 302). Then $M_\varepsilon(\mathscr{P}')$ is formed by attaching the ε-neighbourhood D_ε of a disc bounded by $d_{m+1}d$. By contracting this disc we can "squash" the handle and D_ε down to $M_\varepsilon(\mathscr{P})$, thus showing $M_e(\mathscr{P}')$ to be homeomorphic to $M_\varepsilon(\mathscr{P})$. □

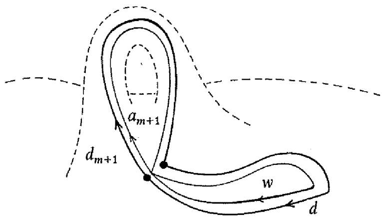


Figure 302

This result finally justifies the reduction of the isomorphism problem to the homeomorphism problem that was outlined in 9.4.2:

$$\text{Group}(\mathscr{P}) = \text{Group}(\mathscr{P}')$$

$$\Leftrightarrow \quad \text{Group}(\mathscr{P} * (n + n' + 1)) = \text{Group}(\mathscr{P}' * (n' + n + 1))$$

$$\Leftrightarrow \quad M_\varepsilon(\mathscr{P} * (n + n' + 1)) \text{ homeomorphic to } M_\varepsilon(\mathscr{P}' * (n' + n + 1)).$$

Since the isomorphism problem is unsolvable by 9.3.7, *the homeomorphism problem is unsolvable.* Readers who wish to carry their study of topology beyond 3-manifolds can therefore rest assured that the subject is not trivial.

EXERCISE 9.4.5.1. Use Exercise 9.4.1.1 to show that the homeomorphism problem is unsolvable for closed 4-manifolds.

Bibliography and Chronology

1752

Euler, L.: Elementa doctrinae solidorum, *Novi comment. acad. sci. Petrop.*, **4** (1752), 109–140.

The Euler polyhedron formula.

1771

Vandermonde, A.-T.: Remarques sur les Problèmes de Situation, *Histoire de l'Académie des Sciences* (*Paris*) 1771, 566–574.

Combinatorial description of knots, with examples from knitting.

1847

Listing, J. B.: Vorstudien zur Topologie, *Göttinger Studien* 1847, 811–875.

Introduces name "topology." States that the two trefoil knots are not the same.

1851

Hermite, C.: Sur les fonctions algebriques, *Comptes rendus, Paris* **32** (1851), 458–461.

The monodromy group.

Riemann, G. F. B.: Grundlagen für eine allgemeine Theorie der Funktionen einer veränderlichen complexen Grösse, *Gesammelte Mathematische Werke*, Dover, New York 1953, 3–43.

Riemann surfaces and connectivity.

1863

Möbius, A. F.: Theorie der elementaren Verwandtschaft, *Werke*, Vol. 2, Martin Sandig, Wiesbaden 1967, 433–471.

Surfaces in R^3 reduced to spheres with handles.

1866

Clebsch, A. and Gordan, P.: *Theorie der Abelschen Funktionen*, Teubner, Leipzig 1866.

Matrix generators for mapping class group of torus.

Jordan, C. (a): Sur la déformation des surfaces, *J. de Math.* (2) **XI** (1866), 105–109.

Attempts to characterize surfaces by connectivity.

Jordan, C. (b): Des contours tracés sur les surfaces, *J. de Math.* (2) **XI** (1866), 110–130.

Generators for π_1 of a surface, but incorrect relations.

1870

Jordan, C.: *Traité des Substitutions et des Équations Algébriques*, Gauthier-Villars, Paris 1870.

Introduces term "abelian," and notions of isomorphism and homomorphism.

Kronecker, L.: Auseinandersetzung einiger Eigenschaften der Klassenanzahl idealen complexer Zahlen, *Werke* Vol. I, Chelsea, New York 1968, 271–282.

The structure theorem for finite abelian groups.

1871

Betti, E.: Sopra gli spazi di un numero qualunque di dimensioni, *Annali di Matematica pura ed applicata* (2) **4** (1871), 140–158.

Betti numbers.

1872

Schwarz, H. A.: Ueber diejenige Fälle, in welchen die Gaussische hypergeometrische Reihe eine algebraische Function ihres vierten Elementes darstellt, *J. reine u. angew. Math.* **75** (1872), 292–335.

Tessellation of the unit disc by curvilinear triangles.

1874

Du Bois-Reymond, P.: Ueber asymptotische Werthe, infinitäre Approximationen und infinitäre Auflösung von Gleichungen, *Math. Ann.* **7** (1874), 363–367.

The diagonal argument.

Klein, F.: Bemerkungen über den Zusammenhang der Flächen, *Math. Ann.* **7** (1874), 549–557.

Sphere as double cover of the projective plane.

1876

Klein, F.: Ueber den Zusammenhang der Flächen, *Math. Ann.* **9** (1876), 476–482.

Intrinsic definition of orientability, using indicatrix.

1877

Clifford, W. K.: On the canonical form and dissection of a Riemann surface, *Proc. Lond. Math. Soc.* (1) **8** (1877), 292–304.

Riemann surface is a sphere with handles.

1878

Cayley, A.: On the theory of groups, *Proc. Lond. Math. Soc.* (1) **9** (1878), 126–133.

Cayley diagrams for finite groups.

1882

Dyck, W.: Gruppentheoretischen Studien, *Math. Ann.* **20** (1882), 1–44.

Generators and relations for Fuchsian groups, including surface groups.

Klein, F. (a): *Über Riemanns Theorie der Algebraischen Funktionen und Ihre Integrale*, Teubner, Leipzig 1882 = *Ges. Math. Abh.* III, 499–573. Translated as *On Riemann's Theory of Algebraic Functions and their Integrals*, Dover, New York 1963.

Sphere with handles, as we know it. The Klein bottle.

Klein, F. (b): Letter to Poincaré, 14 May 1882, *Ges. Math. Abh.* III, 615–616.

The universal covering surface.

Klein, F. (c): Neue Beiträge zur Riemannschen Funktionentheorie, *Math. Ann.* **21** (1882), 141–218.

Generators and relations for Fuchsian groups, including surface groups.

Poincaré, H.: Théorie des Groupes Fuchsiens, *Acta Math.* **1** (1882), 1–62.
Generators and relations for Fuchsian groups, including surface groups.

1883

Poincaré, H.: Mémoire sur les Groupes Kleiniens, *Acta Math.* **3** (1883), 49–92.
Poincaré model of the hyperbolic plane.

1884

Dyck, W.: *On the "Analysis situs" of Three Dimensional Spaces*, Brit. Assoc. Adv. Sci. Report 1884, 648.
Decomposition of 3-manifold into handlebodies.

1887

Appell, P.: Quelques rémarques sur la théorie des potentiels multiforms, *Math. Ann.* **30** (1887), 155–156.
A covering of R^3 branched over the circle.

Jordan, C.: *Cours d'Analyse*, First ed., Gauthier-Villars, Paris 1887.
Statement of Jordan curve theorem, but faulty proof.

1888

Dyck, W.: Beiträge zur Analysis situs I, *Math. Ann.* **32** (1888), 457–512.
Classification of nonorientable surfaces.

1890

Klein, F. and Fricke, R.: *Vorlesungen über die Theorie der elliptischen Modulfunktionen* Vol. I, Teubner, Stuttgart 1890 (vol II, 1892), Johnson Reprint Corp, New York 1966.
Defining relations for the modular group (p. 454).

1891

Cantor, G.: Über eine elementare Frage der Mannigfaltigkeitslehre, *Ges. Math. Abh.*, Georg Olm, Hildesheim 1962, 278–281.
The diagonal argument used to prove any set has more subsets than elements.

1892

Poincaré, H.: Sur l'Analysis situs, *Comptes rendus* **115** (1892), 633–636.
Betti numbers do not completely determine a manifold. Analytical definition of π_1.

1895

Poincaré, H.: Analysis situs, *J. de l'École Polyt.* (2) **1** (1895), 1–123.
Betti numbers, combinatorial definition of π_1, H_1 = abelianized π_1.

1897

Brunn, H.: Ueber verknotete Curven, *Verh. des intern. Math.-Congr.* **1** (1897), 256–259.
Any knot has a projection with a single multiple point.

Fricke, R. and Klein, F.: *Vorlesungen über die Theorie der automorphen Funktionen* vol I, Teubner, Stuttgart 1897 (vol II, 1912), Johnson Reprint Corp., New York 1965.

Hurwitz, A.: Über die Entwicklung der allgemeinen Theorie der analytischen Funktionen in neuerer Zeit, *Verh. des intern. Math.-Congr.* **1** (1897), 97–112.
Statement of the homeomorphism problem.

Sommerfeld, A.: Über verzweigte Potentiale in Raum, *Proc. Lond. Math. Soc.* **28** (1897), 395–429.
Some branched coverings of R^3.

1898

Heegaard, P.: Sur l'Analysis situs (French translation of 1898 Copenhagen dissertation), *Soc. Math. France Bull.* **44** (1916), 161–242.

Heegaard diagrams, some special branched covers of S^3 and computation of their Heegaard diagrams (2-fold, cyclic over trefoil = (3, 1) lens space, 2-fold irregular over trefoil = S^3 itself).

1899

Poincaré, H: 1[r] Complément à l'analysis situs, *Rend. circ. mat. Palermo* **13**, (1899), 285–343.

Simplicial complexes, combinatorial definition of Betti numbers.

1900

Poincaré, H.: 2[d] Complément à l'analysis situs, *Proc. Lond. Math. Soc.* **32** (1900), 277–308.

Torsion, matrix computation of Betti and torsion numbers (structure theorem for finitely generated abelian groups).

1904

Poincaré, H.: Cinquième complément à l'analysis situs, *Rend. circ. mat. Palermo* **18** (1904), 45–110.

Simple curves on surfaces—algorithms to decide which homology and homotopy classes contain them. Construction of homology sphere by Heegaard diagram. The Poincaré conjecture.

1905

Veblen, O.: Theory of plane curves in nonmetrical analysis situs, *Trans. Amer. Math. Soc.* **6** (1905), 83–98.

Rigorous proof of Jordan curve theorem.

1906

Schoenflies, A.: Beiträge zur Theorie der Punktmengen III, *Math. Ann.* **62** (1906), 286–328.

Jordan–Schoenflies theorem.

1907

Dehn, M. and Heegaard, P.: Analysis situs, *Encykl. Math. Wiss.* vol III AB3 Leipzig 1907, 153–220.

Classification theorem for surfaces proved.

1908

Schoenflies, A.: Die Entwicklung der Lehre von den Punktmannigfaltigkeiten II, Jber. *Deutsch. Math. Verein* (2 Ergänzungsband), Leipzig 1908.

Steinitz, E.: Beiträge zur Analysis situs *Sitz.-Ber. Berlin Math. Ges.* **7** (1908), 29–49, also in *Arch. d. Math. u. Phys.* **13** (1908).

States *Hauptvermutung*.

Tietze, H.: Über die topologischen Invarianten mehrdimensionaler Mannigfaltigkeiten, *Monatsh. Math. Phys.* **19** (1908), 1–118.

Tietze transformations, abelianization, Wirtinger presentation of knot groups and enumeration of branched covers of S^3, statement of isomorphism problem for groups, mapping class groups, lens spaces.

1910
Dehn, M.: Über die Topologie des dreidimensional Raumes, *Math. Ann.* **69** (1910), 137–168.
Realization of finitely presented group by 2-complex or 4-manifold, word problem and conjugacy problem, Cayley diagrams for infinite groups, Dehn's lemma (incorrectly proved), knot groups, solution of word problem for trefoil knot group, construction of homology spheres by surgery.

1911
Brouwer, L. E. J.: Beweis der Invarianz der Dimensionzahl, *Math. Ann.* **70** (1911), 161–165.
Topological invariance of dimension.

1912
Brouwer, L. E. J. (a): Über Abbildung von Mannigfaltigkeiten, *Math. Ann.* **71** (1912), 97–115.
Brouwer degree, Brouwer fixed-point theorem.
Brouwer, L. E. J. (b): Beweis des Jordanschen Satzes für den n-dimensionalen Raum, *Math. Ann.* **71** (1912), 314–319.
S^{n-1} separates R^n into two components.
Dehn, M. (a): Über unendliche diskontinuierliche Gruppen, *Math. Ann.* **71** (1912), 116–144.
Dehn's algorithm for solution of the word problem for surface groups, proved using hyperbolic geometry.
Dehn, H. (b): Transformation der Kurven auf zweiseitigen Flächen, *Math. Ann.* **72** (1912), 413–421.
Topological proof of Dehn's algorithm.

1914
Dehn, M.: Die beiden Kleeblattschlingen, *Math. Ann.* **75** (1914), 402–413.
Automorphisms of trefoil knot group used to prove that the two trefoil knots are not the same.

1915
Alexander, J. W.: A proof of the invariance of certain constants of Analysis situs, *Trans. Amer. Math. Soc.* **16** (1915), 148–154.
Topological invariance of homology groups.

1918
Nielsen, J.: Die Isomorphismen der allgemeinen unendlichen Gruppe mit zwei Erzeugenden, *Math. Ann.* **78** (1918), 385–397.

1919
Alexander, J. W. (a): Note on two three-dimensional manifolds with the same group, *Trans. Amer. Math. Soc.* **20** (1919), 339–342.
(5,1) and (5,2) lens spaces are not homeomorphic.
Alexander, J. W. (b): Note on Riemann spaces, *Bull. Amer. Math. Soc.* **26** (1919), 370–372.
3-manifolds are branched covers of $\mathbf{S}^3$.
Nielsen, J.: Über die Isomorphismen unendlicher Gruppen ohne Relation, *Math. Ann.* **79** (1919), 269–272.

1921

Antoine, J.: L'Homéomorphie de deux figures et leurs voisinages, *J. Math. Pure et Appl.* **4** (1921), 221–325.

Antoine's necklace, wild curves, and spheres.

Brahana, T. R.: Systems of circuits on 2-dimensional manifolds, *Ann. Math.* **23** (1921), 144–168.

Nielsen, J.: Om Regning med ikke kommutative Faktorer og dens Anvendelse i Gruppeteorien, *Mat. Tidsskrift B* (1921), 77–94.

Finitely generated subgroups of free groups are free, rank is independent of the presentation, solution of generalized word problem.

1922

Veblen, O.: *Analysis Situs*, American Mathematical Society, New York, 1922.

1923

Alexander, J. W. (a): A proof and extension of the Jordan–Brouwer separation theorem, *Trans. Amer. Math. Soc.* **23** (1923), 333–349.

Alexander duality theorem.

Alexander, J. W. (b): On the deformation of a n-cell, *Proc. Nat. Acad. Sci.* **9** (1923), 406–407.

Simple proof that a homeomorphism fixed on the boundary is isotopic to the identity.

Alexander, J. W. (c): A lemma on systems of knotted curves, *Proc. Nat. Acad. Sci.* **9** (1923), 93–95.

Braid form of a knot or link.

Kerekjarto, B.v.: *Topologie*, Springer-Verlag, Heidelberg 1923.

1924

Nielsen, J. (a): Die Isomorphismengruppe der freien Gruppen, *Math. Ann.* **91** (1924), 169–209.

Finite presentation for automorphism group of free group.

Nielsen, J. (b): Die Gruppe der dreidimensionalen Gitter-transformationen *Kgl. Danske Vidensk. Selsk. Mat. Fys. Meddels.* **12** (1924), 1–29.

Radó, T.: Über den Begriff der Riemannsche Fläche, *Acta Univ. Szeged* **2** (1924), 101–121.

2-manifolds can be triangulated.

Schreier, O.: Über die Gruppen $A^aB^b = 1$, *Abh. Math. Sem. Univ. Hamburg* **3** (1924), 167–169.

Simplified treatment of torus knot groups and their automorphisms by developing some theory of free products.

1925

Hotelling, H.: Three-dimensional manifolds of states of motion, *Trans. Amer. Math. Soc.* **27** (1925), 329–344.

1926

Artin, E.: Theorie der Zöpfe, *Abh. Math. Sem. Univ. Hamburg* **4** (1926), 47–72.

Braid groups, generators and relations, solution of word problem.

Noether, E.: Ableitung der Elementarteilertheorie aus der Gruppentheorie, Jber. *Deutsch. Math. Verein* 34 (1926), 104.

Homology *groups*, as opposed to Betti and torsion numbers.

1927

Alexander, J. W. and Briggs, G. B.: On types of knotted curves, *Ann. Math.* **28** (1927), 562–586.

Knot classification by homology of 2- and 3-fold cyclic covers.
Nielsen, J.: Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen, *Acta Math.* **50** (1927), 189–358.
Automorphisms of surface groups are induced by homeomorphisms.
Reidemeister, K.: Knoten und Gruppen, *Abh. Math. Sem. Univ. Hamburg* **5** (1927), 7–23.
Algorithm for finding presentations of subgroups. Then abelianizes to find homology of cyclic covers of knot complements.
Schreier, O.: Die Untergruppen der freien Gruppen, *Abh. Math. Sem. Univ. Hamburg* **5** (1927), 161–183.
Subgroups of free groups are free, Schreier coset diagrams, Reidemeister–Schreier process.

1928
Baer, R.: Isotopie von Kurven auf orientierbaren geschlossenen Flächen und ihr Zusammenhang mit der topologischen Deformation der Flächen, *J. f. Math.* **159** (1928), 101–116.
Homotopic simple curves are isotopic, surface homeomorphisms are determined by canonical curves.
Brauner, K.: Zur Geometrie der Funktionen zweier komplexer Veränderlicher, *Abh. Math. Sem. Univ. Hamburg* **6** (1928), 1–55.
Wirtinger's application of knot theory to singularities of algebraic functions.
Reidemeister, K.: Fundamentalgruppe und Überlagerungsräume, *Nachrichten, Göttingen* 1928, 69–76.
Covering spaces realize subgroups.

1930
Threlfall, W. and Seifert, H.: Topologische Untersuchungen der Diskontinuitätsbereiche endliche Bewegungsgruppen des dreidimensionalen sphärischen Raumes I, *Math. Ann.* **104** (1930), 1–70, II *Math. Ann.* **107** (1932), 543–586.
Introduces term "lens spaces" (Linsenräume).

1931
Frankl, F.: Zur Topologie des dreidimensional Raumes, *Monatsh. f. Math. u. Phys.* **38** (1931), 357–364.
A triangulation of the ball which cannot be shelled.
Johansson, I.: Topologische Untersuchungen über unverzweigte Überlagerungsflächen, *Skriften norske Vidensk.-Akad. Oslo Math.-natur. Kl.* **1** (1931), 1–69.
π_1 of infinite surface is free.
Magnus, W.: Untersuchungen über einige unendliche diskontinuierliche Gruppen, *Math. Ann.* **105** (1931), 52–74.
Automorphisms of figure eight knot group, commutator subgroup of modular group.
Seifert, H.: Konstruktion dreidimensionaler geschlossener Räume, *Ber. Sächs. Akad. Wiss.* **83** (1931), 26–66.
Seifert–Van Kampen theorem.

1932
Alexander, J. W.: Some problems in topology. *Verh. internat. Math.-Kongr.* 1932, 249–257.
Hilbert, D. and Cohn-Vossen, S.: *Anschauliche Geometrie*, Springer, Berlin, 1932. Translated as *Geometry and the Imagination*, Chelsea, New York 1952.
Reidemeister, K.: *Einführung in die kombinatorische Topologie* Teubner, Leipzig 1932. Reprinted by Chelsea, New York 1950.

Van Kampen, E. R.: Komplexe in euklidischen Räumen, *Abh. Math. Sem. Univ. Hamburg* **9** (1932), 72–78 and 152–153.
An n-complex not embeddable in R^{2n}.

1933
Borsuk, K.: Zur kombinatorischen Eigenschaften der Retrakte, *Fund. Math.* **21** (1933), 91–98.
Retraction induces homomorphism of π_1, deformation retraction induces isomorphism.
Goeritz, L.: Die Abbildungen der Brezelfläche und der Vollbrezel vom Geschlecht **2**, *Abh. Math. Sem. Univ. Hamburg* **9** (1933), 244–259.
Twist homeomorphisms.
Singer, J.: Three-dimensional manifolds and their Heegaard diagrams, *Trans. Amer. Math. Soc.* **35** (1933), 88–111.
Semidecision procedure for homeomorphism of 3-manifolds, using Heegaard diagrams (assumes *Hauptvermutung*).
Van Kampen, E. R.: On the connection between the fundamental groups of some related spaces, *Amer. J. Math.* **55** (1933), 255–260.
Seifert–Van Kampen theorem.
Weber, C. and Seifert, H.: Die beiden Dodekaederräume, *Math. Zeit.* **37** (1933), 237–253.
Equivalence of some forms of the Poincaré homology sphere.

1934
Kurosh, A. G.: Die Untergruppen der freien Produkte von beliebigen Gruppen, *Math. Ann.* **109** (1934), 647–660.
Kurosh subgroup theorem.
Seifert, H.: Über das Geschlecht von Knoten, *Math. Ann.* **110** (1934), 571–592.
Homology of knot covering spaces computed by using nonsingular spanning surfaces.
Seifert, H. and Threlfall, W.: *Lehrbuch der Topologie*, Teubner, Leipzig 1934.

1935
Reidemeister, K.: Homotopieringe und Linsenräume, *Abh. Math. Sem. Univ. Hamburg* **11** (1935), 102–109.
Classification of lens spaces.

1936
Baer, R. and Levi, F.: Freie Produkte und ihre Untergruppen, *Compositio Math.* **3** (1936), 391–398.
Covering space proof of the Kurosh subgroup theorem.
Markov, A. A.: Über die freie Äquivalenz der geschlossenen Zöpfe, *Rec. math. Moscow* **1** (1936), 73–78.
Markov operations on braids.
Post, E. L.: Finite combinatory processes—formulation I, *J. Symbolic Logic* **1** (1936).
Definition of computability.
Todd, J. A. and Coxeter, H. S. M.: A practical method for enumerating cosets of a finite abstract group, *Proc. Edinb. Math. Soc.* **5** (1936), 25–34.
Coset enumeration method.
Turing, A. M.: On computable numbers, with an application to the Entscheidungsproblem, *Proc. Lond. Math. Soc.* **42** (1936), 230–265.
Definition of computability.

Whitehead, J. H. C.: On equivalent sets of elements in a free group, *Ann. Math.* **37** (1936), 782–800.
Algorithm for determining whether elements are equivalent under an automorphism.

1937
Turing, A. M.: Computability and λ-definability, *J. Symbolic Logic* **2** (1937), 153–163.

1943
Papakyriakopoulos, C. D.: A new proof of the invariance of the homology groups of a complex (Greek), *Bull. Soc. Math. Grèce* **22** (1943), 1–154.
Proof of *Hauptvermutung* and solution of homeomorphism problem for 2-complexes.

1946
Gödel, K.: Remarks before the Princeton bicentennial conference on problems in mathematics. In *The Undecidable* (M. Davis, Ed.), Raven Press 84–88, 1965.

1947
Post, E. L.: Recursive unsolvability of a problem of Thue, *J. Symbolic Logic* **12** (1947), 1–11.

1948
Fox, R. H. and Artin, E.: Some wild cells and spheres in three-dimensional space. *Ann. Math.* **49** (1948), 979–990.
Wildness of arcs proved by computation of groups.

1949
Fox, R. H.: A remarkable simple closed curve, *Ann. Math.* **50** (1949), 264–265.
Higman, G., Neumann, B. H., and Neumann, H.: Embedding theorems for groups, *J. Lond. Math. Soc.* **24** (1949), 247–254.

1950
Blankenship, W. A. and Fox, R. H.: Remarks on certain pathological open subsets of 3-space and their fundamental groups, *Proc. Amer. Math. Soc.* **1** (1950), 618–624.

1952
Moise, E. E.: Affine structures in 3-manifolds. V. The triangulation theorem and *Hauptvermutung*, *Ann. Math.* **56** (1952), 96–114.
Proof of triangulation and *Hauptvermutung* for 3-manifolds.

1954
Magnus, W. and Moufang, R.: Max Dehn zum Gedächtnis, *Math. Ann.* **127** (1954), 215–227.

1955
Novikov, P. S.: On the algorithmic unsolvability of the word problem in group theory, (Russian) *Trudy Mat. Inst. Steklov* **44** (1955), 143 pp.
Unsolvability of the word problem for groups.

1956
Griffiths, H. B.: Infinite products of semigroups and local connectivity, *Proc. Lond. Math. Soc.* (3) **6** (1956), 455–480.

1957

Papakyriakopoulos, C. D.: On Dehn's lemma and the asphericity of knots, *Ann. Math.* **66** (1957), 1–26.

Proof of Dehn's lemma.

Sanderson, D. E.: Isotopy in 3-manifolds. I. Isotopic deformations of 2-cells and 3-cells, *Proc. Amer. Math. Soc.* **8** (1957), 912–922.

1958

Markov, A. A.: Unsolvability of the problem of homeomorphy (Russian), *Proc. Int. Cong. Math.*, 1958, 300–306.

Unsolvability of the homeomorphism problem for triangulated 4-manifolds.

Mikhailova, K. A.: The occurrence problem for direct products of groups (Russian), *Dokl. Akad. Nauk SSSR* **119** (1958), 1103–1105.

Rabin, M. O.: Recursive unsolvability of group-theoretic problems, *Ann. Math.* **67** (1958), 172–194.

Unsolvability of isomorphism problem.

Whittlesey, E. F.: Classification of finite 2-complexes, *Proc. Amer. Math. Soc.* **9** (1958), 841–845.

1959

Boone, W. W.: The word problem, *Ann. of Math.* **2** (1959), 70, 207–265.

1960

Fuchs, L.: *Abelian Groups*, Pergamon Press, Oxford 1960.

Wallace, A. H.: Modifications and cobounding manifolds, *Can. J. Math.* **12** (1960), 503–528.

Construction of orientable 3-manifolds by surgery.

1961

Cairns, S. S.: *Introductory Topology*, The Ronald Press Co., New York 1961.

Haken, W.: Theorie der Normalflächen, *Acta Math.* **105** (1961), 245–375.

Algorithm for recognizing the trivial knot.

Milnor, J.: Two complexes which are homeomorphic but combinatorially distinct, *Ann. Math.* **74** (1961), 575–590.

Hauptvermutung is false for 6-complexes.

Schubert, H.: Bestimmung der Primfaktorzerlegung von Verkettungen, *Math. Zeit.* **76** (1961), 116–148.

1962

Hempel, J.: Construction of orientable 3-manifolds in *Topology of 3-manifolds and Related Topics* (ed. M. K. Fort Jr.,), Prentice-Hall, Englewood Cliffs, 1962, 206–212.

Lickorish, W. B. R.: A representation of orientable combinatorial 3-manifolds, *Ann. Math.* **76** (1962), 531–540.

Surgery construction using twist homeomorphisms of surfaces.

Reinhardt, B. L.: Algorithms for Jordan curves on compact surfaces, *Ann. Math.* **75** (1962), 209–222.

Algorithm using hyperbolic geometry.

Stallings, J.: On the recursiveness of sets of presentations of 3-manifold groups, *Fund. Math.* **51** (1962), 191–194.

$\mathbf{Z} \times \mathbf{Z} \times \mathbf{Z} \times \mathbf{Z}$ is not a 3-manifold group, and hence there is no algorithm for recognizing presentations of 3-manifold groups.

1963

Britton, J. L.: The word problem, *Ann. of Math.* **2** (1963), 77, 16–32.

1964

Bing, R. H.: Some aspects of the topology of 3-manifolds related to the Poincaré conjecture in *Lectures on Modern Mathematics* II (ed. T. L. Saaty), Wiley, New York 1964, 93–128.

1965

Zieschang, H., Algorithmen für einfache Kurven auf Flächen, *Math. Scand.* **17** (1965), 17–40.
Combinatorial derivation, based on Whitehead 1936.

1966

Calugareanu, G.: Sur les courbes fermées simples tracées sur une surface fermée orientable, *Mathematica* (Cluj) 8 (1966), 29–38.
Vindication of Poincaré's 1904 algorithm for simple curves.

Magnus, W., Karrass, A., and Solitar, D., *Combinatorial Group Theory*, Wiley Interscience, New York 1966.

1967

Blackett, D. W.: *Elementary Topology*, Academic Press, New York 1967.

Massey, W. S.: *Algebraic Topology: An Introduction*, Harcourt, Brace and World, New York 1967.

Rogers, H. Jr.: *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, New York 1967.

1968

Boone, W. W., Haken, W., and Poénaru, V.: On recursively unsolvable problems in topology and their classification. In *Contributions to Mathematical Logic* (H. A. Schmidt, K. Schütte, and H.-J. Thiele, Eds.), North-Holland, 37–74, 1968.

Waldhausen, F. (a): Heegaard-Zerlegungen der 3-sphäre, *Topology* **7** (1968), 195–203.
All are isotopic to the obvious ones.

Waldhausen, F. (b): The word problem in fundamental groups of sufficiently large irreducible 3-manifolds, *Ann. Math.* **88** (1968), 272–280.
Solution of word problem for knot groups.

1969

Garside, F. A.: The braid group and other groups, *Quart. J. Math. Oxford* **20** (1969), 235–254.
Solution of the conjugacy problem for braid groups.

de Rham, G.: *Lectures on Introduction to Algebraic Topology*, Tata Institute, Bombay 1969.

1971

Hoare, A. H. M., Karrass, A., and Solitar, D.: Subgroups of finite index in Fuchsian groups, *Math. Zeit.* **120** (1971), 289–298.

1972

Bollinger, M.: Geschichtliche Entwicklung des Homologiebegriffs, *Arch. for Hist. Exact Sci.* **9** (1972), 94–166.

1973

Lickorish, W. B. R.: 3-manifolds as branched covers, *Proc. Camb. Phil. Soc.* **74** (1973), 449–451.
With unknotted branch curves.

1974

Hilden, H.: Every closed orientable 3-manifold is a 3-fold branched covering space of S^3, *Bull. Amer. Math. Soc.* **80** (1974), 1243–1244.

Laudenbach, F.: *Topologie de la dimension trois: homotopie et isotopie*, Asterisque # 12, Soc. Math. de France, Paris 1974.

Magnus, W.: *Non-euclidean Tessellations and Their Groups*, Academic Press, New York 1974.

Volodin, I. A., Kuznecov, V. E., and Fomenko, A. T.: The problem of the algorithmic discrimination of the standard three-dimensional sphere. *Russian Math. Surveys* **29** (1974), 71–172.

Sketches S. P. Novikov's proof that there is no algorithm for recognizing S^5.

1975

Birman, J. S.: *Braids, Links, and Mapping Class Groups*, Princeton University Press, Princeton 1975.

Proof of Markov's 1936 theorem on equivalence of closed braids.

Lefschetz, S.: *Applications of Algebraic Topology*, Springer-Verlag, Heidelberg 1975.

1976

Meyerson, M. D.: Representing homology classes of closed orientable surfaces, *Proc. Amer. Math. Soc.* **61** (1976), 181–2.

Montesinos, J.: 3-manifolds as 3-fold branched covers of S^3, *Quart. J. Math. Oxford* **27** (1976), 84–94.

Rolfsen, D.: *Knots and Links*, Publish or Perish Inc., Berkeley 1976.

1977

Giblin, P. J.: *Graphs, Surfaces and Homology*, Chapman and Hall, London 1977.

Moise, E. E.: *Geometric Topology in Dimensions* 2 *and* 3, Springer-Verlag, New York 1977.

Thurston, W.: *Lecture Notes*, Princeton Math. Dept. Princeton 1977.

1978

Waldhausen, F.: Recent results on sufficiently large 3-manifolds, *Proc. Symp. in Pure Math* **32** (1978), vol. 2, 21–38, American Mathematical Society.

Outline of classification for large class of 3-manifolds and their groups, using Hemion 1979 and Haken 1961.

1979

Hemion, G.: On the classification of homeomorphisms of 2-manifolds and the classification of 3-manifolds, *Acta Math.* **142** (1979) 123–155.

Solution of conjugacy problem for mapping class groups.

Weil, A.: Riemann, Betti and the birth of Topology, *Arch. for Hist. Exact Sci.* **16** (1979) 91–96.

1980

Cohen, D. E., and Aanderaa, S.: Modular machines, the word problem for finitely presented groups, and Collins' theorem. In *Word Problems* II (S. I. Adian, W. W. Boone, and G. Higman, Eds.), North-Holland, 1–16, 1980.

1982

Stillwell, J. C.: The word problem and the isomorphism problem for groups, *Bull. Amer. Math. Soc.* **6** (1982), 33–56.

Index

D

Q