Cambridge Studies in Advanced Mathematics 55

Automorphic Forms and Representations

DANIEL BUMP

CAMBRIDGE

more information – www.cambridge.org/9780521550987

This page intentionally left blank

CAMBRIDGE STUDIES IN
ADVANCED MATHEMATICS: 55

AUTOMORPHIC FORMS AND REPRESENTATIONS

CAMBRIDGE STUDIES IN ADVANCED MATHEMATICS 55

Already published

2 K. Petersen *Ergodic theory*
3 P.T. Johnstone *Stone spaces*
4 W.H. Schikhof *Ultrametric calculus*
5 J.-P. Kahane *Some random series of functions, 2nd edition*
7 J. Lambek & P.J. Scott *Introduction to higher-order categorical logic*
8 H. Matsumura *Commutative ring theory*
9 C.B. Thomas *Characteristic classes and the cohomology of finite groups*
10 M. Aschbacher *Finite group theory*
11 J.L. Alperin *Local representation theory*
12 P. Koosis *The logarithmic integral I*
13 A. Pietsch *Eigenvalues and s-numbers*
14 S.J. Patterson *An introduction to the theory of the Riemann zeta-function*
15 H.J. Baues *Algebraic homotopy*
16 V.S. Varadarajan *Introduction to harmonic analysis on semisimple Lie groups*
17 W. Dicks & M. Dunwoody *Groups acting on graphs*
18 L.J. Corwin & F.P. Greenleaf *Representations of nilpotent Lie groups and their applications*
19 R. Fritsch & R. Piccinini *Cellular structures in topology*
20 H. Klingen *Introductory lectures on Siegel modular forms*
21 P. Koosis *The logarithmic integral II*
22 M.J. Collins *Representations and characters of finite groups*
24 H. Kunita *Stochastic flows and stochastic differential equations*
25 P. Wojtaszczyk *Banach spaces for analysts*
26 J.E. Gilbert & M.A.M. Murray *Clifford algebras and Dirac operators in harmonic analysis*
27 A. Frohlich & M.J. Taylor *Algebraic number theory*
28 K. Goebel & W.A. Kirk *Topics in metric fixed point theory*
29 J.F. Humphreys *Reflection groups and Coxeter groups*
30 D.J. Benson *Representations and cohomology I*
31 D.J. Benson *Representations and cohomology II*
32 C. Allday & V. Puppe *Cohomological methods in transformation groups*
33 C. Soule, *et al.* *Lectures on Arakelov geometry*
34 A. Ambrosetti & G. Prodi *A primer of nonlinear analysis*
35 J. Palis & F. Takens *Hyperbolicity, stability and chaos at homoclinic bifurcations*
37 Y. Meyer *Wavelets and operators 1*
38 C. Weibel *An Introduction to Homological Algebra*
39 W. Bruns & J. Herzog Cohen *Macaulay rings*
40 V. Snaith *Explicit Brauer induction*
41 G. Laumon *Cohomology of Drinfeld modular varieties: Part I*
42 E.B. Davies *Spectral theory of differential operators*
43 J. Diestel, H. Jarchow & A. Tonge *Absolutely summing operators*
44 P. Mattila *Geometry of sets and measures in Euclidean spaces*
45 R. Pinsky *Positive harmonic functions and diffusion*
46 G. Tenenbaum *Introduction to analytic and probabilistic number theory*
47 C. Peskine *Complex projective geometry*
48 Y. Meyer & R. Coifman *Wavelets and Operators II*
49 R. Stanley *Enumerative combinatorics*
50 I. Porteous *Clifford algebras and the classical groups*

AUTOMORPHIC FORMS AND REPRESENTATIONS

DANIEL BUMP
Stanford University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1996
First paperback edition 1998

Library of Congress Cataloging-in-Publication Data is available.

A catalog record for this book is available from the British Library.

ISBN 0 521 55098 X hardback
ISBN 0 521 65818 7 paperback

Transferred to digital printing 2004

Contents

Preface

The theory of automorphic forms, rightly or wrongly, has a reputation of being difficult for the student. I felt that there was a need for a book that would present the subject in a style that was accessible yet based on complete proofs and revealed clearly the uniqueness principles that underlie the basic constructions. I have been lecturing on automorphic forms and representation theory at Stanford and the Mathematical Sciences Research Institute since 1990, and this book is the end result.

The level of this book is intermediate between an advanced textbook and a monograph. I hope that it will be interesting to experts as well as graduate students. Its aim is to cover a substantial portion of the theory of automorphic forms on $GL(2)$. Both the "classical" and "representation theoretic" viewpoints are covered.

There are significant omissions from my treatment, most seriously the Selberg trace formula. It has not been my aim to achieve complete coverage of the topics treated or to write a reference book. I feel that the existing reference material is adequate, and that it was not feasible to cover any single topic with the thoroughness I would have liked. I hope that the reader will begin studying the reference material (such as the Corvallis volume (Borel and Casselman, 1979) and above all Jacquet and Langlands (1970)) in the course of reading this book. If I have done my job well, the task of approaching Jacquet and Langlands should be made easier by the current volume.

I can imagine a useful sequel to this book. A second volume is therefore a possibility, but not for several years.

I would like to thank William Banks, Antonia Bluher, Aleksandr Brener, David Cardon, Jim Cogdell, Anton Deitmar, David Feldman, Solomon Friedberg, Masaaki Furusawa, Steve Gelbart, Tom Goetze, David Goldberg, Jiandong Guo, Jeffrey Hoffstein, Özlem Imamoglu, David Joyner, Chris Judge, Par Kurlberg, Annette Klute, David Manderscheid, Greg Martin, Andrei Paraschivescu, Ralph Phillips, Freydoon Shahidi, Tom Shemanske, Trask Stalnaker, Steve Rallis, Ken Ribet, Dinakar Ramakrishnan, Julie Roskies, San Cao Vo, James Woodson – and probably others I've forgotten – for helpful comments, cor-

rections, discussions, or other feedback. Thanks also to Lauren Cowles of Cambridge University Press for her interest in the manuscript and for her guidance, to Ellen Tirpak and the staff at TechBooks for their expert treatment of the manuscript, to Reid Augustin for helping me set up my Linux machine, and to the MSRI for their help and support during 1994–1995. And thank you, my wife Kathi, and my parents Kenneth and Ellen Bump, for your support, which was always there when I needed it most.

Parts of this book were written with the support of the American Mathematical Society Centennial Research Fellowship and grants from the National Science Foundation.

Advice to the Reader

It has not been my intention to write a reference book on automorphic forms. I have cut corners in many places. For example, although I treat the representation theory of $GL(2, \mathbb{R})$ from the viewpoint of $(\mathfrak{g}, K)$-modules with some degree of completeness, I ignore $GL(2, \mathbb{C})$ completely. In laying the foundations for automorphic representations, I have concentrated on the cuspidal representations and given only indications of what happens with the continuous spectrum. I do discuss the theory of Eisenstein series but only in so far as necessary for my limited goals – I want to discuss the Rankin–Selberg method, and I want to show how the intertwining integrals that pervade the local theory arise from the constant terms of the Eisenstein series. My proofs of some foundational results in Section 3.3 are complete only when the ground field is $\mathbb{Q}$. I feel that despite these omissions, I have been able to treat my subject matter with some degree of depth.

In some places I have left details to the reader in the form of exercises. If the result of an exercise is required in the text, I have usually provided enough in the way of hints that the reader will be able to fill the gaps. Hints are enclosed in square brackets, sometimes explicitly labeled as such, sometimes not. Some exercises are trivial and some (those not needed for the text) are genuinely difficult.

Each of the four chapters is itself a complete course. The material of the four chapters is complementary, but each can be studied on its own. However, Chapter 3 should be read after Sections 1–4 of Chapter 2. Also, Chapter 3 makes use of results on the representation theory of $GL(2, F)$, where F is a non-Archimedean local field, whose proofs are postponed until Chapter 4. The reader may take these on faith during a first reading of Chapter 3. Chapter 3 may be more difficult than Chapter 4. Some readers will want to start with Chapter 4.

I have tried to write in a style that encourages the reader to skip around or to start in the middle. Definitions are sometimes repeated, and there is a lot of cross referencing.

Chapter 1 is written in the classical language. It is based on the paper of Doi and Naganuma (1969) which exhibits a rich variety of ideas and techniques. The chapter contains an introduction to the Langlands conjectures, which are taken

up from a more sophisticated point of view later in the book. The phenomenon of *base change*, which was discovered by Doi and Naganuma, is an example of a *lifting* of automorphic forms. Such liftings are systematically predicted by the conjectures of Langlands. The method by which Doi and Naganuma proved their result is also suggestive – it is based on the Rankin–Selberg method and the converse theorem, a pleasant combination of techniques that is at the heart of the present-day project of Piatetski–Shapiro and his collaborators to prove liftings of automorphic forms from classical groups to $GL(n)$.

The reader approaching this subject for the first time might want to read Sections 1.1–1.4, 1.6, and 1.8, skimming Sections 1.5 and 1.7 before attacking Section 1.8. Sections 1.7, 1.9, and 1.10 are more difficult, and the details of how the approach to base change laid out in Section 1.8 are carried out are only important if you care about them. Be aware that the rest of the book is largely independent of Chapter 1.

There is an apparent dichotomy in our field between the "classical" and "representation theoretic" approaches. But in fact this dichotomy is illusory, and it is important for the worker to understand both languages. Although Chapter 1 could be skipped by the reader, it is included for a good reason. The adele group is a large and complex object, and we derive our intuition to a large extent from the example of $GL(2, \mathbb{R})$. Moreover, adelic statements often reduce in the end to classical ones.

Chapter 2 marks the introduction of representation theory into the study of automorphic forms. There are two types of classical automorphic forms, namely, Maass forms and modular forms. This dichotomy reflects the fact that the irreducible representations of $SL(2, \mathbb{R})$ fall into two main classes – the principal series and the discrete series. In Chapter 2, we study the representation theory of $GL(2, \mathbb{R})$ and the spectral theory of compact quotients of the upper half plane and make clear the relationship between these two topics.

Chapter 2 ends with two special results that are needed for Chapter 3, namely, the uniqueness of Whittaker models and a theorem of Harish–Chandra. We do not discuss the representation theory of $GL(2, \mathbb{C})$, though it is in many respects simpler than the representation theory of $GL(2, \mathbb{R})$.

In Chapter 3, we introduce the adeles and the modern approach to automorphic forms. We cover Tate's thesis, the discreteness of the cuspidal spectrum, the tensor product theorem of Flath, and, following Jacquet and Langlands (1970), the implications of the uniqueness of Whittaker models–the strong multiplicity one theorem and the construction of L-functions. We consider the standard L-functions, the Rankin–Selberg L-functions, and (briefly) the triple L-functions. We discuss the Eisenstein series in enough detail for our study of the Rankin–Selberg method, but we do not prove much about the continuous spectrum.

Chapter 4 is independent of the first three chapters, and some readers will want to begin with it. In contrast to Chapter 3, I tried for some measure of completeness in Chapter 4, which is devoted to the representation theory of $GL(2)$

over a local field. I begin the chapter with a long section on the representation theory of $GL(2)$ over a finite field, where complete proofs may be given in just a few pages. The exercises for Section 4.1 emphasize the Weil representation and the philosophy of cusp forms, where representations are built up from cuspidal atoms by parabolic induction. Turning to $GL(2)$ over a non-Archimedean local field, I follow Bernstein and Zelevinsky (1976) in emphasizing sheaves and distributions in the proofs, among other things. I prove the uniqueness of Whittaker models and the fact that the character of a representation is invariant under transpose. As in Bernstein and Zelevinsky (1976), I treat the character as a distribution, without establishing its nature as a locally integrable function. I give fairly complete discussions of the principal series and spherical representations, including the Macdonald formula for the spherical function and the explicit formula for the spherical Whittaker function, which I prove by the method of Casselman and Shalika (1980). I show how the local functional equations follow from a uniqueness principle and how supercuspidal representations may be constructed by means of the Weil representation. I also prove the multiplicativity of the local ϵ-factors by means of the Weil representation. My discussion of the Weil representation is based on Jacquet and Langlands (1970). The chapter ends with a discussion of the local Langlands conjectures.

The reader may take the chapters in any order. Because Chapter 3 contains forward references to Chapter 4 (which is independent of Chapter 3), the *logical* sequence of the chapters is 1, 2, 4, then 3. However, there is a rationale for putting Chapter 3 ahead of Chapter 4. This is that the *motivation* for many of the topics in Chapter 4 comes from automorphic forms.

I would appreciate any comments that you may have. I may be able to take them into account in revising this text some day. Particularly, if you find mistakes – whether typographical errors or more serious mathematical or historical mistakes – I would like to be informed. My e-mail address is `bump@math.stanford.edu`. I intend to maintain a list of errata on my web page at `http://math.stanford.edu/~bump`.

Prerequisites

We assume a basic knowledge of algebraic number theory, the representation theory of finite groups, and of Fourier analysis on locally compact abelian groups (Pontriagin duality). Moreover, we will assume the existence and basic properties of Haar measure on locally compact groups. On the other hand, we develop all the Lie theory and much of the functional analysis that we need from scratch, and the prerequisites from Fourier analysis are reviewed as they arise.

In Chapter 2, Section 9, we use the fact that a solution to an elliptic differential equation is analytic. In Chapter 3, Section 4, we expect the reader to consult Knapp and Vogan (1995) for the properties of a certain Hecke algebra of distributions.

Notations

We have attempted to write in a style that allows the reader to start in the middle if desired. Definitions are sometimes repeated, and there is a lot of cross referencing. We attempt to avoid "global" notations that are defined throughout the book. We did compromise on this point in the matter of matrix transposes. If g is a matrix, ${}^{\top}g$ is its transpose.

We may denote the identity matrix as either I or 1. We will often omit zero entries from a matrix. Thus

$$\begin{pmatrix} a & b \\ & d \end{pmatrix} \quad \text{means} \quad \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}.$$

Occasionally, we will denote an "arbitrary" matrix entry with an asterisk. Thus for example

$$f\begin{pmatrix} y & * \\ & y^{-1} \end{pmatrix} = |y|^s$$

means that this identity is true regardless of the value of $*$.

"Almost all" means for all but finitely many.

If G is an affine algebraic variety, particularly an affine algebraic group, defined over a field F, and if A is a commutative ring containing F, we will denote by $G(A)$ or G_A the points of G with coordinates in A. Some affine varieties, such as $GL(n)$, are really defined over $\mathbb{Z}$. Thus $GL(n, A)$ is defined if A is any commutative $\mathbb{Z}$-algebra, that is, any commutative ring. We will encounter mostly affine algebraic varieties. Our notations for affine algebraic varieties are discussed in more detail at the beginning of Section 3.3.

A *character* of a group is a continuous homomorphism into the group of complex numbers of absolute value 1. A continuous homomorphism into $\mathbb{C}^\times$ is called a *quasicharacter*.

A *global field* F is by definition an algebraic number field, or else a "function field," which is a finitely generated field of transcendence degree 1 over a finite field. If v is a place of F, we will denote by F_v the completion of F at v. If v is non-Archimedean, we will denote the ring of integers in F by $\mathfrak{o}_v$, its

maximal ideal by $\mathfrak{p}_v$, and the cardinality of $\mathfrak{o}_v/\mathfrak{p}_v$ by q_v. By ϖ_v, we will denote an arbitrarily selected generator of $\mathfrak{p}_v$. We will denote by $\mathrm{ord} : F^\times \to \mathbb{Z}$ the valuation, so that $\mathrm{ord}(\epsilon\varpi_v^m) = m$ when $\epsilon \in \mathfrak{o}_v^\times$.

Let F be a global field and A its adele ring. We will use the notation $a = (a_v)$ for elements of A. This notation means that for each place v of F, a_v is the vth component of the adele a. Thus $a_v \in F_v$ and $a_v \in \mathfrak{o}_v$ for almost all v. Similarly, if G is an affine algebraic variety, we will denote elements of $G(A)$ by $g = (g_v)$, where $g_v \in G(F_v)$ and g_v has coordinates in $\mathfrak{o}_v$ for almost all v. This notation requires some explanation. Let us assume that G is the locus in affine n-space of some system of equations. Then *a priori* an element of g is an n-tuple $(g^1, \cdots, g^n)$ with $g^i = (g_v^i) \in A$. For each v, we find that $g_v = (g_v^1, \cdots, g_v^n) \in G(F_v)$, and we write $g = (g_v)$.

If R is a topological ring, then for $a \in R$, we will denote by $|a|$ the module of the endomorphism $x \mapsto ax$ of R, namely, the factor by which this transformation multiplies the additive Haar measure. Thus if $R = \mathbb{R}$, then $|\ |$ is the usual absolute value, while if $R = \mathbb{C}$, then $|\ |$ is the square of the usual absolute value. Also for a topological ring R, we will denote the additive Haar measure by dx and the multiplicative Haar measure by $d^\times x$. These are subject to normalizations that will be discussed when they arise. We will use these notations if R is a local field or the adele ring of a global field.

If F is a local field, we will often denote by $\psi : F \to \mathbb{C}$ a fixed nontrivial character of the additive group F. Alternatively, if A is the adele ring of a global field F, we will denote by ψ a fixed nontrivial character of A that is trivial on F. In either case, we will normalize the additive Haar measure on F or A to be self-dual for Fourier transform with respect to ψ. That is, if f is a compactly supported continuous function on F or A, define its Fourier transform

$$\hat{f}(x) = \int f(y)\,\psi(xy)\,dy.$$

The Fourier inversion formula asserts that

$$\hat{\hat{f}}(x) = f(-x),$$

and this is true for a unique choice of Haar measure. This is the Haar measure that we will usually use. With this normalization in the global case, the Haar volume of the compact quotient A/F is one, as is proved in Proposition 3.1.3.

If G is a topological group, we will always denote the *left* Haar integral by $\int_G dg$.

The end-of-proof symbol is ■, and we will also use □ to indicate the end of the proof of a lemma that is interpolated in the middle of another proof.

1

Modular Forms

In this chapter, we will introduce the study of modular forms through the paper of Doi and Naganuma (1969). This paper gave one of the first historical examples of a *functorial lifting* of automorphic forms, a phenomenon now codified in Langlands' important functoriality conjecture. The paper also uses (following a suggestion of Shimura) a beautiful L-function technique based on the Rankin–Selberg method and the so-called converse theorem for $GL(2)$. Though this method has been somewhat eclipsed by the trace formula, it is still being used to good effect by I. Piatetski-Shapiro and his coworkers in constructing liftings from classical groups to $GL(n)$.

The unifying theme of this chapter and those that follow will be *L-functions*. Briefly, an L-function is a *Dirichlet series*, that is, a series of the form

$$\phi(s) = \sum_{n=1}^{\infty} c(n)\, n^{-s},$$

which has an *Euler product* and a *functional equation*. We may illustrate this with two examples. The first is the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s},$$

which is convergent for $\mathrm{re}(s) > 1$ and has meromorphic continuation to all s, with just a simple pole at $s = 1$. The *Euler product* expresses this as a product over all primes p:

$$\zeta(s) = \prod_{p} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

The *functional equation* asserts that if

$$\xi(s) = \pi^{-s/2}\, \Gamma\left(\frac{s}{2}\right) \zeta(s),$$

then

$$\xi(s) = \xi(1-s).$$

The second example is due to Ramanujan (1916), who defined a certain function $\tau(n)$ by equating the coefficients in the series

$$q \prod (1-q^n)^{24} = \sum_{n=1}^{\infty} \tau(n)\, q^n. \tag{0.1}$$

Thus $\tau(1) = 1$, $\tau(2) = -24$, $\tau(3) = 252$, etc. Ramanujan's intuition had led him to a function with very remarkable properties. To begin with, the coefficients are multiplicative: if $(n, m) = 1$, then $\tau(nm) = \tau(n)\,\tau(m)$. Ramanujan considered the Dirichlet series

$$L(s, \Delta) = \sum_{n=1}^{\infty} \tau(n)\, n^{-s}.$$

He conjectured, and Mordell (1917) later proved, that

$$L(s, \Delta) = \prod_{p} \left(1 - \tau(p)\, p^{-s} + p^{11-2s}\right)^{-1}.$$

This is the *Euler product*. As for the *functional equation*, $L(s, \Delta)$ has analytic continuation to all s, with no poles, and if we define

$$\Lambda(s, \Delta) = (2\pi)^{-s}\, \Gamma(s)\, L(s, \Delta),$$

we have

$$\Lambda(s, \Delta) = \Lambda(12 - s, \Delta).$$

The explanation of these formulas is connected with the fact that the coefficients $\tau(n)$ are the Fourier coefficients of a *modular form*. Let $z = x + iy$ where $x, y \in \mathbb{R}$, $y > 0$. Then with $q = e^{2\pi i z}$, (0.1) becomes a function of z:

$$\Delta(z) = e^{2\pi i z} \prod_{n=1}^{\infty} (1 - e^{2\pi i n z})^{24}.$$

This function, *Ramanujan's discriminant function* is a *modular form of weight 12*, which means that

$$\Delta\left(\frac{az+b}{cz+d}\right) = (cz+d)^{12}\, \Delta(z)$$

for $a, b, c, d \in \mathbb{Z}$, $ad - bc = 1$. It turns out that there are lots of modular forms besides Δ; associated with such modular forms are L-functions having Euler products and functional equations. We will study the modular forms by studying their L-functions.

We will begin by studying a more basic type of L-function than those associated with modular forms; these are the L-functions associated with Dirichlet characters, of which the Riemann zeta function may be regarded as the prototype. Proving the Euler products and functional equations of these L-functions is the subject of Section 1.1.

Sections 1.2–1.4 form a basic course in modular forms, culminating with the Hecke theory. There are many exercises, particularly in Section 1.3, and this part of the book could be studied by an advanced undergraduate. Section 1.5 has as its goal Weil's *converse theorem*, which allows us to assert the existence of an automorphic form if sufficiently many functional equations can be proved, and though this particular theorem is not needed in the remainder of the book, variants of it are applied in Sections 1.9 and 1.10 to prove the existence of automorphic forms. Still, the proof of Weil's theorem could be skipped without loss of continuity. Sections 1.6 and 1.7 introduce tools we need, the Rankin–Selberg method and Hilbert modular forms, and at the end of Section 1.7, the result of Doi and Naganuma is formulated. This result asserts that given a modular form for $SL(2, \mathbb{Z})$, there exists a Hilbert modular form – the *base change lift* – whose L-function is described in terms of the L-function of the given modular form for $SL(2, \mathbb{Z})$. Section 1.8 explains how this result fits into the very general functoriality conjecture of Langlands. To apply the converse theorem, the functional equations of many L-functions must be proved. Sections 1.9 and 1.10, which are more technical than the rest of the chapter, carry out this program by constructing certain nonholomorphic automorphic forms (the first application of a converse theorem). They then form Rankin–Selberg convolutions of these auxiliary forms with the given one in order to prove the functional equations that are needed to conclude (in the second application of a converse theorem) that the base change lift exists.

1.1 Dirichlet L-Functions

The results of this section will be generalized in Section 1.7, and again in Section 3.1.

Let N be an integer. A *Dirichlet character* modulo N is a function $\chi : \mathbb{Z} \to \mathbb{C}$, which is periodic with period N, such that

$$|\chi(n)| = \begin{cases} 1 & \text{if } (n, N) = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and such that $\chi(nm) = \chi(n)\,\chi(m)$. To obtain a Dirichlet character, start with a character of the finite Abelian group $(\mathbb{Z}/N\mathbb{Z})^\times$, and extend it to a function on all of $\mathbb{Z}/N\mathbb{Z}$ by making it zero on the residue classes not prime to N; then compose this function with the canonical map $\mathbb{Z} \to \mathbb{Z}/N\mathbb{Z}$.

If $N_1|N$, there are canonical maps $\mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/N_1\mathbb{Z}$ and $(\mathbb{Z}/N\mathbb{Z})^\times \to (\mathbb{Z}/N_1\mathbb{Z})^\times$. If χ is a Dirichlet character modulo N_1, we may take the corresponding character of $(\mathbb{Z}/N_1\mathbb{Z})^\times$, pull it back to $(\mathbb{Z}/N\mathbb{Z})^\times$, and obtain a

Dirichlet character modulo N. If χ is obtained this way from a character modulo a proper divisor N_1 of N, then χ is called *imprimitive*. If χ is not imprimitive, it is called *primitive*. If χ is primitive modulo N, we say that N is the *conductor* of χ.

If χ is a Dirichlet character, whether primitive or not, we may define an L-function

$$L(s, \chi) = \sum_{n=1}^{\infty} \chi(n) n^{-s}.$$

By comparison with the Riemann zeta function, it is absolutely convergent for $re(s) > 1$. It has an *Euler product*, as we can see as follows. We will prove that for re$(s) > 1$, we have

$$L(s, \chi) = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}. \tag{1.1}$$

Indeed, expanding a geometric series, each individual Euler factor

$$\left(1 - \frac{\chi(p)}{p^s}\right)^{-1} = \sum_{k=0}^{\infty} \chi(p)^k p^{-ks}.$$

Now for any positive integer n, there is exactly one way to write n^{-s} as a product of factors p^{-ks}; so if $n = \prod p_i^{k_i}$, on expanding

$$\prod_p \sum_{k=0}^{\infty} \chi(p)^k p^{-ks},$$

the coefficient of n^{-s} is

$$\prod_i \chi(p_i)^{k_i} = \chi(n).$$

Hence the right side of Eq. (1.1) is equal to the left side, as required.

Dirichlet L-functions have another, deeper property, namely *analytic continuation* and a *functional equation*, which may be regarded as aspects of the interplay between additive and multiplicative Fourier analysis. The functional equation only works well for primitive characters. The remainder of this section will be devoted to developing the functional equations of Dirichlet L-functions.

It is a useful property of primitive characters that there is a convenient way to interpolate the character, originally defined on $\mathbb{Z}$, to a smooth function on $\mathbb{R}$. To this end, we introduce *Gauss sums*.

Let χ be a primitive character modulo N. The *Gauss sum* $\tau(\chi)$ or τ_χ is defined by the formula

$$\tau(\chi) = \sum_{n \bmod N} \chi(n) e^{2\pi i n/N}. \tag{1.2}$$

We will prove that

$$\sum_{n \bmod N} \chi(n)\, e^{2\pi i n m/N} = \overline{\chi(m)}\, \tau(\chi). \tag{1.3}$$

There are two cases. If $(m, N) = 1$, we have $|\chi(m)| = 1$. Making the change of variables $n \to nm$, we see that

$$\tau(\chi) = \sum_{n \bmod N} \chi(nm)\, e^{2\pi i n m/N} = \chi(m) \sum_{n \bmod N} \chi(n)\, e^{2\pi i n m/N},$$

and multiplying by $\chi(m)^{-1} = \overline{\chi(m)}$, we obtain Eq. (1.3).

On the other hand, if $(m, N) > 1$, we have $\chi(m) = 0$, so it is sufficient to show the left side of Eq. (1.3) vanishes. Suppose that $m = dM$ and $N = dN_1$ where $d > 1$. Let us show first that the primitivity of χ implies that there exists $c \equiv 1 \bmod N_1$ such that $(c, N) = 1$ and $\chi(c) \neq 1$. If not, $\chi(c) = 1$ for all c prime to N such that $c \equiv 1 \bmod N_1$, which implies that $\chi(n) = \chi(n')$ whenever n, n' are prime to N and $n \equiv n' \bmod N_1$. Hence χ is well defined modulo N_1, a proper divisor of N, and χ is the pullback of a character under the canonical map $(\mathbb{Z}/N\mathbb{Z})^\times \to (\mathbb{Z}/N_1\mathbb{Z})^\times$, contradicting the primitivity of χ. Now observe that the left side of Eq. (1.3) equals

$$\sum_{n \bmod N} \chi(n)\, e^{2\pi i n M/N_1} = \sum_{r \bmod N_1} \left\{ \sum_{\substack{n \bmod N \\ n \equiv r \bmod N_1}} \chi(n) \right\} e^{2\pi i r M/N_1}. \tag{1.4}$$

Now the substitution $n \to cn$ permutes the residue classes $n \bmod N$ such that $n \equiv r \bmod N_1$ amongst themselves, so

$$\sum_{\substack{n \bmod N \\ n \equiv r \bmod N_1}} \chi(n) = \sum_{\substack{n \bmod N \\ n \equiv r \bmod N_1}} \chi(cn) = \chi(c) \sum_{\substack{n \bmod N \\ n \equiv r \bmod N_1}} \chi(n);$$

as $\chi(c) \neq 1$, this expression must equal zero. Hence Eq. (1.4) vanishes, completing the proof of Eq. (1.3).

Now we need to know that $\tau(\chi) \neq 0$. In fact, we will prove that

$$|\tau(\chi)| = \sqrt{N}. \tag{1.5}$$

Because

$$\overline{\chi(n)\, e^{2\pi i n m/N}} = \overline{\chi(n)}\, e^{-2\pi i n m/N},$$

we have

$$\left| \sum_{n \bmod N} \chi(n)\, e^{2\pi i n m/N} \right|^2 = \sum_{\substack{n_1, n_2 \bmod N \\ (n_1 n_2, N) = 1}} \chi(n_1)\overline{\chi(n_2)} e^{2\pi i (n_1 - n_2) m/N}.$$

By Eq. (1.3), this equals $|\tau(\chi)|^2$ if $(m, N) = 1$ and zero if $(m, N) \neq 1$. Summing over all m modulo N, we obtain

$$\phi(N)\,|\tau(\chi)|^2 = \sum_{m \bmod N} \sum_{\substack{n_1, n_2 \bmod N \\ (n_1 n_2, N) = 1}} \chi(n_1)\overline{\chi(n_2)} e^{2\pi i (n_1 - n_2) m/N},$$

where ϕ is the Euler totient function: $\phi(N)$ is the number of residue classes modulo N prime to N, or in other words, the cardinality of $(\mathbb{Z}/N\mathbb{Z})^\times$. Because

$$\sum_{m \bmod N} e^{2\pi i a m/N} = \begin{cases} N & \text{if } N \mid a, \\ 0 & \text{otherwise,} \end{cases}$$

only terms with $n_1 \equiv n_2 \bmod N$ contribute; for these, $\chi(n_1)\overline{\chi(n_2)} = 1$, so we obtain

$$\phi(N)\,|\tau(\chi)|^2 = \sum_{\substack{n_1, n_2 \bmod N \\ (n_1 n_2, N) = 1 \\ n_1 \equiv n_2 \bmod N}} N = \phi(N)\,N.$$

Hence we have Eq. (1.5).

Now that we know the Gauss sums do not vanish, we may explain how to interpolate a primitive Dirichlet character between the integers. Replace χ by $\overline{\chi}$ and rewrite Eq. (1.3) as follows:

$$\chi(n) = \frac{1}{\tau(\overline{\chi})} \sum_{m \bmod N} \overline{\chi(m)}\, e^{2\pi i n m/N}, \tag{1.6}$$

so by Eq. (1.5) and Exercise 1.1.1, we have

$$\chi(n) = \frac{\chi(-1)\,\tau(\chi)}{N} \sum_{m \bmod N} \overline{\chi(m)}\, e^{2\pi i n m/N}. \tag{1.7}$$

Observe that the right-hand side is defined when n is an arbitrary real number. We have therefore obtained our goal of finding a natural way of interpolating a primitive Dirichlet character to an arbitrary real argument.

We now require the *Poisson summation formula*. Let f be a function on $\mathbb{R}$ that is sufficiently well behaved. For example, it is sufficient if f is piecewise continuous with only finitely many discontinuities, of bounded total variation, satisfies

$$f(a) = \tfrac{1}{2}\left[\lim_{x \to a-} f(x) + \lim_{x \to a+} f(x)\right]$$

for all a, and

$$|f(x)| < c_1 \min(1, x^{-c_2})$$

for some $c_1 > 0, c_2 > 1$. We define the *Fourier transform*

$$\hat{f}(x) = \int_{-\infty}^{\infty} f(y)\, e^{2\pi i x y}\, dy. \tag{1.8}$$

The Poisson summation formula asserts that

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \hat{f}(n). \tag{1.9}$$

To prove this, let

$$F(x) = \sum_{n=-\infty}^{\infty} f(x+n).$$

Then $F(x)$ is periodic with period 1, is of bounded variation, and satisfies

$$F(a) = \tfrac{1}{2}\left[\lim_{x\to a-} F(x) + \lim_{x\to a+} F(x)\right].$$

It is a standard theorem from Fourier analysis that $F(x)$ has a Fourier expansion that represents it for all values of x:

$$F(x) = \sum_{-\infty}^{\infty} a_m\, e^{2\pi i m x}.$$

(This follows, for example, from Whittaker and Watson (1927, 9.42 on p. 175)) The constants a_m are computed in the usual way, by orthogonality:

$$a_m = \int_0^1 F(x)\, e^{-2\pi i m x}\, dx = \int_0^1 \sum_{n=-\infty}^{\infty} f(x+n)\, e^{-2\pi i m x}\, dx.$$

Because $e^{2\pi i m x} = e^{2\pi i m(x+n)}$, this equals

$$\sum_{n=-\infty}^{\infty} \int_0^1 f(x+n)\, e^{-2\pi i m(x+n)}\, dx = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i m x}\, dx,$$

so $a_m = \hat{f}(-m)$. Thus

$$\sum_{n=-\infty}^{\infty} f(n) = F(0) = \sum_{n=-\infty}^{\infty} a_n = \sum_{n=-\infty}^{\infty} \hat{f}(n),$$

as required.

Actually, we require a slight generalization of the Poisson summation formula, which we call *twisted Poisson summation*; the formula is like the usual

Poisson formula, except that it is "twisted" by a Dirichlet character. Let χ be a primitive character modulo N. We will prove that

$$\sum_{n=-\infty}^{\infty} \chi(n) f(n) = \frac{\chi(-1)\tau(\chi)}{N} \sum_{n=-\infty}^{\infty} \overline{\chi(n)} \hat{f}(n/N). \tag{1.10}$$

To prove this, let us observe that by Eq. (1.7), the left side of Eq. (1.10) equals $\sum f_1(n)$, where

$$f_1(x) = \frac{\chi(-1)\tau(\chi)}{N} \sum_{m \bmod N} \overline{\chi(m)} e^{2\pi i x m/N} f(x).$$

We may thus apply the Poisson summation formula. It is easy to check that

$$\hat{f}_1(x) = \frac{\chi(-1)\tau(\chi)}{N} \sum_{m \bmod N} \overline{\chi(m)} \hat{f}\left(\frac{Nx+m}{N}\right).$$

Thus the left side of Eq. (1.10) equals

$$\frac{\chi(-1)\tau(\chi)}{N} \sum_{m \bmod N} \sum_{n=-\infty}^{\infty} \overline{\chi(m)} \hat{f}\left(\frac{Nn+m}{N}\right).$$

Because $\chi(m) = \chi(Nn+m)$, and because $Nn+m$ runs uniquely through $\mathbb{Z}$ when m runs through a set of residue classes modulo N and n runs through $\mathbb{Z}$, this equals the right side of Eq. (1.10). This completes the proof of the twisted Poisson summation formula.

Now we need to compute the Fourier transform of the *Gaussian distribution*. Let t have positive real part, and let

$$f_t(x) = e^{-\pi t x^2}. \tag{1.11}$$

Then f_t is of faster-than-polynomial decay as $x \to \pm\infty$. We will prove that

$$\hat{f}_t = \frac{1}{\sqrt{t}} f_{1/t}. \tag{1.12}$$

Because both sides are analytic functions of t defined when the real part of t is positive, it is sufficient to prove Eq. (1.12) when t is real, which we now assume. Completing the square

$$\hat{f}_t(x) = \int_{-\infty}^{\infty} e^{-\pi(ty^2 - 2ixy)}\, dy = e^{-\pi x^2/t} \int_{-\infty}^{\infty} e^{-\pi(\sqrt{t}y - ix/\sqrt{t})^2}\, dy.$$

Now we may use Cauchy's theorem to shift the path of integration with respect to y vertically by a constant distance depending on x, amounting to replacing

y by $y + ix/t$. Thus

$$\hat{f}_t(x) = e^{-\pi x^2/t} \int_{-\infty}^{\infty} e^{-\pi t y^2}\, dy = \frac{c}{\sqrt{t}}\, e^{-\pi x^2/t}, \tag{1.13}$$

where

$$c = \int_{-\infty}^{\infty} e^{-\pi y^2}\, dy.$$

It is well known that $c = 1$; in fact, we may prove that right now by applying Eq. (1.13) twice to obtain

$$\hat{\hat{f}}_t = \frac{c}{\sqrt{t}} \hat{f}_{1/t} = c^2 f_t.$$

We recall the *Fourier inversion formula*

$$\hat{\hat{f}}(x) = f(-x),$$

which is valid if f is any continuous function such that both f and $\hat{f}$ are in $L^1(\mathbb{R})$. (See Katznelson, 1976, VI.1.12.) Applying this to the even function f_t, we have $\hat{\hat{f}}_t = f_t$. Therefore $c^2 = 1$. Evidently, $c > 0$, and so $c = 1$, from which we get Eq. (1.13).

Now we may construct some *theta functions*. Let χ be a primitive character modulo N. First assume that $\chi(-1) = 1$, and define

$$\theta_\chi(t) = \tfrac{1}{2} \sum_{n=-\infty}^{\infty} \chi(n)\, e^{-\pi n^2 t} = \tfrac{1}{2}\chi(0) + \sum_{n=1}^{\infty} \chi(n)\, e^{-\pi n^2 t} \tag{1.14}$$

when t has positive real part. (Here $\chi(0) = 0$ unless $N = 1$ because χ is primitive.) Using Eq. (1.12) and the twisted Poisson summation formula, we obtain the functional equation:

$$\theta_\chi(t) = \frac{\tau(\chi)}{N\sqrt{t}}\, \theta_{\overline{\chi}}\left(\frac{1}{N^2 t}\right). \tag{1.15}$$

Now suppose that $\chi(-1) = -1$. In this case, we cannot use the definition Eq. (1.14); the terms for n and $-n$ cancel, so Eq. (1.14) is zero in this case. So we must do something else. Let

$$g_t(x) = x\, e^{-\pi t x^2} = -\frac{1}{2\pi t}\left(\frac{df_t}{dx}\right)(x). \tag{1.16}$$

We will prove that

$$\hat{g}_t(x) = \frac{i}{t^{3/2}}\, g_{1/t}. \tag{1.17}$$

We have, by definition of the Fourier transform

$$\hat{g}_t(x) = -\frac{1}{2\pi t} \int_{-\infty}^{\infty} \left(\frac{df_t}{dy}\right)(y)\, e^{2\pi i x y}\, dy.$$

Integrating by parts, this equals

$$\frac{ix}{t} \int_{-\infty}^{\infty} f_t(y)\, e^{2\pi i x y}\, dy,$$

which by Eq. (1.12) equals the right side of Eq. (1.17), which is now proved. Now if $\chi(-1) = -1$, we define

$$\theta_\chi(t) = \tfrac{1}{2} \sum_{n=-\infty}^{\infty} n\, \chi(n)\, e^{-\pi n^2 t} = \sum_{n=1}^{\infty} n\, \chi(n)\, e^{-\pi n^2 t}. \tag{1.18}$$

Applying twisted Poisson summation to g_t, we obtain

$$\theta_\chi(t) = \frac{-i\,\tau(\chi)}{N^2 t^{3/2}}\, \theta_{\overline{\chi}}\left(\frac{1}{N^2 t}\right). \tag{1.19}$$

Now we may prove the functional equations of Dirichlet L-functions.

Theorem 1.1.1 *Let χ be a primitive Dirichlet character with conductor N, and let $\epsilon = 0$ or 1 be chosen so that $\chi(-1) = (-1)^\epsilon$. Let*

$$\Lambda(s, \chi) = \pi^{-(s+\epsilon)/2}\, \Gamma\left(\frac{s+\epsilon}{2}\right) L(s, \chi).$$

Then $\Lambda(s, \chi)$ has meromorphic continuation to all s; indeed, if $\chi \neq 1$, it is entire, while if $\chi = 1$, it is analytic for all s except $s = 1$ or $s = 0$, where it has simple poles. We have the functional equation

$$\Lambda(s, \chi) = (-i)^\epsilon\, \tau(\chi)\, N^{-s}\, \Lambda(1 - s, \overline{\chi}). \tag{1.20}$$

Proof We will consider the case where $\chi \neq 1$. If χ is the trivial character, then the primitivity of χ implies that $N = 1$, and $L(s, \chi)$ is just the Riemann zeta function; we leave this case to the reader (Exercise 1.7). Because $\chi(0) = 0$, the series $\theta_\chi(t)$ is a sum of terms of the form $n^\epsilon\, \chi(n)\, e^{-\pi n^2 t}$ with $|n| \geq 1$, each of which is of very rapid decay as $t \to \infty$; combining Eqs. (1.15) and (1.19), we have

$$\theta_\chi(t) = \frac{(-i)^\epsilon\, \tau(\chi)}{N^{1+\epsilon}\, t^{\epsilon+1/2}}\, \theta_{\overline{\chi}}\left(\frac{1}{N^2 t}\right);$$

this implies that $\theta_\chi(t)$ is also of very rapid decay as $t \to 0$. Therefore the Mellin transform

$$\int_0^\infty \theta_\chi(t)\, t^{(s+\epsilon)/2}\, \frac{dt}{t} \tag{1.21}$$

is convergent for all s. If the real part of $s > 1$, we may use the identity

$$\int_0^\infty e^{-\pi t n^2}\, t^{(s+\epsilon)/2}\, \frac{dt}{t} = \pi^{-(s+\epsilon)/2}\, \Gamma\left(\frac{s+\epsilon}{2}\right) n^{-s-\epsilon}$$

to see that Eq. (1.21) equals $\Lambda(s, \chi)$. Because the integral Eq. (1.21) is convergent for all s, and clearly defines an analytic function, this gives the analytic continuation of $\Lambda(s, \chi)$. Now substituting Eq. (1.15) or Eq. (1.19) into Eq. (1.21) and making the change of variables $t \mapsto 1/N^2 t$, Eq. (1.21) equals

$$(-i)^\epsilon\, \tau(\chi)\, N^{-s} \int_0^\infty \theta_{\overline{\chi}}(t)\, t^{(1-s+\epsilon)/2}\, \frac{dt}{t}.$$

Hence we obtain Eq. (1.20). ■

Exercises

Exercise 1.1.1 Let χ be a primitive character modulo N. Show that $\tau(\overline{\chi}) = \chi(-1)\,\overline{\tau(\chi)}$.

Exercise 1.1.2: Dirichlet (a) Show that the identity

$$\sum_{n=1}^\infty \frac{x^n}{n} = -\log(1-x),$$

valid if $|x| < 1$, remains true if $|x| = 1$ and $x \neq 1$, in which case the series is conditionally convergent.
(b) Let χ be a nontrivial primitive character modulo N. Assume $N > 1$, so that χ is nontrivial. Use Eq. (1.7) to prove that

$$L(1, \chi) = \sum_{n=1}^\infty \frac{\chi(n)}{n} = -\frac{\chi(-1)\,\tau(\chi)}{N} \sum_{m \bmod N} \overline{\chi(m)}\, \log(1 - e^{2\pi i m/N}). \tag{1.22}$$

From this, deduce that

$$L(1,\chi) = \begin{cases} -\dfrac{\tau(\chi)}{N} \displaystyle\sum_{m \bmod N} \overline{\chi(m)} \log|1 - e^{2\pi i m/N}| & \text{if } \chi(-1) = 1; \\ \dfrac{i\pi\tau(\chi)}{N^2} \displaystyle\sum_{m=1}^{N} \overline{\chi(m)}\, m & \text{if } \chi(-1) = -1. \end{cases}$$

Recall that for the character χ modulo N to be *quadratic* means that $\chi(n) = \pm 1$ for all $(n, N) = 1$, but that χ is not identically one.

Exercise 1.1.3 Let p be an odd prime.

(a) Prove that there is a unique quadratic character χ modulo p.
(b) Prove that the number of solutions to $x^2 \equiv a \bmod p$ equals $1 + \chi(a)$.
(c) Show that

$$\tau(\chi) = \sum_{n=0}^{p-1} e^{2\pi i n^2/p}.$$

Exercise 1.1.4 Let $\tau = x + iy$, where $x, y \in \mathbb{R}$ and $y > 0$. Let k be an integer greater than or equal to 2. Define

$$f(u) = (u - \tau)^{-k}.$$

Use the residue theorem to show that

$$\hat{f}(v) = \begin{cases} 2\pi i \ \operatorname{res}\left(e^{2\pi i u v}\,(u-\tau)^{-k}\right)|_{u=\tau} & \text{if } v > 0; \\ 0 & \text{if } v \le 0. \end{cases}$$

Hence

$$\hat{f}(v) = \begin{cases} \frac{(2\pi i)^k}{(k-1)!}\, v^{k-1}\, e^{2\pi i v \tau} & \text{if } v > 0; \\ 0 & \text{if } v \le 0. \end{cases}$$

Conclude that

$$\sum_{n=-\infty}^{\infty} (n - \tau)^{-k} = \frac{(2\pi i)^k}{(k-1)!} \sum_{n=1}^{\infty} n^{k-1}\, e^{2\pi i n \tau}.$$

The Quadratic Reciprocity Law I will state without proof this fundamental theorem of Gauss. If p is an odd prime, the *Legendre symbol*

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{if } x^2 \equiv a \bmod p \text{ has two solutions;} \\ -1 & \text{if } x^2 \equiv a \bmod p \text{ has no solutions;} \\ 0 & \text{if } a \equiv 0 \bmod p. \end{cases}$$

The definition is extended so that $\left(\frac{a}{b}\right)$ is defined whenever b is an odd positive number by the rule

$$\left(\frac{a}{\prod_{i=1}^{n} p_i}\right) = \prod_{i=1}^{n} \left(\frac{a}{p_i}\right)$$

where $p_1, \cdots, p_n$ are primes. Extended in this way, the symbol is called the *Jacobi symbol*. The basic properties of the Jacobi symbol are as follows:

(i) $$\left(\frac{a}{b}\right) = \left(\frac{a'}{b}\right) \qquad \text{if } a \equiv a' \bmod b.$$

(ii) $$\left(\frac{aa'}{b}\right) = \left(\frac{a}{b}\right)\left(\frac{a'}{b}\right).$$

(iii) $$\left(\frac{a}{bb'}\right) = \left(\frac{a}{b}\right)\left(\frac{a}{b'}\right).$$

(iv) $$\left(\frac{-1}{b}\right) = (-1)^{(b-1)/2} = \begin{cases} 1 & \text{if } b \equiv 1 \bmod 4; \\ -1 & \text{if } b \equiv -1 \bmod 4. \end{cases}$$

(v) $$\left(\frac{2}{b}\right) = (-1)^{(b^2-1)/8} = \begin{cases} 1 & \text{if } b \equiv \pm 1 \bmod 8; \\ -1 & \text{if } b \equiv \pm 3 \bmod 8. \end{cases}$$

If a and b are *both* odd positive integers, we have

(vi) $$\left(\frac{a}{b}\right) = (-1)^{\frac{1}{2}(a-1)\cdot\frac{1}{2}(b-1)}\left(\frac{b}{a}\right) = \begin{cases} -\left(\frac{b}{a}\right) & \text{if } a \equiv b \equiv -1 \bmod 4; \\ \left(\frac{b}{a}\right) & \text{otherwise.} \end{cases}$$

Part (vi) is the *quadratic reciprocity law*.

Exercise 1.1.5: Quadratic Fields If K is a quadratic extension of $\mathbb{Q}$, let o_K denote the ring of integers in K. Then $o_K \cong \mathbb{Z} \oplus \mathbb{Z}$ as an Abelian group. Let α, β be a $\mathbb{Z}$-basis of o_K. The *discriminant* of K is by definition

$$D_K = \begin{vmatrix} \alpha & \beta \\ \alpha' & \beta' \end{vmatrix}^2,$$

where $x \mapsto x'$ denotes conjugation, that is, the nontrivial Galois automorphism of K over $\mathbb{Q}$. Show that this definition is independent of the choice of basis and that $D_K \in \mathbb{Z}$.

Exercise 1.1.6: Fundamental Discriminants Part (c) of this exercise assumes the quadratic reciprocity law, and part (d) assumes the definition of a discriminant of a quadratic field.

(a) Prove that if q is a prime power, then there exists a primitive quadratic character modulo q if and only if q equals 4, 8, or is an odd prime; in each of these cases there is precisely one primitive quadratic character, except that if $q = 8$, there are two, one satisfying $\chi(-1) = 1$ and one satisfying $\chi(-1) = -1$.

(b) Show that if $(m, n) = 1$, then

$$(\mathbb{Z}/mn\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times,$$

and deduce that there exists a primitive quadratic Dirichlet character modulo a positive integer d if and only if d is the product of relatively prime factors, each of which is an odd prime, or else equals 4 or 8.
(c) Show that if D is an integer, positive or negative, and if there exists a quadratic character χ that is primitive modulo $|D|$ such that the sign of D is equal to $\chi(-1)$, then

$$\chi(n) = \left(\frac{D}{n}\right)$$

for odd positive integers n.

The integers D satisfying the condition of (c) are called *fundamental discriminants*. They are in one-to-one correspondence with the primitive quadratic characters.

The restriction to odd n in (c) is undesirable; it is sometimes removed by employing Kronecker's modification of the Jacobi symbol, in which $\left(\frac{a}{b}\right)$ is sometimes defined even when b is even. Using the Kronecker symbol, one may say that the unique quadratic character modulo $|D|$, where D is a fundamental discriminant, is $n \mapsto \left(\frac{D}{n}\right)$. This has some disadvantages; for example, (i) above is no longer true for the Kronecker symbol. (Shimura (1973) has proposed yet another modification of the quadratic symbol.) We will avoid using the Kronecker symbol. If D is a fundamental discriminant, we will denote the unique primitive quadratic character modulo $|D|$ such that $\chi(-1)$ has the same sign as D by χ_D.
(d) Prove if K is a quadratic extension of $\mathbb{Q}$, there exists a unique fundamental discriminant D such that $K = \mathbb{Q}(\sqrt{D})$; thus the fundamental discriminants are in one-to-one correspondence with quadratic fields.
(e) Prove that if D is a fundamental discriminant, then $D \equiv 0$ or $1 \bmod 4$, and that the ring of integers in $K = \mathbb{Q}(\sqrt{D})$ is $\mathbb{Z} \oplus \mathbb{Z}\tau$ where

$$\tau = \begin{cases} \frac{1}{2}\sqrt{D} & \text{if } D \equiv 0 \bmod 4; \\ \frac{1}{2}(\sqrt{D}+1) & \text{if } D \equiv 1 \bmod 4. \end{cases}$$

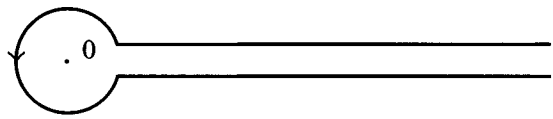
Conclude that D is the discriminant of K. Hence the fundamental discriminants are precisely the discriminants of quadratic fields.
(f) Let D be a fundamental discriminant, p a prime, and $K = \mathbb{Q}(\sqrt{D})$. Show that

$$\begin{cases} p \text{ splits in } K & \text{if and only if } \chi_D(p) = 1; \\ p \text{ remains prime in } K & \text{if and only if } \chi_D(p) = -1; \\ p \text{ ramifies in } K & \text{if and only if } \chi_D(p) = 0. \end{cases}$$

(Use the quadratic reciprocity law. The case $p = 2$ must be handled separately.)

Exercise 1.1.7 Explain how to modify the proof of Theorem 1.1 to handle the case where $N = 1$, so that $\chi = 1$ and $L(s, \chi)$ is the Riemann zeta function.

Exercise 1.1.8: Riemann (1892) Riemann gave two proofs of the functional equation of ζ, each important in its own way. The proof based on taking the Mellin transform of a theta function as in the proof of Theorem 1.1 is Riemann's second proof. (It was extended to L-functions by Hecke (1918) and (1920)). This exercise, based on Riemann's first proof of the functional equation, leads to a determination of the values of $\zeta(s)$ at the negative odd integers or equivalently, at the positive even integers. Riemann's paper is discussed at length in Edwards (1974). For the extension to L-functions, see Chapter 4 of Washington (1982).
(a) The *Hankel Contour C* begins and ends at ∞, circling the origin counterclockwise:



Prove that if re(s) is large

$$\int_C (-x)^{s-1} e^{-x}\, dx = -2i \sin(\pi s) \int_0^\infty t^{s-1} e^{-t}\, dt$$
$$= -2i \sin(\pi s)\, \Gamma(s).$$

In this integration, we define $(-x)^{s-1}$ to be $e^{(s-1)\log(-x)}$, where we choose the branch of log that is real when $(-x)$ is real and positive. In view of the well-known identity

$$\Gamma(s)\,\Gamma(1-s) = \frac{\pi}{\sin(\pi s)},$$

this may be rewritten

$$\frac{1}{\Gamma(1-s)} = \frac{i}{2\pi} \int_C (-x)^{s-1} e^{-x}\, dx.$$

Although we proved this only for re(s) large, observe that the integral is convergent for all s, so by analytic continuation, this formula is valid for all s.
(b) Use the geometric series identity

$$\frac{1}{e^x - 1} = \sum_{n=1}^{\infty} e^{-nx},$$

valid if $\mathrm{re}(x) > 0$, and adapt the calculation of (a) to show that

$$\zeta(s) = -\frac{\Gamma(1-s)}{2\pi i} \int_C \frac{(-x)^{s-1}}{e^x - 1}\, dx.$$

This formula is valid for all s.

The *Bernoulli numbers* are defined by the identity

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} B_n \frac{t^n}{n!}.$$

We have $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, and $B_4 = -1/30$. It is not hard to see that $B_n = 0$ if n is odd and greater than 1; if n is even, it is clear that B_n is rational, and it will follow from (d) below that the sign of B_n is $-(-1)^{n/2}$.

(c) Use the functional equation to show that ζ vanishes at the negative even integers. Use the residue theorem to show that if n is a positive even integer, then $\zeta(1-n) = -B_n/n$.

(d) Use the functional equation to deduce that if n is a positive even integer

$$\zeta(n) = -\frac{2^{n-1}\pi^n (-1)^{n/2} B_n}{n!}.$$

Exercise 1.1.9 We return to the setting of Exercise 1.2(b). Assume that χ is quadratic, so its conductor $N = |D|$, where D is a fundamental discriminant. Then χ is the quadratic character attached to the quadratic extension $K = \mathbb{Q}(\sqrt{D})$. We recall the factorization of the Dedekind zeta function $\zeta_K(s) = \zeta(s)\, L(s, \chi)$ (see Lang (1970, Theorem XII.1, p. 230)). Thus $L(1, \chi)$ is the residue at $s = 1$ of ζ_K, which is computed classically as in Lang (1970, Theorem XIII.2, p. 259). Suppose that $\chi(-1) = -1$ so that K is imaginary quadratic. Then

$$L(1, \chi) = \frac{2\pi\, h}{w\sqrt{|D|}},$$

where D is the discriminant of K, h is its class number, and w is the number of roots of unity in K (two unless $D = -4$ or -3.) Thus by Exercise 1.2(b),

$$h = i\tau(\chi)\, w\, |D|^{-3/2} 2^{-1} \sum_{m=1}^{D} \chi(m)\, m.$$

But $\tau(\chi) = i\sqrt{D}$. (See Washington (1982, Corollary 4.6, p. 35) for the evaluation of quadratic Gauss sums. Also, compare Eq. (9.15) in section 1.9.) We obtain *Dirichlet's class number formula*

$$h = -\frac{w}{2|D|} \sum_{m=1}^{D} \chi(m)\, m. \tag{1.23}$$

Exercise 1.1.10 (a) Let χ be a primitive character modulo N. Prove, using the functional equation, that $L(s, \chi)$ has a simple zero at $s = 0$ if $\chi(-1) = 1$ and is nonzero at $s = 0$ if $\chi(-1) = -1$.
(b) Stark (1971), (1975), (1976) and (1980) has conjectured that if $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL(n, \mathbb{C})$ is a Galois representation such that the Artin L-function $L(s, \rho)$ has a zero of order r at $s = 0$, the leading coefficient in its Taylor expansion is essentially an $r \times r$ "Stark regulator" of units in some number field. The simplest open cases of the conjecture are when $r = 1$. Artin's reciprocity law allows us to consider χ to be a Galois character, namely, it gives a character of $\mathrm{Gal}\big(\mathbb{Q}(\zeta)/\mathbb{Q}\big) \cong (\mathbb{Z}/N\mathbb{Z})^\times$, $\zeta = e^{2\pi i/N}$, where $a \in (\mathbb{Z}/N\mathbb{Z})^\times$ corresponds to $\sigma_a \in \mathrm{Gal}\big(\mathbb{Q}(\zeta)/\mathbb{Q}\big)$, $\sigma_a(\zeta) = \zeta^a$. With this identification, reinterpret (a) to show that $r = 1$ if $\chi(-1) = 1$, and $r = 0$ if $\chi(-1) = -1$.
(c) Assume that $\chi(-1) = 1$. In this case Exercise 1.2(b) verifies the Stark conjecture because it shows that

$$L(1, \chi) = -\frac{\tau(\chi)}{N} \sum_{m \bmod N} \overline{\chi(m)} \log |\epsilon_m|, \tag{1.24}$$

where $\epsilon_m = (1 - e^{2\pi i m/N})/(1 - e^{2\pi i/N})$. Note that if m and N are coprime, ϵ_m is a unit in $\mathbb{Z}[\zeta]$.

1.2 The Modular Group

In this section, let $G = SL(2, \mathbb{R})$ and let $\mathcal{H}$ be the *Poincaré upper half plane* consisting of $z = x + iy$ where $x, y \in \mathbb{R}$, and $y > 0$. G acts on $\mathcal{H}$ via linear fractional transformations:

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \to g(z) = \frac{az + b}{cz + d}. \tag{2.1}$$

It is easy to check that this is a bona fide group action, that is, $(g_1 g_2)(z) = g_1\big(g_2(z)\big)$. This action is not quite faithful because $-I$ acts trivially, I being the identity matrix. If we wish to work with a group having a faithful action, we may pass to the group $PSL(2, \mathbb{R}) = G/\{\pm I\}$, which may be identified with a group of transformations of $\mathcal{H}$. If $\Gamma \subset G$ is any group, we will denote by $\overline{\Gamma}$ its image in $\overline{G} = PSL(2, \mathbb{R})$. Thus $\overline{\Gamma} = \Gamma/\{\pm I\}$ if $-I \in \Gamma$, or $\overline{\Gamma} \cong \Gamma$ if $-I \notin \Gamma$. We may sometimes extend the action of $SL(2, \mathbb{R})$ to the group $GL(2, \mathbb{R})^+$ of 2×2 nonsingular matrices with positive determinant by the formula (2.1). Of course, the scalar matrices act trivially.

More generally, we allow $SL(2, \mathbb{C})$ (or $GL(2, \mathbb{C})$) to act on the Riemann sphere $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$ by linear fractional transformations using the same formula (2.1). The subgroup of $SL(2, \mathbb{C})$ that maps the subspace $\mathcal{H} \subset \mathbb{P}^1(\mathbb{C})$ onto itself is just $SL(2, \mathbb{R})$.

The action of G on $\mathcal{H}$ is transitive because in fact the subgroup B of upper

triangular matrices acts transitively. Indeed,

$$\begin{pmatrix} y^{1/2} & xy^{-1/2} \\ & y^{-1/2} \end{pmatrix} : i \to x + iy,$$

so every element of $\mathcal{H}$ is in the orbit of i. The stabilizer of i is the subgroup

$$SO(2) = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \middle| a^2 + b^2 = 1 \right\}.$$

Recall that if X is any set, and if G is a group acting transitively on X, we may identify X with the set of cosets G/G_x, where G_x is the stabilizer of some fixed point $x \in X$. In the particular case where $G = SL(2, \mathbb{R})$, $X = \mathcal{H}$, and $x = i$, we see that we may identify $\mathcal{H}$ with the space of cosets $G/SO(2)$. Because, as we have just seen, B acts transitively on $\mathcal{H}$, it acts transitively on $G/SO(2)$, and so we obtain a geometric proof that $G = B \cdot SO(2)$. This relation is known as the *Iwasawa decomposition* for $SL(2, \mathbb{R})$.

We will be particularly interested in the subgroup $\Gamma(1) = SL(2, \mathbb{Z})$ of $SL(2, \mathbb{R})$ and certain subgroups that are called *congruence subgroups*. Let

$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(1) \middle| a \equiv d \equiv 1 \bmod N, b \equiv c \equiv 0 \bmod N \right\}$$

Note that $\Gamma(N)$ is the kernel of the canonical map $\Gamma(1) \to SL(2, \mathbb{Z}/N\mathbb{Z})$. Because this is a finite group, we see that $\Gamma(N)$ is normal in $\Gamma(1)$ and of finite index. A subgroup of $SL(2, \mathbb{Z})$ is called a *congruence subgroup* if it contains $\Gamma(N)$ for some N.

The identity

$$\operatorname{im}\big(g(z)\big) = |cz + d|^{-2}\, y \tag{2.2}$$

$$\text{for } g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R}),\ z = x + iy \in \mathcal{H}$$

is readily confirmed (Exercise 1.2.1). If Γ is a subgroup of G, we say that the action of Γ on $\mathcal{H}$ is *discontinuous* if for any two compact subsets $K_1, K_2 \subset \mathcal{H}$, the set

$$\{\gamma \in \Gamma | K_2 \cap \gamma(K_1) \neq \emptyset\}$$

is finite. As a first application of Eq. (2.2), let us prove the following:

Proposition 1.2.1 *The group $\Gamma(1)$ acts discontinuously on $\mathcal{H}$.*

It may be shown more generally that a subgroup $\Gamma \subset SL(2, \mathbb{R})$ acts discontinuously on $\mathcal{H}$ if Γ is discrete in the topology that it inherits from $SL(2, \mathbb{R})$. For our purposes, however, Proposition 2.1 is sufficient.

Proof Let K_1 and K_2 be compact subsets of $\mathcal{H}$. There exists a constant $\epsilon > 0$ such that $\text{im}(w) > \epsilon$ for all $w \in K_2$. Now for fixed $z = x + iy \in K_1$, note that $(c, d) \mapsto |cz + d|^2$ is a positive definite quadratic form. Applying Eq. (2.2), there is a constant $R(z)$ such that $\text{im}\big(g(z)\big) = |cz+d|^{-2}\, y < \epsilon$ unless $|c|, |d| < R(z)$. Because K_1 is compact, $R = \max\{R(z) | z \in K_1\} < \infty$, and $K_2 \cap \gamma(K_1) = \emptyset$ unless $|c|, |d| < R$. This proves that there are only a finite number of possible bottom rows of $\gamma \in \Gamma(1)$ such that $K_2 \cap \gamma(K_1) \neq \emptyset$. We must therefore show that given c, d, there are only a finite number of possible γ with given bottom row (c, d) with $K_2 \cap \gamma(K_1) \neq \emptyset$. If γ_1 and γ_2 have the same bottom row, then $\gamma_2 = \gamma_0 \gamma_1$ where γ_0 has the form $\left(\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right)$. The effect of the matrix γ_0 is therefore translation by an integer distance n: $\gamma_0(z) = z + n$. For fixed γ_1, there can clearly be only finitely many γ_0 such that $\gamma_0\big(\gamma_1(K_1)\big) \cap K_2 \neq \emptyset$.

To summarize, there are only finitely many possible bottom rows (c, d) of γ such that $\gamma(K_1) \cap K_2 \neq \emptyset$, and for each (c, d) there are only finitely many possible γ with the prescribed bottom row such that $\gamma(K_1) \cap K_2 \neq \emptyset$. Hence the action of $\Gamma(1)$ is discontinuous. ■

As a second application of Eq. (2.2), we will determine a fundamental domain for Γ. If $\Gamma \subset SL(2, \mathbb{R})$ is a subgroup acting discontinuously on $\mathcal{H}$, a *fundamental domain* for Γ will be an open subset $F \subset \mathcal{H}$ such that (i) for every $z \in \mathcal{H}$, there exists $\gamma \in \Gamma$ such that $\gamma(z)$ is in the closure $\overline{F}$; and (ii) if $z_1, z_2 \in F$, and $\gamma(z_1) = z_2$ for some $\gamma \in \Gamma$, then $z_1 = z_2$, and $\gamma = \pm I$.

Let $F = \{z = x + iy \in \mathcal{H} | -\frac{1}{2} < x < \frac{1}{2},\ |z| > 1\}$.

Proposition 1.2.2 *The set F is a fundamental domain for $SL(2, \mathbb{Z})$.*

Proof Let $z \in \mathcal{H}$. Because $(c, d) \mapsto |cz + d|^2$ is a positive definite quadratic form, it has a minimum value as (c, d) runs through the pairs of relatively prime integers. It follows from Eq. (2.2) that $\text{im}\big(\gamma(z)\big)$ has a maximum with $\gamma \in SL(2, \mathbb{Z})$, so let $\gamma \in \Gamma(1)$ maximize $\text{im}\big(\gamma(z)\big)$. Now we can find $n \in \mathbb{Z}$ so that $\gamma(z) + n$ has a real part with absolute value $\leq 1/2$; replacing γ by $\left(\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right)\gamma$, we see that there exists γ such that $|\,\text{re}\,\big(\gamma(z)\big)| \leq 1/2$ and $\text{im}\big(\gamma(z)\big)$ is maximal for $\gamma \in SL(2, \mathbb{Z})$. This implies that $|\gamma(z)| \geq 1$, because otherwise the imaginary part of $\gamma_1(z)$ would be larger, where

$$\gamma_1 = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}\gamma, \qquad \text{im}\,\big(\gamma_1(z)\big) = \frac{\text{im}\ \gamma(z)}{|\gamma(z)|^2}.$$

This shows that every $\Gamma(1)$ orbit intersects the closure of F establishing property (i) in the definition of a fundamental domain.

Now suppose that $z = x + iy \in F$ and that $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma(1)$ such that $w = \gamma(z) \in F$. If $c = 0$, then $\gamma = \pm\left(\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right)$ for some $n \in \mathbb{Z}$, and $z, \gamma(z) \in F$ implies that $n = 0$, so $z = \gamma(z)$, in accordance with (ii) in the definition of a fundamental domain. Thus we may assume that $c \neq 0$. Observe that every element of F has imaginary part greater than $\sqrt{3}/2$. Also, clearly $|cz+d| \geq cy$.

We therefore have the inequalities

$$\frac{\sqrt{3}}{2} < \operatorname{im}(\gamma(z)) = \frac{y}{|cz+d|^2} \leq \frac{1}{c^2 y} < \frac{2}{c^2\sqrt{3}}.$$

Hence $c^2 < 4/3$, implying that $c = \pm 1$. Suppose that $c = \pm 1$. Because γ and $-\gamma$ have the same action on $\mathcal{H}$, we may assume without loss of generality that $c = 1$. Then $ad - bc = 1$ implies that

$$\gamma = \begin{pmatrix} a & b \\ 1 & d \end{pmatrix} = \begin{pmatrix} 1 & a \\ & 1 \end{pmatrix} \begin{pmatrix} & -1 \\ 1 & \end{pmatrix} \begin{pmatrix} 1 & d \\ & 1 \end{pmatrix}.$$

Now let $z_1 = z + d$ and $w_1 = w - a$. Because $|\operatorname{re}(z)| < 1/2$, we have $|z_1| \geq |z| > 1$, and similarly $|w_1| > 1$, yet $w_1 = \left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right)(z_1)$. This is a contradiction. This proves that F satisfies (ii) in the definition of a fundamental domain. ■

It is often convenient to have generators for the group $SL(2, \mathbb{Z})$. Let

$$T = \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}, \qquad S = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}. \tag{2.3}$$

Proposition 1.2.3 *$SL(2, \mathbb{Z})$ is generated by S and by T.*

The method of proof generalizes easily to give generators for other discontinuous groups. See Exercise 1.2.3(b) for an example.

Proof Let Γ be the subgroup of $\Gamma(1)$ generated by S and T. We will show that $\Gamma = \Gamma(1)$. Because $-I = S^2 \in \Gamma$, it is sufficient to show that the images $\overline{\Gamma} = \overline{\Gamma(1)}$ in $PSL(2, \mathbb{R})$. Let $\gamma \in \Gamma(1)$; we will describe a process by which γ may be reduced to a product of elements of the form S, T, and T^{-1}. We do not distinguish now between matrices A and $-A$ because these are equal in $PSL(2, \mathbb{R})$ and have the same effect on $\mathcal{H}$.

Because F is a fundamental domain for $\Gamma(1)$, $\mathcal{H}$ is the union of the closure $\overline{\gamma(F)}$ with $\gamma \in \Gamma(1)$, and these sets have disjoint interiors. We may therefore find a sequence $\gamma_1, \cdots, \gamma_n \in \Gamma(1)$ such that $\gamma_1(F) = F$ and $\gamma_n(F) = \gamma(F)$, and each $\gamma_k(F)$ is adjacent to $\gamma_{k+1}(F)$. Of course, this implies that $\gamma_1 = I$ and $\gamma_n = \gamma$. Observe that the domains γF that are adjacent to F are precisely $T(F)$, $T^{-1}(F)$, and $S(F)$ (cf. Figure 1).

Because $\gamma_k(F)$ is adjacent to $\gamma_{k+1}(F)$, we must have $\gamma_k^{-1}\gamma_{k+1}(F)$ adjacent to F, and so $\gamma_k^{-1}\gamma_{k+1}$ equals S, T, or T^{-1}. Thus $\gamma = \gamma_1^{-1}\gamma_n = \prod \gamma_k^{-1}\gamma_{k+1} \in \overline{\Gamma}$, as required. ■

What is the "boundary" of the Poincaré upper half plane? If we embed $\mathcal{H}$ in the Riemann sphere $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$, the topological boundary is $\mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\}$. The point ∞ should be regarded as no different from the

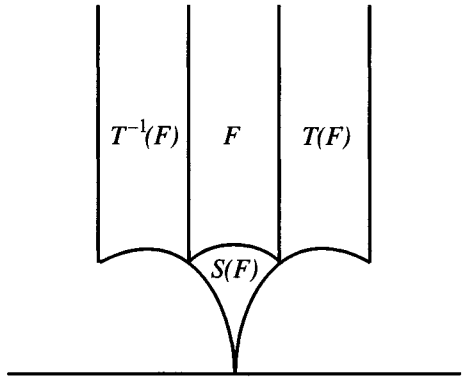


Figure 1

other boundary points. For example, the group $SL(2, \mathbb{R})$ acts transitively on $\mathbb{R} \cup \{\infty\}$. Another way of seeing this is to use the *Cayley transform* $\mathcal{C}$, which maps $\mathcal{H}$ to $\mathcal{D}$, the unit disk, defined by

$$\mathcal{C}(z) = \frac{z - i}{z + i}. \tag{2.4}$$

(Clearly $z \in \mathcal{H}$ if and only if z is closer to i than to $-i$, i.e., if and only if $\mathcal{C}(z) \in \mathcal{D}$.) Then $\mathcal{C}$ maps $\mathbb{R} \cup \{\infty\}$ onto the unit circle. This shows again that the points of $\mathbb{R} \cup \{\infty\}$ should be considered equivalent to each other. To give an example, the image under the Cayley transform of the fundamental domain F for $SL(2, \mathbb{Z})$ described above looks like that shown in Figure 2.

If Γ is a discontinuous group acting on $\mathcal{H}$, let $\Gamma\backslash\mathcal{H}$ be the quotient space consisting of the orbits of elements of $\mathcal{H}$ under the action of Γ. We topologize $\Gamma\backslash\mathcal{H}$ as a quotient: This means that a subset of $\Gamma\backslash\mathcal{H}$ is open if and only if its preimage in $\mathcal{H}$ under the canonical map $\mathcal{H} \to \Gamma\backslash\mathcal{H}$ is open.

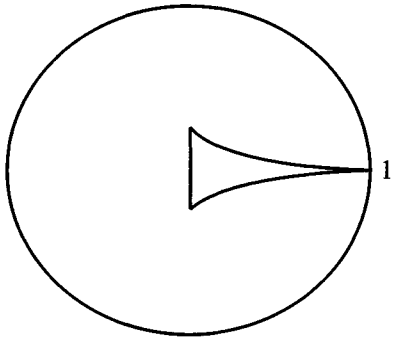


Figure 2

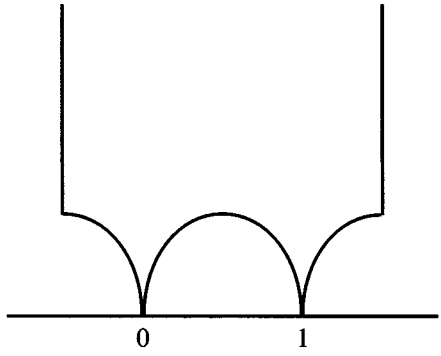


Figure 3

We now consider the *cusps* of a congruence group Γ. (We will extend this discussion shortly to the case of a discontinuous group Γ such that $\Gamma\backslash\mathcal{H}$ has finite volume with respect to the measure introduced in Exercise 1.2.6.) Intuitively, the cusps are the places where a fundamental domain for Γ touches the boundary of $\mathcal{H}$. For example, in Figure 2, we see that $SL(2, \mathbb{Z})$ has one cusp. On the other hand, $\Gamma(2)$ has three. Indeed, its fundamental domain looks like that shown in Figure 3 (Exercise 1.2.3).

Its image under the Cayley transform looks like the region shown in Figure 4. Evidently, there should be three cusps, if we can give the correct definition.

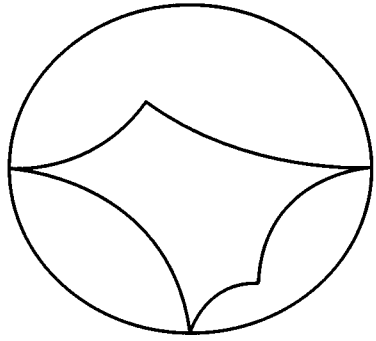


Figure 4

Let $\mathbb{P}^1(\mathbb{Q}) = \mathbb{Q} \cup \{\infty\}$ be the projective line over $\mathbb{Q}$; $SL(2, \mathbb{Z})$ acts transitively on $\mathbb{P}^1(\mathbb{Q})$, so a subgroup of finite index can have only finitely many orbits on this set. An orbit of Γ in $\mathbb{P}^1(\mathbb{Q})$ is called a *cusp* of Γ. In practical terms, these are the points where a fundamental domain for Γ must touch the boundary of $\mathcal{H}$.

More generally, if Γ is not assumed to be a congruence subgroup, but only a discontinuous group acting on $\mathcal{H}$ with $\Gamma\backslash\mathcal{H}$ having finite volume, the term *cusp* refers to either (i) a point of $a \in \mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\}$ such that Γ contains a

parabolic element $\gamma \neq I$ with $\gamma(a) = a$, or (ii) an orbit of such points under the action of Γ. (We recall from Exercise 1.2.7(c) that $\gamma \neq I$ is called *parabolic* if $|\operatorname{tr}(\gamma)| = 2$.) See Exercise 1.2.10 for the relationship of this definition to the one previously given for congruence subgroups.

We now show how $\Gamma\backslash\mathcal{H}$, for Γ a congruence group, may be compactified to give a compact Riemann surface by adjoining a finite number of points; indeed, by adjoining precisely one point for each cusp. Again, the discussion generalizes easily to a discontinuous group Γ where $\Gamma\backslash\mathcal{H}$ has finite volume. We start with the topological space $\mathcal{H}^* = \mathcal{H} \cup \mathbb{Q} \cup \{\infty\}$. (For a general discontinuous group Γ, we would take $\mathcal{H}^*$ to be the union of $\mathcal{H}$ and the cusps of Γ in $\mathbb{P}^1(\mathbb{R})$.) We topologize $\mathcal{H}^*$ as follows. The set $\mathcal{H}$ is to be an open set with its usual topology. We must describe the topology in the neighborhood of a point $a \in \mathbb{Q} \cup \{\infty\}$. If $a = \infty$, we take as a neighborhood base at a the sets of the form $\{\infty\} \cup \{z \mid \operatorname{im}(z) > C\}$ for $0 \leq C \in R$. On the other hand, if $a \in \mathbb{Q}$, we take as a neighborhood base at a the sets $\{a\} \cup U$, where U is the interior of a circle contained in $\mathcal{H}$, tangent to the real line at the point a. With $\mathcal{H}^*$ topologized in this way, we give $\Gamma\backslash\mathcal{H}^*$ the quotient topology. This is a manifold. We will now specify charts around each point, which will give it a complex structure.

Around "most" points $a \in \Gamma\backslash\mathcal{H}$, we may simply take a neighborhood of a in $\mathcal{H}$ to be a chart. Certain points must be treated carefully: These are the elliptic points. We call a point $a \in \mathcal{H}$ an *elliptic point* if there exists a nontrivial subgroup $\overline{\Gamma_a}$ of the image $\overline{\Gamma}$ of Γ in $SL(2, \mathbb{R})/\{\pm I\}$ that stabilizes a. Such a group is necessarily cyclic (Exercise 1.2.4). Its order is called the *order* of a.

For example, if $\Gamma = \Gamma(1)$, the Γ-orbits of elliptic points are represented by i, with $\overline{\Gamma_i} = \left\langle \left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right) \right\rangle$, and by $\rho = e^{2\pi i/3}$, with $\overline{\Gamma_\rho} = \left\langle \left(\begin{smallmatrix} 1 & -1 \\ 1 & \end{smallmatrix}\right) \right\rangle$. The orders of these points are 2 and 3, respectively. Most congruence subgroups do not have elliptic points.

Suppose that a is an elliptic point. How are we to construct a chart in a neighborhood of a? We use the modified Cayley transform

$$z \mapsto \frac{z - a}{z - \overline{a}}.$$

This maps $\mathcal{H}$ to the unit disk $\mathcal{D}$ but maps a to zero. Conjugation by this modified Cayley transform maps $\overline{\Gamma_a}$ to the group of rotations of the unit disk in angles that are a multiple of $2\pi/n$, where n is the order of a (Exercise 1.2.5(c)). If w is the coordinate function on $\mathcal{D}$, it is easy to see that $z \mapsto w^n$ maps a neighborhood of a in $\Gamma\backslash\mathcal{H}^*$ homeomorphically onto a neighborhood of the origin in $\mathbb{C}$, and we take this map to be a coordinate chart near a. This takes care of the remaining points in $\mathcal{H}$.

As for the cusps, if $a \in \mathbb{Q} \cup \{\infty\}$, let $\rho \in SL(2, \mathbb{Z})$ such that $\rho(a) = \infty$. Let $\overline{\Gamma_a}$ be the stabilizer of a in $\overline{\Gamma}$. Now $\rho\Gamma\rho^{-1}$ is a subgroup of finite index in $\overline{\Gamma(1)}$, and the stabilizer of ∞ in this group is $\rho\overline{\Gamma_a}\rho^{-1}$. Hence this is a subgroup of finite index in the stabilizer of infinity in $\Gamma(1)$. The image of this stabilizer in

$PSL(2,\mathbb{R})$ is the infinite cyclic group generated by $z \mapsto z+1$. Thus $\overline{\rho\Gamma_a\rho^{-1}}$ is an infinite cyclic group generated by $z \mapsto z+n$ for some n. It is not hard to see then that $z \mapsto e^{2\pi i\rho(z)/n}$ maps a neighborhood of a in $\Gamma\backslash\mathcal{H}^*$ homeomorphically onto a neighborhood of the origin in $\mathbb{C}$, and we take this map to be a coordinate chart near a.

We have specified a coordinate chart near each point of $\Gamma\backslash\mathcal{H}^*$, which thus becomes a compact Riemann surface.

Exercises

Exercise 1.2.1 Prove Eq. (2.2).

Exercise 1.2.2 Let Γ be a discontinuous subgroup of $SL(2,\mathbb{R})$, and let Γ' be a subgroup. Let F be a fundamental domain for Γ, and let $\gamma_1, \cdots, \gamma_n$ be a set of coset representatives for $\Gamma'\backslash\Gamma$; that is, $\Gamma = \bigcup \Gamma'\gamma_i$ disjointly. Prove that $\bigcup \gamma_i(F)$ is a fundamental domain for Γ'.

Exercise 1.2.3 (a) Prove that a fundamental domain for $\Gamma(2)$ consists of $x+iy$ such that $-1/2 < x < 3/2$, $|z+1/2| > 1/2$, $|z-1/2| > 1/2$ and $|z-3/2| > 1/2$ (cf. Figure 3). [HINT: construct first a fundamental domain by means of Exercise 2.2, then modify it to obtain the domain in question.]
(b) Use the method of Proposition 2.3 to prove that $\Gamma(2)$ is generated by $\left(\begin{smallmatrix}1&2\\1&\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}1&\\2&1\end{smallmatrix}\right)$.

Exercise 1.2.4 Prove that the stabilizer of an elliptic point is cyclic.

Exercise 1.2.5 (a) Let $SL(2,\mathbb{C})$ act on $\mathbb{P}^1(\mathbb{C})$ by linear fractional transformations as in Eq. (2.1). Prove that the subgroup that maps the unit disk $\mathcal{D}$ onto itself is

$$SU(1,1) = \left\{ \begin{pmatrix} a & b \\ \bar{b} & \bar{a} \end{pmatrix} \;\middle|\; |a|^2 - |b|^2 = 1 \right\}.$$

(b) Prove that the group $SU(1,1)$ is conjugate to $SL(2,\mathbb{R})$ in $SL(2,\mathbb{C})$. [HINT: use the Cayley transform.]
(c) Prove that the subgroup of $SU(1,1)$ fixing $0 \in \mathcal{D}$ is the group of rotations

$$\begin{pmatrix} e^{i\theta/2} & \\ & e^{-i\theta/2} \end{pmatrix}.$$

Exercise 1.2.6 (a) *Bruhat decomposition:* Prove that if B is the *Borel subgroup* of $SL(2,\mathbb{R})$ consisting of upper triangular matrices and $S = \left(\begin{smallmatrix}&-1\\1&\end{smallmatrix}\right)$, then

$$SL(2,\mathbb{R}) = B \cup BSB,$$

and the union is disjoint. Thus $SL(2,\mathbb{R})$ is generated by matrices of the

following types:

$$\begin{pmatrix} a & \\ & a^{-1} \end{pmatrix}, \quad \begin{pmatrix} 1 & b \\ & 1 \end{pmatrix}, \quad \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}.$$

(Making use of the identity (IV.1.16) below, it is possible to dispense with the diagonal matrices here.)
(b) Show that the measure $|y|^{-2}\,dx\,dy$ is invariant under the action of $SL(2,\mathbb{R})$ by checking that it is invariant under generators in part (a).
(c) Show that the volume of $\Gamma(1)\backslash\mathcal{H}$ is finite with respect to this invariant measure.

Exercise 1.2.7 Let $\pm I \neq \gamma \in SL(2,\mathbb{R})$ acting on the Riemann sphere $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$.
(a) If $|\operatorname{tr}(\gamma)| < 2$, show that γ has two fixed points in $\mathbb{P}^1(\mathbb{C})$: one in $\mathcal{H}$ and its complex conjugate. Such an element is called *elliptic*.
(b) If $|\operatorname{tr}(\gamma)| > 2$, show that γ has two fixed points in $\mathbb{P}^1(\mathbb{R})$ and no other fixed points in $\mathbb{P}^1(\mathbb{C})$. Such an element is called *hyperbolic*.
(c) If $|\operatorname{tr}(\gamma)| = 2$, show that γ has a single fixed point in $\mathbb{P}^1(\mathbb{R})$ and no other fixed points in $\mathbb{P}^1(\mathbb{C})$. Such an element is called *parabolic*. If $\operatorname{tr}(\gamma) = 2$, then both eigenvalues of γ are one, in which case the matrix γ is called *unipotent*. If γ is parabolic, then either γ or $-\gamma$ is unipotent.

Exercise 1.2.8 (a) Let $\Gamma \subset SL(2,\mathbb{R})$ be a discontinuous group, and let $\pm I \neq \gamma \in \Gamma$. Show that γ is elliptic if and only if it has finite order. In this case, we call the fixed point of γ in $\mathcal{H}$ an *elliptic fixed point* for Γ.
(b) Show that there are only two orbits of elliptic fixed points for $SL(2,\mathbb{Z})$, represented by i and $e^{2\pi i/3}$, respectively. [Prove this by examining the fundamental domain.]

Exercise 1.2.9 Let $\Gamma \subset SL(2,\mathbb{R})$ be a discontinuous group such that the quotient $\Gamma\backslash\mathcal{H}$ has finite volume (cf. Exercise 2.6(c)). Show that $\Gamma\backslash\mathcal{H}$ is compact if and only if Γ contains no parabolic elements.

Exercise 1.2.10 Let Γ be a congruence subgroup of $SL(2,\mathbb{Z})$. Prove that if $a \in \mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\}$, then there exists a parabolic element $\gamma \in \Gamma$ such that $\gamma(a) = a$ if and only if $a \in \mathbb{P}^1(\mathbb{Q}) = \mathbb{Q} \cup \{\infty\}$.

Exercise 1.2.11 Let $\gamma \in \Gamma(1)$ be a hyperbolic element. Show that there exists a real quadratic field $K = \mathbb{Q}(\sqrt{D})$ with $D > 0$ such that the fixed points and eigenvalues of γ lie in K. Show that the eigenvalues of γ are a conjugate pair of units of norm one in K. *Make the following assumption about K: assume that the ring generated by the units of norm one in K is the full ring of integers.* This may or may not be true. Let ϵ, ϵ' be the eigenvalues of γ. If $\mathfrak{a}$ is a fractional ideal of K, then $\mathfrak{a}$ is a free $\mathbb{Z}$-module of rank 2; let $\{a_1, a_2\}$ be a basis. Then there exists an element $\gamma \in SL(2,\mathbb{Z})$ such that $\epsilon(a_1, a_2) = (a_2, a_2)\gamma$. Show that the $GL(2,\mathbb{Z})$-conjugacy class of γ depends only on the ideal class of $\mathfrak{a}$, and that the $GL(2,\mathbb{Z})$-conjugacy classes of hyperbolic elements with eigenvalues

ϵ and ϵ' are thus in bijection with the ideal classes of K. For a fuller discussion of the hyperbolic conjugacy classes in $SL(2, \mathbb{Z})$, see the references in Volume I of Terras (1985, p. 273).

1.3 Modular Forms for $SL(2, \mathbb{Z})$

Modular forms are certain holomorphic functions on $\mathcal{H}$ that have in common with Dirichlet characters the remarkable property of being associated with Euler products having functional equations. We will consider first the case of a modular form for $\Gamma(1) = SL(2, \mathbb{Z})$.

Let k be an even nonnegative integer. A *modular form of weight* k for $SL(2, \mathbb{Z})$ is a holomorphic function f on $\mathcal{H}$, which satisfies the identity

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z) \qquad \text{for } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(1) \tag{3.1}$$

and which is *holomorphic at the cusp* ∞. The latter condition requires some discussion. Recall from the end of the previous section that we may choose the quantity $q = e^{2\pi i z}$ as a coordinate function near ∞ in $\Gamma(1)\backslash\mathcal{H}^*$. Because $\left(\begin{smallmatrix} 1 & 1 \\ 1 \end{smallmatrix}\right) \in \Gamma(1)$, Eq. (3.1) implies that $f(z+1) = f(z)$, and thus any function satisfying Eq. (3.1) has a Fourier expansion

$$f(z) = \sum_{n=-\infty}^{\infty} a_n e^{2\pi i n z} = \sum_{n=-\infty}^{\infty} a_n q^n. \tag{3.2}$$

If for some sufficiently large N the coefficients a_n are zero for $n < -N$, we say the function f is *meromorphic* at ∞. If $a_n = 0$ for $n < 0$, we say that f is *holomorphic* at ∞. If f is holomorphic at ∞ and furthermore $a_0 = 0$, we say f *vanishes* or is *cuspidal at* ∞.

A modular form for $SL(2, \mathbb{Z})$ that vanishes at ∞ is called a *cusp form*. We denote the space of modular forms of weight k for $\Gamma(1) = SL(2, \mathbb{Z})$ as $M_k\big(\Gamma(1)\big)$, and the space of cusp forms as $S_k\big(\Gamma(1)\big)$. Our first objective is to prove that these spaces are finite dimensional.

There is a related notion of an *automorphic function*. We call f an *automorphic function* for Γ if

$$f\left(\frac{az+b}{cz+d}\right) = f(z)$$

and f is meromorphic on $\mathcal{H}$ and at ∞. Hence f may be regarded as a meromorphic function on the compact Riemann surface $\Gamma(1)\backslash\mathcal{H}^*$. Note that an automorphic function is allowed to have poles, while a modular form is not. It is a consequence of the maximum modulus principle that an automorphic function with no poles is constant. (An automorphic function with no poles is the same as a modular form of weight zero, so we may equally well state that a modular form of weight zero is constant.) This simple fact, together with the

observation that if f_1, $f_2 \in M_k(\Gamma(1))$ then f_1/f_2 is an automorphic function, is quite a powerful tool that we will use systematically in determining the spaces $M_k(\Gamma(1))$. A first application of this principle is Proposition 1.3.2 below.

Proposition 1.3.1 *Let X be a compact Riemann surface, $P_1, \cdots, P_n \in X$, and let $r_1, \cdots, r_n$ be positive integers. Let V be the vector space of meromorphic functions on X, which are holomorphic except possibly at the points P_m, and which are holomorphic or else have poles of order at most r_m at P_m. Then the space V has dimension at most $r_1 + \ldots + r_m + 1$.*

More precise information about the dimension of this space is contained in the Riemann–Roch theorem.

Proof We will denote $r = r_1 + \ldots + r_m$. Let us choose a coordinate function $t = t_j$ in a neighborhood of P_j with respect to which P_j is the origin. If $\phi \in V$, it has a Laurent expansion:

$$\phi(t) = a_{j,-r_j} t^{-r_j} + a_{j,-r_j+1} t^{-r_j+1} + \ldots$$

We associate with ϕ the vector $A(\phi) \in \mathbb{C}^r$ whose coordinates are the r Taylor coefficients $a_{j,-h}$, $1 \leq h \leq r_j$. If $\phi_1, \cdots, \phi_N \in V$, and if $N > r$, we may find coefficients $c_1, \cdots, c_N$, not all zero, such that $\sum c_j A(\phi_j) = 0$. This means that $\sum c_j \phi_j$ has no poles. It is a consequence of the maximum modulus principle that any meromorphic function on a compact Riemann surface having no poles is automatically constant. Hence any vector subspace of V having dimension greater than r contains a nonzero constant function. This implies that $\dim V \leq r + 1$. ∎

Proposition 1.3.2 *The space $M_k(\Gamma(1))$ is finite dimensional.*

Proof Let f_0 be a nonzero element of $M_k(\Gamma(1))$. Let X be the compactification of $\Gamma(1)\backslash\mathcal{H}$ described in the previous section. Let $P_1, \cdots, P_m$ be the zeros of f_0, and let $r_1, \cdots, r_m$ be the orders of vanishing of f_0 at these points. (Actually, we must count the order of vanishing at an elliptic fixed point carefully. The order of vanishing of a function on $\mathcal{H}$ at an elliptic point a whose stabilizer $\overline{\Gamma}_a$ has order e will be e times the order of vanishing of the corresponding function on X.) If $f \in M_k(\Gamma(1))$, then f/f_0 is an automorphic function, and indeed, $f \mapsto f/f_0$ is an isomorphism of $M_k(\Gamma(1))$ with the vector space V in Proposition 3.1. Thus $M_k(\Gamma(1))$ has dimension at most $r + 1$. ∎

We would like to know, on the other hand, that modular forms do exist. A convenient construction is by means of *Eisenstein series*. Let us assume that k is an even integer ≥ 4. Define

$$E_k(z) = \tfrac{1}{2} \sum_{\substack{m, n \in \mathbb{Z} \\ (m, n) \neq (0, 0)}} (mz + n)^{-k}. \tag{3.3}$$

The series is absolutely convergent (Exercise 1.3.1). Let us show that $E_k(z)$ is a modular form of weight k. We have

$$E_k\left(\frac{az+b}{cz+d}\right) = (cz+d)^k \tfrac{1}{2} \sum_{\substack{m,n \in \mathbb{Z} \\ (m,n) \neq (0,0)}} (m(az+b)+n(cz+d))^{-k}$$

$$= (cz+d)^k \tfrac{1}{2} \sum_{\substack{m,n \in \mathbb{Z} \\ (m,n) \neq (0,0)}} ((ma+nc)z+(mb+nd))^{-k}.$$

Observe that because c and d are coprime, $(m,n) \mapsto (ma+nc, mb+nd)$ permutes the nonzero elements of $\mathbb{Z} \times \mathbb{Z}$ amongst themselves, so we see that E_k satisfies Eq. (3.1). To show that it is analytic at ∞, let us compute its Fourier expansion. Firstly, the sum of the terms with $m = 0$ is clearly just $\zeta(k)$. For the terms with $m \neq 0$, because k is even, the terms ± 1 contribute equally, and we may consider only $m > 0$. By Exercise 1.1.4, these contribute

$$\frac{(2\pi i)^k}{(k-1)!} \sum_{n=1}^{\infty} n^{k-1} e^{2\pi i n m z}.$$

If r is a complex number, let us define the *divisor sum*

$$\sigma_r(n) = \sum_{d|n} d^r.$$

We see that the Fourier expansion of E_k has the form

$$E_k(z) = \zeta(k) + \frac{(2\pi)^k (-1)^{k/2}}{(k-1)!} \sum_{n=1}^{\infty} \sigma_{k-1}(n)\, q^n, \qquad q = e^{2\pi i z}. \tag{3.4}$$

Note that by Exercise 1.1.8, the Fourier coefficients of $G_k(z) = \zeta(k)^{-1} E_k(z)$ are rational numbers.

For given k, either $S_k\big(\Gamma(1)\big) = M_k\big(\Gamma(1)\big)$ or else $\dim M_k\big(\Gamma(1)\big) = \dim S_k\big(\Gamma(1)\big) + 1$, because if there exists a modular form of weight k with nonvanishing constant Fourier coefficient, we may subtract a suitable multiple of that from any given modular form to obtain a cusp form. Because there exist Eisenstein series with nonvanishing constant coefficient for $k \geq 4$, we see that

$$\dim M_k\big(\Gamma(1)\big) = \dim S_k\big(\Gamma(1)\big) + 1 \qquad \text{for } k \geq 4. \tag{3.5}$$

Although we have now constructed some modular forms, we still have not constructed *cusp forms*. This may be accomplished as follows. The modular

forms form a graded ring, because if $f \in M_k(\Gamma(1))$, $g \in M_l(\Gamma(1))$, then $fg \in M_{k+l}(\Gamma(1))$. One may thus construct a large number of modular forms by ring operations. For example, we see that

$$G_4(z) = 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) q^n, \qquad G_6(z) = 1 - 504 \sum_{n=1}^{\infty} \sigma_5(n) q^n.$$

From these we may compute the Fourier coefficients of $G_4^3 - G_6^2$, a modular form of weight 12. We find that

$$\tfrac{1}{1728}(G_4^3 - G_6^2) = q - 24q^2 + 252q^3 - 1472q^4 + 4830q^5 + \dots .$$

We have constructed a nontrivial cusp form of weight 12. This modular form is denoted $\Delta(z)$.

We will obtain another formula for $\Delta(z)$, due to Ramanujan (1916). It will be useful to have at our disposal a famous formula from the theory of elliptic functions, *Jacobi's triple product formula*:

$$\sum_{n=-\infty}^{\infty} q^{n^2} x^n = \prod_{n=1}^{\infty} (1 - q^{2n})(1 + q^{2n-1}x)(1 + q^{2n-1}x^{-1}), \tag{3.6}$$

valid if $0 < |q| < 1$ and $x \neq 0$. A proof of this is sketched in Exercise 1.3.2.

Now in Eq. (3.6), substitute $q^{3/2}$ for q and $-q^{-1/2}$ for x. (Because $q = e^{2\pi i z}$, a fractional power q^r is naturally interpreted as $e^{2\pi i r z}$.) We see that

$$\sum_{n=-\infty}^{\infty} (-1)^n q^{(3n^2+n)/2} = \prod_{n=1}^{\infty} (1 - q^{3n})(1 - q^{3n-1})(1 - q^{3n-2}) = \prod_{n=1}^{\infty} (1 - q^n).$$

Now completing the square in this identity, we see that

$$\sum_{n=-\infty}^{\infty} (-1)^n q^{(6n+1)^2/24} = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n). \tag{3.7}$$

The function $\eta(z) = q^{1/24} \prod (1 - q^n)$ is known as the *Dedekind eta function*. Grouping the terms with n positive and negative together, we may rewrite this formula:

$$\eta(z) = \sum_{n=1}^{\infty} \chi(n)\, q^{n^2/24}, \tag{3.8}$$

where χ is the primitive quadratic character with conductor 12 (Exercise 1.1.6). We have

$$\chi(n) = \begin{cases} 1 & \text{if } n \equiv \pm 1 \bmod 12; \\ -1 & \text{if } n \equiv \pm 5 \bmod 12; \\ 0 & \text{otherwise.} \end{cases}$$

Now we show that if $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma(1)$, there exists a 24th root of unity $\epsilon(\gamma)$ such that

$$\eta\left(\frac{az+b}{cz+d}\right) = \epsilon(\gamma)\,(cz+d)^{1/2}\,\eta(z). \tag{3.9}$$

(There is an ambiguity in sign in the choice of square root $(cz+d)^{1/2}$, but because we are only asserting that $\epsilon(\gamma)$ lies in the group of 24th roots of unity, this is not a problem.)

From Proposition 1.2.3, it is sufficient to check this when $\gamma = T = \left(\begin{smallmatrix} 1 & 1 \\ & 1 \end{smallmatrix}\right)$ and $\gamma = S = \left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right)$. The transformation property for $\gamma = T$ is clear from the right-hand side of Eq. (3.7), because under this transformation $q^{1/24} \mapsto e^{2\pi i/24} q^{1/24}$. On the other hand, if $\gamma = S$, Eq. (3.8) gives $\eta(z) = \theta_\chi(-iz/12)$ with θ_χ as in Eq. (1.14). In Eq. (1.15), we have $\tau(\chi) = 2\sqrt{3}$ and $N = 12$, so

$$\sqrt{-iz}\,\eta(z) = \eta\left(-\frac{1}{z}\right), \tag{3.10}$$

so we have Eq. (3.9) in this case also. This completes the proof of Eq. (3.9) for all $\gamma \in \Gamma(1)$.

Next, we raise Eq. (3.9) to the 24th power to get rid $\epsilon(\gamma)$. We see that if

$$\Delta(z) = \eta(z)^{24} = q\prod_{n=1}^{\infty}(1-q^n)^{24},$$

then Δ is a cusp form of weight 12. Observe that Δ is defined by a convergent infinite product, each of whose factors has no zero on $\mathcal{H}$. Consequently,

$$\Delta(z) \neq 0 \qquad \text{for } z \in \mathcal{H}.$$

Proposition 1.3.3 *The space $S_{12}\big(\Gamma(1)\big)$ is one dimensional and spanned by Δ. In particular*

$$\Delta = \tfrac{1}{1728}(G_4^3 - G_6^2). \tag{3.11}$$

Proof If $f \in S_{12}\big(\Gamma(1)\big)$, then f/Δ is an automorphic function. It clearly has no poles in $\mathcal{H}$. It is also holomorphic at the cusp because Δ has only a first-order zero there, while f also vanishes. Because an automorphic function without poles is constant, f/Δ is constant. In particular, $\frac{1}{1728}(G_4^3 - G_6^2) = c\Delta$ for some c; examining the Fourier coefficients, we see that $c = 1$. ■

In general, one may give formulas for the dimensions of spaces of modular forms starting with either the Riemann–Roch theorem or the Selberg trace formula. For the group $SL(2, \mathbb{Z})$, however, we will obtain complete information using *ad hoc* tools.

Proposition 1.3.4 *Suppose that k is an even nonnegative integer. Let $k = 12j + r$ where $0 \leq r \leq 10$. Then*

$$\dim M_{12j+r}\big(\Gamma(1)\big) = \begin{cases} j+1 & \text{if } r = 0, 4, 6, 8 \text{ or } 10; \\ j & \text{otherwise.} \end{cases} \tag{3.12}$$

The ring $\bigoplus_{k=0}^{\infty} M_k\big(\Gamma(1)\big)$ of modular forms is generated by G_4 and G_6.

Proof First let us show that $M_k\big(\Gamma(1)\big)$ is one dimensional and generated by E_k if $k = 4, 6, 8$, or 10. Let $h = 6(12-k)$. If $f \in M_k\big(\Gamma(1)\big)$ is not in the one-dimensional space spanned by E_k, we may subtract a multiple of E_k to cancel the constant Fourier coefficient, and so we may assume that f is a nonzero element of $S_k\big(\Gamma(1)\big)$. We consider $E_h(f/\Delta)^6$. This is an automorphic function with no poles, and hence is constant. Therefore $E_h = c\Delta^6/f^6$ for some c. We see that E_h can have no zeros on $\mathcal{H}$. Now $h = 12H$ where $H = 1, 2, 3$, or 4. We consider Δ^H/E_h. This is a nonzero automorphic function with no poles but with a zero of order H at ∞, which is a contradiction. This shows that each of the spaces $M_k\big(\Gamma(1)\big)$ is one dimensional, spanned by E_k if $k = 4, 6, 8$, or 10.

Now let us show that $M_2\big(\Gamma(1)\big)$ is zero. Suppose that f is a nonzero element of this space. Then $fE_4 \in M_6\big(\Gamma(1)\big)$, so $fE_4 = cE_6$ for some nonzero constant c. Because $E_4(\rho) = 0$ with $\rho = e^{2\pi i/3}$ (Exercise 1.3.3), we see that $E_6(\rho) = 0$. Now Eq. (3.11) implies that $\Delta(\rho) = 0$, a contradiction. Hence $M_2\big(\Gamma(1)\big) = 0$. Of course $M_0\big(\Gamma(1)\big)$ is one dimensional, comprising the constant functions. Thus Eq. (3.12) is proved if $k < 12$.

If $k \geq 12$, we show that multiplication by Δ is an isomorphism of $M_{k-12}(\Gamma(1))$ with $S_k(\Gamma(1))$. Indeed, it is an injection of $M_{k-12}\big(\Gamma(1)\big)$ into $S_k\big(\Gamma(1)\big)$, and if $f \in S_k\big(\Gamma(1)\big)$, then f/Δ has no poles, and hence lies in $M_{k-12}\big(\Gamma(1)\big)$. Formula (3.12) now follows from Eq. (3.5).

As for the fact that G_4 and G_6 generate the ring of modular forms, let R be the subring generated by these. It follows from the one-dimensionality of M_8 and M_{10} that E_8 and E_{10} are constant multiples of E_4^2 and E_4E_6; M_k lies in R for $k \leq 10$. Also, Eq. (3.11) implies that $\Delta \in R$. Let k be the first even positive integer such that $M_k\big(\Gamma(1)\big)$ is not contained in R; we see that $k \geq 12$. Now $R \supset \Delta M_{k-12}\big(\Gamma(1)\big) = S_k\big(\Gamma(1)\big)$. Moreover, R contains a noncuspidal modular form in $M_k\big(\Gamma(1)\big)$, namely, $E_4^r E_6^s$, where r and s are chosen so that $4r + 6s = k$. Hence by Eq. (3.5), R contains M_k. ∎

There exists a natural inner product on $S_k\big(\Gamma(1)\big)$, known as the *Petersson inner product*. If $f(z)$, $g(z) \in S_k\big(\Gamma(1)\big)$, then it is a consequence of Eq. (2.2) that

$$f(z)\,\overline{g(z)}\,y^k$$

is invariant under $z \mapsto \frac{az+b}{cz+d}$. Hence by Exercise 1.2.6(b), the integral

$$\langle f, g \rangle = \int_{\Gamma(1)\backslash\mathcal{H}} f(z)\,\overline{g(z)}\,y^k\,\frac{dx\,dy}{y^2} \tag{3.13}$$

is well defined. Because if $n > 0$, $q^n \to 0$ very rapidly as $z \to \infty$, and because a cusp form has a Fourier expansion $\sum a_n q^n$ with $a_n \neq 0$ only for $n > 0$, a cusp form $f(z)$ decays very rapidly as $y \to \infty$. Hence the integrand in Eq. (3.13) is very small near the cusp, and the integral is very rapidly convergent. Evidently, $\langle f, g \rangle$ is a positive definite Hermitian inner product.

Let $f(z) = \sum_{n=0}^{\infty} a_n q^n$ be an element of $M_k(\Gamma(1))$. Let

$$L(s, f) = \sum_{n=1}^{\infty} a_n n^{-s}.$$

This is known as the *L-function* of f. We need to know that this series is convergent for s sufficiently large. For this, the following estimate is sufficient.

Proposition 1.3.5 *If f is cuspidal, its Fourier coefficients satisfy $a_n \leq C\,n^{k/2}$ for some constant C independent of n.*

This estimate, called the *trivial estimate*, is due to Hardy (1927) and (more simply) Hecke (1937). The correct estimate $a_n \leq C\,n^{(k-1)/2+\epsilon}$ for any $\epsilon > 0$, was conjectured (for $f = \Delta$) by Ramanujan (1916); this famous statement, the *Ramanujan conjecture*, was finally proved around 1970 by Deligne (1971) using difficult techniques from algebraic geometry. See Section 3.5 for further discussion of this conjecture.

Proof It follows from Eqs. (2.2) and (3.1) that $|f(z)\,y^{k/2}|$ is $\Gamma(1)$ invariant. Because f is cuspidal, this function decays rapidly as z approaches the cusp, and so it is bounded on the fundamental domain; consequently, there exists a constant C_1 such that $|f(z)\,y^{k/2}| < C_1$ for all $z \in \mathcal{H}$. Now for fixed y,

$$|a_n|\,e^{-2\pi n y} = \left|\int_0^1 f(x+iy)\,e^{-2\pi i n x}\,dx\right| \leq \int_0^1 |f(x+iy)|\,dx < C_1\,y^{-k/2}.$$

This estimate is independent of n. We choose $y = 1/n$ and obtain

$$a_n < e^{2\pi}\,C_1\,n^{k/2}$$

as required. ■

If f is not a cusp form, this estimate is no longer valid. If f is the Eisenstein series E_k, the nth Fourier coefficient of f is $\sigma_{k-1}(n)$, which is bounded by a

constant times $\log(n)\, n^{k-1}$. In any case, the L-series $L(s, f)$ is convergent for $\mathrm{re}(s)$ sufficiently large.

Proposition 1.3.6 *The L-function $L(s, f)$ has meromorphic continuation to all s and satisfies a functional equation. In fact, if*

$$\Lambda(s, f) = (2\pi)^{-s}\, \Gamma(s)\, L(s, f) \tag{3.14}$$

then $\Lambda(s, f)$ extends to an analytic function of s if f is a cusp form; if f is not cuspidal, then it has simple poles at $s = 0$ and $s = k$. It satisfies

$$\Lambda(s, f) = (-1)^{k/2}\, \Lambda(k - s, f). \tag{3.15}$$

Proof In this proof we will assume that f is a cusp form, leaving the remaining case to the reader (Exercise 1.3.5). Because f is cuspidal, $f(iy) \to 0$ very rapidly as $y \to \infty$. When $\gamma = S$, Eq. (3.1) implies that

$$f(iy) = (-1)^{k/2}\, y^{-k}\, f(i/y), \tag{3.16}$$

so $f(iy) \to 0$ very rapidly as $y \to 0$ also. Hence the integral

$$\int_0^\infty f(iy)\, y^s\, \frac{dy}{y} \tag{3.17}$$

is convergent for all s and clearly defines an analytic function of s. If $\mathrm{re}(s)$ is large, we may substitute the Fourier expansion for f. Noting that

$$\int_0^\infty e^{-2\pi n y}\, y^s\, \frac{dy}{y} = (2\pi)^{-s}\, \Gamma(s)\, n^{-s},$$

we see that Eq. (3.17) equals $\Lambda(s, f)$. Now substituting Eq. (3.16) into Eq. (3.17) and substituting $1/y$ for y, we see that Eq. (3.17) equals

$$(-1)^{k/2} \int_0^\infty f(iy)\, y^{k-s}\, \frac{dy}{y},$$

from which we get Eq. (3.15). ■

The first historical hint that a Euler product should be associated with the L-series of a modular form came from Ramanujan's investigation of Δ. The Fourier coefficients of Δ comprise Ramanujan's tau function: $\Delta(z) = \sum \tau(n)\, q^n$. Ramanujan (1916) conjectured, and Mordell (1917) proved shortly afterward, that

$$\sum_{n=1}^{\infty} \tau(n)\, n^{-s} = \prod_p \left(1 - \tau(p)\, p^{-s} + p^{11-2s}\right)^{-1}. \tag{3.18}$$

The true explanation of this identity requires the theory of *Hecke operators*, which is our next topic. We will give the proof of Eq. (3.18) at the end of Section 4.

Exercises

Exercise 1.3.1 Verify that the Eisenstein series Eq. (3.3) is absolutely convergent if $k \geq 4$.

Exercise 1.3.2 This exercise outlines a proof of Jacobi's triple product formula (3.7). Let z and w be complex parameters such that $z \in \mathcal{H}$. Let $\Lambda \subset \mathbb{C}$ be the lattice $\{2mz + n | m, n \in \mathbb{Z}\}$. We will also let $q = e^{2\pi i z}$ and $x = e^{2\pi i w}$.
(a) An *elliptic function* with respect to the lattice Λ, is meant a meromorphic function f such that $f(u + \lambda) = f(u)$ for $\lambda \in \Lambda$. Use the maximum modulus principle to show that if f is an elliptic function that has no poles, then f is constant.
(b) Define

$$\vartheta(z, w) = \sum_{n=-\infty}^{\infty} q^{n^2} x^n,$$

and let

$$P(z, w) = \prod_{n=1}^{\infty} (1 + q^{2n-1} x)\,(1 + q^{2n-1} x^{-1}).$$

Prove that

$$\vartheta(z, w + 2z) = (qx)^{-1}\, \vartheta(z, w)$$

and that

$$P(z, w + 2z) = (qx)^{-1}\, P(z, w).$$

Hence for fixed z, $f(w) = \vartheta(z, w)/P(z, w)$ is an elliptic function.
(c) Prove for fixed z that if $P(z, w) = 0$ then either $w = \frac{1}{2} + z + \lambda$ or else $w = \frac{1}{2} - z + \lambda$ for some $\lambda \in \Lambda$. Show that these values of w are also zeros of $\vartheta(z, w)$, and conclude that $f(w)$ has no poles, and hence by (a) is constant. This shows that

$$\vartheta(z, w) = \phi(q)\, P(z, w)$$

where $\phi(q)$ is independent of w.
(d) The Jacobi triple-sum formula will follow if we know that

$$\phi(q) = \prod_{n=1}^{\infty} (1 - q^{2n}).$$

To this end, show that

$$\vartheta(4z, 1/2) = \vartheta(z, 1/4),$$

whereas

$$P(4z, 1/2)/P(z, 1/4) = \prod_{n=1}^{\infty}(1 - q^{4n-2})(1 - q^{8n-4}).$$

Then

$$\phi(q) = \frac{P(4z, 1/2)}{P(z, 1/4)}\,\phi(q^4).$$

Now show that $\phi(q) \to 1$ as $q \to 0$, and thus evaluate $\phi(q)$.

Exercise 1.3.3 Show that if $\rho = e^{2\pi i/3} \in \mathcal{H}$, and if $3 \nmid k$, then $f(\rho) = 0$ for any modular form of weight k. [HINT: Observe that $\gamma(\rho) = \rho$ where $\gamma = \left(\begin{smallmatrix} 1 & 1 \\ -1 & \end{smallmatrix}\right)$, and apply Eq. (3.1).]

Exercise 1.3.4 Show that G_4 and G_6 are algebraically independent.

Exercise 1.3.5 Prove Proposition 1.3.6 in the case where f is not necessarily cuspidal.

Exercise 1.3.6 Show that the inner product Eq. (3.13) is defined if only one of f and g is cuspidal and the other is an arbitrary modular form. Prove that the Eisenstein series E_k is orthogonal to the cusp forms (cf. Exercise 1.6.4).

Exercise 1.3.7 (a) Let M be a compact Riemann surface, and let $f : M \to \mathbb{C}$ be a meromorphic function. Assume that f has only one pole, at $m \in M$, which is simple. Extend f to a mapping $M \to \mathbb{P}^1(\mathbb{C})$ by $f(m) = \infty$. Prove that f is an isomorphism of Riemann surfaces.
(b) Define a function $j : \mathcal{H} \to \mathbb{C}$ by $j(z) = G_4^3/\Delta$. Show that j is an automorphic function for $SL(2, \mathbb{Z})$ with a Fourier expansion

$$j(z) = \frac{1}{q} + 744 + 196884\,q + 21493760\,q^2 + 864299970\,q^3 + \dots .$$

Prove that $j(i) = 1728$ and $j(e^{2\pi i/3}) = 0$. Use part (a) to conclude that j is a bijection of the compactified space

$$SL(2, \mathbb{Z})\backslash\mathcal{H} \cup \{\infty\} \cong \mathbb{P}^1(\mathbb{C}).$$

We now recall some elementary facts from the topology of surfaces and the theory of compact Riemann surfaces, particularly the notions of genus and ramification. For further information, see Siegel (1969, 1971, and 1973), Lang (1982), and Gunning (1966).

If X is a (connected) compact Riemann surface, then as a topological space, X is a compact orientable surface, which is homeomorphic to a sphere with g handles attached, where the *genus* g of X is half the rank of the first homology group $H_1(X, \mathbb{Z}) \cong \mathbb{Z}^{2g}$. The second homology group $H_2(X, \mathbb{Z}) \cong \mathbb{Z}$. If $f : X \to Y$ is a holomorphic mapping of compact Riemann surfaces, then the *topological degree* of f is defined to be the positive integer n such that the map

$$\mathbb{Z} \cong H_2(X, \mathbb{Z}) \to H_2(Y, \mathbb{Z}) \cong \mathbb{Z}$$

induced by f is multiplication by n. Equivalently, f induces an injection of the field F_Y of meromorphic functions on Y into the field F_X of meromorphic functions on F_X, and n is the field degree $[F_X : F_Y]$. If y is a point of Y in general position, then the cardinality of the fiber $f^{-1}(y)$ is usually n. However, there can be a finite number of points y such that the cardinality of $f^{-1}(y)$ is strictly less than n; we say that these points *ramify*. Intuitively, we think of the mapping f as a covering of Y by X; however, if ramification occurs, it is not strictly a covering in the topological sense but a *ramified covering*. Here *ramification* means that some of the points of the fiber can coalesce when a is specialized to a point that ramifies.

Suppose that $y \in Y$ and that $x_1, \cdots, x_r$ are the points of $f^{-1}(y)$. Let e_i be the number of points of $f^{-1}(\eta)$ that are near x_i when $\eta \in Y$ is a nonramified point of Y that is near y. Then e_i is called the *ramification index* of x_i, and if $e_i > 1$, we say that x_i is *ramified*. It is clear that $\sum e_i = n$. Note that with this definition, $e_i = 1$ if x_i is not ramified. We also denote $e_i = e(x_i|y)$. There is a strong analogy between ramification in this geometric setting and the ramification of primes in a number field.

We have the *Hurwitz genus formula* (Lang (1982)). Let g_X and g_Y be the genera of X and Y, respectively, let n be the topological degree of f, let $P_1, \cdots, P_u$ be the finite number of ramified points in X, and let $Q_i = f(P_i)$, their images in Y. Then the genus formula asserts that

$$(2g_X - 2) = n(2g_Y - 2) + \sum \big(e(P_i|Q_i) - 1\big). \tag{3.19}$$

Exercise 1.3.8 Apply this now in the case of the canonical map $f : \Gamma(N)\backslash\mathcal{H}^* \to SL(2, \mathbb{Z})\backslash\mathcal{H}^*$, with $N \geq 2$. Show that the degree n of f is 6 if $N = 2$, and $\frac{1}{2}N^3 \prod_{p|N}(1-p^{-2})$ if $N > 2$. Show that the points of $SL(2, \mathbb{Z})\backslash\mathcal{H}^*$ that ramify are i, $e^{2\pi i/3}$, and ∞. Show that there are $n/2$ points in the fiber over i, each with ramification index 2, $n/3$ points in the fiber over $e^{2\pi i/3}$, each with ramification index 3, and n/N points in the fiber over ∞, each with ramification index N. Hence show that when $N = 2$ or 3, $\Gamma(N)\backslash\mathcal{H}^*$ has genus zero. Confirm this when $N = 2$ by examining the fundamental domain in Exercise 1.2.3.
(Note: The fact that all the points in the fiber over the three points that ramify all have the same ramification index is due to the fact that $\Gamma(N)$ is a normal subgroup of $SL(2, \mathbb{Z})$. This phenomenon does not occur for subgroups that are not normal.)

Exercise 1.3.9 Show that $\Gamma_0(11)\backslash\mathcal{H}^*$ has genus one.

Exercise 1.3.10: Picard's theorem Prove that if ϕ is an entire function on $\mathbb{C}$ such there are two complex numbers a and b such that $a, b \notin \phi(\mathbb{C})$, then ϕ is constant.

[HINT: By Exercise 1.3.8, $\Gamma(2)$ has three cusps and $\Gamma(2)\backslash\mathcal{H}^*$ has genus zero. Consequently, $\Gamma(2)\backslash\mathcal{H}$ is equivalent to the Riemann sphere minus three points, or $\mathbb{C} - \{a, b\}$. Making this identification, we may regard f as taking values

in $\Gamma(2)\backslash\mathcal{H}$. Now mapping $\mathcal{H}$ onto the unit disk by the Cayley transform, we obtain a bounded entire function, which is therefore constant.]

Now we require some basic facts about (nonramified) covering spaces. This well-known and important theory has applications to ramified coverings because if $f : X \to Y$ is a ramified covering, and $P_1, \cdots, P_r$ are the points of Y that ramify, and if Y' is the (noncompact) space $Y - \{P_i\}$ and $X' = f^{-1}(Y')$, then X' is a *bona fide* covering space of Y'. For the theory of covering spaces, see Spanier (1966) and Hilton and Wylie (1960), or other standard references on topology.

Let U be a topological space, $x, y \in U$. A *path* from x to y is a continuous map t of the unit interval $[0, 1]$ to U with $t(0) = x$ and $t(1) = y$. x is called the *left endpoint* and y is called the *right endpoint*. U is called *path connected* if any two points may be joined by a path. The space U is called *contractible* if the identity map $U \to U$ is homotopic to a constant map.

We will consider topological spaces U satisfying the following axiom:

Axiom 1.3.1 *U is path connected and every point of U has a contractable neighborhood.*

For example, (connected) manifolds have this property. Let U and V satisfy Axiom 1.3.1, and let $p : V \to U$ be a continuous map. We say that p is a *covering* or *covering map* if the fibers $p^{-1}(u)$ are discrete and if every point $u \in U$ has a neighborhood N such that $p^{-1}(N)$ is homeomorphic to a direct product $N \times p^{-1}(u)$ in such a way that the composition

$$p^{-1}(N) \cong N \times p^{-1}(u) \to N,$$

where the second map is projection, coincides with p. Covering maps have the following important *path-lifting property*:

Property 1.3.1 *If $u : [0, 1] \to U$ is a path, and $v \in p^{-1}\big(u(0)\big)$, then there exists a unique path $\widetilde{u} : [0, 1] \to V$ such that $u = p \circ \widetilde{u}$ and $\widetilde{u}(0) = v$.*

This property is crucial in supplying proofs in the theory that we now describe. One easy consequence of the path-lifting property is that the fibers $p^{-1}(u)$ all have the same cardinality. Indeed, if u' is another point, we choose a path from u to u'. Now for every $v \in p^{-1}(u)$, by lifting this path to V with v as the left endpoint, the right endpoint of the lifted path is an element of $p^{-1}(u')$, defining a bijection between the fibers. The cardinality of the fibers is called the *degree* of the covering.

We say a topological space V is *simply connected* if it is path connected and if every homeomorphism of the circle into V is homotopic to a constant map. If U satisfies Axiom 1.3.1, then U admits a simply connected cover $\widetilde{U}$, called the *universal covering space*. To construct it, we fix a base point $u_0 \in U$; let $\widetilde{U}$ be the space of all paths $h : [0, 1] \to U$ such that the left endpoint

$h(0) = u_0$, modulo the identification of two paths when they can be deformed one into the other by a homotopy fixing both endpoints. The projection map $\widetilde{p} = \widetilde{p}_U : \widetilde{U} \to U$ is given by $\widetilde{p}(h) = h(1)$. We may topologize $\widetilde{U}$ in a natural way (we leave this to the reader), and $\widetilde{p}$ is a local homeomorphism.

Defining the *fundamental group* $\pi_1(U)$ also requires fixing a base point $u_0 \in U$. Then $\pi_1(U)$ may be defined to be the set of all paths $h : [0, 1] \to U$ with *both* endpoints equal to u_0, modulo the identification of paths that can be deformed one into the other by a homotopy fixing both endpoints. Thus $\pi_1(U)$ is precisely the fiber $\widetilde{p}^{-1}(u_0)$ in the map $\widetilde{p} : \widetilde{U} \to U$. To make $\pi_1(U)$ a group, if $\gamma_1, \gamma_2 \in \pi_1(U)$, we define the product to be the homotopy class of paths obtained by gluing the right endpoint of γ_1 to the left endpoint of γ_2. Similarly, if $\gamma \in \pi_1(U)$ and $h \in \widetilde{U}$, we define γh by gluing the right endpoint of γ to the left endpoint of h, and we obtain an action of $\pi_1(U)$ on $\widetilde{U}$, and we may identify the quotient space $\Gamma\backslash\widetilde{U} = U$.

We say that two covering maps $p : V \to U$ and $p' : V' \to U$ are *equivalent* if there exists a homeomorphism $\phi : V \to V'$ such that $p = p' \circ \phi$.

Exercise 1.3.11 Prove that there is a bijection between equivalence classes of coverings of U and conjugacy classes of subgroups of the fundamental group $\pi_1(U)$, which associates with the subgroup $\Gamma \subset \pi_1(U)$ the covering map $\Gamma\backslash\widetilde{U} \to U$ induced by projection $\widetilde{p}_U : \widetilde{U} \to U$. Show that $\pi_1(\Gamma\backslash\widetilde{U}) \cong \Gamma$.

[HINT: using the the path-lifting property, show that any covering map $p : V \to U$ lifts to an isomorphism $\widetilde{V} \to \widetilde{U}$; that is, the existence of the covering p implies that $\widetilde{V}$ may be identified with $\widetilde{U}$ in such a way that $\widetilde{p}_U = p \circ \widetilde{p}_V$. We assume the base points $u_0 \in U$ and $v_0 \in V$ are chosen so that $u_0 = p(v_0)$; then the fundamental group $\pi_1(V) = \widetilde{p}_V^{-1}(v_0)$ is a subgroup of $\pi_1(U) = \widetilde{p}_U^{-1}(u_0)$. The flexibility in this construction is that we may change the base point v_0 to another element of the fiber $p^{-1}(u_0)$; this has the effect of replacing $p_V^{-1}(v_0)$ by another conjugate subgroup. Conversely, given a subgroup Γ of $\pi_1(U)$, we define a covering space of U as the quotient space $\Gamma\backslash\widetilde{U}$, where the action of Γ is inherited from the natural action of $\pi_1(U)$ on $\widetilde{U}$. These two constructions are inverses of each other.]

Let $p : V \to U$ and $p' : V' \to U$ be covering maps. We say that p *dominates* p' if there exists a covering map $q : V \to V'$ such that $p = p' \circ q$.

Exercise 1.3.12 Let $p : V \to U$ and $p' : V' \to U$ be covering maps, and let $\Gamma, \Gamma' \subset \pi_1(U)$ be the subgroups associated with these covering maps by Exercise 1.3.11. Show that p dominates p' if and only if Γ is conjugate in $\pi_1(U)$ to a subgroup of Γ'.

A covering $p : V \to U$ is called *regular* if the group $\Gamma \subset \pi_1(U)$ associated with the covering by Exercise 3.11 is normal. In this case, the quotient group $\pi_1(U)/\Gamma$ acts on V. Indeed, identifying V with $\Gamma\backslash\widetilde{U}$, if $\gamma \in \Gamma, \widetilde{u} \in \widetilde{U}$, let the coset $\overline{\gamma}$ of γ in $\pi_1(U)/\Gamma$ act by $\overline{\gamma}\Gamma\widetilde{u} = \Gamma\gamma\widetilde{u}$. The action of $\pi_1(U)/\Gamma$ commutes with the map p, and hence preserves the fiber $p^{-1}(u_0)$ and is transitive on the fiber.

Exercise 1.3.13 Conversely, show that if $p : V \to U$ is a covering, and if there exists a group G of automorphisms of V that commute with p such that G is transitive on the fiber $p^{-1}(u_0)$, then the covering p is regular, and if Γ is the subgroup of $\pi_1(U)$ associated with p by Exercise 1.3.11, then $G \cong \pi_1(U)/\Gamma$.

Exercise 1.3.14 Show that every covering is dominated by a regular covering.

A regular covering should be thought of as analogous to a Galois field extension, and the covering group $\pi_1(U)/\Gamma$ should be thought of as analogous to the Galois group.

Exercise 1.3.15 shows that there is a close connection between the topology of covering spaces and holomorphic mappings of compact Riemann surfaces. The covering map $p : V \to U$ is called *finite* if the fibers $p^{-1}(u)$ are finite for $u \in U$.

Exercise 1.3.15 Let X and Y be compact Riemann surfaces and let $f : X \to Y$ be a holomorphic mapping. Let $P_1, \cdots, P_r \in Y$ be the points that ramify. Let $U = Y - \{P_1, \cdots, P_r\}$, and let $V = f^{-1}(U)$. Then the restriction of f to V is a finite covering of U. Conversely, show that if $f' : V' \to U$ is any finite covering of U, then V' may be identified with an open subset of a compact Riemann surface X', and f' may be extended to a holomorphic mapping $X' \to Y$.

[HINT: V' inherits a complex structure from U by the requirement that f' be a holomorphic mapping. The problem is how to compactify V' by adjoining points to make up the fiber $f'^{-1}(P)$ when P is one of the exceptional points P_i. This is a purely local question. First solve the topological problem of constructing the fiber; what remains then is the analytic problem of imposing a complex structure in the neighborhood of a point $Q \in f'^{-1}(P)$. Let $e = e(Q|P)$ be the ramification index of Q. Let (U, y) be a chart near P so U is a small neighborhood of P and $y : U \to \mathbb{C}$ is a holomorphic equivalence of U with a domain in $\mathbb{C}$; assume that $y(P) = 0$. Show on purely topological grounds that $y \circ f' = x^e$ for a function x defined in the connected component of $f'^{-1}(U - P)$ whose closure contains Q. Now use a theorem on removable singularities, such as Rudin's Theorem 10.20 (Rudin, 1974), to extend x to a chart near Q, making X' a complex manifold.]

Exercise 1.3.16 In the setting of Exercise 1.3.15, the holomorphic mapping $f : X \to Y$ induces an inclusion $F_Y \to F_X$ of the fields of meromorphic functions. Show that the field degree $[F_X : F_Y]$ equals the degree of the cover $V \to U$ and that the cover is regular if and only if F_X/F_Y is a Galois extension, in which case the group $\pi_1(\Gamma\backslash\widetilde{U}) \cong \Gamma$ of Exercise 1.3.11 is isomorphic to the Galois group $\mathrm{Gal}(F_X/F_Y)$.

Exercise 1.3.17 Let $Y = \mathbb{P}^1(\mathbb{C})$, let y_0, y_1, and y_∞ be three distinct points of Y, and let $U = Y - \{y_0, y_1, y_\infty\}$. Prove that there exists a regular cover of degree six of U, which can be extended to a holomorphic mapping $f : X \to Y$ of compact Riemann surfaces, and such that $f^{-1}(y_0)$ and $f^{-1}(y_1)$ each consist of three points, with ramification index two, and $f^{-1}(y_\infty)$ consists of two points,

each with ramification index three. Use the genus formula (3.19) to show that X has genus zero. Let $p : Z \to Y$ be any holomorphic map from another Riemann surface to Y. Assume that only y_0, y_1, and y_∞ ramify, and that the ramification index of any point in the fiber over y_0 or y_1 is either 1 or 2 and that the ramification index of any point in the fiber over y_∞ is either 1 or 3. Prove that there exists a holomorphic mapping $q : X \to Z$ such that $f = p \circ q$.

[HINTS: For the first part, note that $\pi_1(U)$ is a free group with two generators γ_0 and γ_1. Here γ_i is a loop issuing out of the base point and circling the point y_i once counterclockwise before returning to the base point. Let Γ be the smallest normal subgroup of $\pi_1(U)$ containing by γ_0^2, γ_1^2, and $(\gamma_0\gamma_1)^3$, and let $V = \Gamma\backslash\widetilde{U}$. Note that $\pi_1(U)/\Gamma$ is the group with two generators g_0 and g_1 subject to the relations

$$g_0^2 = g_1^2 = (g_0 g_1)^3 = 1.$$

This group is isomorphic to the symmetric group S_3. Extend the cover $V \to U$ to a ramified cover of Y by a Riemann surface X by using Exercise 1.3.15. For the other part, first construct q over $f^{-1}(U)$ by Exercise 1.3.12, then extend it by Exercise 1.3.15.]

We now show how these ideas can be applied to the construction of automorphic functions for various groups.

Exercise 1.3.18 Prove that there exists an automorphic function z on $\Gamma(2)\backslash\mathcal{H}$ that satisfies the polynomial

$$z^3 - zj - 16j = 0.$$

[HINTS: By identifying $SL(2,\mathbb{Z})\backslash\mathcal{H}$ with $\mathbb{P}^1(\mathbb{C})$ by means of the map j as in Exercise 1.3.7 (b), we may take $y_0 = 0$, $y_1 = 1728$, and $y_\infty = \infty$ in Exercise 1.3.17. If $X = \Gamma(2)\backslash\mathcal{H}$, the projection

$$f : \Gamma(2)\backslash\mathcal{H} \to SL(2,\mathbb{Z})\backslash\mathcal{H}$$

has the ramification described for the map $X \to \mathbb{P}^1(\mathbb{C})$ in Exercise 1.3.17, so we may identify the covering space X with $\Gamma(2)\backslash\mathcal{H}$. Check that if $j_0 \in \mathbb{C}$, the polynomial $z^3 - zj_0 - 16j_0$ has no multiple roots unless $j_0 = 0$ or $j_0 = 1728$. (Compute the discriminant of this cubic polynomial.) Define, therefore, a threefold covering $Z \to SL(2,\mathbb{Z})\backslash\mathcal{H}$ by taking

$$Z = \{(z_0, \tau) \in \mathbb{C} \times SL(2,\mathbb{Z})\backslash\mathcal{H} | z_0^3 - z_0 j(\tau) - 16j(\tau) = 0\},$$

with the covering map $p : Z \to SL(2,\mathbb{Z})\backslash\mathcal{H}$ being the projection on the second component. Check that the hypotheses of Exercise 1.3.17 are satisfied, and conclude that there exists a holomorphic mapping $q : X \to Z$ such that $f = p \circ q$. Composing q with the projection on the first component gives the required holomorphic mapping.]

Exercise 1.3.19 Prove that there exists an automorphic function on $\Gamma(3)\backslash\mathcal{H}$ whose cube equals j.

1.4 Hecke Operators

Hecke (1937) introduced a certain ring of operators acting on modular forms. The commutativity of this ring leads to Euler products associated with modular forms. In the modern viewpoint, the Hecke ring is seen as a convolution ring of functions on $GL(2, A_f)$, where A_f is the ring of "finite adeles," which we introduce in Section 3.1. We will encounter Hecke operators in various forms throughout the book.

We are influenced in our treatment of this subject by the discussion in Shimura (1971). Let us fix a weight k, which is a positive integer. It may be *even* or *odd*.

If f is a holomorphic function on $\mathcal{H}$, and $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in GL(2, \mathbb{R})^+$, we will denote by $f|\gamma$ the function

$$(f|\gamma)(z) = (\det \gamma)^{k/2} (cz + d)^{-k} f\left(\frac{az + b}{cz + d}\right). \tag{4.1}$$

That $(f|\gamma)|\gamma' = f|(\gamma\gamma')$ may be checked, so this is a bona fide right action on holomorphic functions on $\mathcal{H}$. Note that if k is even, scalar matrices act trivially; on the other hand, if k is odd,

$$f\left|\begin{pmatrix} \lambda & \\ & \lambda \end{pmatrix}\right. = \begin{cases} f & \text{if } \lambda > 0; \\ -f & \text{if } \lambda < 0. \end{cases}$$

It will be convenient to immediately generalize the definition of modular forms. Let Γ be a discontinuous subgroup of $SL(2, \mathbb{R})$ such that $\Gamma\backslash\mathcal{H}$ has finite volume with respect to the measure defined in Exercise 1.2.6, for example, a congruence subgroup of $\Gamma(1) = SL(2, \mathbb{Z})$. We say that a holomorphic function f on $\mathcal{H}$ is a *modular form* with respect to Γ if it satisfies Eq. (3.1) for all $\gamma \in \Gamma$ and is *holomorphic* at the cusps of $\Gamma\backslash\mathcal{H}$. Furthermore, it *vanishes* at all the cusps, we say it is a *cusp form*. The notion of holomorphicity or vanishing at the cusp $c \in \mathbb{R} \cup \{\infty\}$ is made precise as follows. Choose $\rho \in SL(2, \mathbb{R})$ such that $\rho(c) = \infty$. Then $f|\rho^{-1}$ is modular with respect to the group $\rho\Gamma\rho^{-1}$, which contains a translation $z \mapsto z + t$ for some real $t > 0$. Hence $f|\rho^{-1}$ has a Fourier expansion $\sum a_{\rho,n} e^{2\pi i n z/t}$. We say f is *meromorphic* at c if the coefficients $a_{\rho,n} = 0$ for $n < -C$ for some constant C; we say it is *holomorphic* at c if $a_{\rho,n} = 0$ for $n < 0$, and that it *vanishes* at c if the coefficients $a_{\rho,n} = 0$ for $n \leq 0$.

We note that if $-I \in \Gamma$, then Eq. (3.1) is impossible unless k is even.

Lemma 1.4.1 *Let Γ be a congruence subgroup of $SL(2, \mathbb{Z})$, and let $\alpha \in GL(2, \mathbb{Q})^+$. Then there exists an integer M such that $\alpha^{-1}\Gamma\alpha \supseteq \Gamma(M)$. Consequently, $\alpha^{-1}\Gamma\alpha \cap \Gamma(1)$ is a congruence group.*

Proof Let N be such that $\Gamma(N) \subseteq \Gamma$. We may find positive integers M_1, M_2 such that $M_1\alpha$, $M_2\alpha^{-1} \in \mathrm{Mat}_2(\mathbb{Z})$. Let $M = M_1M_2N$. If $\gamma \in \Gamma(M)$, write

$\gamma = I + Mg$, where I is the 2×2 identity matrix, and $g \in \mathrm{Mat}_2(\mathbb{Z})$. Then $\alpha\gamma\alpha^{-1} = I + N(M_1\alpha)\, g\, (M_2\alpha^{-1})$. This is clearly an element of $\Gamma(N)$. ■

Now if f is a modular form for a congruence subgroup Γ, and $\alpha \in GL(2, \mathbb{Q})^+$, then $f|\alpha$ is modular with respect to $\alpha^{-1}\Gamma\alpha \cap \Gamma(1)$, which, we see, is a congruence group. Let us say that f is a *congruence modular form* or a *congruence cusp form* if it is a modular or cusp form, respectively, for some congruence subgroup of $\Gamma(1)$. We see that the action of $GL(2, \mathbb{Q})^+$ preserves the property of being a congruence modular form or cusp form.

If H is a group acting on the left on a set X, we will denote by $H\backslash X$ the set of orbits of X under this action. If X is a topological space, then $H\backslash X$ is given the *quotient topology* in which a subset is open if and only if its preimage under the natural map $X \to H\backslash X$ is open. The set $H\backslash X$ is variously known as a *quotient space, homogeneous space,* or *orbit space* – these terminologies are especially appropriate if X is a topological space but may be used in any case.

For example, if $G \supset H$ is a bigger group, H acts on G by left translation and $H\backslash G$ is the set of right cosets Hg for $g \in G$. Similarly, if H acts on X by right translation, we denote the set of orbits by X/H, so if $G \supset H$ is a bigger group, G/H is the set of left cosets gH. Of course, this is a group if H is normal, but otherwise it is just a set. If H_1 and H_2 are groups acting on X on the left and right, respectively, such that the actions are compatible

$$(h_1 x)h_2 = h_1(xh_2) \qquad \text{for } h_1 \in H_1, x \in X, h_2 \in H_2,$$

we again have a set of orbits; x and y will lie in the same orbit if $x = h_1 y h_2$ for some $h_1 \in H_1$ and $h_2 \in H_2$. The set of orbits in this situation is denoted $H_1\backslash X/H_2$. As a special case, if H_1 and H_2 are subgroups of a group G, $H_1\backslash G/H_2$ is the set of double cosets $H_1 g H_2$. One way to think of this is that H_1 acts on G/H_2 by left translation, and $H_1\backslash G/H_2$ is simply the set of orbits under this action; equivalently, H_2 acts on $H_1\backslash G$ by right translations, and $H_1\backslash G/H_2$ may be equally regarded as the set of orbits under this right action.

We will describe Hecke operators for $\Gamma(1) = SL(2, \mathbb{Z})$, leaving Hecke operators for congruence subgroups to the exercises. Because $-I \in SL(2, \mathbb{Z})$, as we have already noted, Eq. (3.1) requires that the weight k must be even, and we assume this for the remainder of this section, excluding the exercises, where we consider Hecke operators for congruence subgroups.

Proposition 1.4.1 *Let $\alpha \in GL(2, \mathbb{Q})^+$. Then the double coset $\Gamma(1)\alpha\Gamma(1)$ is a finite union of right cosets:*

$$\Gamma(1)\alpha\Gamma(1) = \bigcup_{i=1}^{N} \Gamma(1)\alpha_i, \qquad \alpha_i \in GL(2, \mathbb{Q})^+. \tag{4.2}$$

Indeed, the number of right cosets in this decomposition equals $[\Gamma(1) : \alpha^{-1}\Gamma(1)\alpha \cap \Gamma(1)]$, which is finite.

Proof We will show the cardinality of $\Gamma(1)\backslash\Gamma(1)\alpha\Gamma(1)$ is equal to $[\Gamma(1) : \alpha^{-1}\Gamma(1)\alpha \cap \Gamma(1)]$. (This cardinality is finite by Lemma 1.4.1 because $\alpha^{-1}\Gamma(1)\alpha \cap \Gamma(1)$ is a congruence subgroup.) Right translation by α^{-1} is a bijection of $GL(2, \mathbb{Q})^+$ onto itself, which induces a bijection of this set with

$$\Gamma(1)\backslash\Gamma(1)\alpha\Gamma(1)\alpha^{-1} \cong \big(\Gamma(1) \cap \alpha\Gamma(1)\alpha^{-1}\big)\backslash\alpha\Gamma(1)\alpha^{-1}.$$

Conjugating by α, this quotient has the same cardinality as $\big(\alpha^{-1}\Gamma(1)\alpha \cap \Gamma(1)\big)\backslash\Gamma(1)$. ■

If $\alpha \in GL(2, \mathbb{Q})^+$, we define the *Hecke operator* $T_\alpha = T(\alpha)$ on $M_k\big(\Gamma(1)\big)$ by

$$f|T_\alpha = \sum f|\alpha_i, \tag{4.3}$$

with the α_i as in Eq. (4.2). Observe that $f|T_\alpha$ is independent of the choice of representatives α_i because f is modular. Moreover $f|T_\alpha$ is modular, because if $\gamma \in \Gamma(1)$, it follows from Eq. (4.2) that the cosets $\Gamma(1)\,\alpha_i\gamma$ are the same as the $\Gamma(1)\,\alpha_i$ permuted, so there exist $\gamma_i \in \Gamma(1)$ such that the $\alpha_i\gamma$ are the $\gamma_i\alpha_i$ permuted, and then

$$(f|T_\alpha)|\gamma = \sum f|\alpha_i\gamma = \sum f|\gamma_i\alpha_i = \sum f|\alpha_i = f|T_\alpha.$$

Thus $f|T_\alpha$ is a modular form for $\Gamma(1)$, and T_α is a linear transformation of $M_k\big(\Gamma(1)\big)$. The space $S_k\big(\Gamma(1)\big)$ is clearly an invariant subspace.

If $\alpha, \beta \in GL(2, \mathbb{Q})^+$, let α_i be as in Eq. (4.2) and also $\Gamma(1)\beta\Gamma(1) = \bigcup \Gamma(1)\beta_i$ (disjoint). We have

$$f|T_\alpha|T_\beta = \sum f|\alpha_i\beta_j = \sum_{\sigma \in \Gamma(1)\backslash GL(2,\mathbb{Q})^+} m(\alpha, \beta; \sigma)\, f|\sigma,$$

where σ runs through a set of representatives for $\sigma \in \Gamma(1)\backslash GL(2, \mathbb{Q})^+$, and $m(\alpha, \beta; \sigma)$ is the cardinality of the set of indices (i, j) such $\sigma \in \Gamma(1)\alpha_i\beta_j$. We see easily that $m(\alpha, \beta; \sigma)$ depends only on the double coset $\Gamma(1)\sigma\Gamma(1)$, so we may rewrite this

$$f|T_\alpha|T_\beta = \sum f|\alpha_i\beta_j = \sum_{\sigma \in \Gamma(1)\backslash GL(2,\mathbb{Q})^+/\Gamma(1)} m(\alpha, \beta; \sigma)\, f|T_\sigma, \tag{4.4}$$

where σ runs through a set of representatives for $\Gamma(1)\backslash GL(2, \mathbb{Q})^+/\Gamma(1)$. This prompts us to introduce a certain ring $\mathcal{R}$. Let $\mathcal{R}$ be the free Abelian group generated by the symbols $T_\alpha = T(\alpha)$ as α runs through a complete set for $\Gamma(1)\backslash GL(2, \mathbb{Q})^+/\Gamma(1)$. We define a multiplication in $\mathcal{R}$ by

$$T_\alpha \cdot T_\beta = \sum_{\sigma \in \Gamma(1)\backslash GL(2,\mathbb{Q})^+/\Gamma(1)} m(\alpha, \beta; \sigma)\, T_\sigma. \tag{4.5}$$