# 2

# Modular stuff

This chapter introduces modular functions and forms, a subject central to the remainder of the book. Some earlier parts of this chapter are beautifully covered in [**414**].

Section 2.1 supplies the underlying geometry, but can be skimmed on a first reading. In spite of this background material, the theory of modular forms and functions discussed in Sections 2.2 and 2.3 will probably appear as somewhat arbitrary to the uninitiated reader. Section 2.4.1 addresses some of this apparent artificiality, by developing the broader context of automorphic forms.

As explained in the introductory chapter, Moonshine involves unexpected occurrences of modularity. The modularity of Moonshine functions follows from Zhu's Theorem (Theorem 5.3.8). However, the complexity of the underlying mathematics begs the question: Can modularity be established in a more elementary way? The simplest example of Moonshine involves theta functions. Hence we explore the limits and potentials of four classical strategies for proving the modularity of theta functions: Poisson summation, Dirichlet series, the heat kernel and representations of Heisenberg groups (Sections 2.2.3, 2.3.1, 2.3.4 and 2.4.2, respectively).

Moonshine has really only been worked out in genus 1,[1] but conformal field theory tells us that there is an analogue for every genus (Section 6.3.1). It will be much more complicated, but it will be more rewarding because the number theoretic side is much less developed. In other words, we will find traces of, for example, the Monster in automorphic forms for the higher mapping class groups $\Gamma_{g,n}$ and $\mathrm{Sp}_{2n}(\mathbb{Z})$. We include Sections 2.1.4 and 2.3.5 in anticipation of this most natural and significant future development.

## 2.1 The underlying geometry

### *2.1.1 The hyperbolic plane*

The birth of hyperbolic geometry is one of the most remarkable and instructive in the history of mathematics. Euclid's Fifth Postulate[2] was noticeably more complicated than the other axioms, looking more like a theorem than a self-evident proposal. Indeed, its converse was a theorem proved by Euclid. For example, compare it with Euclid's First

[1] There are two possible meanings of 'genus' in a phrase like 'higher genus Moonshine'. Ordinary Monstrous Moonshine is genus 0 in the sense that the $j$-function is a Hauptmodul, i.e. a function on a sphere. It is genus 1 in the sense that the argument $\tau$ of $j$ parametrises different tori. In this paragraph we are anticipating Moonshine's extension to higher genus in this second sense.

[2] Also called the Parallel Postulate, it is equivalent to the simpler statement: *Given any line* L *and a point* p *not on* L, *there is a unique line parallel to* L *that passes through* p.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

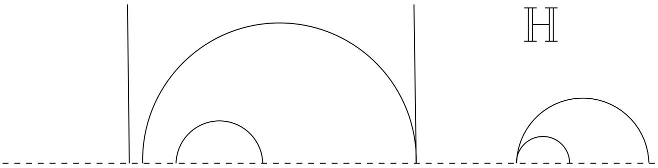

Fig. 2.1 Several parallel lines in the hyperbolic plane $\mathbb{H}$.

Postulate: *There is a unique line passing through any two points*, or Euclid's Fourth Postulate: *All right angles are equal*. For centuries, starting with Archimedes, mathematicians (both professional and amateur) tried to prove it from the other axioms. Finally in 1868 Beltrami established its independence by finding models for the hyperbolic plane, proving the conjecture of Gauss, Bolyai and Lobachevski as to the existence (i.e. internal consistency) of this non-Euclidean geometry. (More precisely, Beltrami's models reduced the question of the consistency of hyperbolic geometry to the consistency of Euclidean geometry.) Far from being an artificial construct, we've now learned that hyperbolic geometry is far more important than Euclidean geometry, at least in two and three dimensions.

In place of the Euclidean plane $\mathbb{R}^2$, consider the upper half-plane

$$\mathbb{H} := \{(x, y) \in \mathbb{R}^2 \mid y > 0\} = \{\tau \in \mathbb{C} \mid \mathrm{Im}\, \tau > 0\}. \tag{2.1.1}$$

The angles between intersecting curves in $\mathbb{H}$ are measured as in $\mathbb{R}^2$ (namely, take the angle between the two Euclidean lines tangent to the curves at the point of intersection). However, the hyperbolic lines consist of all half-lines perpendicular to the $x$-axis, together with all semi-circles with centre on the $x$-axis (see Figure 2.1). All axioms of Euclidean geometry hold here (e.g. between any two distinct points there passes a unique line), except for the Parallel Postulate: there are always infinitely many hyperbolic lines parallel to a given hyperbolic line $L$ and passing through a given point $p \notin L$.

It is possible to prove from the other axioms that the remaining possibility (namely that there are *no* lines parallel to line $L$ through point $p$) cannot occur. Nevertheless, there is a second kind of non-Euclidean geometry, called *spherical* geometry. In place of $\mathbb{R}^2$ we have the sphere $S^2$, and lines now are great circles. If we identify antipodal points $\pm p \in S^2$, then we get a geometry satisfying most of Euclid's axioms. The exceptions are that we can't speak unambiguously of the portion of a line between two points, and the Parallel Postulate (there are no parallel lines). Spherical geometry is older than Euclid – we needed it, for example, in our study of the night sky.

In Euclidean $\mathbb{R}^2$ the metric (infinitesimal length-squared) is given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$, and so the arc-length of a curve $\gamma : [0, 1] \to \mathbb{R}^2$ is

$$\mathrm{length}(\gamma) := \int_0^1 \sqrt{\gamma_1'(t)^2 + \gamma_2'(t)^2}\, \mathrm{d}t.$$

On $\mathbb{H}$ the arc-length of a curve $\gamma : [0, 1] \to \mathbb{H}$ becomes

$$\mathrm{length}_{\mathbb{H}} := \int_0^1 \frac{\sqrt{\gamma_1'(t)^2 + \gamma_2'(t)^2}}{\gamma_2(t)}\, \mathrm{d}t = \int_0^1 \frac{|\gamma'(t)|}{\mathrm{Im}\, \gamma(t)}\, \mathrm{d}t. \tag{2.1.2}$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Define the hyperbolic distance $\mathrm{dist}_{\mathbb{H}}(p,q)$ between two points $p, q \in \mathbb{H}$ to be the infimum $\inf_\gamma \mathrm{length}_{\mathbb{H}}(\gamma)$ of the arc-lengths of all paths $\gamma$ between $p = \gamma(0)$ and $q = \gamma(1)$. Just as the shortest path (geodesic) between two points in Euclidean geometry is the line segment between them, so in hyperbolic geometry it is the hyperbolic line segment.

The 'boundary' for $\mathbb{R}^2$ can be thought of as the circular horizon of 'points at infinity', parametrised by angle, and every line touches this circle at two points. Likewise, the boundary of $\mathbb{H}$ can be thought of as the circle $\mathbb{R} \cup \{\infty\}$, and again every line touches this circle at two points. This circle will appear as the infinitely distant horizon to beings living in $\mathbb{H}$. The point '$\infty$' here is often written $\mathrm{i}\infty$ to emphasise its relation to the vertical lines. The difference is that in $\mathbb{R}^2$, all parallel lines share the same two points at infinity; in $\mathbb{H}$, parallel lines share at most one point at infinity.

The most compelling model of the hyperbolic plane is perhaps the Poincaré disc

$$\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}.$$

Here, angles are again as in $\mathbb{R}^2$, but lines consist of diameters of the boundary circle $|z| = 1$, together with the intersection of $\mathbb{D}$ with circles hitting the boundary $|z| = 1$ at right angles. The metric is $|\mathrm{d}z|^2/(1-|z|^2)^2$, and the 'points at infinity' form the boundary circle $|z| = 1$. The equivalence with $\mathbb{H}$ is given by the isometry $\tau \mapsto \frac{\tau-\mathrm{i}}{\tau+\mathrm{i}}$ taking $\mathbb{H}$ onto $\mathbb{D}$.

It may seem strange that both models $\mathbb{H}$ and $\mathbb{D}$ of hyperbolic geometry have a distorted notion of length and line. Is there any way to realise hyperbolic geometry, using a surface embedded in $\mathbb{R}^3$ inheriting the usual metric and angle of $\mathbb{R}^3$? Hilbert proved the answer is No: *There is no complete surface in* $\mathbb{R}^3$ *with constant negative curvature* (see e.g. page 51 of [**527**]). Nash's Theorem (footnote 5 in chapter 1) implies though that there will be an embedding of the hyperbolic plane in some $\mathbb{R}^n$ ($n = 5$ works). 'Complete' means that any Cauchy sequence converges, so there aren't any points missing. To find the curvature of a surface at a point, first find the smallest and largest circles hugging the surface the closest at that point; the curvature is the inverse product $r^{-1}R^{-1}$ of their radii. For example, a sphere of radius $r$ has constant curvature $r^{-2}$. A surface with 0 curvature is (locally) flat in one direction – for example, a cylinder or torus has constant curvature 0. The small and large circles for a surface $\Sigma$ with negative curvature have centres on opposite sides of the tangent plane $T_p\Sigma$, like a saddle curving up from front to back, but curving down from side to side. The hyperbolic plane has constant negative curvature (Theorem 2.1.4(b)).

What is the significance of the word 'hyperbolic' here? It was chosen by Klein, partly because sinh and cosh appear in many formulae, but also because of another model of $\mathbb{H}$. Consider the hyperboloid $x_1^2 + x_2^2 - x_3^2 = -1$, embedded in Minkowski space $\mathbb{R}^{2,1}$ (so it is a Minkowski sphere of radius i). It consists of two sheets; let's focus on the upper one (where $x_3 \geq 1$). As a surface in $\mathbb{R}^{2,1}$, it inherits its notions of angle and metric $\mathrm{d}s^2 = \mathrm{d}x_1^2 + \mathrm{d}x_2^2 - \mathrm{d}x_3^2$ – in particular this induced geometry is equivalent to the hyperbolic plane. The lines here consist of the intersection of planes through the origin with the upper sheet (when those intersections are non-empty). Stereographic projection from the point $(0, 0, -1)$ conformally maps the upper sheet onto the Poincaré disc $\mathbb{D} \times \{0\}$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Just as the area of a region $R \subset \mathbb{R}^2$ is given by the double integral $\int_R \mathrm{d}x\, \mathrm{d}y$, so is the hyperbolic area of region $R \subset \mathbb{H}$ given by

$$\mathrm{area}_{\mathbb{H}}(R) := \int_R \frac{\mathrm{d}x\, \mathrm{d}y}{y^2}. \tag{2.1.3a}$$

This just says that the hyperbolic area of the infinitesimal rectangle $[x, x + \mathrm{d}x] \times [y, y + \mathrm{d}y]$ is the product $\frac{\mathrm{d}x}{y} \times \frac{\mathrm{d}y}{y}$ of hyperbolic length with hyperbolic height. This area formula fails for macroscopic rectangles, if for no other reason than that there are *no* macroscopic rectangles! In fact, one of the most remarkable formulae of geometry must be the expression, originally due to Lambert (1766),[3] for the area of a triangle $T$ in terms of its interior angles $\alpha_1, \alpha_2, \alpha_3$:

$$\mathrm{area}_{\mathbb{H}}(T) = \pi - \alpha_1 - \alpha_2 - \alpha_3. \tag{2.1.3b}$$

More generally, the area of an $n$-sided hyperbolic polygon is $(n-2)\pi - \sum_i \alpha_i$. From this we obtain the non-existence of rectangles. These formulae apply even in the limiting case where some vertices lie on the boundary $\mathbb{R} \cup \{\mathrm{i}\infty\}$. In particular, the area of any hyperbolic triangle is bounded above (even though $\mathbb{H}$ itself has infinite area)!

Klein proposed to study geometry using the group of symmetries of whichever geometric quantities are important to the context (Section 1.2.2). The group $\mathrm{Isom}(\mathbb{R}^2)$ of isometries (i.e. distance-preserving maps) of $\mathbb{R}^2$ consists of all translations $x \mapsto x + a$, all orthogonal maps (rotations and reflections) $x \mapsto xA$ where $AA^t = I$, and all combinations $xA + b$ thereof. Likewise, the group $\mathrm{Isom}(\mathbb{H})$ of hyperbolic isometries consists of all *Möbius*, or *fractional linear*, transformations

$$z \mapsto \frac{az+b}{cz+d}, \qquad \forall a, b, c, d \in \mathbb{R} \text{ with } ad - bc = 1, \tag{2.1.4a}$$

together with the reflection $z \mapsto -\overline{z}$, and all combinations thereof. As in the Euclidean case, $\mathrm{Isom}(\mathbb{H})$ is a three-dimensional real Lie group, with two connected components; the component $\mathrm{Isom}^+(\mathbb{H})$ containing the identity consists of (2.1.4a), and is isomorphic to

$$\mathrm{PSL}_2(\mathbb{R}) := \mathrm{SL}_2(\mathbb{R}) / \left\{ \pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}. \tag{2.1.4b}$$

As in the Euclidean case, isometries preserve the absolute value $|\theta|$ of angles; maps $\alpha \in \mathrm{Isom}^+(\mathbb{H})$ preserve the angles themselves and so are conformal. Isometries preserve area and send hyperbolic lines to hyperbolic lines. $\mathrm{PSL}_2(\mathbb{R})$ preserves everything of geometric significance and is thus the group of symmetries of the hyperbolic plane.

Likewise, the group $\mathrm{Isom}^+(S^2)$ of symmetries of spherical geometry is $\mathrm{PSL}_2(\mathbb{C})$, acting on the Riemann sphere $\mathbb{P}^1(\mathbb{C})$ by Möbius transformations. The symmetries $\mathrm{PSL}_2(\mathbb{R})$ of $\mathbb{H}$ are precisely those transformations in $\mathrm{PSL}_2(\mathbb{C})$ that send $\mathbb{H}$ to itself. The only reason this action by Möbius transformations of the $2 \times 2$ matrices on $\mathbb{P}^1(\mathbb{C})$ or $\mathbb{H} \cup \{\mathrm{i}\infty\}$ might not look strange to us, is because familiarity breeds numbness. Much more natural is

[3] This is the same Lambert who proved the irrationality of $\pi$ and $e$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

the action of $n \times n$ matrices on $\mathbb{C}^n$, and this induces their action on $\mathbb{C}^{n-1}$ (together with a codimension-2 set of 'points at infinity') by interpreting $\mathbb{C}^n$ as homogeneous coordinates for $\mathbb{C}^{n-1}$ (Section 1.2.2). Specialising to $n = 2$ gives us the action (2.1.4a). In Section 2.4.1 we interpret (2.1.4a) using the multiplication of matrices in $\mathrm{SL}_2(\mathbb{R})$.

A model for $n$-dimensional hyperbolic geometry is the upper half-space $\mathbb{H}^n := \{(x_i) \in \mathbb{R}^n \mid x_n > 0\}$, which is conformally equivalent to the interior of the unit $n$-ball, or to the upper (i.e. $x_{n+1} > 0$) sheet of the hyperboloid $x_1^2 + \cdots + x_n^2 - x_{n+1}^2 = -1$. Euclidean angle is used, but the metric is $\mathrm{d}s^2 = (\mathrm{d}x_1^2 + \cdots + \mathrm{d}x_n^2)/x_n^2$. Hyperbolic lines consist of half-lines and semi-circles perpendicular to the boundary hyperplane $x_0 = 0$; hyperbolic planes in $\mathbb{H}^n$ consist of half-planes and half-spheres perpendicular to the boundary hyperplane $x_0 = 0$. The hyperboloid model makes it clear that the isometries $\mathrm{Isom}(\mathbb{H}^n)$ of hyperbolic $n$-space is isomorphic to the group of those matrices $A \in \mathrm{O}_{n,1}(\mathbb{R})$ with $A_{n+1,n+1} \geq 1$. The group $\mathrm{Isom}^+(\mathbb{H}^n)$ of *conformal* isometries is the Lorentz group $\mathrm{SO}_{n,1}(\mathbb{R})^+$, obeying in addition the condition $\det(A) = 1$. Of course the Lorentz group $\mathrm{SO}_{3,1}(\mathbb{R})^+$ is more famous in its incarnation as the symmetry of special relativity (Section 4.1.2). By identifying the boundary plane of $\mathbb{H}^3$ with $\mathbb{C}$, the group $\mathrm{Isom}^+(\mathbb{H}^3) \cong \mathrm{SO}_{3,1}(\mathbb{R})^+$ can be naturally identified with the Möbius transformations $\mathrm{PSL}_2(\mathbb{C})$.

Recall Hilbert's theorem from a few paragraphs ago. Although no surface embedded in $\mathbb{R}^3$ can provide a model of the full hyperbolic plane, they can provide a model of a piece of that plane (i.e. be 'incomplete'). This is accomplished by any surface of constant negative curvature. For example, consider the 'tractrix' – the path traced by a stone, initially placed at (0,1), pulled ('tractored') by a string of length 1 as we walk along the $x$-axis. Take the tractrix in the $xy$-plane and rotate it about the $x$-axis; the result is called the 'pseudo-sphere', and is a surface of constant negative curvature in $\mathbb{R}^3$. More generally, by a *hyperbolic surface* we mean a surface that is also a metric space (i.e. it has a notion of distance between points, and of arc-length), which is locally isometric to $\mathbb{H}$ (i.e. the open sets $V_\alpha$ in Definition 1.2.3 are taken to be in $\mathbb{H} \subset \mathbb{R}^2$, and the transition functions $\varphi_{\alpha\beta}$ are in $\mathrm{Isom}(\mathbb{H})$). The pseudo-sphere is an example of a hyperbolic surface different from the hyperbolic plane; crocheting constructs several other examples [**284**]. Similarly, we can define hyperbolic manifolds of arbitrary dimension. We conclude this subsection with the classification of all hyperbolic surfaces. But first we need the notion of a Fuchsian group.

As was discussed in Section 1.2.2, tori $S^1 \times S^1$ arise from the quotient $\mathbb{R}^2/L$ of the plane by a two-dimensional lattice. This construction is equivalent to the familiar depiction of a torus as a parallelogram with opposite sides identified. We discuss the Riemann surfaces in more detail next subsection, but a genus-$g$ surface can be depicted by identifying appropriate sides in a $4g$-gon (see Figure 2.2 for the situation with a genus 2 surface). This arises from making $2g$ circular cuts into the surface and flattening it out. But can we also interpret that $4g$-gon as corresponding to some quotient of $\mathbb{R}^2$, generalising the $\mathbb{R}^2/L$ construction of a torus? The answer is no – the group $\mathrm{Isom}(\mathbb{R}^2)$ doesn't have a rich enough supply of discrete subgroups. We *can* interpret the $4g$-gon as a quotient, but of the *hyperbolic* plane and not the Euclidean one.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

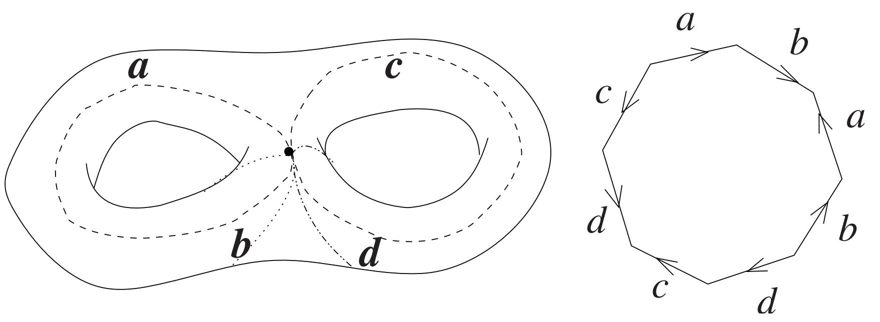

Fig. 2.2 A genus 2 surface and its octagon.

**Definition 2.1.1** *A* Fuchsian group *is a discrete subgroup* $\Gamma$ *of* $SL_2(\mathbb{R})$*, i.e. one with* $\inf\{(a-1)^2+b^2+c^2+(d-1)^2\} > 0$*, where the infimum is over all* $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq I$ *in* $\Gamma$.

We identify a subgroup $\Gamma$ of $\mathrm{SL}_2(\mathbb{R})$ with its canonical projection $\overline{\Gamma}$ into $\mathrm{PSL}_2(\mathbb{R})$, since these give rise to identical surfaces. Examples of Fuchsian subgroups are

$$G_N = \left\{ \begin{pmatrix} \cos(\pi k/N) & \sin(\pi k/N) \\ -\sin(\pi k/N) & \cos(\pi k/N) \end{pmatrix} \,|\, 0 \leq k < N \right\}, \qquad \forall N = 1, 2, \ldots,$$
$$G_{\mathbb{Z}} = \left\{ \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} \,|\, k \in \mathbb{Z} \right\},$$

and the modular group $\mathrm{SL}_2(\mathbb{Z})$. The latter is certainly the most interesting of these.

Let $\Gamma$ be a Fuchsian group. Most points $z \in \mathbb{H}$ (i.e. all but at most countably many) are fixed only by the identity in $\Gamma$ (why?). Let $z_0 \in \mathbb{H}$ be any of those generic points. Define the set

$$D_\Gamma(z_0) := \{w \in \mathbb{H} \,|\, \mathrm{dist}_{\mathbb{H}}(z_0, w) < \mathrm{dist}_{\mathbb{H}}(\gamma.z_0, w) \text{ for all } \gamma \in \Gamma,\ \gamma \neq \pm I\}.$$

So $D_\Gamma(z_0)$ is the intersection of a number of hyperbolic half-planes. This set $D = D_\Gamma(z_0)$ is called a *fundamental domain* of $\Gamma$, as it satisfies the following properties: (i) it is open; (ii) each orbit $\Gamma.z$ intersects $D$ in at most one point, and every orbit intersects the closure of $D$ in at least one point; (iii) the boundary $\partial D$ of $D$ in $\mathbb{H}$ consists of at most countably many hyperbolic line segments. (In fact, as long as $\Gamma$ is finitely generated, $D$ can be chosen with boundary consisting of only finitely many segments.)

For example, a fundamental domain for $G_N$ consists of the points lying between any pair of hyperbolic lines intersecting at i with angle $2\pi/N$. A fundamental domain for $G_{\mathbb{Z}}$ is $\{z \in \mathbb{H} \,|\, -\frac{1}{2} < \mathrm{Re}\, z < \frac{1}{2}\}$. Choosing $z_0 = 2\mathrm{i}$, we get the fundamental domain $D$ for $\mathrm{SL}_2(\mathbb{Z})$ depicted in Figure 2.3: the vertical sides are $\mathrm{Re}\, z = \pm\frac{1}{2}$, and the circle is $|z| = 1$.

Applying $\Gamma$ to a fundamental domain $D$ will tile the hyperbolic plane – see Escher's *Circle Limit I,II*, ... for examples. Since $\Gamma \subset \mathrm{Isom}^+(\mathbb{H})$, each of these tiles is an identical copy (a congruent translate) of $D$. All this holds as well in hyperbolic $n$-space – for example, an analogue of $\mathrm{SL}_2(\mathbb{Z})$ for $\mathbb{H}^3$ is $\mathrm{SL}_2(\mathbb{Z} + \mathrm{i}\mathbb{Z})$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

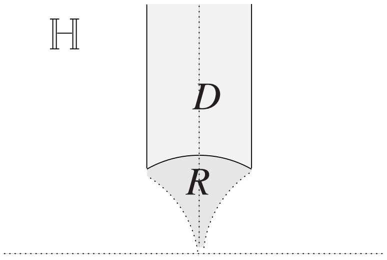

Fig. 2.3 Two fundamental domains for $\mathrm{SL}_2(\mathbb{Z})$.

Just as we constructed the torus by identifying opposite sides of the parallelogram, so we can obtain a surface by identifying the appropriate sides of the fundamental domain of a Fuchsian group $\Gamma$. This surface will be a realisation of the orbit space $\Gamma\backslash\mathbb{H}$ (we write $\Gamma$ on the left because it acts on the left). Provided no $\gamma \in \Gamma$ has fixed points in $\mathbb{H}$ (except for the trivial maps $\gamma = \pm I$), the orbit space $\Gamma\backslash\mathbb{H}$ will inherit the hyperbolic geometry of $\mathbb{H}$ and be a hyperbolic surface.

**Theorem 2.1.2** *Any complete hyperbolic surface $\Sigma$ is isometric to a surface of the form $\overline{\Gamma}\backslash\mathbb{H}$ where $\overline{\Gamma}$ is a torsion-free Fuchsian subgroup of* $\mathrm{PSL}_2(\mathbb{R})$. *Two such subgroups $\overline{\Gamma_1}, \overline{\Gamma_2}$ define isometric surfaces $\overline{\Gamma_1}\backslash\mathbb{H}$ and $\overline{\Gamma_2}\backslash\mathbb{H}$ iff $\alpha\overline{\Gamma_1}\alpha^{-1} = \overline{\Gamma_2}$ for some $\alpha \in$* $\mathrm{PSL}_2(\mathbb{R})$.

'Torsion-free' means that all nontrivial elements of $\overline{\Gamma}$ have infinite order – see Question 2.1.2(b). Almost all surfaces with a conformal or metric or complex structure are $\Gamma\backslash\mathbb{H}$ for some Fuchsian subgroup $\Gamma$. An unexpected revelation of Thurston's Programme is that something similar happens in three dimensions – see the review [**497**]. Any surface of genus $g \geq 2$ supports uncountably many different hyperbolic structures. By contrast, the Mostow Rigidity Theorem (1973) tells us that a connected compact oriented manifold of dimension $n \geq 3$ supports only one.

### *2.1.2 Riemann surfaces*

Manifolds $M, N$ are homeomorphic if there is a continuous map $M \to N$ with continuous inverse. Compact connected orientable surfaces are characterised, up to homeomorphism, by the *genus* $g \in \mathbb{N}$. A sphere is genus 0, a torus genus 1, and the double-torus of Figure 2.2 is genus 2. The surface of a wine glass or fork is topologically a sphere, while coffee mugs and keys are (usually) tori. A ladder with $n$ rungs has genus $n - 1$. The surface of a pair of pants is genus 2, while that of a sweater is genus 3.

A torus can be realised in many different ways. One is the Cartesian product $S^1 \times S^1$ of circles (lay one circle horizontally, then from each point on it place a vertical circular rib perpendicular to it, filling out the torus's surface). A complex curve of the form $y^2 = ax^3 + bx^2 + cx + d$ is a torus (at least if the points at infinity are included), as is the quotient $\mathbb{C}/L$ of the complex plane with a two-dimensional lattice $L$ (Section 1.2.1).

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

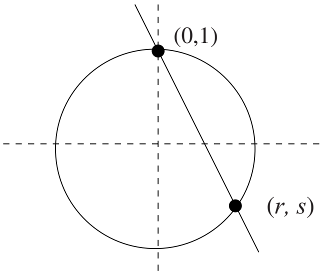

Fig. 2.4 Diophantus' argument.

If we drop the requirement that our surface be *compact*, then up to homeomorphism it is uniquely specified by two numbers: the genus $g$ as above, and the number of punctures (or boundary components) $n$. For instance, a sphere with one puncture is homeomorphic to an open disc or equivalently the plane $\mathbb{C}$. We see this when we pop a balloon: the sphere becomes a rather jagged-edged disc. A sphere with two punctures is a cylinder or annulus.

The *non-orientable* surfaces have a very similar classification. For example, if we could create a $\mathbb{P}^2(\mathbb{R})$-shaped balloon, then popping it would create a jagged-edged Möbius band. We always require orientability in this book.

The surfaces we encounter have more structure than mere topology. If the surface $\Sigma$ is in fact *smooth* (Section 1.2.2), then we are interested in their classification up to *diffeomorphism*. In this case though nothing changes, the surface is again parametrised by the genus and number of punctures: any surface $\Sigma$ has a unique differential structure compatible with its topology. In order to obtain a finer distinction between the surfaces, we need to further enrich their structure. The easiest way to do this is by introducing a metric onto the tangent spaces, or give the surface a complex or conformal structure. More on the resulting *Riemann surfaces* shortly. Nevertheless, the genus remains the single most important invariant distinguishing Riemann surfaces. There are many qualitative differences captured by genus – we will give three of them.

Diophantus [**45**] was a mathematical giant who lived in Alexandria in the second or third century A.D. He seems to have been the first Greek to regard fractions as legitimate numbers, and he was the first to use negative numbers (though only in intermediate arithmetical calculations, so probably didn't believe their ontological reality), and the first to invent an abstract symbolism for algebra. The following (expressed in modern language) is how Diophantus found all Pythagorean triples, that is the integer solutions to $a^2 + b^2 = c^2$.

First, it's enough to look for all rational solutions to the circle $x^2 + y^2 = 1$. Then the integers $a, b, c$ can be recovered by clearing denominators. Consider a line through the point $(0, 1)$ that intersects the circle at another rational point $(r, s)$ (see Figure 2.4). Clearly this line must have rational (or infinite) slope $\frac{s-1}{r}$. Conversely, consider any line through (0,1) with rational slope $u$: its equation will be $y = ux + 1$. Where does it intersect the circle? We get $1 = x^2 + (ux + 1)^2 = (u^2 + 1)x^2 + 2ux + 1$, i.e. $x\left((u^2 + 1)x + 2u\right) = 0$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

So apart from our original point (0, 1), it will also intersect the circle at

$$(x, y) = \left( \frac{-2u}{u^2+1}, \frac{1-u^2}{u^2+1} \right).$$

As long as $u$ is rational, so will be this point. Thus Diophantus found a parametrisation of all rational points on the circle, and hence all Pythagorean triples.

His method is far more general than this, as he knew. In fact, consider any nondegenerate conic. To find all rational points on it, we first find one rational point, and then consider all lines with rational slope through that point. This will exhaust all rational points on the curve. Thus if a conic has one rational point (it might have none), then it will have infinitely many, and all can be found explicitly.

Why won't this trick work for other equations of this sort? For example, Fermat's Last Theorem challenges us to find a nontrivial rational solution to $x^n + y^n = 1$, for $n > 2$. If we draw a line through the obvious solution $(x, y) = (0, 1)$, we simply get a mess. What's so special, geometrically, about conics?

The modern way (due to Bezout in the eighteenth century) to think about this is to regard the given equation, say $x^2 + y^2 = 1$, as an equation relating two complex numbers $(x, y) \in \mathbb{C}^2$. The result will be a complex curve, that is a real surface. To which complex curve does $x^2 + y^2 = 1$ correspond? The *real* curve (a circle) is parametrised by $x = \cos\theta$ and $y = \sin\theta$, and a moment's deliberation will convince oneself that permitting $\theta$ to take complex values will exhaust all points on the *complex* curve. So write $x = \frac{1}{2}(w + w^{-1})$ and $y = \frac{i}{2}(w - w^{-1})$ for any $w \in \mathbb{C}$ except $w = 0$; this identifies the complex curve $x^2 + y^2 = 1$ with the complex plane punctured at 0, that is a cylinder. The unit circle in $\mathbb{R}^2$ is merely the slice of this cylinder in $\mathbb{C}^2$ by the plane passing through the two real axes of $\mathbb{C}^2$. A different slice will produce, for instance, an hyperbola.

More generally, any polynomial in $x, y$ defines a noncompact surface in $\mathbb{C}^2$. For example, a nondegenerate cubic $y^2 = x^3 + ax^2 + bx + c$ is a once-punctured torus – explicitly, the quotient $\mathbb{C}'/(\mathbb{Z} + \tau\mathbb{Z})$, where $\mathbb{C}'$ means deleting from $\mathbb{C}$ the lattice points $\mathbb{Z} + \tau\mathbb{Z}$, is equivalent in every sense one could want (e.g. conformally) to the cubic

$$y^2 = 4x^3 - 60G_4(\tau)x - 140G_6(\tau),$$

where the Eisenstein series $G_k(\tau)$ is defined in (0.1.5). Similarly, the complex curve $x^3 + y^3 = 1$ corresponds to the torus $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ with three points removed.

In any case, we can now answer our question: What is so special geometrically about the conics, that Diophantus' method works for them? The answer: They are (punctured) spheres, that is have genus 0.

It will always seem that some points 'at infinity' are missing from these complex curves. Kepler back in 1604 knew that adding such points simplifies the geometry. We do this by *projectifying* the given equation (Section 1.2.2). For example, $x^2 + y^2 = 1$ corresponds to the homogeneous equation $x^2 + y^2 = z^2$, where we identify $(x, y, z)$ and $(\lambda x, \lambda y, \lambda z)$ for $\lambda \neq 0$. The two 'infinite' points, that is the points with $z = 0$, are then $(1, \pm 1, 0)$. Similarly, the three missing points on the Fermat curve $x^3 + y^3 = 1$ have homogeneous coordinates $(x, y, z) = (1, -\xi, 0)$ for any third root of unity $\xi$. We see that in homogeneous coordinates the 'infinite points' don't look so bad.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

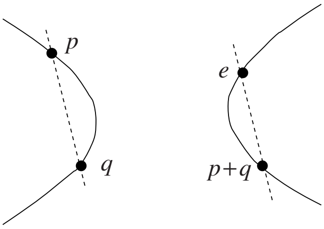

Fig. 2.5 Addition of points on a hyperbola.

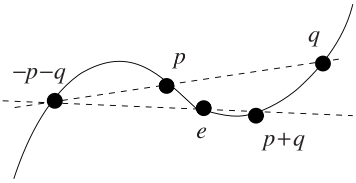

Fig. 2.6 Addition of points on a cubic.

Another special property of conics (avoiding the infinite points) is that they are additive groups. Fix any point $e$ on the conic $C$ (it will be the identity); given any two finite points $p, q$ on the conic, the sum $p + q \in C$ is defined to be the intersection with $C$ of the line through $e$ parallel to the line through $p$ and $q$ (Figure 2.5). Associativity follows from Pascal's Theorem concerning hexagons inscribed in conics. For example, choosing the identity $e = (1, 0)$ and the parametrisation $(x(t), y(t)) = (\cos(t), \sin(t))$ of the circle $x^2 + y^2 = 1$, this addition of points corresponds to addition of angle $t$. The same conclusion holds for the hyperbola $x^2 - y^2 = 1$, with $e = (1, 0)$ and parametrisation $t \mapsto (\cosh(t), \sinh(t))$ of the $x > 0$ branch. See Question 2.1.3.

Better known is the addition of points on a nondegenerate (projective) cubic $C$. Fix any $e \in C$ (again it will play the role of identity), and choose any points $p, q \in C$. Let $r \in C$ be the intersection with $C$ of the line through $p, q$; the sum $p + q$ is defined to be $-r$, that is the intersection with $C$ of the line through $r$ and $e$ (see Figure 2.6). This also is commutative and associative, provided we include the points at infinity. Addition continues to work when the cubic is complexified, and that's how to make sense of it: the resulting surface is a torus, equivalent to one of the form $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ for some $\tau \in \mathbb{C}$, and this addition on the cubic lifts to ordinary addition on $\mathbb{C}$. Incidentally, the addition of points is only one of a number of senses in which conics are toy models for the much richer theory of elliptic curves (i.e. cubics with a marked point $e$) [**372**].

The simplest quantitative distinction between surfaces of different homeomorphism type $(g, n)$ is the fundamental group $\pi_1$, defined in Section 1.2.3. For example, $\pi_1(S^2) = 1$ since $S^2$ is simply connected, and $\pi_1$ of a torus is $\mathbb{Z} \oplus \mathbb{Z}$. Let $\Sigma_g$ be a compact genus $g > 0$ surface. Then $\pi_1(\Sigma_g)$ has presentation

$$\pi_1(\Sigma_g) \cong \left\langle \alpha_1, \ldots, \alpha_g, \beta_1, \ldots, \beta_g \,|\, \alpha_1\beta_1\alpha_1^{-1}\beta_1^{-1} \cdots \alpha_g\beta_g\alpha_g^{-1}\beta_g^{-1} = 1 \right\rangle. \qquad (2.1.5a)$$

The generators $\alpha_i, \beta_j$ are chosen as in Figure 2.2 ($\alpha_1 = a, \beta_1 = b$, etc.). The easiest way to read off the genus from (2.1.5a) is to compute the abelianisation $\pi_1/[\pi_1, \pi_1]$ (which

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

equals incidentally the first homology group $H_1(\Sigma_g, \mathbb{Z})$); as is clear from (2.1.5a), it is the abelian group $\mathbb{Z}^{2g}$ generated by $\alpha_i, \beta_j$. On the other hand, the fundamental group of a genus-$g$ surface $\Sigma_{g,n}$ with $n > 0$ punctures is free (see e.g. page 64 of [**103**]):

$$\pi_1\left(\Sigma_{g,n}\right) \cong \mathcal{F}_{2g+n-1}. \tag{2.1.5b}$$

The preceding discussion indicates the significance of genus. Now let's impose more structure. A *Riemann surface* is a connected orientable surface with a *conformal structure*, together with a choice of orientation. Equivalently, a Riemann surface can be defined as a complex analytic curve: any polynomial equation in $x, y \in \mathbb{C}$ inherits the conformal and differential structure of $\mathbb{C}$. This is because locally the conformal maps in $\mathbb{R}^2$ are precisely the locally holomorphic maps in $\mathbb{C}$ with nonvanishing derivative (theorem 14.2 of [**481**]). A third possible definition is that Riemann surfaces consist of those connected 2-manifolds with a complete metric with constant curvature. As mentioned above, its homeomorphism class is given by its genus $g$ and number of punctures $n$, and the surface is compact iff $n = 0$. We are primarily interested in compact Riemann surfaces.

Any topological surface can be made into a Riemann surface, usually in a continuum of inequivalent ways (Section 2.1.4). We identify two Riemann surfaces if they are conformally equivalent, or holomorphically equivalent, or isometric. In Section 2.1.4 we discuss the classification of Riemann surfaces up to conformal equivalence.

The basic example of a Riemann surface is the complex plane $\mathbb{C}$. Also important is the complex projective line $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$; stereographic projection verifies that it is topologically a sphere, called the *Riemann sphere*. Now, a *meromorphic function* $f : D \to \mathbb{C}$ by definition is holomorphic everywhere except for isolated poles; if $f$ has poles at $z_i$, then defining $f(z_i) = \infty$ gives a *conformal* map $f : D \to \mathbb{P}^1(\mathbb{C})$ between Riemann surfaces (perhaps it is this picture, in which $z_i$ is sent to the 'north pole' $\infty$, which is the origin of the term 'pole'). Likewise, we can extend the *domain* of a function $f$ on $\mathbb{C}$ to $\mathbb{P}^1(\mathbb{C})$, provided it is meromorphic at $\infty$. For example, if $p$ is a polynomial of degree $n$, then $p$ has a pole of degree $n$ at $\infty$, and we obtain a holomorphic map $p : \mathbb{P}^1(\mathbb{C}) \to \mathbb{P}^1(\mathbb{C})$. By comparison, the functions $e^z$ and $\cos(z)$ have essential singularities at $\infty$ and so cannot be extended to $\mathbb{P}^1(\mathbb{C})$.

Historically, Riemann surfaces were introduced by Riemann to supply the maximal domain (via analytic continuation) of a holomorphic function. The problem is that many of the most natural complex functions are multivalued, for example $f(z) = \sqrt{z}$ or $g(z) = \log z$ or other inverses of nice functions. As we move counterclockwise along the unit circle $|z| = 1$, starting at $z = 1$, the value $f(z) = \sqrt{z}$ changes continuously from $f(1) = 1$ to $f(1) = -1$, and the value of $g(z) = \log z$ changes continuously from $g(1) = 0$ to $g(1) = 2\pi\mathrm{i}$. To Riemann, we should regard $f(z)$ as a holomorphic function on a double cover $D = D^b \cup D^t$ of the complex plane, and $g(z)$ is holomorphic on a helix. As we move along the circle, the argument $z$ of $f(z)$ moves from the bottom sheet $D^b \cong \mathbb{C}$ to the top sheet $D^t \cong \mathbb{C}$, and if we continue a second time around the circle, we return from the sheet $D^t$ to $D^b$. To identify $D$ homeomorphically, cut both $D^b$ and $D^t$ from 0 to $\infty$, and glue the $\theta = 0^+$ slit of $D^b$ to the $\theta = 0^-$ slit of $D^t$ and vice versa. The result is homeomorphic to a sphere with one puncture, corresponding to the point at infinity.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Note that $f : D \to \mathbb{C}$ is well-defined and holomorphic; it is an example of what we will shortly call a cover of $\mathbb{C}$, ramified at $z = 0$.

The remainder of this subsection describes an important realisation (called *uniformisation*) of any Riemann surface. The idea is simple. There are two different connected real curves, up to homeomorphism, and they are the line $\mathbb{R}$ and the circle $S^1$. The circle can be realised as $S^1 \cong \mathbb{R}/\mathbb{Z}$. We call $\mathbb{R}$ the 'universal cover' $\widetilde{S^1}$ of $S^1$, because it is simply-connected; $\mathbb{Z}$ here is the fundamental group $\pi_1(S^1)$. See also Theorem 1.4.3.

The same works with surfaces. For example, the sphere with two punctures (a cylinder) and a torus both have universal cover homeomorphic to $\mathbb{C}$; the cylinder itself is homeomorphic to $S^1 \times \mathbb{R}$ and the torus to $\mathbb{C}/(\mathbb{Z} + \mathrm{i}\mathbb{Z})$, where $\mathbb{Z}$ and $\mathbb{Z} + \mathrm{i}\mathbb{Z}$ are isomorphic to their fundamental groups. Let's make these ideas more precise, and incorporate as well the conformal structure.

**Definition 2.1.3** *Let $\Sigma^*$, $\Sigma$ be two Riemann surfaces. We say that $\Sigma^*$* covers *$\Sigma$ by $f$ if $f : \Sigma^* \to \Sigma$ is a holomorphic map from $\Sigma^*$ onto $\Sigma$. If in addition $f$ is locally conformal, we call $f$ a* conformal *or* unramified cover. *If $f : \Sigma^* \to \Sigma$ is a conformal cover, and $\Sigma^*$ is simply-connected, then we call $\Sigma^*$ a* universal cover *of $\Sigma$.*

Let $U_\alpha \subset \Sigma$, $\varphi_\alpha : U_\alpha \to V_\alpha \subset \mathbb{C}$ be a family of coordinate charts for $\Sigma$ (Definition 1.2.3); by *local coordinates* we mean the complex numbers $z \in V_\alpha$. In local coordinate $z$ about point $p^* \in \Sigma^*$, a cover $f$ sends a neighbourhood of $p^*$ to one of $f(p^*) \in \Sigma$ with local coordinates $a + cz^n +$ higher terms, for some constants $a$ and $c \neq 0$. To be conformal, this order $n$ must always be 1 (otherwise we say $f$ is ramified at $p^*$).

If $f : \Sigma^* \to \Sigma$ is a conformal cover, then the fundamental group $\pi_1(\Sigma^*)$ is naturally isomorphic to a subgroup of $\pi_1(\Sigma)$ (Section 1.7.2). In this way, the covers $\Sigma^*$ of $\Sigma$ (up to homeomorphism) are in one-to-one correspondence with conjugacy classes of subgroups of $\pi_1(\Sigma)$. A universal cover $\widetilde{\Sigma}$ is the 'largest' and most important cover, and is unique up to conformal equivalence. It can be identified as the space of all homotopy-equivalence classes of paths on $\Sigma$ with fixed initial point $p \in \Sigma$. For example, visualise a 'point' $\tilde{p}$ on $\widetilde{S^1}$ as a curve starting at $1 \in S^1$ and ending at $e^{\mathrm{i}\theta}$ $(0 \leq \theta < 2\pi)$, and wrapping around the circle (i.e. crossing $1 \in S^1$) $n$ times; the identification of $\widetilde{S^1}$ with $\mathbb{R}$ comes from identifying this path with the number $\theta + 2\pi n \in \mathbb{R}$.

We are now ready to state the basic result of this subsection.

**Theorem 2.1.4 (Uniformisation Theorem)**

(a) *Up to conformal equivalence, the only simply-connected Riemann surfaces (i.e. the only candidates for a universal cover) are the sphere $S^2 = \mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$, the plane $\mathbb{C}$ and the upper half-plane $\mathbb{H}$.*

(b) *Let $\Sigma$ be any Riemann surface, and let $\widetilde{\Sigma}$ be its universal cover. Then $\Sigma$ is conformally equivalent to $\widetilde{\Sigma}/\Gamma$, where $\Gamma \cong \pi_1(\Sigma)$ is a subgroup of the automorphisms of $\widetilde{\Sigma}$ that act on $\widetilde{\Sigma}$ without fixed points. A metric can be chosen for $\Sigma$ with constant curvature $+1, 0, -1$, respectively, if $\widetilde{\Sigma} = S^2, \mathbb{C}, \mathbb{H}$, respectively. Two surfaces $\widetilde{\Sigma}/\Gamma$, $\widetilde{\Sigma}'/\Gamma'$ are conformally equivalent iff the universal covers $\widetilde{\Sigma}$ and $\widetilde{\Sigma}'$ are the same, and $\Gamma$ and $\Gamma'$ are conjugate subgroups in* $\mathrm{Aut}(\widetilde{\Sigma})$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Table 2.1. *The universal covers of the genus g surfaces with n punctures*

| $g\backslash n$ | 0 | 1 | 2 | $\geq 3$ |
|---|---|---|---|---|
| 0 | $S^2$ | $\mathbb{C}, \mathbb{H}$ | $\mathbb{C}, \mathbb{H}$ | $\mathbb{H}$ |
| 1 | $\mathbb{C}$ | $\mathbb{H}$ | $\mathbb{H}$ | $\mathbb{H}$ |
| $\geq 2$ | $\mathbb{H}$ | $\mathbb{H}$ | $\mathbb{H}$ | $\mathbb{H}$ |

Of course $\mathbb{H}$ and $\mathbb{C}$ are homeomorphic, but they aren't conformally equivalent (replacing $\mathbb{H}$ with the disc $\mathbb{D}$, this follows from *Liouville's Theorem*: a bounded holomorphic function on $\mathbb{C}$ must be constant). Part (a) is due to Klein, Poincaré and Koebe. These three possibilities for $\widetilde{\Sigma}$ correspond respectively to the three geometries: spherical, Euclidean and hyperbolic. The group of automorphisms of $\widetilde{\Sigma}$ is just $\mathrm{Isom}^+$. The condition that $\Gamma$ acts without fixed points (apart from the identity in $\Gamma$) is significant – fixed points change the geometry. A famous example of an orbit space with fixed points is $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, which has conical singularities at $\mathrm{i}$ and $e^{\pi \mathrm{i}/3}$.

Table 2.1 gives the universal cover of any Riemann surface, as a function of the genus and number of punctures. We see there that almost every surface is hyperbolic: *the generic geometry in two dimensions is hyperbolic.*

The Uniformisation Theorem easily proves *Picard's Theorem* ('the range $f(\mathbb{C})$ of any holomorphic nonconstant function $f : \mathbb{C} \to \mathbb{C}$ can omit at most one point from $\mathbb{C}$'). The proof, which the reader can fill in, uses Liouville's Theorem together with the fact that the universal cover of the twice-punctured plane is $\mathbb{D}$.

### *2.1.3 Functions and differential forms*

The last subsection gives several equivalent notions of a Riemann surface. Here we see that any compact Riemann surface is the locus of a homogeneous polynomial $f(a, b, c) = 0$ in the complex projective plane $\mathbb{P}^2(\mathbb{C})$.

We study a manifold through the functions living on it. Two manifolds differing merely by a single point can have a completely different family of functions. For instance, we all know many examples of holomorphic functions on $\mathbb{C}$. But the only functions holomorphic on $\mathbb{C}$ and also holomorphic at $\infty$ are the constants. More generally, any noncompact Riemann surface $\Sigma$ has several functions $f : \Sigma \to \mathbb{C}$ holomorphic everywhere, while if $\Sigma$ is compact, the only holomorphic functions $f : \Sigma \to \mathbb{C}$ are the constants. We are more interested in compact $\Sigma$.

Given any Riemann surface $\Sigma$, let $\mathcal{K}(\Sigma)$ denote all the meromorphic functions $f : \Sigma \to \mathbb{C}$ – equivalently, all holomorphic functions $f : \Sigma \to \mathbb{P}^1(\mathbb{C})$ (by convention we discard the constant function $f \equiv \infty$). Let $U_\alpha \subset \Sigma$, $\varphi_\alpha : U_\alpha \to V_\alpha \subset \mathbb{C}$ be a family of coordinate charts for $\Sigma$. Then $f \in \mathcal{K}(\Sigma)$ iff each $f \circ \varphi_\alpha^{-1}$ is a meromorphic function of the local coordinate $z \in V_\alpha$.

For example, $\mathcal{K}(\mathbb{P}^1(\mathbb{C}))$ consists of all rational functions $f(z) = \frac{\mathrm{poly}(z)}{\mathrm{poly}(z)}$, while $\mathcal{K}(\mathbb{C})$ is much larger. This space $\mathbb{K}(\Sigma)$ is in fact always a field; its algebraic structure determines the surface $\Sigma$ (up to conformal equivalence) and naturally mirrors all aspects of $\Sigma$. A

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

compact Riemann surface $\Sigma$ has genus 0 iff $\mathbb{K}(\Sigma) \cong \mathbb{C}(z)$, the field of rational functions in some variable $z$. For positive genus, two generators are needed.

**Theorem 2.1.5** *Let $\Sigma$ be a compact Riemann surface of genus $g > 0$. Choose any nonconstant function $f \in \mathcal{K}(\Sigma)$. Then there exists another nonconstant function $g \in \mathcal{K}(\Sigma)$, such that $\mathcal{K}(\Sigma) = \mathbb{C}(f)[g]$, i.e. for some $n \in \mathbb{N}$, any $h \in \mathcal{K}(\Sigma)$ can be written in the form $h = \sum_{i=0}^{n-1} a_i(f)\, g^i$, where $a_i(z)$ are rational. Moreover, there is an irreducible polynomial $P(z, w)$ such that $P(f, g) = 0$, and such that $\mathcal{K}(\Sigma)$ is isomorphic as a field to the quotient $\mathbb{C}(z, w)/(P(z, w))$ of the algebra of rational functions in $z, w$ by the ideal generated by polynomial $P$. Moreover, writing $P$ as a homogeneous polynomial in three variables, $\Sigma$ is conformally equivalent to the complex curve $P = 0$ in the complex projective plane $\mathbb{P}^2(\mathbb{C})$.*

For a proof and more material on Riemann surfaces, see [**180**]. It is nontrivial that we can embed any Riemann surface into the complex projective plane. In fact, most complex $n$-tori $\mathbb{C}^n/L$ (where $L \subset \mathbb{C}^n$ is a $2n$-dimensional lattice), for $n > 1$, *cannot* be embedded in *any* projective space (Section 6.3.2). The plane curve $P = 0$ will typically have ‘singularities’, that is points where all three partial derivatives vanish, where the curve self-intersects transversely. These singularities can be ‘blown up’, that is the two intersecting ‘complex strands’ (i.e. open discs in $\mathbb{C}$) can be separated, but this requires the complex curve to be embedded in $\mathbb{P}^3$, not $\mathbb{P}^2$.

Every geometric feature (except the choice of orientation) of the surface $\Sigma$ has an algebraic analogue in $\mathcal{K}(\Sigma)$, and hence the geometry of $\Sigma$ can be studied via algebra. For example, a $\mathbb{C}$-algebra homomorphism $F : \mathcal{K}(\Sigma') \to \mathcal{K}(\Sigma)$ lifts to a holomorphic map $\widetilde{F} : \Sigma \to \Sigma'$. This general observation is the starting point of both algebraic geometry and noncommutative geometry. For example, the space of smooth complex-valued functions on a manifold $M$ will be an infinite-dimensional commutative algebra, since the target $\mathbb{C}$ is a commutative algebra. Connes suggests that we study a noncommutative algebra as if it too is the algebra of functions on some manifold. The hope is that this should be directly relevant to quantum theories, since we access space-time only indirectly, via the functions (‘quantum fields’) living on it. We seem to get into problems in quantum field theory when we take too literally the (naive and improbable) intuition that space-time is anything like a manifold. In any case calculus in noncommutative geometry formally resembles quantum mechanics (e.g. the role of coordinates is played by self-adjoint operators – observables – and infinitesimal distance d$s$ by the fermion propagator).

For a concrete example of Theorem 2.1.5, consider the torus $T_\tau = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$. A meromorphic function $f : T_\tau \to \mathbb{C}$ lifts to a meromorphic function (which we also call $f$) on $\mathbb{C}$, with periods 1 and $\tau$. That is, $f \in \mathcal{K}(T_\tau)$ iff $f : \mathbb{C} \to \mathbb{C}$ is meromorphic and $f(z + m + n\tau) = f(z)\ \forall z \in \mathbb{C}, \forall m, n \in \mathbb{Z}$. Any such doubly-periodic meromorphic function is called an *elliptic function*, for fairly obscure reasons.[4] We know

[4] One of the more carefree creative outlets for mathematicians is through their happy role as nomenclators. Elliptic functions first arose historically as the functional inverse of a certain class of integrals called ‘elliptic integrals’. This class got its name since it included the integral computing arc-lengths of ellipses. Likewise, the name ‘elliptic curve’ for a genus-1 complex curve arose since the functions living on it are those elliptic functions. There is however no direct relation between ellipses and elliptic curves.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

any nonconstant $f \in \mathcal{K}(T_\tau)$ must have at least one pole in the 'fundamental parallelogram' $P_\tau$ with corners at $0, \tau, 1, 1+\tau$. Moreover, the contour integral $\int_C f$ about the parallelogram $C = \partial P_\tau$ vanishes by periodicity, so the sum of residues of $f$ inside $P_\tau$ must vanish. Hence any nonconstant elliptic function must have at least two poles in $P_\tau$.

We can construct an elliptic function by averaging $f(z) = \sum_{m,n} g(z + m + n\tau)$ for any function $g$ over each orbit $z + \mathbb{Z} + \tau\mathbb{Z}$. As the simplest possibility for a nonconstant elliptic function would have a single pole of order 2 at the lattice points, it is tempting to take $g(z) = z^{-2}$. Unfortunately, for large $m, n$, $(z + m + n\tau)^{-2}$ is close to $(m + n\tau)^{-2}$, and so its sum over all $m, n$ won't converge. Thus we are led to consider its 'regularisation'

$$\mathfrak{p}(z) := z^{-2} + \sum_{m,n=-\infty}^{\infty}{}'\{(z+m+n\tau)^{-2} - (m+n\tau)^{-2}\} \tag{2.1.6a}$$

function, called the *Weierstrass function* (although Eisenstein knew of it years earlier), where $\sum'$ here means to avoid $m = n = 0$. Its derivative

$$\mathfrak{p}'(z) = -2 \sum_{m,n=-\infty}^{\infty} (z+m+n\tau)^{-3} \tag{2.1.6b}$$

is also elliptic. Being meromorphic functions on a compact Riemann surface, $\mathfrak{p}$ and $\mathfrak{p}'$ must be polynomially related: we find

$$\mathfrak{p}'(z)^2 = 4(\mathfrak{p}(z) - e_1)(\mathfrak{p}(z) - e_2)(\mathfrak{p}(z) - e_3), \tag{2.1.6c}$$

where $e_1 = \mathfrak{p}(1/2)$, $e_2 = \mathfrak{p}(\tau/2)$ and $e_3 = \mathfrak{p}((1+\tau)/2)$. This is shown by verifying that $(\mathfrak{p} - e_1)(\mathfrak{p} - e_2)(\mathfrak{p} - e_3)/\mathfrak{p}'$ has no poles and hence must be constant. Together, $\mathfrak{p}$ and $\mathfrak{p}'$ generate $\mathcal{K}(T_\tau)$: we can write any elliptic function $f \in \mathcal{K}(T_\tau)$ as $R_1(\mathfrak{p}) + \mathfrak{p}' R_2(\mathfrak{p})$, where $R_1(\mathfrak{p}(z))$ is the even part $(f(z) + f(-z))/2$ of $f$ and $\mathfrak{p}'(z)\,R_2(\mathfrak{p}(z))$ the odd part. $T_\tau$ is conformally equivalent to the projective curve with 'finite' points $(\mathfrak{p}(z), \mathfrak{p}'(z), 1) \in \mathbb{P}^2(\mathbb{C})$, together with the 'infinite' point $(0, 1, 0)$ corresponding to the pole of $\mathfrak{p}$ and $\mathfrak{p}'$ at $z = 0$.

One way to embed Riemann surfaces into projective space uses *theta functions*:

$$\theta_{r,s}(\tau, z) := \sum_{m \in \mathbb{Z}} \exp[\pi \mathrm{i} \tau\,(m+r)^2 + 2\pi \mathrm{i}\,(m+r)(z+s)], \tag{2.1.7a}$$

for any $r, s \in \mathbb{Q}$. These functions and their generalisations are central to Moonshine, but for now note that they converge for all $(\tau, z) \in \mathbb{H} \times \mathbb{C}$ to a function holomorphic in both $\tau$ and $z$. These $\theta_{r,s}$ are nearly doubly-periodic in $z$: if $r, s \in \frac{1}{N}\mathbb{Z}$ then

$$\theta_{r,s}(\tau, z + Nm + \tau Nn) = \exp[-\pi \mathrm{i} N^2 n^2 \tau - 2\pi \mathrm{i} N n z]\,\theta_{r,s}(\tau, z), \tag{2.1.7b}$$

for all $m, n \in \mathbb{Z}$. Apart from a constant root of unity, $\theta_{r,s}$ depends only on the values of $r$ and $s$ mod 1. Enumerate the $N^2$ pairs $(r_i, s_i) \in \frac{1}{N}\mathbb{Z}_N \times \frac{1}{\mathbb{Z}_N}\mathbb{Z}_N$. Then for any $N$ and any

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

$\tau \in \mathbb{H}$, the map from $T_\tau$ to $\mathbb{P}^{N^2-1}(\mathbb{C})$ defined in homogeneous coordinates by

$$z \mapsto (\theta_{r_1,s_1}(\tau, Nz), \theta_{r_2,s_2}(\tau, Nz), \ldots) \in \mathbb{P}^{N^2-1}(\mathbb{C})$$

is well defined (to see that this $N^2$-tuple can never be the 0-vector, find explicitly the zeros of $\theta_{r,s}$). This map is one-to-one, that is it embeds the torus $T_\tau$ as a complex submanifold of $\mathbb{P}^{N^2-1}(\mathbb{C})$. We can specify this submanifold more explicitly (in the simplest case, namely $N = 2$) by the homogeneous polynomials

$$\theta_{0,0}(\tau)^2 z_1^2 = \theta_{0,1/2}(\tau)^2 z_2^2 + \theta_{1/2,0}(\tau)^2 z_3^2, \qquad \theta_{0,0}(\tau)^2 z_4^2 = \theta_{1/2,0}(\tau)^2 z_2^2 - \theta_{0,1/2}(\tau)^2 z_3^2,$$

where $(z_1, z_2, z_3, z_4) \in \mathbb{P}^3(\mathbb{C})$ are homogeneous coordinates and $\theta_{r,s}(\tau) = \theta_{r,s}(\tau, 0)$. The fact that the image of $T_\tau$ satisfies those equations follows from the Riemann theta identities. Moreover, any elliptic function $f : T_\tau \to \mathbb{C}$ can be written in the form

$$f(z) = c \prod_{1 \le i \le \ell} \frac{\theta_{0,0}(\tau, z - a_i)}{\theta_{0,0}(\tau, z - b_i)},$$

for arbitrary complex numbers $a_i, b_i, c$ subject to the relation $\sum_i a_i = \sum_i b_i$. The Weierstrass $\mathfrak{p}$-function can be written

$$\mathfrak{p}(z) = -\frac{\mathrm{d}^2}{\mathrm{d}z^2}\theta_{1/2,1/2}(\tau, z) - \frac{\pi^2}{3}.$$

For any $k \in \mathbb{Z}$, a holomorphic (respectively meromorphic) $k$-form $\omega$ (Section 1.2.2) on a complex curve $\Sigma$ looks like $f\,\mathrm{d}z^k$ in local coordinates, where $f$ is holomorphic (respectively meromorphic). If we change local coordinates $z_1 \mapsto \varphi_2(\varphi_1^{-1}(z_1))$, then (1.2.4b) becomes

$$f_\beta(z_\beta) = \frac{\mathrm{d}^k z_\alpha}{\mathrm{d}z_\beta^k} f_\alpha(z_\alpha). \tag{2.1.8}$$

For example, $\mathrm{d}z$ is a meromorphic (but not holomorphic) 1-differential on $\mathbb{P}^1(\mathbb{C})$ (it has a pole of order 2 at $\infty$). Let $\mathcal{H}^k(\Sigma)$ be the vector space of holomorphic $k$-forms, and $\mathcal{M}^k(\Sigma)$ be the space of meromorphic ones. Given any $\omega, \omega' \in \mathcal{M}^k(\Sigma)$, $\omega'$ not identically 0, the ratio $\omega/\omega'$ lies in the function field $\mathcal{K}(\Sigma)$. Of course, as vector spaces $\mathcal{M}^0(\Sigma) = \mathcal{K}(M)$. For any surface $\Sigma$ and integer $k$, $\mathcal{M}^k(\Sigma)$ is infinite-dimensional, but for any compact surface $\Sigma$ and any integer $k$, the Riemann–Roch theorem implies that $\mathcal{H}^k(\Sigma)$ is always finite-dimensional and may be 0.

### *2.1.4 Moduli*

In physics, the *phase space* lets us consider all possible states of a physical system; the actual time-evolution of a given instance of that system will be a curve in phase space. Likewise, we often want to consider simultaneously families of manifolds, rather than fix a single manifold. For example, last subsection we treated all tori $T_\tau$ simultaneously. The role of phase space is played by a *moduli space*, the space of orbits of a group of diffeomorphisms of a geometric structure placed on a manifold. A path on the moduli

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

space connecting orbits $[p]$ and $[q]$ is a continuous deformation from the geometric structure on $p$ to that on $q$.

The notion of moduli space for surfaces is due to Riemann, who also computed its dimension. The idea is to consider the space $\mathfrak{M}(\Sigma_0)$ of all conformal equivalence classes of Riemann surfaces homeomorphic to a given surface $\Sigma_0$. As $\Sigma_0$ is completely characterised by the genus $g$ and number $n$ of punctures, we also denote this by $\mathfrak{M}_{g,n}$. With a few exceptions mentioned shortly, $\mathfrak{M}_{g,n}$ has complex dimension $3g - 3 + n$. However, these moduli spaces usually aren't manifolds (they have conical singularities). It was for this reason that Teichmüller introduced a cover, now called the *Teichmüller space* $\mathfrak{T}_{g,n}$. The moduli space is recovered by the quotient $\mathfrak{M}_{g,n} = \mathfrak{T}_{g,n}/\Gamma_{g,n}$, where $\Gamma_{g,n}$ is a discrete group called the *mapping class group* (see Definition 2.1.6). Teichmüller space is much better behaved than the moduli space – it is a complex manifold (except for certain small $g, n$), and as a real manifold is diffeomorphic to $\mathbb{R}^{6g-6+2n}$.

As we shall see, there's a small number of pairs $(g, n)$ that don't behave completely generically for one reason or another: namely, (0, 0), (0, 1), (0, 2), (0, 3), (0, 4), (1, 0), (1, 1) and (2,0). We mention some of their individual peculiarities below.

In order to anticipate the definitions, consider a torus $T$ (so $g = 1, n = 0$). For concreteness (this doesn't lose any generality), restrict to tori coming from a parallelogram in the complex plane $\mathbb{C}$, with one pair of opposite sides labelled '1', and the other pair labelled '2'; the torus is recovered by first identifying the opposite sides labelled '1', and then identifying the opposite sides labelled '2' (changing this order changes the shape – though not the conformal class – of the torus). By translating, rotating and rescaling this parallelogram, we can put the vertices at $0, 1, \tau$ and $\tau + 1$, for some $\tau \in \mathbb{H}$, where the horizontal sides are labelled '1', which continuously deforms the torus without changing its conformal equivalence class. This is the best we can do, if we restrict to continuous deformations. The resulting parameter space, namely the upper half-plane $\mathbb{H}$, is the Teichmüller space $\mathfrak{T}_{1,0}$ for the torus. The torus corresponding to $\tau \in \mathbb{H}$ is $T_\tau = \mathbb{C}/(\mathbb{Z} + \mathbb{Z}\tau)$.

However, different points $\tau$ in $\mathbb{H}$ can correspond to conformally equivalent tori. For example, we can cut the torus open along the seam '2', twist the open arm $m$ complete turns, and then sew it back up. This amounts to replacing parameter $\tau$ with $\tau + m$. As long as $m$ is an integer, this is a conformal diffeomorphism of the torus (if $m$ isn't an integer, this map isn't even continuous). Thus the points $\tau + \mathbb{Z}$ all correspond to the same conformal structure. Similarly, cutting open seam '1' and giving the upper cap $n$ complete twists before resewing corresponds to replacing the parallelogram $0$, $1$, $\tau$ and $\tau + 1$ with the parallelogram $0$, $1 + n\tau$, $\tau$ and $(n + 1)\tau + 1$ – after putting it into canonical form, this replaces $\tau$ with $\tau/(n\tau + 1)$. Both these twists are called *Dehn twists*. We can also switch the roles of sides '1' and '2', which replaces $\tau$ with $-1/\tau$ (why?). More generally, the tori corresponding to parameters $\tau$ and $\frac{a\tau+b}{c\tau+d}$ are conformally equivalent, for any $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$. This accounts for all redundancies in the parametrisation by $\mathbb{H}$ of the conformal equivalence classes of tori. The orbit space $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ is the 'moduli space' $\mathfrak{M}_{1,0}$ for the torus. Note that $\mathfrak{M}_{1,0}$ has conical singularities at the orbits

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

$[\tau] = [\mathrm{i}]$ and $[e^{2\pi \mathrm{i}/3}]$, corresponding to those tori with additional automorphisms. This happens in higher genus too. Indeed, any finite group $G$ is the automorphism group of some surface of sufficiently high genus. For example, there will be a compact Riemann surface with automorphism group exactly the Monster $\mathbb{M}$, though it will have genus at least $9.6 \times 10^{51}$.

**Definition 2.1.6** *Let $\Sigma_0$ be a fixed Riemann surface. Consider all pairs $(\Sigma, f)$, where $f$ is an orientation-preserving homeomorphic map of $\Sigma_0$ onto $\Sigma$. Write $(\Sigma, f) \sim (\Sigma', f')$ if there exists a conformal homeomorphism $h : \Sigma \to \Sigma'$ such that the homeomorphism $f'^{-1} \circ h \circ f : \Sigma_0 \to \Sigma_0$ is homotopic to the identity. The set of these equivalence classes is the* Teichmüller space $\mathfrak{T}(\Sigma_0)$. *The* mapping class group $\Gamma(\Sigma_0)$ *is the quotient* $\mathrm{Homeo}_+(\Sigma_0)/\mathrm{Homeo}_0(\Sigma_0)$ *of the group of orientation-preserving self-homeomorphisms $f$ of $\Sigma_0$, by the (normal) subgroup consisting of those homotopic to the identity.*

For example, $\Gamma_{1,0} = \mathrm{SL}_2(\mathbb{Z})$ and $\mathfrak{T}_{1,0} = \mathbb{H}$; as we explain in Section 2.2.4, the moduli space $\mathfrak{M}_{1,0}$ is a punctured sphere. Because $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ can also be interpreted as a torus with a special point, namely the additive identity 0, we also have $\mathfrak{T}_{1,1} = \mathbb{H}$ and $\Gamma_{1,1} = \mathrm{SL}_2(\mathbb{Z})$. For a different reason, we also have $\mathfrak{T}_{0,4} = \mathbb{H}$ and $\Gamma_{0,4} = \mathrm{SL}_2(\mathbb{Z})$.

The basic idea, illustrated above, is that the Teichmüller space $\mathfrak{T}_{g,n}$ accounts for ‘continuous’ conformal equivalences, while the mapping class group $\Gamma_{g,n}$ contains the left-over ‘discontinuous’ ones. To help make this important but abstract definition more accessible, consider the following artificial example. Let $X = \mathbb{R}^2$, and suppose the additive group $G = \mathbb{Z} \times \mathbb{R}$ acts on $X$ by addition. Then $G$ is a disconnected Lie group with connected components $G_n := \{n\} \times \mathbb{R}$ for each $n \in \mathbb{Z}$; the component $G_0$ is the one containing the identity $(0, 0)$. The group $\pi_0 = G/G_0 \cong \mathbb{Z}$ interchanges the components in the obvious way. We can mod out first by the continuous part $G_0$ of $G$ (which should be relatively easy), then by the discontinuous $\pi_0$: the orbit space $X/G$ is then $(X/G_0)/\pi_0 = \mathbb{R}/\mathbb{Z} = S^1$. Of course, here $X$ plays the role of the infinite-dimensional space of all conformal structures, $G$ plays the role of all conformal homeomorphisms, and $X/G$ is the moduli space. The identity component $G_0$ corresponds to the homeomorphisms homotopic to the identity, $\pi_0$ is the mapping class group and $X/G_0$ is the Teichmüller space.

The mapping class groups are central to our story, so we’ll try to make them more accessible. More details and proofs are provided in [**56**], [**270**], [**60**] and chapter 4 of [**59**]. A simple presentation of the mapping class group $\Gamma_{g,n}$ for $n = 0, 1$ – the cases of greatest interest to us – is given in [**550**].

$\Gamma_{g,n}$ acts like a braid group. For example, any $f \in \mathrm{Homeo}_+(\Sigma)$ permutes the $n$ punctures, so the same is true of $\gamma \in \Gamma_{g,n}$; the ‘pure’ mapping class group $\mathrm{P}\Gamma_{g,n}$ consists of those $\gamma \in \Gamma_{g,n}$ that fix each puncture. Then $\mathrm{P}\Gamma_{g,n}$ is normal in $\Gamma_{g,n}$ and has quotient $\Gamma_{g,n}/\mathrm{P}\Gamma_{g,n} = \mathcal{S}_n$.

A braid group $\mathcal{B}_n(\Sigma)$ can be associated with any surface $\Sigma$ in the obvious way [**59**]. For genus $g \geq 2$ and any $n \geq 0$, the group $\Gamma_{g,n}$ is an extension of $\mathcal{B}_n(\Sigma_g)$, by the group $\Gamma_{g,0}$. For genus $g = 1$ and $n \geq 2$, $\Gamma_{1,n}$ is an extension of the quotient $\mathcal{B}_n(\Sigma_1)/Z(\mathcal{B}_n(\Sigma_1))$ by $\mathrm{PSL}_2(\mathbb{Z})$, where the centre $Z(\mathcal{B}_n(\Sigma_1)) \cong \mathbb{Z}^2$. For genus $g = 0$ and $n \geq 3$, the group

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

$\Gamma_{0,n}$ is isomorphic to the quotient $\mathcal{B}_n(S^2)/Z(\mathcal{B}_n(S^2))$, where $Z(\mathcal{B}_n(S^2)) \cong \mathbb{Z}_2$. For any $n$, $\Gamma_{0,n}$ is a homomorphic image (i.e. a quotient) of the braid group $\mathcal{B}_n$.

Let $\Sigma$ be a compact Riemann surface. To any simple closed loop $\gamma$ in $\Sigma$, we can define the *Dehn twist* about $\gamma$, by cutting out from $\Sigma$ a neighbourhood of the loop homeomorphic to a cylinder, giving one end of this cylinder an integral twist, and gluing it back. The Dehn twists about the $2g$ elementary loops $a_i, b_j$ defined in Section 2.1.2 generate the mapping class group of $\Sigma$.

Teichmüller space need not be connected. In particular, there are three different kinds of twice-punctured spheres: one is flat and has conformal structure given by the cylinder $\mathbb{C}/\mathbb{Z}$; one is the punctured disc $0 < |z| < 1$ and corresponds to the half-cylinder $\mathbb{H}/\langle z \mapsto z+1\rangle$; and finally, we have the family of annuli $A_r := \{r < |z| < 1\}$, which are all of the form $\mathbb{H}/\langle z \mapsto \lambda z\rangle$ for $\lambda > 1$. Thus $\mathfrak{T}_{0,2}$ and $\mathfrak{M}_{0,2}$ consist of two isolated points and an open line segment $(0, 1)$ say. $\Gamma_{0,2} \cong \mathbb{Z}_2$ consists of the identity, and the inversion through 0 that exchanges the two boundary circles. Similarly, both $\mathfrak{T}_{0,1}$ and $\mathfrak{M}_{0,1}$ consist of two isolated points.

The mapping class group usually (but not always) acts *faithfully* on Teichmüller space (a faithful action means that the only group element that acts trivially is the identity element). $\Gamma_{1,0} = \Gamma_{1,1} = \Gamma_{0,4}$ are exceptions: $-I \in \mathrm{SL}_2(\mathbb{Z})$ acts trivially on $\mathbb{H}$. Also, consider the thrice-punctured sphere $\mathbb{P}^1(\mathbb{C})/\{z_1, z_2, z_3\}$. As is well known, $\mathrm{Aut}(S^2) \cong \mathrm{PSL}_2(\mathbb{C})$ can conformally move any three points to any other three points, so we can send $z_1, z_2, z_3 \in \mathbb{P}^1(\mathbb{C})$ respectively to $0, 1, \infty$. Thus $\mathfrak{T}_{0,3}$ consists of a single point. However, we could have moved, for example, $z_2, z_1, z_3$ instead to $0, 1, \infty$, respectively. A total of six different choices could have been made, corresponding to the mapping class group $\Gamma_{0,3} = \mathcal{S}_3$, which acts trivially on Teichmüller space.

$\mathfrak{M}_{g,n}$ is simultaneously the moduli space of: (i) conformal equivalence classes of real surfaces; (ii) complete Riemannian metrics of constant negative curvature on real surfaces; and (iii) complex-analytic structures on complex curves. This is an accident of small dimensions, for example the Mostow Rigidity Theorem says that in three dimensions the moduli space of (ii) consists of a single point.

A different approach to moduli spaces ties in with Sections 2.3.5 and 6.3.2. First, by the *Siegel upper half-space* $\mathbb{H}_g$ we mean the space of all symmetric $g \times g$ complex matrices $\Omega$ whose imaginary part $\mathrm{Im}(\Omega)$ is positive-definite – that is, $v^t\,\mathrm{Im}(\Omega)\,v > 0$ for any nonzero column vector $v \in \mathbb{R}^g$. $\mathbb{H}_g$ is a higher-genus generalisation of $\mathbb{H}$. The role of the group $\mathrm{SL}_2(\mathbb{Z})$ here is played by the symplectic group $\mathrm{Sp}_{2g}(\mathbb{Z})$, that is the group of all determinant 1 $2g \times 2g$ matrices $M$ satisfying $M^t \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} M = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$, where $I = I_g$ and 0 are, respectively, the $g \times g$ identity and $g \times g$ zero matrices. The familiar action $\begin{pmatrix} a & b \\ c & d \end{pmatrix}.\tau = \frac{a\tau+b}{c\tau+d}$ is replaced by the action

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}.\Omega = (A\Omega + B)(C\Omega + D)^{-1}, \qquad \forall \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathrm{Sp}_{2g}(\mathbb{R}),\ \forall \Omega \in \mathbb{H}_g. \tag{2.1.9a}$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

The generalisation of the Jacobi theta function (2.1.7a) is *Siegel's theta function*

$$\theta(\Omega, z) = \sum_{n \in \mathbb{Z}^g} \exp(\pi \mathrm{i}\, n^t \Omega n + 2\pi \mathrm{i}\, n \cdot z), \tag{2.1.9b}$$

which converges for all $\Omega \in \mathbb{H}_g$ and $z \in \mathbb{C}^g$.

Where does $\mathbb{H}_g$ come from? Associate with a compact genus-$g$ surface $\Sigma_g$ its Jacobian variety, as follows. The space $\mathcal{H}^1(\Sigma_g)$ of holomorphic 1-forms is $g$-dimensional, so let $\{\omega_1, \ldots, \omega_g\}$ be a basis. Fix any base-point $p \in \Sigma_g$; then we get a map from $\Sigma_g \times \cdots \times \Sigma_g$ to $\mathbb{C}^g$ by integrating:

$$(q_1, \ldots, q_g) \mapsto \sum_{i=1}^{g} \left( \int_{C_i} \omega_1, \int_{C_i} \omega_2, \ldots, \int_{C_i} \omega_g \right),$$

where $C_i$ is any path on $\Sigma_g$ from $p$ to $q_i$. Of course the result depends on which paths $C_i$ are chosen, and so isn't well defined as a function of $q_i$'s alone. However, consider the set $L$ of all possible values $(\int_C \omega_1, \ldots, \int_C \omega_g) \in \mathbb{C}^g$, where $C$ runs over all possible closed loops in $\Sigma_g$ passing through $P$. Then our ill-defined map $\Sigma_g \times \cdots \times \Sigma_g \to \mathbb{C}^g$ will become well-defined (i.e. independent of the choice of path $C_i$) if we replace the target $\mathbb{C}^g$ with $\mathbb{C}^g/L$. It isn't hard to show that $L$ is a $2g$-dimensional lattice (in fact a basis is given by the values on the $2g$ loops we call $\alpha_i, \beta_j$ in (2.1.5a)), and so $\mathbb{C}^g/L$ is a $2g$-dimensional torus, called the *Jacobian variety* $\mathrm{Jac}(\Sigma_g)$. This map $\Sigma_g \times \cdots \times \Sigma_g \to \mathbb{C}^g/L$ is holomorphic and surjective ('Jacobi Inversion'). Restricting it to the diagonal embedding $q \mapsto (q, \ldots, q) \in \Sigma_g \times \cdots \times \Sigma_g$, we get a one-to-one conformal embedding $q \mapsto F(C, \ldots, C)$ of $\Sigma_g$ into $\mathrm{Jac}(\Sigma_g)$. When $g = 1$, $\Sigma_1$ and $\mathrm{Jac}(\Sigma_1)$ are identical; when $g > 1$ the embedding is into a proper submanifold of the Jacobian (check dimensions).

Now, we can select our basis $\omega_i$ of 1-forms so that the integral $\int_{\alpha_i} \omega_j$ equals $\delta_{ij}$. This choice means that our lattice $L$ contains $\mathbb{Z}^g$. The remaining basis vectors of $L$ are $(\int_{\beta_i} \omega_1, \ldots, \int_{\beta_i} \omega_g) \in \mathbb{C}^g$, and it can be shown (the 'Riemann bilinear relations') that these basis vectors will be column vectors of a symmetric $g \times g$ matrix $\Omega$ whose imaginary part is positive-definite – that is, the *period matrix* $\Omega$ lies in $\mathbb{H}_g$. So the lattice $L$ becomes $\mathbb{Z}^g + \Omega\mathbb{Z}^g$ and the Jacobian becomes $T_\Omega := \mathbb{C}^g/(\mathbb{Z}^g + \Omega\mathbb{Z}^g)$, where we regard vectors in $\mathbb{Z}^g$ and $\mathbb{C}^g$ as column vectors. Different choices of bases correspond to the $\mathrm{Sp}_{2g}(\mathbb{Z})$-orbit of $\Omega$.

So every surface $\Sigma_g$ corresponds to an $\mathrm{Sp}_{2g}(\mathbb{Z})$-orbit in $\mathbb{H}_g$. The *Schottky Problem* asks which points in $\mathbb{H}_g$ arise as period matrices. Call this subset $\mathfrak{C}_g$. Our moduli space $\mathfrak{M}_{g,0}$ can be identified with $\mathfrak{C}_g/\mathrm{Sp}_{2g}(\mathbb{Z})$ and $\mathrm{Sp}_{2g}(\mathbb{Z})$ is a homomorphic image (or quotient) of $\Gamma_{g,0}$. Since the symplectic group $\mathrm{Sp}_{2g}(\mathbb{Z})$ is much more accessible than the mapping class group $\Gamma_{g,0}$, the main difficulty is to find a nice characterisation of $\mathfrak{C}_g$ and the kernel of $\Gamma_{g,0} \to \mathrm{Sp}_{2g}(\mathbb{Z})$. For a formal solution to the Schottky problem, see e.g. [**12**].

The moduli space $\mathfrak{M}_{g,n}$ is rarely compact. A very naturalway to compactify $\mathfrak{M}_{g,n}$, due to Deligne and Mumford, is fundamental to conformal field theory. Consider first the complex curve $w^2 = (z-2)(z+1-\alpha)(z-1-\alpha)$, where $\alpha$ is a parameter. Provided $\alpha \neq 0, \pm 1$, this is a genus-1 nonsingular curve, conformally equivalent to the torus

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

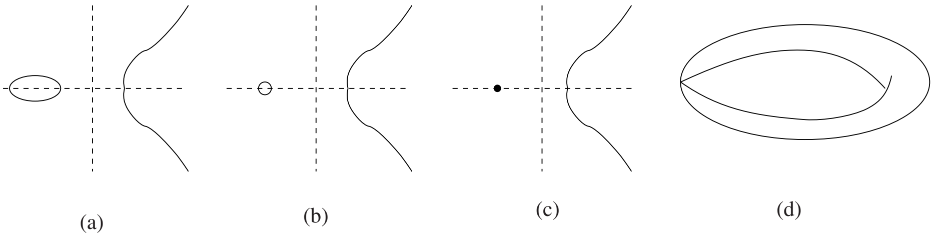

Fig. 2.7 The surface $w^2 = (z-2)(z+1-\alpha)(z+1+\alpha)$.

$\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$ where

$$j(\tau) = \frac{(\alpha^2+3)^2 - (2\alpha^2-2)^3}{(2-1+\alpha)^2(2+1-\alpha)^2(1-\alpha-1+\alpha)^2}.$$

We know that $\mathfrak{M}_{1,0}$ is real-diffeomorphic to a sphere with one point removed. As we vary $\alpha$, we move through $\mathfrak{M}_{1,0}$, and as $\alpha \to 0$ we approach the missing point. What happens to the curve in that limit? In Figure 2.7(a)–(c) we intersect our curve, for $\alpha = 1/2, 1/4, 0$, respectively, with the plane $\mathbb{R}^2 \subset \mathbb{C}^2$. Figure 2.7(d) gives a picture of the complex curve at $\alpha = 0$: it is a pinched torus. We call the nonsmooth point $(z, w) = (-1, 0)$ a *node*. This is the surface to which the boundary point of $\mathfrak{M}_{1,0}$ corresponds. Including it, compactifies $\mathfrak{M}_{1,0}$ to $\overline{\mathfrak{M}_{1,0}} \cong S^2$.

More generally, we add to each moduli space $\mathfrak{M}_{g,n}$ the surfaces $\Sigma$ with nodes. These are connected compact spaces where the neighbourhood of any point either looks like $\mathbb{C}$ (i.e. $\Sigma$ is smooth there) or like $zw = 0$ at $(0, 0)$ (these are the nodes). We say $\Sigma$ has type $(g, n)$ if unpinching each node results in a genus-$g$ surface with $n$ punctures – for example, Figure 2.7(d) has type (1,0). We require these surfaces to have the following property: when you delete all nodes and the surface falls into connected pieces, none of those pieces is a sphere with one or two punctures (the only exception is that we also allow a torus with one node). These surfaces are called *stable*, because they have a finite automorphism group (this terminology is explained by visualising a marble versus a dice on a tabletop). As we know, the larger the automorphism group, the worse the singularity is in moduli space.

The moduli space $\mathfrak{M}_{g,n}$ is compactified if we include the conformal equivalence classes of stable type $(g, n)$ surfaces with nodes. The resulting space $\overline{\mathfrak{M}_{g,n}}$ is called the *moduli space of stable surfaces*. A nice review is given in [**447**]. For example, the moduli space $\mathfrak{M}_{0,4}$ is also a sphere with one missing point. That missing point corresponds to pinching a sphere with four punctures into two spheres, each with two punctures.

Moduli spaces of curves seem first to have been introduced into string theory and conformal field theory by Polyakov in 1981, and have played an important role there ever since. We are actually more interested in an enhanced moduli space, obtained by decorating Riemann surfaces with additional structure. Many more or less equivalent alternatives have appeared in the literature. In particular, let $\Sigma$ be a compact genus-$g$ surface, possibly with nodes, with $n$ marked points $p_i \in \Sigma$ (none of which are at a node). About each point $p_i$ is chosen a local coordinate $z_i$, vanishing at $p_i$, identifying a neighbourhood

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

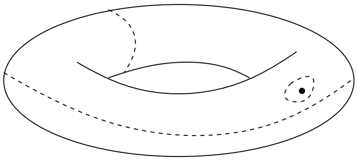

Fig. 2.8 The Dehn twists on the torus with one marked point.

of $p_i$ with a neighbourhood of 0 in $\mathbb{C}$ (see section 2.1 of [**530**] for details). We call this data $(\Sigma, \{p_i\}, \{z_i\})$ an *enhanced surface* of type $(g, n)$. It is essentially equivalent to removing a disc from $\Sigma$ about $p_i$ and choosing a parametrisation about the boundary circle. We call two enhanced surfaces $(\Sigma, \{p_i\}, \{z_i\})$ and $(\Sigma', \{p'_i\}, \{z'_i\})$ equivalent if there is a conformal equivalence $h : \Sigma \to \Sigma'$ such that $h(p_i) = p'_i$ and $z'_i(hx) = z_i(x)$ locally about $p_i$. The resulting moduli space $\widehat{\mathfrak{M}}_{g,n}$ will be infinite-dimensional, but the mapping class group $\widehat{\Gamma}_{g,n}$ will be an extension of the usual $\Gamma_{g,n}$ by $\mathbb{Z}^n$.

These groups $\widehat{\Gamma}_{g,n}$ are of great interest to us – for example, a rational conformal field theory gives a projective finite-dimensional representation of each of them. This yields the braid group representations in quantum groups or Jones subfactor theory, as well as the modularity of Moonshine. They are discussed, with many examples, in section 5.1 of [**32**] (where they are denoted $\Gamma_{g,n}$, and what we call $\Gamma_{g,n}$ is denoted there $\Gamma^n_g$). For example, $\widehat{\Gamma}_{1,1}$ is the braid group $\mathcal{B}_3$, a central extension of $\mathrm{SL}_2(\mathbb{Z})$ by $\mathbb{Z}$. It is generated by the Dehn twists depicted in Figure 2.8. We return to this in Sections 4.3.3, 5.3.4 and 7.2.4.

The main reason we prefer extended surfaces to ordinary Riemann surfaces is that there are canonical ways to sew them together. This sewing operation is fundamental in conformal field theory, because it permits us to decompose a higher-genus surface into discs and ‘pairs-of-pants’ (Section 4.4.1).

Question 2.1.1. How would a hyperbolic mathematician model the Euclidean plane?

Question 2.1.2. (a) Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$, $\gamma \neq \pm I$. We can regard $\gamma$ as a map from the extended upper-half plane $\mathbb{H} \cup \mathbb{R} \cup \{\infty\}$ to itself. Show that:

(i) $|a + d| = 2$ iff $\gamma$ has exactly one fixed point on the boundary $\mathbb{R} \cup \{\infty\}$, iff $\gamma$ can be conjugated in $\mathrm{SL}_2(\mathbb{R})$ to the translation $z \mapsto z + t$;

(ii) $|a + d| > 2$ iff $\gamma$ has exactly two distinct fixed points on the boundary $\mathbb{R} \cup \{\infty\}$, iff $\gamma$ can be conjugated in $\mathrm{SL}_2(\mathbb{R})$ to the dilation $z \mapsto \lambda z$;

(iii) $|a + d| < 2$ iff $\gamma$ has exactly one fixed point in $\mathbb{H}$, iff $\gamma$ can be conjugated in $\mathrm{SL}_2(\mathbb{R})$ to the rotation $z \mapsto \frac{\cos(\theta)z+\sin(\theta)}{-\sin(\theta)z+\cos(\theta)}$ about i with fixed point i.

(b) Suppose $\Gamma$ is a Fuchsian group. Prove that $\gamma \in \Gamma$ has a fixed point in $\mathbb{H}$ iff $\gamma$ has finite order.

Question 2.1.3. Explain how the addition of points on a conic is a degenerate case of the addition of points on a cubic.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Question 2.1.4. Find all rational solutions $(r, s)$ to $r^2 - 2rs + r + 2s - s^2 = 0$. Verify that, for the choice of identity $e = (0, 0)$ and addition defined as in Figure 2.6, the rational points form a subgroup. As an abstract group, what is this subgroup isomorphic to?

Question 2.1.5. Using the conformal map $z \mapsto (x, y) = (\mathfrak{p}(z), \mathfrak{p}'(z))$ between $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ and the cubic $y^2 = 4(x - e_1)(x - e_2)(x - e_3)$, verify that the addition of points on the cubic corresponds to the addition $z_1 + z_2 \pmod{\mathbb{Z} + \tau\mathbb{Z}}$ in $\mathbb{C}$.

Question 2.1.6. Identify $\mathcal{M}_{0,4}$ with a space of $\mathcal{S}_3$-orbits in $\mathbb{C} \setminus \{0, 1\}$.

Question 2.1.7. Let $G$ be a finite group. Define

$$K(g, h) = \frac{1}{\|G\|} \sum_\rho \dim(\rho)\, \mathrm{ch}_\rho(gh^{-1}),$$

for $g, h \in G$, where the sum is over all irreducible representations $\rho$ of $G$.
(a) Verify that $K(g, h) = \delta_{g,h}$.
(b) For any $\gamma \in \mathbb{N}$, take $f : G^{2\gamma} \to G$ by

$$f(g_1, h_1, \ldots, g_\gamma, h_\gamma) = g_1 h_1 g_1^{-1} h_1^{-1} g_2 h_2 g_2^{-1} h_2^{-1} \cdots g_\gamma h_\gamma g_\gamma^{-1} h_\gamma^{-1}.$$

Define $I = \sum_{(g_i, h_i) \in G^{2\gamma}} K(f(g_i, h_i), e)$. By evaluating $I$ in two ways, obtain the formula

$$\|\mathrm{Hom}(\pi_1(\Sigma_\gamma), G)\| = \|G\|^{2\gamma - 1} \sum_\rho \dim(\rho)^{2-2\gamma},$$

where $\Sigma_\gamma$ is a compact genus-$\gamma$ surface.

## 2.2 Modular forms and functions

Number theory, at its most elemental level, is concerned with finding integer solutions to various (systems of) equations. It is truly remarkable how this seemingly pedestrian pursuit has resulted in the creation of the richest and deepest mathematics. Indeed, it is tempting to suspect that beneath any spot on the mathematical turf, no matter how remote or seemingly barren, is a gemstone merely requiring hard work and discerning fingertips to unearth.

### *2.2.1 Definition and motivation*

As we saw in several different contexts in Section 2.1, the group $\mathrm{SL}_2(\mathbb{R})$ of $2 \times 2$ matrices with real entries and determinant 1 acts on the upper half-plane $\mathbb{H} = \{\tau \in \mathbb{C} \mid \mathrm{Im}(\tau) > 0\}$ by Möbius transformations (2.1.4a). For example, the matrices $s := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $t := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ correspond to the functions $\tau \mapsto -1/\tau$ and $\tau \mapsto \tau + 1$, respectively.

Consider $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, the subgroup of $\mathrm{SL}_2(\mathbb{R})$ consisting of the matrices with integer entries. It is generated by $s$ and $t$:

$$\mathrm{SL}_2(\mathbb{Z}) = \left\langle \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right\rangle = \langle s, t \mid s^4 = e, (st)^3 = s^2 \rangle. \qquad (2.2.1a)$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Because $-I \in \mathrm{SL}_2(\mathbb{Z})$ yields the trivial map in $\mathbb{H}$, weare also interested in the group

$$\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\} = \langle s, t \,|\, s^2 = (st)^3 = e\rangle =: \mathbb{Z}_2 * \mathbb{Z}_3, \tag{2.2.1b}$$

the free product of $\mathbb{Z}_2$ with $\mathbb{Z}_3$. Groups like $\Gamma$ act on the extended upper half-plane $\overline{\mathbb{H}} := \mathbb{H} \cup \{\mathrm{i}\infty\} \cup \mathbb{Q}$ in the obvious way (e.g. $s$ interchanges 0 and $\mathrm{i}\infty$). The extra points $\{\mathrm{i}\infty\} \cup \mathbb{Q}$ are called *cusps* because of the hyperbolic triangle $R$ in Figure 2.3, which points at one of them. Cusps correspond to tori with a single node (Figure 2.7(d)), and compactify the moduli space $\mathfrak{M}_{1,0}$.

Recall Definition 0.1: a modular function for $\Gamma$is a meromorphic function $f : \overline{\mathbb{H}} \to \mathbb{C}$, symmetric with respect to $\Gamma$. A related definition is:

**Definition 2.2.1** *A modular form $f$ for $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ of weight $k \in \mathbb{Q}$ and multiplier $\mu : \Gamma \to \mathbb{C}$, $|\mu| = 1$ is a holomorphic function $f : \mathbb{H} \to \mathbb{C}$, which is also holomorphic at the cusps $\mathbb{Q} \cup \{\mathrm{i}\infty\}$ and obeys the transformation law*

$$f\left(\frac{a\tau + b}{c\tau + d}\right) = \mu\begin{pmatrix} a & b \\ c & d \end{pmatrix}(c\tau + d)^k\, f(\tau), \qquad \forall \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma. \tag{2.2.2}$$

For fractional $k$ we choose the branch of the $k$th power to be the principal one (so $x^k > 0$ when $x > 0$). For number-theoretic purposes, we require the values of $\mu$ to be roots of unity. Writing $\mu(t) = e^{2\pi \mathrm{i} h}$, we can expand $f$ in powers of $q$: $f(\tau) = q^h \sum_{n=-\infty}^{\infty} a_n q^n$. By ‘meromorphic at $\mathrm{i}\infty$’ we mean that all but finitely many negative $n$ have $a_n = 0$, so $f$ has a pole of finite order at $q = 0$; by ‘holomorphic at $\mathrm{i}\infty$’ we mean $h \geq 0$ and $a_n = 0$ for all negative $n$. Meromorphicity or holomorphicity at the other cusps is implied by that at $\mathrm{i}\infty$, because of (2.2.2) and the fact that all cusps lie in the same $\mathrm{SL}_2(\mathbb{Z})$-orbit.

For the significance, which is considerable, of the condition that $f$ be meromorphic at the cusps, see Question 2.2.1. The moduli spaces $\overline{\mathfrak{M}_{1,0}}$, $\overline{\mathfrak{M}_{1,1}}$ and $\overline{\mathfrak{M}_{0,4}}$ all are $\mathrm{SL}_2(\mathbb{Z})\backslash\overline{\mathbb{H}}$. The cusps $\mathbb{Q} \cup \{\mathrm{i}\infty\}$ of $\mathbb{H}$ correspond to pinched tori or spheres (Section 2.1.4). Meromorphicity at the cusps says $f$ respects this surface degeneration in the appropriate way.

If the weight $k$ is an integer, the multiplier $\mu$ will necessarily be a one-dimensional representation of $\Gamma$; when $k$ is rational, $\mu$ will be a projective representation. We define projective representations, and explain what to do with them, in Section 3.1.1. An intriguing implication for fractional $k$ is described in Section 2.4.3.

The function $f$ is called a modular *form* because $f(\tau)\,\mathrm{d}^{-k/2}\tau$ is a holomorphic $(-k/2)$-form on the space $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$; by contrast, a modular *function* $f$ is a meromorphic function on the space $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$.

The easiest examples of modular forms of weight $k \geq 4$ ($k$ even) are the Eisenstein series $G_k$ defined in equation (0.1.5). It is conventional to normalise them as follows:

$$E_k(\tau) := \frac{1}{2\zeta(k)} G_k(\tau) = 1 - \frac{2k}{B_k}\sum_{n=1}^{\infty} \sigma_{k-1}(n)\, q^n \in \mathbb{Z}[q], \tag{2.2.3a}$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

where $B_k$ are the Bernoulli numbers, defined by the generating function $\frac{x}{e^x-1} = \sum_{k=0}^{\infty} B_k \frac{x^k}{k!}$, and where $\sigma_{k-1}(n)$ and the Riemann zeta function $\zeta(s)$ are defined by

$$\sigma_m(n) = \sum_{d|n} d^m, \tag{2.2.3b}$$

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_{p \text{ prime}} (1-p^{-s})^{-1} \tag{2.2.3c}$$

(see Section 2.3.1). The $E_k$ and $G_k$ have multiplier $\mu \equiv 1$.

Indeed, the Eisenstein series generate all modular forms for $\mathrm{SL}_2(\mathbb{Z})$ with trivial multiplier $\mu$. More specifically, the span of all such modular forms (over all $k$) is a ring graded by $k$ (i.e. the product of modular forms of weight $k$ and $k'$ is one of weight $k+k'$). This ring is generated (over $\mathbb{C}$) by the Eisenstein series $E_4(\tau)$ and $E_6(\tau)$ – that is, any level $k$ modular form $f$ can be written as a polynomial (homogeneous in the obvious sense) in $E_4$ and $E_6$. Moreover, $E_4$ and $E_6$ are algebraically independent, so that polynomial is unique. Using this we can readily compute the dimension of (and find a basis for) the space of weight $k$ modular forms. For instance, a basis for the weight 24 modular forms is $\{E_6^4, E_6^2E_4^3, E_4^6\}$.

The definition of modular forms seems fairly arbitrary. For example, one may ask how significant the upper half-plane $\mathbb{H}$ is, or where the factor $(c\tau+d)^k$ in (2.2.2) comes from. We confront this in Section 2.4.1. But for now note that Definition 2.2.1 (like Definition 0.1 before it) also makes perfect sense if $\mathrm{SL}_2(\mathbb{Z})$ is replaced by any Fuchsian group $\Gamma$ that sends the cusps $\mathbb{Q}\cup\{\mathrm{i}\infty\}$ to themselves. The only (minor) complication is that the cusps may not lie in the same orbit. See, for example, [**352**] for the proper definition. We are interested in $\Gamma$ *commensurable* with $\mathrm{SL}_2(\mathbb{Z})$, that is, $\Gamma\cap\mathrm{SL}_2(\mathbb{Z})$ has finite index in both $\Gamma$ and $\mathrm{SL}_2(\mathbb{Z})$. Typical choices for $\Gamma$ are the *congruence subgroups*

$$\Gamma(N) := \left\{\begin{pmatrix} a & b \\ c & d\end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,\middle|\, \begin{pmatrix} a & b \\ c & d\end{pmatrix} \equiv \pm\begin{pmatrix} 1 & 0 \\ 0 & 1\end{pmatrix} \ (\mathrm{mod}\ N)\right\}, \tag{2.2.4a}$$

$$\Gamma_0(N) := \left\{\begin{pmatrix} a & b \\ c & d\end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,\middle|\, c \equiv 0 \ (\mathrm{mod}\ N)\right\}, \tag{2.2.4b}$$

for any $N\in\mathbb{N}$. Incidentally, for $N>1$, $\Gamma(N)/\{\pm 1\}$ is always free (i.e. isomorphic to some $\mathcal{F}_m$), while $\Gamma_0(N)$ may or may not be free.

It is not at all obvious that modular forms and functions should be interesting, but in fact they are unavoidable in modern number theory. For example, consider the question of writing numbers as sums of squares. We can write $5 = 1^2+(-2)^2 = (-1)^2+1^2+0^2+1^2+(-1)^2$, to give a couple of trivial examples. Let $N_n(k)$ be the number of ways we can write the integer $n$ as a sum of $k$ squares, counting order and signs. For example, $N_5(1)=0$ (since 5 is not a perfect square), $N_5(2)=8$ (since $5=(\pm1)^2+(\pm2)^2=(\pm2)^2+(\pm1)^2$), $N_5(3)=24$, etc. Their generating functions are:[5]

$$\sum_{n=0}^{\infty} N_n(k)\,x^n = \theta(x)^k,$$

[5] A fundamental principle in mathematics is: whenever you have a subscript with an infinite range, make a power series (called a *generating function*) out of it.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

where

$$\theta(x) = 1 + 2x + 2x^4 + \cdots = \sum_{n\in\mathbb{Z}} x^{n^2}$$

is called a *theta function*. It turns out that $\theta$ transforms nicely with respect to $\mathrm{SL}_2(\mathbb{Z})$, once we make the change-of-variables $x = \exp[\pi \mathrm{i}\tau]$ (what we usually call $\sqrt{q}$). Write $\theta_3(\tau)$ for $\theta(x)$. Then $\theta_3$ is clearly invariant under the action of $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, and a little work (done next subsection) shows that $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ takes $\theta_3(\tau)$ to $\sqrt{\frac{\tau}{\mathrm{i}}}\,\theta_3(\tau)$. Together those two modular transformations generate the group

$$\Gamma_\theta := \left\langle \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \right\rangle = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,|\, ac \equiv bd \equiv 0 \ (\mathrm{mod}\ 2) \right\}. \tag{2.2.5}$$

$\theta_3$ is a modular form of weight $\frac{1}{2}$ and nontrivial multiplier for $\Gamma_\theta$.

Jacobi introduced that important change-of-variables $x = \exp[\pi \mathrm{i}\tau]$ two centuries ago, in his analysis of elliptic integrals. His theory is poorly remembered today, which is very disheartening considering how much of modern mathematics is touched by it. Have a look at the book [**94**], written over a century ago; the style of mathematics in our time is rather different from that in Jacobi's, and we've lost a little in innocence what we've gained in power. See also the beautiful book [**414**]. Let's briefly sketch Jacobi's theory.

Just as we could develop a theory of 'circular functions' (i.e. sine, etc.) starting from the integral $s(a) = \int_0^a \frac{\mathrm{d}x}{\sqrt{1-x^2}}$, so we can develop a theory of 'elliptic functions' starting from the elliptic integral $F(k,a) = \int_0^a \frac{\mathrm{d}x}{\sqrt{(1-x^2)(1-k^2x^2)}}$. Inverting $s(a)$ gives a function both more useful and with nicer properties than $s(a)$: we call it $\sin(u)$. Similarly, for any $k$ the elliptic function $\mathrm{sn}(k,u)$ is defined by $u = F(k, \mathrm{sn}(k,u))$. Just as we can define a numerical constant $\pi$ by $\sin(\frac{1}{2}\pi) = 1$ (i.e. $\frac{1}{2}\pi = \int_0^1 \frac{\mathrm{d}x}{\sqrt{1-x^2}}$), we get a function $K(k) = \int_0^1 \frac{\mathrm{d}x}{\sqrt{(1-x^2)(1-k^2x^2)}}$. Just as $\sin(u)$ has period $4(\frac{1}{2}\pi)$, so sn has $u$-period $4K(k)$. sn also turns out to have $u$-period $4\mathrm{i}\,K(k')$ where $k' = \sqrt{1-k^2}$ – today we take this as the starting point and define an elliptic function to be doubly periodic or, what is the same thing, to be a function on a torus (Section 2.1.3).

Theta functions aren't elliptic functions, but they are closely related, as we see in Section 2.1.3. In Jacobi's language, we have

$$\theta_3\left(\frac{\mathrm{i}\,K(k')}{K(k)}\right) = \sqrt{\frac{2K(k)}{\pi}}.$$

The 'modular transformation' $\tau \mapsto \frac{-1}{\tau}$ interchanges the 'modulus' $k$ with the 'complementary modulus' $k'$, and is completely natural in Jacobi's theory. The important formula $\theta_3(\frac{-1}{\tau}) = \sqrt{\frac{\tau}{\mathrm{i}}}\,\theta_3(\tau)$ is trivial here. Closely related to this is Poincaré's remarkable path to modular functions (Section 3.2.4).

Surprisingly, many seemingly innocent questions can be dragged (usually with effort) into the richly developed realm of elliptic curves and modular forms, where they are often

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

solved. For instance, we all know the ancient Greeks were interested in Pythagorean triples: integer solutions $a, b, c$ to $a^2 + b^2 = c^2$, or equivalently right-angle triangles with rational side-lengths (Section 2.1.1).

There are two ways of pushing this. One is to ask which $n \in \mathbb{Z}$ can arise as areas of these rational right-angle triangles. It turns out $n = 5$ is the smallest one: $a = \frac{3}{2}$, $b = \frac{20}{3}$, $c = \frac{41}{6}$ works ($5 = \frac{1}{2}(\frac{3}{2})(\frac{20}{3})$ and $(\frac{3}{2})^2 + (\frac{20}{3})^2 = (\frac{41}{6})^2$). This is a hard problem – just try to show $n = 1$ cannot work. The number $n = 157$ works, though the simplest triangle has $a$ and $b$ as quotients of integers of size around $10^{25}$, and $c$ as the quotient of integers around $10^{47}$. Although this problem was studied by the ancient Greeks and also by the Arabs in the tenth century, it was finally cracked in the 1980s by first translating it into the question of whether the elliptic curve $y^2 = x^3 - n^2x$ has infinitely many rational points.

The other extension of Pythagorean triples is more famous: find all integer solutions to $a^n + b^n = c^n$. 350 years ago Fermat wrote in the margin of a book he was reading (the book was describing at that point Diophantus'classification of Pythagorean triples) that he had found a 'truly marvelous' proof that for $n > 2$ there are no nontrivial solutions, but that the margin was too narrow to contain it. This result came to be known as 'Fermat's Last Theorem'[6] and despite considerable effort no one has succeeded in rediscovering his proof. Most people believe that Fermat soon realised his 'proof' wasn't valid, otherwise he would have alluded to it in later letters. In any case, a very long and complicated proof was finally achieved in the 1990s: the 'Taniyama–Shimura conjecture' says that a certain function associated with any elliptic curve over $\mathbb{Q}$ will be modular; if $a^n + b^n = c^n$ for some $n > 2$ and nonzero integers $a, b, c$, then the elliptic curve $y^2 = x^3 + (a^n - b^n)x^2 - a^nb^n$ will violate that conjecture; finally, Wiles proved the Taniyama–Shimura conjecture.

A certain interpretation of modular functions also indicates their usefulness, and explains the adjective 'modular'. The moduli space of tori is $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ (Section 2.1.4). So if we have a complex-valued function $F$ on the set of all tori, which associates the same value to conformally equivalent tori (an example is the genus-1 partition function (4.3.8b) in conformal field theories), then $F$ is a function $F : \mathbb{H} \to \mathbb{C}$, symmetric with respect to $\mathrm{SL}_2(\mathbb{Z})$.

Likewise, suppose we are interested in meromorphic functions $f : \Sigma \to \mathbb{C}$ living on some surface $\Sigma$.We know from the last section that almost every surface $\Sigma$ is a quotient $\Sigma = \Gamma\backslash\mathbb{H}$, for some Fuchsian group $\Gamma$. Then $f$ can be lifted to a meromorphic function on $\mathbb{H}$ with symmetry $\Gamma$.

What is the meaning of the Fourier expansion? Think of the parameter $q$ as the local coordinate about the cusp $\mathrm{i}\infty$. The Fourier expansion is simply the local expansion of $f$ about that cusp. There is a similar expansion about any other cusp $x \in \mathbb{Q}$. In the case of $\mathrm{SL}_2(\mathbb{Z})$, all cusps are equivalent, but for smaller groups the cusps typically fall into

[6] It was called his 'Last Theorem' because it was the last of his 48 margin notes to be proved by other mathematicians – a different margin note is discussed in Section 1.7. The story of Fermat's Last Theorem is a fascinating one, but alas this footnote is too small to do it credit. See for instance the excellent book [**508**].

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

several distinct orbits, and the corresponding expansions carry independent information. These coefficients are often quite interesting (e.g. they may give the numbers of solutions to various equations, or the dimensions of certain subspaces). The modular form $f$ is a holomorphic interpolation between this local information.

### 2.2.2 Theta and eta

Two modular forms that appear throughout the following are the Jacobi theta function $\theta_3$ and the Dedekind eta function $\eta$:

$$\theta_3(\tau) := 1 + 2\sum_{m=1}^{\infty} q^{n^2/2} = \prod_{n=1}^{\infty}\left(1+q^{(2n-1)/2)}\right)^2 \prod_{n=1}^{\infty}(1-q^n), \tag{2.2.6a}$$

$$\eta(\tau) := q^{1/24}\prod_{n=1}^{\infty}(1-q^n) = q^{1/24}\sum_{m=-\infty}^{\infty}(-1)^m q^{(3m^2+m)/2}. \tag{2.2.6b}$$

The equality in (2.2.6a) comes from the denominator identity (3.4.5b) for $A_1^{(1)}$, while that in (2.2.6b) comes from Euler’s pentagonal identity; in both cases the first expressions are more important. We saw $\theta_3$ last subsection. Unlike the Eisenstein series, its modularity is not obvious. It can be established though in a number of ways, the most familiar perhaps being *Poisson summation*. This says that for any rapidly decreasing smooth function $g : \mathbb{R} \to \mathbb{C}$ ($g$ is in the Schwartz space $\mathcal{S}(\mathbb{R})$ of Section 1.3.1),

$$\sum_{n\in\mathbb{Z}} g(n) = \sum_{m\in\mathbb{Z}} \widehat{g}(m), \tag{2.2.7a}$$

where $\widehat{g}$ is the Fourier transform of $g$:

$$\widehat{g}(y) = \int_{-\infty}^{\infty} e^{-2\pi \mathrm{i} xy}\, g(x)\,\mathrm{d}x. \tag{2.2.7b}$$

Choose $g(x) = e^{-\pi t x^2}$ with $t \in \mathbb{R}$, so $\tau = \mathrm{i}t \in \mathbb{H}$; then $\widehat{g}(y) = \sqrt{1/t}\, e^{-\pi y^2/t}$ and we obtain (by analytic continuation to all $\tau \in \mathbb{H}$) the transformation formula for $\theta_3$ under $\tau \mapsto -1/\tau$:

$$\theta_3\left(\frac{-1}{\tau}\right) = \sqrt{\frac{\tau}{\mathrm{i}}}\,\theta_3(\tau). \tag{2.2.7c}$$

$\theta_3$ is a modular form for $\Gamma_\theta$ (2.2.5) of weight $1/2$ and nontrivial multiplier. Both Poisson summation and its application to (2.2.7c) are due to Gauss. In Question 2.2.4 you are asked to prove Poisson summation, and next subsection we try to understand what it is saying. In Sections 2.3.1, 2.3.4 and 2.4.2 we give alternate proofs of (2.2.7c).

The modularity of $\eta$ can be summarised by

$$\eta(\tau+1) = \xi_{24}\,\eta(\tau), \tag{2.2.8a}$$

$$\eta\left(\frac{-1}{\tau}\right) = \sqrt{\frac{\tau}{\mathrm{i}}}\,\eta(\tau), \tag{2.2.8b}$$

where $\xi_{24} = \exp[2\pi \mathrm{i}/24]$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

More generally, we get the complicated transformation law

$$\eta\left(\frac{a\tau+b}{c\tau+d}\right)=\mu(a,b,c,d)\sqrt{c\tau+d}\,\eta(\tau),\qquad \forall\begin{pmatrix}a & b\\ c & d\end{pmatrix}\in\mathrm{SL}_2(\mathbb{Z}),\tag{2.2.8c}$$

where, for $c>0$, $\mu(a,b,c,d)=\exp(\pi\mathrm{i}\,(\frac{a+d}{12c}-\frac{1}{2}-s(d,c)))$ for the *Dedekind sum*

$$s(d,c)=\sum_{i=1}^{c-1}\frac{i}{c}\left(\frac{di}{c}-\left\lfloor\frac{di}{c}\right\rfloor-\frac{1}{2}\right).\tag{2.2.8d}$$

For $c=0$, the transformation follows immediately from (2.2.8a), while for $c<0$ an analogue to (2.2.8c) holds. The denominator of the rational number $s(d,c)$ will always divide $6c$; $\mu$ will always be a 24th root of 1. Although Dedekind sums have many special properties [**468**], we find in Section 2.4.3 a much cleaner way to write (2.2.8c). In any case, $\eta$ is a modular form for $\mathrm{SL}_2(\mathbb{Z})$ of weight $\frac{1}{2}$ and nontrivial multiplier.

Once again, (2.2.8a) is immediate from the definition (2.2.6b) and isn't deep. There are several arguments in the literature that establish (2.2.8b), including Poisson summation applied to the series in (2.2.6b). Here is another, which is instructive for other reasons. In the following paragraph, let's not be distracted by mere analytic concerns, like convergence or interchanging integrals and infinite sums.

Fix $\tau=\mathrm{i}t$, $t>0$. The expression

$$-\frac{1}{4}\int(\theta_3(\mathrm{i}st)-1)(\theta_3(\mathrm{i}s/t)-1)\,\mathrm{d}s\tag{2.2.9a}$$

is manifestly invariant under the transformation $t\mapsto 1/t$. Applying the transformation (2.2.7c) to $\theta_3(\mathrm{i}s/t)$ and expanding out both $\theta_3$'s, we get

$$\begin{aligned}&-\frac{1}{4}\int\left(\sum_{\ell=1}^{\infty}2e^{-\pi st\ell^2}\right)\left(\sqrt{\frac{t}{s}}\left(1+2\sum_{n=1}^{\infty}e^{-\pi tn^2/s}\right)-1\right)\mathrm{d}s\\ &=-\sum_{\ell=1}^{\infty}\sum_{n=1}^{\infty}\int\sqrt{\frac{t}{s}}e^{-\pi st\ell^2-\pi tn^2/s}\mathrm{d}s+\frac{1}{2}\sum_{\ell=1}^{\infty}\int e^{\pi st\ell^2}\mathrm{d}s-\frac{1}{2}\sum_{\ell=1}^{\infty}\int\sqrt{\frac{t}{s}}e^{-\pi st\ell^2}\mathrm{d}s.\end{aligned}\tag{2.2.9b}$$

Now, replace the indefinite integral here with $\int_0^\infty$. The third term in the right-side of (2.2.9b) is independent of $t$ (to see this, change variables: $y=ts$) and so is a constant. The second term can be evaluated explicitly:

$$\frac{1}{2}\sum_{\ell=1}^{\infty}\int_0^\infty e^{-\pi st\ell^2}\mathrm{d}s=\frac{1}{2}\sum_{\ell=1}^{\infty}\frac{1}{\pi t\ell^2}=\frac{1}{2\pi t}\frac{\pi^2}{6}=\frac{\pi}{12t}.\tag{2.2.9c}$$

To simplify the first term of (2.2.9b), replace $s$ with $x^2$ and apply the identity

$$e^{-2\sqrt{ab}}=2\sqrt{\frac{a}{\pi}}\int_0^\infty e^{-ax^2-bx^{-2}}\mathrm{d}x$$

(this is identity 3.325 of [**258**]) with $a=\pi t\ell^2$, $n=\pi tn^2$. The first term becomes

$$-\sum_{\ell=1}^{\infty}\sum_{n=1}^{\infty}\frac{1}{\ell}e^{-2\pi t\ell n}=-\sum_{n=1}^{\infty}\log(1-e^{-2\pi tn}).$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Putting these together, we get

$$-\frac{1}{4}\int_0^\infty (\theta_3(\mathrm{i}st)-1)(\theta_3(\mathrm{i}s/t)-1)\,\mathrm{d}s = \log\eta(\mathrm{i}t)+\frac{\pi t}{12}+\frac{\pi}{12t}+C$$

for some constant $C$.

Two unfortunate remarks should probably be made regarding this calculation. First, it would imply (2.2.8b) holds without the prefactor $\sqrt{\tau/\mathrm{i}}$. Second, the constant $C$ diverges, as does the integral in (2.2.9a). Calculations like this mellow somewhat one's disdain for analysis. The way to proceed is to 'regularise' (2.2.9a) by subtracting from the integrand near $s=0$ the term $s^{-1}$ responsible for the divergence. This results in the identity

$$\log\eta(\mathrm{i}t) = -\frac{1}{4}\int_1^\infty (\theta_3(\mathrm{i}st)-1)(\theta_3(\mathrm{i}s/t)-1)\,\mathrm{d}s - \frac{1}{4}\int_0^1 (\theta_3(\mathrm{i}st)-1)(\theta_3(\mathrm{i}s/t)-1) -s^{-1}\,\mathrm{d}s - \frac{\pi t}{12}-\frac{\pi}{12t}-\frac{1}{4}\log t. \tag{2.2.9d}$$

In Question 2.2.5 the reader is asked to fill in the details, proving (2.2.9d) and thus (2.2.8b). We see from this argument that the mysterious power $1/24$ in (2.2.6b), required for the modularity of $\eta$, in fact equals $\zeta(2)/(2\pi)^2$.

At least in spirit, this calculation is reminiscent of the regularisation of Feynman integrals in quantum field theory (Section 4.2.3). For example, the Dedekind eta arises in the calculation of the one-loop partition function of a boson compactified on a circle (see e.g. section 8 of [**246**]). The normalisation factor there involves the product of the nonzero eigenvalues of the Laplacian $\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}$ on the torus $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$: namely the modulus-squared $|D|^2$ of

$$D(\tau)=\prod_{(m,n)\neq(0,0)}\frac{\pi}{\tau_2}(n-\tau m), \tag{2.2.10a}$$

where $\tau_2=\mathrm{Im}(\tau)>0$. This expression diverges enthusiastically, but it is to be interpreted using the substitutions (*zeta-function regularisation*)

$$\prod_{n=1}^\infty a = a^{\zeta(0)} = a^{-\frac{1}{2}},\quad \prod_{n=1}^\infty n^\alpha = e^{-\alpha\zeta'(0)}=(2\pi)^{\alpha/2},\quad \prod_{n=1}^\infty a^n = a^{\zeta(-1)}=a^{-\frac{1}{12}}, \tag{2.2.10b}$$

where $\zeta$ here is the Riemann zeta function (2.2.3c). It is found that

$$D(\tau)=2\tau_2\,\eta(\tau)^2. \tag{2.2.10c}$$

In this 'derivation' of $\eta$, the exponent $1/24$ in (2.2.6b) equals $-\zeta(-1)/2$. Since the values $\zeta(-1)$ and $\zeta(2)$ are related by the functional equation (2.3.2), they are indeed equivalent. Also, note that (2.2.10a) obeys $D(\tau+1)=D(\tau)$ and $D(-1/\tau)=D(\tau)/\overline{\tau}$, while (2.2.10c) obeys $D(\tau+1)=e^{\pi\mathrm{i}/6}D(\tau)$ and $D(-1/\tau)=-\mathrm{i}D(\tau)/\overline{\tau}$. Thus the identifications (2.2.10b) don't preserve modular behaviour. It is somewhat reminiscent of the $-s^{-1}$ regularisation in (2.2.9), which breaks the $t \leftrightarrow 1/t$ symmetry.

Prefactors $q^m$ as in (2.2.6b) are very common, as we shall see later with the characters of Kac–Moody algebras or vertex algebras. In Monstrous Moonshine, this is the $q^{-1}$ with

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

which the $j$-function begins. These factors are a little mysterious – for example, why start the grading in (0.3.1) at $-1$ rather than 0 – and there are several explanations (Sections 3.1.2, 3.2.3 and 5.3.4). The point of our little digression into string theory is to introduce its term *conformal anomaly* for this factor $q^m$. In physics, an anomaly is a symmetry of a classical system that is broken in its quantisation. Here, the $\tau \mapsto \tau + 1$ symmetry (an aspect of conformal invariance) of $D(\tau)$ is broken by regularisation, an anomaly.

We see in (2.2.3a) that the coefficients of the $q$-expansion of Eisenstein series are interesting. In fact, we are usually more interested in the coefficients of a modular form than in the function itself. A classic example of this is the theta series of a lattice. Let $L \subset \mathbb{R}^n$ be any $n$-dimensional positive-definite lattice (Section 1.2.1), and choose any vector $t \in \mathbb{R}^n$. Define

$$\Theta_{t+L}(\tau) := \sum_{x \in t+L} q^{x \cdot x/2}. \tag{2.2.11a}$$

In words, the coefficient of $q^r$ is the number of vectors in $t + L$ with length $\sqrt{2r}$. For example, $\Theta_{\mathbb{Z}} = \theta_3$. Let $L$ be rational (i.e. for all $u, v \in L$ we have $u \cdot v \in \mathbb{Q}$) and $t$ have finite order $m$ in $L$ (i.e. $mt \in L$). Then Poisson summation again yields

$$\Theta_{t+L}\left(\frac{-1}{\tau}\right) = \frac{(\tau/\mathrm{i})^{n/2}}{\sqrt{|L|}} \sum_{k=0}^{m-1} \xi_m^k \, \Theta_{ks+L_0}(\tau), \tag{2.2.11b}$$

where (as always) $\xi_m := \exp[2\pi\mathrm{i}/m]$, $s \in L^*$ satisfies $s \cdot t \equiv \frac{1}{m}$ (mod 1) (why must such a vector $s$ always exist?) and where $L_0 = \{u \in L^* \mid u \cdot t \in \mathbb{Z}\}$. In particular,

$$\Theta_L\left(\frac{-1}{\tau}\right) = \frac{(\tau/\mathrm{i})^{n/2}}{\sqrt{|L|}} \Theta_{L^*}(\tau). \tag{2.2.11c}$$

**Definition 2.2.2** *Let $\mathcal{I}$ be a finite set, and suppose for each $i \in \mathcal{I}$ we have a function $f_i(\tau)$ meromorphic in $\mathbb{H}$ and with $q$-expansion $f_i(\tau) = \sum_{r \in \mathbb{Q}} a_{r,i} q^r$, such that for each $N$ only finitely many $r < N$ have nonzero coefficients $a_{r,i}$. We call the set $\{f_i(\tau)\}_{i \in \mathcal{I}}$ a* vector-valued modular function for $\mathrm{SL}_2(\mathbb{Z})$ with multiplier $\rho : \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{GL}_{\mathcal{I}}(\mathbb{C})$ *if, for each $A \in \mathrm{SL}_2(\mathbb{Z})$ and $i \in \mathcal{I}$, we have*

$$f_i\left(\frac{a\tau + b}{c\tau + d}\right) = \sum_{j \in \mathcal{I}} \rho(A)_{ij} \, f_j(\tau).$$

The strange condition on the $a_{r,i}$ simply says that each $f_i$ is meromorphic at $\tau = \mathrm{i}\infty$. Vector-valued modular forms are studied in, for example, [**350**]. By the usual argument, $\rho$ will be a $\|\mathcal{I}\|$-dimensional representation of $\mathrm{SL}_2(\mathbb{Z})$. We are interested in the case when the matrices $\rho(A)$ are unitary. In this case, at least when the functions $f_i(\tau)$ are linearly independent, a vector-valued modular function for $\mathrm{SL}_2(\mathbb{Z})$ defines a flat, holomorphic, Hermitian vector bundle over $\mathfrak{M}_{1,0}$: namely, the diagonal quotient $(\mathbb{H} \times \mathrm{span}\{f_i(\tau)\})/\mathrm{PSL}_2(\mathbb{Z})$. The fibre above any point in $\mathfrak{M}_{1,0}$ will be $\|\mathcal{I}\|$-dimensional, except possibly for the singular points [i] and $[e^{\pi\mathrm{i}/3}]$. The $f_i$ are holomorphic sections of this bundle.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

A classical property of theta functions, apparently due in this generality to Hecke in 1940, anticipates beautifully what we see later in this book in more and more generality.

**Theorem 2.2.3** *Let $L \subset \mathbb{R}^n$ be any $n$-dimensional positive-definite lattice.*

(a) *Suppose for all $v \in L$ that $v \cdot v \in \mathbb{Q}$. Let $t \in \mathbb{R}^n$ be any vector with finite order in $L$: i.e. $mt \in L$ for some nonzero $m \in \mathbb{Z}$. Then the theta series $\Theta_{L+t}(\tau)$, divided by $\eta(\tau)^n$, is a modular function for some $\Gamma(N)$.*

(b) *Suppose further that $L$ is an even lattice (i.e. all $v \cdot v$ lie in $2\mathbb{Z}$), and let $L^*$ be its dual. Write $t_i + L$, $i = 1, \ldots, M$, for the finitely many cosets in $L^*/L$. Define a column vector $\vec{\chi}_L(\tau)$ with $i$th component $\Theta_{t_i+L}(\tau)/\eta(\tau)^n$. Then $\vec{\chi}_L$ forms a vector-valued modular function for $SL_2(\mathbb{Z})$.*

For the proof of part (a), see theorem 20 of [**456**]. Part (b) follows quickly from (2.2.11b) and (2.2.8). This theorem can be interpreted as being a special case of Theorem 3.2.3 below, when $\mathfrak{g}$ is the affinisation of the reductive (abelian) Lie algebra $\mathbb{C}^n$. Note, however, that the functions in (2.2.11a) are linearly dependent, and so the matrices $\rho(A)$ are not uniquely defined by (b). The easiest way to get linear independence is by adding variables (Section 2.3.2).

The Leech lattice $\Lambda$ (Section 1.2.1) is to lattices much as the Moonshine module $V^\natural$ is to VOAs (see Section 7.2.1 below). It has no length-squared 2-vectors, and has precisely 196 560 length-squared 4-vectors – a number remarkably close to the monstrous 196 883. Indeed its theta function $\Theta_\Lambda(\tau)$, when divided by $\eta(\tau)^{24}$, equals $J(\tau) + 24$. Is this another example of Moonshine, on par with McKay'sequation (0.2.1a)?

Indeed it is. However, for the Leech lattice $\Lambda$, we can quickly identify $\Theta_\Lambda(\tau)$ in terms of $J(\tau)$ (see Question 2.2.7). Although the 196 560 $\approx$ 196 884 coincidence is thus easy to explain, it nevertheless turns out to be an instructive example of Moonshine.

### 2.2.3 Poisson summation

Theta series (2.2.11a) are sums, over periodic sets, of the exponential of a quadratic polynomial. According to the argument given last subsection, two ingredients go into their modularity: together with Poisson summation (2.2.7a), we also needed the fact that the Fourier transform of the Gaussian $e^{-tx^2}$ is essentially itself. Poisson summation requires the infinite periodic sum. There are many other simple functions $f$ that are likewise nearly invariant under Fourier transform: for example, the Fourier transform over $\mathbb{R}^2$ of $f(x, y) = e^{ix^3/y} y^{-2/3} \mathrm{sign}(y)$ is $\mathrm{i} f(x, -y/27)$. For several other examples, see [**176**]. To see how to use this to get ‘cubic’ analogues of theta functions (which will transform nicely with respect to $\mathrm{SL}_3(\mathbb{Z})$), as well as possible applications to physics, see the intriguing review [**462**] and references therein.

What is the other ingredient, Poisson summation, really saying? Meaning arises from a natural embedding of the particular into a more general context, so let’s try to generalise Poisson summation.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

First, let $G$ be a group – we require it to be a topological group (separable and locally compact). As defined in Section 1.5.5, its unitary dual $\widehat{G}$ consists of all unitary irreducible representations. For example, the unitary duals of $\mathbb{R}$ and $\mathbb{Z}$ can be identified with $\mathbb{R}$ and $S^1$, respectively, while the unitary dual of compact groups (like finite $G$ or $G = S^1$) consists of a discrete set of points. When the group is abelian, the representations $\pi \in \widehat{G}$ are all one-dimensional; the dual $\widehat{G}$ itself forms an abelian group, and *Pointrjagin duality* says that the double-dual $\widehat{\widehat{G}}$ is isomorphic to $G$. For example, the representations in $\widehat{\mathbb{R}}$ look like $\psi_\lambda(x) = e^{2\pi \mathrm{i}\lambda x}$ for each $\lambda \in \mathbb{R}$, so $\widehat{\mathbb{R}} \cong \mathbb{R}$. When $G$ is non-abelian, Pointrjagin duality becomes the more abstract Tannaka–Krein duality of Section 1.6.2.

Let us begin with abelian groups. Let $\Gamma$ be a (discrete) subgroup of an abelian group $G$, such that the quotient $\Gamma\backslash G$ is compact. The theta series modularity arguments last subsection correspond to the choices $G = \mathbb{R}$ and $\Gamma = \mathbb{Z}$ and, more generally, $G = \mathbb{R}^n$ and $\Gamma = L$; of course the circle $\mathbb{Z}\backslash\mathbb{R}$ and the $n$-torus $L\backslash\mathbb{R}^n$ are compact.

The Fourier transform $f \mapsto \widehat{f}$ for the group $G$ – explicitly, $\widehat{f}(\psi) = \int_G f(x)\,\overline{\psi(x)}\,\mathrm{d}x$ – is a unitary map taking Schwartz functions on $G$ to Schwartz functions on the dual $\widehat{G}$. Incidentally, the integrals here and below are with respect to the invariant Haar measure (Section 1.5.4). Then the classical Poisson summation (2.2.7a) becomes

$$\int_\Gamma f(\gamma)\,\mathrm{d}\gamma = \int_{\Gamma^\perp} \widehat{f}(\psi)\,\mathrm{d}\psi, \tag{2.2.12}$$

where $\Gamma^\perp$ consists of all $\psi \in \widehat{G}$ such that $\psi(\gamma) = 1$ for all $\gamma \in \Gamma$. The integrals here reduce to sums, thanks to discreteness. It is through $\Gamma^\perp$ that the dual lattice $L^*$ enters into (2.2.11c). Since $\mathbb{Z}^\perp = \mathbb{Z}$, we find that (2.2.12) is indeed a generalisation of (2.2.7a).

(2.2.12) is too easy a generalisation to help us much.The meaning of Poisson summation, and of (2.2.12), becomes a little clearer when we generalise to non-abelian groups. Let $\Gamma$ now be an arbitrary discrete closed subgroup of a separable locally compact group $G$. $G$ and $\Gamma$ may or may not be abelian. For simplicity we assume that the coset space $\Gamma\backslash G$ is compact. Then $\Gamma\backslash G$ has a finite invariant measure, and the space $L^2(\Gamma\backslash G)$ of square-integrable functions forms a Hilbert space (Section 1.3.1). The regular representation $R$ of $G$ on $L^2(\Gamma\backslash G)$ is defined by $(R(x)f)(y) = f(yx)$, as usual, and is unitary. This representation decomposes as a direct sum of irreducible unitary representations:

$$L^2(\Gamma\backslash G) = \oplus_{\pi\in\widehat{G}} m_\pi \pi,$$

where the numbers $m_\pi \geq 0$ are the (finite) multiplicities.

Even though $R$ is infinite-dimensional, we can define a character for it as follows. For any sufficiently nice function $\phi$ on $G$ (e.g. $\phi$ smooth and of compact support), define the operator $R(\phi) = \int_G \phi(y)\,R(y)\,\mathrm{d}y$ on $L^2(\Gamma\backslash G)$ by

$$(R(\phi)f)(x) = \int_G \phi(y)\,f(xy)\,\mathrm{d}y.$$

This assignment $\phi \mapsto R(\phi)$ forms a representation of the algebra of smooth functions with compact support, with multiplication given by convolution $\phi * \phi'$. The trace of an operator is defined to be the sum of its eigenvalues. It can be shown that the trace tr $R(\phi)$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

exists, and in fact equals

$$\sum_{\pi\in\widehat{G}} m_\pi \operatorname{tr}\pi(\phi) = \sum_{\gamma\in T} \operatorname{vol}(\Gamma_\gamma\backslash G_\gamma)\int_{G_\gamma\backslash G}\phi(x^{-1}\gamma x)\,\mathrm{d}x, \tag{2.2.13}$$

where $T$ is a set of conjugacy class representatives in $\Gamma$, and $\Gamma_\gamma$ and $G_\gamma$ are the stabilisers of $\gamma$ in $\Gamma$ and $G$, respectively (e.g. $\Gamma_\gamma = \{g \in \Gamma \mid g\gamma g^{-1} = \gamma\}$). The left side of (2.2.13) is obviously spectral, that is involves eigenvalues. The right side is geometric; the integral over $G_\gamma\backslash G$ is called an ‘orbital integral’. Equation (2.2.13) has an immediate generalisation: replace the regular representation $R$ of $G$ on $L^2(\Gamma\backslash G)$ with any representation of $G$ induced from a finite-dimensional unitary representation $\rho$ of $\Gamma$. The trivial representation of $\Gamma$ yields the regular representation $R$. [**20**] gives the straightforward proof of (2.2.13) as well as other generalisations.

In the abelian case (e.g. $G = \mathbb{R}^n$, $\Gamma = L$), all $m_\pi = 0$ or 1 and $\Gamma^\perp$ consists of all $\pi \in \widehat{G}$ with $m_\pi = 1$, and (2.2.13) reduces to (2.2.12). In effect we have reinterpreted the Fourier transform $\widehat{f}(\psi)$ by fixing $\psi \in \widehat{G}$ and varying the function $f$, as a sort of character value for the (possibly infinite-dimensional) irreducible representation $\psi$. Another special case of (2.2.13) is to take the group $G$ to be finite, in which case it reduces to Frobenius reciprocity. Interesting finite group applications are described in chapters 22–25 of [**522**].

Equation (2.2.13) is called the *Selberg trace formula*; there is a more complicated version (due in fuller generality to Arthur) when $\Gamma\backslash G$ is noncompact (in which case there are continuous parts to the spectrum). Selberg (a 1950 Fields medalist) was most interested in the case where $G = \mathrm{SL}_2(\mathbb{R})$ and, for example, $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, which has noncompact quotient. For this $G$ he found explicit expressions for the orbital integrals, and the resulting trace formula has powerful consequences.

The Selberg trace formula (2.2.13) can be thought of as an expression for the character of the regular representation of $G$ on $L^2(\Gamma\backslash G)$. This expression is geometric in the sense that for typical groups, the quantities on the right-side typically have geometric interpretations (e.g. for $G = \mathrm{SL}_2(\mathbb{R})$, and $\Gamma$ a Fuchsian group acting without fixed points, the orbital integrals can be expressed using lengths of closed geodesics on the compact Riemann surface $\Gamma\backslash\mathbb{H}$). Of course these orbital integrals, and hence much of the potential geometry, are trivial in the abelian group case used last section.

Although Poisson summation, and its generalisations like the Selberg trace formula, play a central role in the theory of automorphic forms and Langlands programme, they have only played sporadic roles so far in Moonshine and conformal field theory. For example, [**130**] applies the Selberg trace formula to string theory, to find the trace of the heat kernel. Orbital integrals also play a fundamental role in the approach [**346**] to understand group representations via coadjoint orbits; I. Frenkel extended this method to express the characters of affine Kac–Moody algebras as orbital integrals [**198**], and in this way obtained new proofs of the Macdonald identities. It seems unlikely though that Poisson’s and Selberg’s formulae can provide a unified explanation of all modularity proofs in Moonshine. A rigorous proof in mathematics may be too slick, much as a painting can be too photographic. It seems to this author that, although Poisson

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

summation permits a quick proof of theta function modularity, it doesn't tell us *why* it's true. A conceptual proof should open the door to natural generalisations of the given theorem, by underscoring the confluence of properties needed for that theorem to hold.

### *2.2.4 Hauptmoduls*

Let's identify the orbit space $\mathrm{SL}_2(\mathbb{Z})\backslash\overline{\mathbb{H}}$, by studying the fundamental domain $D$ of Figure 2.3. Apart from the boundary of $D$, every $\mathrm{SL}_2(\mathbb{Z})$-orbit will intersect $D$ in one and only one point. But what should we do about the boundary? Well, the edge $\mathrm{Re}(\tau) = -\frac{1}{2}$ gets mapped by the translation $T : \tau \mapsto \tau + 1$ to the edge $\mathrm{Re}(\tau) = \frac{1}{2}$, so we should identify these, i.e. glue them together. The result is a cylinder running off to infinity, with a strange lip at the bottom. The inversion $S : \tau \mapsto -1/\tau$ tells us how we should close that lip: identify $\mathrm{i}e^{\mathrm{i}\theta}$ and $\mathrm{i}e^{-\mathrm{i}\theta}$. This seals the bottom of the cylinder, so we get an infinitely tall cup with a strangely puckered base. In fact the top of this cup is also capped off, by the cusp $\mathrm{i}\infty$. So what we have (topologically speaking) is a *sphere*. It inherits the smoothness of $\mathbb{H}$ except for conical singularities at the fixed points i and $e^{\pi \mathrm{i}/3}$. The cusps are responsible for compactness. This interpretation of $\mathrm{SL}_2(\mathbb{Z})\backslash\overline{\mathbb{H}}$ means that a modular function for $\mathrm{SL}_2(\mathbb{Z})$ can be reinterpreted as a meromorphic complex-valued function on this sphere. There is a canonical sphere in complex analysis, namely the Riemann sphere $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$. The meromorphic functions on the Riemann sphere must be *rational*, that is of the form $f(w) = \frac{\text{some polynomial } P(w)}{\text{some polynomial } Q(w)}$, where $w$ is the complex parameter on the Riemann sphere. So a modular function $f(\tau)$ for $\mathrm{SL}_2(\mathbb{Z})$ is simply some rational function $P/Q$ evaluated at the change-of-local-parameters, or at the uniformising function $w = c(\tau)$ that maps us from our sphere $\Gamma\backslash\overline{\mathbb{H}}$ to the Riemann sphere. There are many different choices for this function $c(\tau)$, but the standard one is the $j$-function:[7]

$$j(\tau) := \frac{\left(1 + 240\sum_{n=1}^{\infty} \sigma_3(n)\, q^n\right)^3}{q\prod_{n=1}^{\infty}(1-q^n)^{24}} = \frac{\Theta_{E_8}(\tau)^3}{\eta(\tau)^{24}} = q^{-1} + 744 + 196\,884\, q + \cdots \tag{2.2.14}$$

(see also (0.1.8)), where $\sigma_3$ is in (2.2.3b), $\Theta_{E_8}$ is the theta series of the $E_8$ root lattice (2.2.11a) and $\eta$ is the Dedekind eta (2.2.6b). Thus, any modular function for $\mathrm{SL}_2(\mathbb{Z})$ can be written as a rational function $f(\tau) = P(j(\tau))/Q(j(\tau))$ in the $j$-function. Conversely, any such function is modular.

This is analogous to (and much stronger than) saying that any function $g(x)$ periodic under $x \mapsto x + 1$ is really a function on the unit circle $S^1 \subset \mathbb{C}$ evaluated at the uniformising function $x \mapsto e^{2\pi \mathrm{i} x}$, and hence has a Fourier expansion $\sum_n g_n \exp[2\pi \mathrm{i} n x]$.

We can generalise the argument that led to $j$. Recall (2.2.4).

**Definition 2.2.4** *Call a discrete subgroup* $\Gamma$ *of* $\mathrm{SL}_2(\mathbb{R})$ *a* congruence subgroup *if it contains some* $\Gamma(N)$. *Call it of moonshine-type if it contains some* $\Gamma_0(N)$*, and obeys*

$$\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \in \Gamma \;\Rightarrow\; t \in \mathbb{Z}. \tag{2.2.15}$$

[7] Historically, $j$ was the standard choice, but in Monstrous Moonshine the preferred choice would be the function $J = j - 744$ with zero constant term.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

The congruence subgroups are relatively rare among finite index subgroups of $\mathrm{SL}_2(\mathbb{Z})$, but their theory is much better developed. Let $f$ be a modular function for a congruence subgroup $\Gamma$. Then we can expand $f$ as a Laurent series in $q^{1/N}$. We analyse this as before: look at the orbit space $\Sigma = \Gamma\backslash\overline{\mathbb{H}}$; because $\Gamma$ is not too big, $\Sigma$ will be a Riemann surface; because $\Gamma$ is not too small, $\Sigma$ will be compact.

We call $\Gamma$ 'genus $g$' if its surface $\Sigma$ has genus $g$. If $\Gamma$ is a subgroup of $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$, and without loss of generality we have $-I \in \Gamma$, then the genus is given by

$$g = 1 + \frac{n}{12} - \frac{n_2}{4} - \frac{n_3}{3} - \frac{n_\infty}{2}, \tag{2.2.16}$$

where $n$ is the index $\|\Gamma(1)/\Gamma\|$ of $\Gamma$ in $\Gamma(1)$, and where $n_k$ ($k = 2, 3, \infty$) is the number of $\Gamma$-orbits of order-$2k$ fixed points. For the easy proof from the Hurwitz formula, see proposition 1.40 of [**505**]. Note that $n_\infty$ is the number of punctures of $\Gamma\backslash\mathbb{H}$. For example, for $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ we have $n = 1 = n_2 = n_3 = n_\infty$ and we recover our result that the genus is 0. The values $n, n_2, n_3, n_\infty$ for all $\Gamma(N)$ and $\Gamma_0(N)$ are given in Section 1.6 of [**505**].

For example, $\Gamma = \Gamma_0(2)$ and $\Gamma = \Gamma_0(25)$ are both genus 0 (with 2, respectively 6, punctures), while $\Gamma_0(50)$ is genus 2 with 12 punctures and $\Gamma_0(24)$ is genus 1 with 7 punctures. Once again, we are interested here in the genus-0 case. As before, this means that there is a uniformising function $J_\Gamma$ that is a modular function for $\Gamma$, and all other modular functions for $\Gamma$ can be written as a rational function in it. Because of (2.2.15), we can choose $J_\Gamma$ to look like

$$J_\Gamma(\tau) = q^{-1} + a_1(\Gamma)\,q + a_2(\Gamma)\,q^2 + \cdots$$

So $J_\Gamma$, the *Hauptmodul* for $\Gamma$, plays exactly the same role for $\Gamma$ that $J := j - 744$ plays for $\mathrm{SL}_2(\mathbb{Z})$. For example, $\Gamma_0(2)$, $\Gamma_0(13)$ and $\Gamma_0(25)$ are all genus 0, with Hauptmoduls

$$J_2(\tau) = q^{-1} + 276\,q - 2048\,q^2 + 11202\,q^3 - 49152\,q^4 + 184024\,q^5 + \cdots, \tag{2.2.17a}$$

$$J_{13}(\tau) = q^{-1} - q + 2\,q^2 + q^3 + 2\,q^4 - 2\,q^5 - 2\,q^7 - 2\,q^8 + q^9 + \cdots, \tag{2.2.17b}$$

$$J_{25}(\tau) = q^{-1} - q + q^4 + q^6 - q^{11} - q^{14} + q^{21} + q^{24} - q^{26} + \cdots \tag{2.2.17c}$$

The smaller (sparser) the modular group, the smaller the coefficients of the Hauptmodul. In this sense, the $j$-function is optimally bad among the Hauptmoduls: for example, for it $a_{23} \approx 10^{25}$.

In Theorem 2.1.5 we see what happens in genus $> 0$: two generators, not one, are needed, although they will be polynomially related.

As is mentioned in Chapter 0, Monstrous Moonshine is interested directly in genus-0 groups. We construct certain functions associated with the Monster, and it turns out unexpectedly that these functions are actually Hauptmoduls.

An obvious question is, how many genus-0 groups (equivalently, how many Hauptmoduls) are there? It turns out that $\Gamma_0(p)$ is genus 0, for a prime $p$, iff $p - 1$ divides 24. Thompson [**526**] proved that for any $g$, there are only finitely many genus-$g$ groups of moonshine type. Cummins [**121**] has shown that there are in fact exactly 6486 genus-0

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

groups of moonshine type. 616 of these have Hauptmoduls with integer coefficients $a_i(\Gamma)$, and all of the remainder have $q$-coefficients in some cyclotomic field.

Question 2.2.1. How important are the conditions at the cusps for the definition of modular functions or forms? For example, describe all functions $f$ holomorphic on $\mathbb{C}$, symmetric with respect to $\mathrm{SL}_2(\mathbb{Z})$ (i.e. $f(\gamma.\tau) = f(\tau)$ for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$), but which need not be holomorphic or even meromorphic at the cusps (i.e. $f$ may have an essential singularity there).

Question 2.2.2. Show that if $f$ is a modular form of weight $k$, and 3 doesn't divide $k$, then $f(e^{2\pi \mathrm{i}/3}) = 0$.

Question 2.2.3. Suppose $f$ is a modular form, not identically 0, for some $\Gamma$, with multiplier $\mu$ and *integral* weight $k$. Prove that $\mu$ must be a one-dimensional representation of $\Gamma$. Where does the proof go wrong if $k$ is fractional?

Question 2.2.4. Prove Poisson summation (2.2.7a). (*Hint*: $x \mapsto \widetilde{f}(x) = \sum_{n\in\mathbb{Z}} f(n+x)$ is periodic, so can be Fourier expanded. Compute $\widetilde{f}(0)$ in two different ways.)

Question 2.2.5. By modifying slightly the argument beginning with (2.2.9a), prove (2.2.9d) and thus (2.2.8b).

Question 2.2.6. Let $L$ be any self-dual positive-definite lattice. Then $\Theta_L(\tau)$ is a polynomial in $\theta_3(\tau)$ and $\Theta_{E_8}(\tau)$ (you can assume this, which is proved for instance in [**503**]). Using this fact, show that the theta function for any self-dual positive-definite lattice of dimension $< 24$ is uniquely determined by the numbers $N_1, N_2$ of norm-squared 1- and 2-vectors.

Question 2.2.7. Let $L$ be a positive-definite 24-dimensional even self-dual lattice. Prove that $\Theta_L(\tau)/\eta(\tau)^{24} = J(\tau) + c_L$ for some constant $c_L$. Find that constant.

Question 2.2.8. Find the genus of $\Gamma(2)$, using (2.2.16).

## 2.3 Further developments

### *2.3.1 Dirichlet series*

One of the most remarkable formulae in science is surely

$$1 + 2 + 3 + 4 + \cdots = -\frac{1}{12}. \tag{2.3.1}$$

Of course the right side is the value at $s = -1$ of the Riemann zeta function (2.2.3c). The expressions in (2.2.3c) converge absolutely when $\mathrm{Re}(s) > 1$, where $\zeta$ is then holomorphic, and $\zeta$ has a unique holomorphic extension to all of $\mathbb{C}$, except for a simple pole at $s = 1$ (the harmonic series). Equation (2.3.1) is used in quantum field theory in the context of zeta function regularisation (2.2.10); it is related to the $q^{1/24}$ in the Dedekind eta function (2.2.6b) and the normalisation $C/12$ in Lie brackets (3.1.5a) of the Virasoro algebra.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

The equality of the infinite sum and product in (2.2.3c) is merely an analytic reformulation of unique factorisation in $\mathbb{Z}$, but it shows crucially the relation between $\zeta(s)$ and the primes. For a trivial example, taking logs of (2.2.3c) quickly gives the divergence of $\sum 1/p$.

As important as analytic continuation and the product expansion are, more important for us is the *functional equation*

$$\Lambda(1-s) = \Lambda(s), \tag{2.3.2}$$

where $\Lambda(s) := \pi^{-s/2}\Gamma(s/2)\,\zeta(s)$, using the Gamma function

$$\Gamma(s) := (2\pi)^s \int_0^\infty e^{-2\pi y} y^{s-1} \mathrm{d}y.$$

Indeed, Hecke discovered that (2.3.2) is equivalent to modularity (2.2.7c).

**Theorem 2.3.1 (Hecke, 1936)** *Let* $f(\tau) = \sum_{n=0}^{\infty} a_n e^{2\pi \mathrm{i} n\tau/d}$ *and* $\phi(s) = \sum_{n=1}^{\infty} a_n n^{-s}$, *where* $|a_n| < Cn^c$ *for some constants* $d, C, c$. *Define* $\Phi(s) = (2\pi/d)^{-s}\Gamma(s)\,\phi(s)$. *Then the following two statements are equivalent:*

(i) $f(\frac{-1}{\tau}) = \left(\frac{\tau}{\mathrm{i}}\right)^k f(\tau)$;

(ii) $\Phi(k-s) = \Phi(s)$, *and* $\Phi(s) + \frac{a_0}{s} + \frac{a_0}{k-s}$ *is holomorphic and bounded in each vertical strip in* $\mathbb{H}$.

*Proof:* The key idea of the proof is that $\Phi(s)$ and $f(\tau)$ are related by the Mellin transform:

$$\Phi(s) = \int_0^\infty x^{s-1}\,(f(\mathrm{i}x) - a_0)\,\mathrm{d}x, \tag{2.3.3a}$$

$$f(\mathrm{i}x) - a_0 = \frac{1}{2\pi\mathrm{i}} \int_{\mathrm{Re}(s)=a} x^{-s}\,\Phi(s)\,\mathrm{d}s, \tag{2.3.3b}$$

for any constant $a > 0$ sufficiently large.

To prove (ii) from (i), write $\int_0^\infty = \int_0^1 + \int_1^\infty$ in (2.3.3a), so we get the sum $\Phi(s) = \Phi_0 + \Phi_\infty$. Note that $\Phi_\infty(s)$ is clearly holomorphic everywhere, and $\phi_0(s)$ is holomorphic when $\mathrm{Re}(s)$ is sufficiently large. Then, using (i), for those $s$

$$\begin{aligned}\Phi_0(s) &= \int_0^1 x^{s-1}\,(f(\mathrm{i}x) - a_0)\,\mathrm{d}x = \int_1^\infty x^{-s-1}x^k f(\mathrm{i}x)\,\mathrm{d}x - \frac{a_0}{s}\\ &= \Phi_\infty(k-s) - \frac{a_0}{s} - \frac{a_0}{k-s}.\end{aligned}$$

Therefore $\Phi_0(s)$ extends holomorphically everywhere, except for simple poles at $s = 0$ and $s = k$, and $\Phi_0(s) = \Phi_\infty(k-s) - a_0 s^{-1} - a_0(k-s)^{-1}$ holds $\forall\, s \neq 0, k$. Thus

$$\begin{aligned}\Phi(k-s) &= \Phi_0(k-s) + \Phi_\infty(k-s)\\ &= \left(\Phi_\infty(s) - \frac{a_0}{k-s} - \frac{a_0}{s}\right) + \left(\Phi_0(s) + \frac{a_0}{s} + \frac{a_0}{k-s}\right) = \Phi(s).\end{aligned}$$

To prove (i) from (ii), shift the vertical contour $\mathrm{Re}(s) = a > 0$ in (2.3.3b) to the left, to $\mathrm{Re}(s) = b < 0$, and pick up residues $-a_0$ at $s = 0$ and $x^{-k}a_0$ at $s = k$:

$$\begin{aligned} f(\mathrm{i}x) - a_0x^{-k} &= \frac{1}{2\pi\mathrm{i}}\int_{\mathrm{Re}(s)=b} x^{-s}\,\Phi(s)\,\mathrm{d}s = \frac{1}{2\pi\mathrm{i}}\int_{\mathrm{Re}(s)=k-b} x^{-(k-s)}\Phi(s)\,\mathrm{d}s \\ &= x^{-k}\,(f(\mathrm{i}/x) - a_0). \end{aligned}$$

Therefore $f(\mathrm{i}/x) = x^k f(\mathrm{i}x)$, and (i) follows by analytic continuation. ■

When $f$ is a modular form, we call $\phi$ the *Dirichlet series* or *L-function* corresponding to $f$ (the term L-function is usually reserved for those $\phi$ which also have product expansions as in (2.2.3c)). The modular form corresponding to the Riemann zeta function $\zeta(s)$ is $f(\tau) = \frac{1}{2}\theta_3(\tau)$. Theorem 2.3.1 applies with $k = \frac{1}{2}$, $d = 2$ and $\Lambda(2s) = \Phi(s)$, and relates (2.3.2) directly to (2.2.7c). Another famous example, due to Ramanujan, is $f = \eta^{24}$. Its $\Phi$ is holomorphic everywhere and its $\phi$ has a product form $\prod_p(1 - \tau(p)p^{-s} + p^{11-2s})^{-1}$, where $\tau$ here is the so-called *Ramanujan tau-function* (see e.g. (3.4.6)).

Mysteriously, we can associate Dirichlet series to many of the basic objects of arithmetic – modular forms, number fields, algebraic varieties, etc. – in such a way that basic operations performed on, and relations between, the Dirichlet series correspond to natural operations on, and relations between, the arithmetic objects. In its most general form, this is Langlands functoriality. For a famous special case, given an elliptic curve $E$ defined over $\mathbb{Q}$, its L-function keeps track of the number of points on $E$ as we vary its field of definition from $\mathbb{Q}$ to the finite fields. The Taniyama–Shimura Conjecture states that $E$ is modular, i.e. that this L-function is the Dirichlet series of a modular form of weight 2. As we know, Wiles *et al.* proved Taniyama–Shimura and hence Fermat's Last Theorem.

See [**456**] for a clear treatment of the material of this subsection. We have been hurried since there is at this point no evidence for its direct relevance to Moonshine. There are many generalisations of Theorem 2.3.1. Let us mention one. Generators for the groups $\mathrm{SL}_2(\mathbb{Z})$ and $\Gamma_\theta$ are given in (2.2.1a) and (2.2.5), so Theorem 2.3.1 gives a Dirichlet series characterisation for $f$ to be a modular form for those groups. When $\Gamma$ is smaller (say $\Gamma = \Gamma(N)$), to which Dirichlet series conditions does the modularity of $f$ translate? The list of generators is far more complicated. An answer is provided by Weil's Converse Theorem (Section 2.3.3).

### *2.3.2 Jacobi forms*

The general quadratic polynomial in one variable $x$ looks like $ax^2 + bx + c$, so we might try to generalise $\theta_3(\tau)$ by replacing $n^2\tau$ with $an^2\tau + bnz + cu$. Consider then the function

$$\theta_3(\tau, z, u) = \sum_{n\in\mathbb{Z}} e^{\pi\mathrm{i}\tau n^2 + 2\pi\mathrm{i}zn + 2\pi\mathrm{i}u}, \tag{2.3.4}$$

where $\tau, z, u \in \mathbb{C}$. We've seen these kinds of functions before in (2.1.7a). The $2\pi\mathrm{i}$'s in front of $z$ and $u$ are conventional. As before, convergence requires $\tau \in \mathbb{H}$. Obviously, the $u$-dependence is rather trivial and is retained only for book-keeping.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Fix $\tau \in \mathbb{H}$ and $u \in \mathbb{C}$, and consider this as a function of $z \in \mathbb{C}$. It has period 1 and quasi-period $\tau$:

$$\theta_3(\tau, z + m\tau + \ell, u + mz + m^2\tau/2) = \theta_3(\tau, z, u), \qquad \forall m, \ell \in \mathbb{Z}, \tag{2.3.5a}$$

and thus is a function living (projectively) on the torus $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$.

Next, fix $z, u \in \mathbb{C}$ and consider $\theta_3$ as a function of $\tau \in \mathbb{H}$. Completing the square $\tau n^2 + 2nz = \tau\,(n + \frac{z}{\tau})^2 - \frac{z^2}{\tau}$ and restricting $\tau, z$ to the imaginary axis, Poisson summation (2.2.7a) and analytic continuation gives us

$$\theta_3(\tau, z, u) = \sqrt{\frac{\mathrm{i}}{\tau}}\,\theta_3\left(\frac{-1}{\tau}, \frac{z}{\tau}, u - \frac{z^2}{2\tau}\right), \tag{2.3.5b}$$

valid for all $\tau \in \mathbb{H}$ and $z, u \in \mathbb{C}$.

**Definition 2.3.2 [170]** *By a* Jacobi form *for $SL_2(\mathbb{Z})$ of weight $k$ and index $m$ we mean a holomorphic function $f : \mathbb{H} \times \mathbb{C} \to \mathbb{C}$ satisfying*

$$f\left(\frac{a\tau + b}{c\tau + d}, \frac{z}{c\tau + d}\right) = (c\tau + d)^k \exp\left[2\pi\mathrm{i}\frac{mcz}{c\tau + d}\right] f(\tau, z), \tag{2.3.6a}$$

$$f(\tau, z + \ell\tau + n) = \exp[-2\pi\mathrm{i}m\,(\ell^2\tau + 2\ell z)]\, f(\tau, z), \tag{2.3.6b}$$

*for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ and $\ell, n \in \mathbb{Z}$. Moreover, $f$ must have a Fourier expansion of the form*

$$f(\tau, z) = \sum_{n \in \mathbb{N}} \sum_{r \in \mathbb{Z},\, r^2 \leq 4mn} c_{n,r}\, e^{2\pi\mathrm{i}\,(n\tau + rz)}. \tag{2.3.6c}$$

Similarly, we call $\theta_3(\tau, z, 0)$ a Jacobi form of weight $\frac{1}{2}$ and index 0 for $\Gamma_\theta$. The Weierstrass $\mathfrak{p}$-function $\mathfrak{p}(\tau, z)$ in (2.1.6a) is a Jacobi form for $\mathrm{SL}_2(\mathbb{Z})$ of level 2 and index 1 (Question 2.3.1). A Jacobi form is a natural blend of the notions of modular form and elliptic function: the parameter $\tau \in \mathbb{H}$ tells us where on the moduli space of tori we are, and the parameter $z$ lives on that torus. Given such classical examples, it is hard to understand why their theory was developed only in the 1980s. The introduction of the index $m$ in Definition 2.3.2 may be somewhat unexpected, but is explained in Section 2.4.1.

We can generalise the example (2.3.4) to lattices (and in fact to translates of lattices). Let $L$ be an $n$-dimensional lattice in $\mathbb{R}^n$. Define

$$\Theta_L(\tau, z, u) = \sum_{v \in L} \exp[\pi\mathrm{i}\tau\, v \cdot v + 2\pi\mathrm{i}\, z \cdot v + 2\pi\mathrm{i}u], \tag{2.3.7}$$

where $z \in \mathbb{C}^n$, $u \in \mathbb{C}$ and $\tau \in \mathbb{H}$. The $z$-periods of $\Theta_L$ fill out the dual lattice $L^*$, and the $z$-quasi-periods fill out $\tau L^*$. Provided $L$ is a rational lattice, we get the obvious analogue of (2.2.11c), again from Poisson summation. To make $\Theta_L$ into a Jacobi form for some $\Gamma(N)$ at weight $n$ and index 0, it suffices to embed $z \in \mathbb{C}$ into $\mathbb{C}^n$ along any nonzero dual weight vector $u^* \in L^*$: i.e. $\Theta_L(\tau, zu^*, 0)$ will be a Jacobi form.

As any string theorist knows, there are several different lattices $L, L'$ that have the same theta function: $\Theta_L(\tau) = \Theta_{L'}(\tau)$. Perhaps the most famous example of this is the

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

pair of even self-dual lattices of dimension 16 (namely, $D_{16}^{+}$ and $E_8 \oplus E_8$ [**113**]). Actually there are lattice examples in every dimension $\geq 3$ [**108**]. However, their Jacobi forms are unique in the strongest form possible (see Question 2.3.2).

Writing theta functions as Jacobi forms is crucial to their interpretation as heat kernels, or using Heisenberg groups, as we see in Sections 2.3.4 and 2.4.2. In Theorem 3.2.3 we find that the characters of affine Kac–Moody algebras are Jacobi forms of weight and index 0. Indeed, they are rational functions of lattice Jacobi forms (2.3.7).

An obvious question to ask is, to any modular form $f(\tau)$, is there a Jacobi form $f(\tau, z)$ for the same group and at the same weight such that $f(\tau, 0) = f(\tau, z)$? And if so, is this Jacobi form unique? It turns out that every weight-$k$ modular form $f$, at least for $\mathrm{SL}_2(\mathbb{Z})$, *can* be lifted to a Jacobi form for the same weight and group, at index $m = 1$. This Jacobi form is far from unique, even at $m = 1$. In fact, the redundancy has the same dimension as the space of weight-$k + 2$ cusp forms for $\mathrm{SL}_2(\mathbb{Z})$. This fact is a consequence of theorem 3.5 in [**170**].

### *2.3.3 Twisted #2: shifts and twists*

Recall the classical Jacobi theta functions $\theta_1 = \theta_{\frac{1}{2},\frac{1}{2}}$, $\theta_2 = \theta_{\frac{1}{2},0}$, $\theta_3 = \theta_{0,0}$, $\theta_4 = \theta_{0,\frac{1}{2}}$, using the notation of (2.1.7a). These obey simple modular transformation rules, most concisely stated in vector notation as

$$\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix}(\tau + 1, z) = \begin{pmatrix} e^{\pi \mathrm{i}/4} & 0 & 0 & 0\\ 0 & e^{\pi \mathrm{i}/4} & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\end{pmatrix}\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix}(\tau, z), \tag{2.3.8a}$$

$$\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix}\left(\frac{-1}{\tau}, \frac{z}{\tau}\right) = e^{\pi \mathrm{i} z^2/\tau}\sqrt{\frac{\tau}{\mathrm{i}}}\begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\end{pmatrix}\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix}(\tau, z). \tag{2.3.8b}$$

That is, these $\theta_i$ define a vector-valued Jacobi form for $\mathrm{SL}_2(\mathbb{Z})$ (Definition 2.2.2). The $q$-expansions of $\theta_1$ and $\theta_4$ have negative coefficients; we can make 'positive' combinations of these theta functions that have almost as nice transformations under $\mathrm{SL}_2(\mathbb{Z})$:

$$\theta_{[0]}(\tau, z) = \frac{\theta_3(\tau, z) + \theta_4(\tau, z)}{2} = 1 + q^2(r^2 + r^{-2}) + q^8(r^4 + r^{-4}) + \cdots, \tag{2.3.9a}$$

$$\theta_{[1]}(\tau, z) = \frac{\theta_1(\tau, z) + \theta_2(\tau, z)}{2} = q^{1/8}r^{1/2}(1 + qr^{-2} + q^3r^2 + q^6r^{-4} + \cdots), \tag{2.3.9b}$$

$$\theta_{[2]}(\tau, z) = \frac{\theta_3(\tau, z) - \theta_4(\tau, z)}{2} = q^{1/2}((r + r^{-1}) + q^4(r^3 + r^{-3}) + \cdots), \tag{2.3.9c}$$

$$\theta_{[3]}(\tau, z) = \frac{\theta_2(\tau, z) - \theta_1(\tau, z)}{2} = q^{1/8}r^{1/2}(r^{-1} + qr + q^3r^{-3} + q^6r^3 + \cdots), \tag{2.3.9d}$$

where $r = e^{2\pi \mathrm{i} z}$. Note that $\theta_{[i]}$ has the geometric interpretation as the theta series (2.2.11a) of the translate $2\mathbb{Z} + \frac{i}{2}$.

We regard $\theta_1, \theta_2, \theta_4$ as $\mathbb{Z}_2$-twists and -shifts of $\theta_3$. More generally, the parameter $r \in \frac{1}{N}\mathbb{Z}$ in $\theta_{r,s}$ corresponds to a $\mathbb{Z}_N$-shift, and $s \in \frac{1}{N}\mathbb{Z}$ to a $\mathbb{Z}_N$-twist. A far-reaching generalisation of this simple construction is studied in Section 5.3.6; the analogue there of the positive combinations (2.3.9) is the characters for a vertex operator algebra. In Monstrous Moonshine the twists of $J(\tau)$ are the McKay–Thompson series $J_g(\tau)$, and its more general shifts and twists are the Norton series of Maxi-Moonshine (Section 7.3.2). Physically, this corresponds to the orbifold construction (Section 4.3.4). There, the positive linear combinations have the direct interpretation as graded dimensions of sectors of the conformal field theory.

As always, the clearest example is provided by lattices (Section 1.2.1). Let $L$ be an integral positive-definite lattice and let $r, s$ be two vectors in $\mathbb{Q} \otimes L$. As in (2.1.7a), write

$$\Theta_{L;r,s}(\tau, z) = \sum_{x \in L} e^{\pi \mathrm{i} \tau\, (x+r)\cdot(x+r)} e^{\pi \mathrm{i}\, (z+s)\cdot(2x+r)}, \tag{2.3.10a}$$

where as before $z \in \mathbb{C} \otimes L$. Then $\Theta_{L;r,s}$ will be a Jacobi form for some subgroup of $\mathrm{SL}_2(\mathbb{Z})$, as is (2.3.7). In fact, if $L$ is even and self-dual, we can be much more explicit. For any $r, s \in \mathbb{Q} \otimes L$, and any $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, we have

$$\Theta_{L;r,s}\left(\frac{a\tau + b}{c\tau + d}, \frac{z}{c\tau + d}\right) = (c\tau + d)^{n/2} \exp\left[\pi \mathrm{i} \frac{cz}{c\tau + d}\right] \Theta_{L;ar+cs,br+ds}(\tau, z), \tag{2.3.10b}$$

where $n$ is the dimension of $L$.

As usual, certain positive combinations of these $\Theta_{L;r,s}$ have a direct (geometric) interpretation. Again let $L$ be self-dual, and suppose the vector $s \in \mathbb{Q} \otimes L$ has order $m$ in $L$ (so $ms \in L$). Then there will be a vector $s' \in L$ such that $s \cdot s' \equiv \frac{1}{m}$ (mod 1). For any integer $k$, and vector $r \in \mathbb{Q} \otimes L$, we get this generalisation of (2.3.9):

$$\frac{1}{m} \sum_{j=0}^{m-1} \exp\left[2\pi \mathrm{i} j \left(s \cdot r - \frac{k}{m}\right)\right] \Theta_{L;r,js} = \Theta_{L_0 + r + ks'}, \tag{2.3.10c}$$

the theta series of a translate of the lattice $L_0 = \{v \in L \mid v \cdot s \in \mathbb{Z}\}$.

In the orbifold construction of vertex operator algebras and *chiral* conformal field theory, the role of vectors $r, s$ is played by automorphisms $g, h$ in some group $G$, and the role of the sublattice $L_0$ in (2.3.10c) is played by the vertex operator subalgebra $\mathcal{V}^G$ fixed by $G$. However, as we see in Section 4.3, *full* conformal field theory or string theory involves the interplay of two vertex operator algebras; the orbifold construction there involves in addition a reconstruction of a new full conformal field theory from $\mathcal{V}^G$. We address this further in Sections 4.3.4 and 5.3.6.

This reconstruction is again beautifully illustrated by lattices. Let $L$ be any rational lattice and $T = \{t_i\}$ be a finite set of vectors in $\mathbb{Q} \otimes L$. Then by $L\{T\}$ we mean the set

$$L\{T\} = \left\{x + \sum_i \ell_i t_i \mid \ell_i \in \mathbb{Z},\ x \in L,\ \left(x + \sum_i \ell_i t_i\right) \cdot t_j \in \mathbb{Z}\ \forall j\right\}.$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Then $L\{T\}$ is a lattice *rationally equivalent* to $L$ (i.e. there is an orthogonal transformation $T : \mathbb{Q} \otimes L\{T\} \to \mathbb{Q} \otimes L$). Conversely, if $L_1$ and $L_2$ are rationally equivalent integral lattices, then there is a finite set $T = \{t_1, \ldots, t_m\} \subset \mathbb{Q} \otimes L_1$ such that $L_1\{T\}$ is isomorphic to $L_2$ [**238**]. Clearly the theta series of $L\{T\}$ is the average of $\Theta_{L;r,s}$ for a finite number of $r, s$ in the $\mathbb{Z}$-span of $T$. The important special case is when $L$ is self-dual; then $L\{T\}$ will also be self-dual provided all $t_i \cdot t_j \in \mathbb{Z}$. In this case,

$$L\{T\} = \bigcup_{\ell_i \in \mathbb{Z}} \left( L_0 + \sum_i \ell_i t_i \right),$$

where $L_0 = \{x \in L \mid x \cdot t_i \in \mathbb{Z}\}$. Call two self-dual lattices $L_1, L_2$ *neighbours* if there is some vector $t$ with integer length-square $t \cdot t$ such that $2t \in L_1$, and $L_2$ and $L_1\{t\}$ are isomorphic. Then any two self-dual lattices, with equal dimensions $n_+ + n_-$ and signature $n_+ - n_-$, will be neighbours of neighbours of $\cdots$ of neighbours of each other [**238**].

Another way to collect some of these results is through *Dirichlet characters*, which are important in the classical theory of modular forms. A Dirichlet character is a function $\chi : \mathbb{Z} \to \mathbb{C}$, with some period $N$, such that $\chi(a) \neq 0$ iff $a$ is coprime to $N$, and for all $a, b \in \mathbb{Z}$ $\chi(ab) = \chi(a)\chi(b)$. Dirichlet introduced these $\chi$ in his proof that there are infinitely many primes in any arithmetic series $a, a+b, a+2b, \ldots,$ provided only that $a$ and $b$ are coprime (clearly a necessary condition). He proved this by twisting the Riemann zeta function (2.2.3c) by $\chi$:

$$L(\chi, s) = \sum_{i=1}^{\infty} \chi(n) n^{-s} = \prod_p (1 - \chi(p) p^{-s})^{-1}. \tag{2.3.11}$$

Given the lesson of Section 2.3.1, it should also be interesting to Dirichlet-twist modular forms.

Modular forms and functions for the principal congruence subgroup $\Gamma(N)$ can be defined as in Definitions 2.2.1 and 0.1, except now there are several orbits of cusps, and we have invariance under only $\begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix}$, so the $q$-expansion takes the form

$$f(\tau) = \sum_{n \in \mathbb{Z}} a_n e^{2\pi \mathrm{i} n \tau / N} = \sum_{n \in \mathbb{Z}} a_n q^{n/N}. \tag{2.3.12}$$

Given any Dirichlet character $\chi$, we can twist this function $f$ and obtain

$$f_\chi(\tau) = \sum_{n \in \mathbb{Z}} \chi(n)\, a_n q^{n/N}. \tag{2.3.13}$$

Then if $f$ is a modular form for $\Gamma(N)$, $f_\chi$ will be a modular form of the same weight for some $\Gamma(M)$. It isn't very deep that modularity should be preserved – see Question 2.3.4 for one such argument. Theorem 14 in [**456**] provides a generalisation. The Dirichlet twist takes on a clear algebraic significance in the context of automorphic representations (Section 2.4.1).

A deeper use of Dirichlet twists is *Weil's Converse Theorem* (see e.g. theorem 17 of [**456**] or page 64 of [**90**]), which characterises modular forms for $\Gamma(N)$ by generalising

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Theorem 2.3.1, using infinitely many Dirichlet twists. It is a 'converse' in that it generalises the converse of (i) ⇒ (ii). Applications of this are given in sections 1.9 and 1.10 of [**89**].

A more surprising example of twisting is by Galois automorphisms. Let $F_N$ be the space (in fact field) of all modular functions for $\Gamma(N)$, with $q$-expansion as in (2.3.12), where each coefficient $a_i$ lies in the cyclotomic field $\mathbb{Q}[\xi_N]$ (recall Section 1.7.3). This field $F_N$ is explicitly constructed in section 6.2 of [**505**]. Clearly, $j(\tau)$ lies in each $F_N$. It can be shown that $F_N$ is a Galois extension over $\mathbb{Q}(j)$, with Galois group

$$\mathrm{Gal}(F_N/\mathbb{Q}(j(\tau))) \cong \mathrm{GL}_2(\mathbb{Z}_N)/\{\pm 1\} \tag{2.3.14a}$$

(see Section 1.7.2 for definitions). For any matrix $\overline{A} \in \mathrm{GL}_2(\mathbb{Z}_N)$, we can find an integer $\ell \in \mathbb{Z}_N^\times$ (namely $\ell = \det(A)$) and a matrix $B = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ such that $\overline{A} = B \begin{pmatrix} 1 & 0 \\ 0 & \ell \end{pmatrix}$ (mod $N$). Then the action of $\overline{A} \in \mathrm{GL}_2(\mathbb{Z}_N)$ on a modular function $f(\tau)$ is given by

$$\overline{A}.f(\tau) = (\sigma_\ell f)\left(\frac{a\tau + c}{b\tau + d}\right), \tag{2.3.14b}$$

$$\sigma_\ell \sum_{n\in\mathbb{Z}} a_n q^{n/N} = \sum_{n\in\mathbb{Z}} \sigma_\ell(a_n)\, q^{n/N}, \tag{2.3.14c}$$

where $\sigma_\ell \in \mathrm{Gal}(\mathbb{Q}[\xi_N]/\mathbb{Q})$ sends $\xi_N$ to $\xi_N^\ell$. This Galois action plays a technical but important role in both Moonshine (e.g. Question 7.3.3) and rational conformal field theory (e.g. Section 6.1); see Section 6.3.3 for some speculation.

### *2.3.4 The remarkable heat kernel*

Various topological proofs of modularity, inspired by conformal field theory, have arisen in recent years. For instance [**24**], [**203**] and section 6 of [**502**] all provide proofs for $\eta(\tau)$. These suggest the thought that, more generally, modularity – hence Moonshine – may be a topological effect (Section 7.2.4). The oldest and perhaps most fundamental observation along these lines is the relation between theta function modularity and the heat kernel.

Fourier determined that the rate of flow of heat energy in a material is proportional to the gradient of the temperature, and thus wrote down the *diffusion* or *heat equation*, which in one dimension looks like

$$\frac{\partial}{\partial t}u(t,x) = \frac{1}{4\pi}\frac{\partial^2}{\partial x^2}u(t,x), \qquad \forall x \in \mathbb{R}, \forall t > 0 \tag{2.3.15a}$$

(the harmless normalisation $1/4\pi$ is introduced for later convenience). Suppose that the initial distribution of heat in the infinite rod is $f(x) = \lim_{t\to 0} u(t,x)$. Then Fourier analysis tells us how to find a solution $u(t,x)$ for all times $t$. Letting

$$\widehat{u}(t,\alpha) = \frac{1}{2\pi}\int_{-\infty}^{\infty} u(t,y)\, e^{-\mathrm{i}\alpha y}\, \mathrm{d}y, \qquad \widehat{f}(\alpha) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(y)\, e^{-\mathrm{i}\alpha y}\, \mathrm{d}y,$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

the equation to be solved has been transformed to $\partial\widehat{u}/\partial t = -\alpha^2\widehat{u}/4\pi$, with initial condition $\widehat{f}$, which has the solution $\widehat{u}(t,\alpha) = \widehat{f}(\alpha)\, e^{-\alpha^2 t/4\pi}$. We can now find $u$ by using the inverse transform:

$$u(t,x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\mathrm{i}\alpha x - \alpha^2 t/4\pi}\int_{-\infty}^{\infty} f(y)\, e^{-\mathrm{i}\alpha y}\,\mathrm{d}y\,\mathrm{d}\alpha.$$

But

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\mathrm{i}\alpha z - \alpha^2 t/4\pi}\,\mathrm{d}\alpha = t^{-1/2}e^{-\pi z^2/t} =: K(t,z).$$

Thus $u(t,x)$ is given by the convolution

$$u(t,x) = \int_{-\infty}^{\infty} K(t, x-y)\, f(y)\,\mathrm{d}y. \tag{2.3.15b}$$

We see that $K(t,x)$ is itself a solution to the heat equation, with initial condition $f(x) = \delta(x)$, the Dirac delta. Physically, $K$ corresponds to an infinitely hot spot placed at position $x = 0$ at time $t = 0$, on an otherwise uniform, infinitely long rod. This fundamental solution $K(t,x)$ is called the *heat kernel* or *propagator* for $\mathbb{R}$.

What has this to do with the theta function? Consider the specialisation $\theta_3(\mathrm{i}t, x)$, where $t, x \in \mathbb{R}$, $t > 0$. Note that

$$\frac{\partial}{\partial t}\theta_3(\mathrm{i}t,x) = \frac{1}{4\pi}\frac{\partial^2}{\partial x^2}\theta(\mathrm{i}t,x),$$

so $\theta_3$ is a solution to the heat equation. Also, in the $t \to 0$ limit, $\theta_3(0,x)$ becomes the distribution $\sum_{n=-\infty}^{\infty}\delta(x-n)$ (this is proved by evaluating $\lim_{t\to 0}\int_0^1 \theta_3(\mathrm{i}t,x)\, f(x)\,\mathrm{d}x$, but is merely the statement that $\sum_n e^{2\pi\mathrm{i}mx} = \sum_m \delta(x-m)$). Thus $\theta_3$ plays the same role on the circle $\mathbb{R}/\mathbb{Z}$ that $K(t,x)$ played on the line $\mathbb{R}$: $\theta_3$ is the heat kernel for the circle. But we can obtain this kernel in another way, by averaging the heat kernel $K(t,x)$ for $\mathbb{R}$:

$$\sum_{n=-\infty}^{\infty} t^{-1/2}e^{-\pi(x-n)^2/t} = t^{-1/2}e^{-\pi x^2/t}\theta_3\left(\frac{\mathrm{i}}{t}, \frac{x}{\mathrm{i}t}\right).$$

Equating this to $\theta_3(\mathrm{i}t,x)$ recovers (2.3.15b).

As with Poisson summation, the notion of heat kernel can be generalised considerably. For example, let $M$ be a compact $n$-dimensional Riemannian manifold and let $\Delta$ be the Laplacian. In local coordinates,

$$\Delta(x) = -\sum_{i,j=1}^{n} g^{ij}(x)\,\frac{\partial^2}{\partial x^i\,\partial x^j},$$

where $g^{ij}(x)$ is the metric. The heat equation on $M$ is

$$\frac{\partial}{\partial t}u(t,x) = -\Delta u(t,x), \quad x \in M,\ t > 0,$$

with initial condition $f(x) = \lim_{t\to 0}u(t,x)$. This can be solved formally by the expression $u(t,x) = e^{-t\Delta}f(x)$. In fact $e^{-t\Delta}$ makes sense as an operator on $L^2(M)$, for any $t \in \mathbb{C}$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

with $\mathrm{Re}(t) > 0$. By the heat kernel $K(t, x, y)$ for $M$ we mean as before the solution to the heat equation with initial condition $\delta(x, y)$, or equivalently $K(t, x, y)$ generates the solution $(e^{-t\Delta} f)(x) = \int_M K(t, x, y) f(y)\, dy$ to the heat equation with arbitrary initial condition $f$. The heat kernel always exists and is unique, and is analytic for $t > 0$. In fact the heat kernel can be expressed as

$$K(t, x, y) = \sum_n e^{-\lambda_n t} \phi_n(x) \overline{\phi_n(y)},$$

where $\lambda_n \geq 0$ are the (discrete) eigenvalues of the Laplacian $\Delta$ with (orthonormal) eigenfunctions $\phi_n \in C^\infty(M) \subset L^2(M)$. Incidentally, $K$ is the kernel of the operator $e^{-t\Delta}$ in the sense of the Schwartz kernel theorem. For $t$ small,

$$K(t, x, y) = (4\pi t)^{-n/2} e^{-d(x,y)^2/4t} \sum_{i=0}^{\infty} t^i f_i(x, y)$$

where $d(x, y)$ is the distance between $x, y \in M$, and $f_i$ are certain functions. In the language of quantum field theory, the heat kernel $K(t, x, y)$ equals $\langle x|e^{-t\Delta}|y\rangle$. The heat kernel stores geometric information on $M$, and interpolates between the identity operator of $L^2(M)$ at $t = 0$ and the projection onto the kernel of $\Delta$ as $t \to \infty$.

For example, for $M = \mathbb{R}^n$ the heat kernel is $K(t, x, y) = (4\pi t)^{-n/2} \exp[-|x - y|^2/4t]$, so for any $n$-dimensional lattice $L \subset \mathbb{R}^n$ the heat kernel of the $n$-torus $\mathbb{R}^n/L$ is

$$(4\pi t)^{-n/2} \sum_{v \in L} \exp[-|x - y - v|^2/4t].$$

But it also equals (normalising the arguments appropriately) $\frac{1}{\sqrt{|L|}} \Theta_{L^*}$, and so we recover the modularity of (2.3.7).

The natural generalisation of the $M = \mathbb{R}^n$ calculation is performed by [**231**]. In particular, let $G$ be a connected, noncompact reductive Lie group, let $K$ be a maximal compact subgroup, and let $\Gamma$ be a discrete subgroup of $G$ such that the quotient $\Gamma\backslash G$ is compact. Then two expressions for the heat kernel, and its trace, on the space $\Gamma\backslash G/K$ are obtained. In the special case of $G = \mathbb{R}^n$ and $\Gamma$ being a lattice, the trace formula reduces to the usual formula expressing $\Theta_L(-1/\tau)$. The naturality of this construction $\Gamma\backslash G/K$ will be clear after reading Section 2.4.1. Moreover, [**181**] proves the Macdonald identities using the heat equation on compact Lie groups.

Further generalisations are possible (see e.g. [**52**]). For example, degree 1 and 0 terms can be added to the Laplacian $\Delta$, and we can consider more generally differential operators on sections of line bundles over $M$, rather than on $M$. Heat kernel techniques can be used to prove various formulations of the Atiyah–Singer Index Theorem, and equivariant analogues of the theory yield the Atiyah–Bott fixed-point theorem. The strategy typically followed by these applications is to consider the integral $I(t) = \int_M K(t, f(y), x)\, dy$ for some map $f : M \to N$, where $K$ is the heat kernel on $N$. The $t \to 0$ limit collapses the integral to an integral or sum over $f^{-1}(x)$. But a global expression for $I(t)$ can often be found, for example using representation theory or geometry; taking its $t \to 0$ limit

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

yields an identity between the local integral $\int_{f^{-1}(x)}$ and some global data of $M$ and $N$. See, for example, [**389**]. Question 2.1.7 is essentially an example of this strategy – what we call $K(g, h)$ there is the heat kernel at $t = 0$ of the finite group $G$.

Some of the many applications and occurrences of the heat kernel are collected in [**320**]. But can the heat kernel be directly relevant to Moonshine? This seems very possible. After all, the Atiyah–Bott fixed-point theorem yields an elegant proof of the Weyl character formula for compact Lie groups. In the conformal field theories associated with Lie groups (namely, the Wess–Zumino–Witten models), the heat kernel is used to explicitly construct the flat Knizhnik–Zamolodchikov connection on spaces of chiral blocks [**288**] (more on this starting in Section 3.2.4). This is significant because, according to conformal field theory, it is the monodromy of the Knizhnik–Zamolodchikov equation that is responsible (in genus 1) for the modularity of the affine algebra characters.

To this author's knowledge, heat kernel methods have never been used directly in the context of Monstrous Moonshine, but surely they can be used to prove at minimum the modularity of the McKay–Thompson series, and to help us understand a little better the geometry of Monstrous Moonshine. It seems possible that equivariant heat kernel methods could provide a geometric umbrella under which herd the more interesting examples of Moonshine.

### *2.3.5 Siegel forms*

Vaughn Jones considered how one von Neumann algebra can be embedded in another (e.g. itself), and the result – subfactor theory – is profoundly interesting. This success suggests the following analogue of Galois theory:

**The Jones Programme** *Study the ways in which one infinite beast can be embedded in another.*

Let's probe this thought with the simplest infinite beast this author can think of: lattices (Section 1.2.1). Let $L \subset \mathbb{R}^n$, $L' \subset \mathbb{R}^{n'}$ be lattices of dimension $n$ and $n'$, respectively. Fix bases $\{x^{(1)}, \ldots, x^{(n)}\}$, $\{y^{(1)}, \ldots, y^{(n')}\}$ and construct the $n \times n$ matrix $M$, whose columns are the $x^{(i)}$. An embedding of $L'$ into $L$ is a linear map $\varphi : L' \to L$ that preserves all inner-products. It is determined by the values $\varphi(y^{(j)}) = \sum_i \varphi_{ji} x^{(i)}$. The coefficients $\varphi_{ji}$ all lie in $\mathbb{Z}$ and form an $n' \times n$ matrix $(\varphi)$. Now, $\varphi$ preserves all inner-products, iff $\varphi(y^{(i)}) \cdot \varphi(y^{(j)}) = y^{(i)} \cdot y^{(j)}$ $\forall i, j$, iff

$$(\varphi)\, M^t M\, (\varphi)^t = M^t M. \tag{2.3.16}$$

Let $N(L', L)$ be the number of these embeddings, i.e. the number of $n' \times n$ $\mathbb{Z}$-matrices $(\varphi)$ satisfying (2.3.16). This number will be 0 unless $n' \le n$.

For example, $N(\mathbb{Z}, L)$ equals the number of unit vectors in $L$. Thus, if $L$ is integral, the generating function $\sum_{k=0}^{\infty} N(\sqrt{k}\mathbb{Z}, L)\, x^k$ is the theta function $\Theta_L(\tau)$, for $x = e^{\pi i \tau}$. We might hope that the numbers $N(L', L)$ are coefficients of some other modular-like function.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Construct a multi-variable generating function as follows. Fix an $n$-dimensional integral lattice $L$. Let $x_{ij}$, $1 \leq i, j \leq n$, be variables. Consider

$$\mathrm{Th}_L(x_{ij}) := \sum_{n'=0}^{n} \sum_{[L']} \sum_{\mathbb{Z}\{\beta_1,\ldots,\beta_n\}=L'} \frac{N(L',L)}{\mathrm{Aut}(L')} \prod_{1\leq i,j\leq n} x_{ij}^{\beta_i\cdot\beta_j}. \tag{2.3.17a}$$

The sum over $[L']$ is of all isomorphism classes of $n'$-dimensional even lattices. For each of these classes, fix a representative $L' \subset \mathbb{R}^n$. The $\{\beta_i\}$ run over all possible ordered $n$-tuples of lattice vectors that span $L'$. There is an equivalent but cleaner way to write (2.3.17a). Let $\mathcal{A}_n$ be the set of all $n \times n$ positive semidefinite matrices $A$ with integer entries and even integers down the diagonal. These are precisely the matrices $A_{ij} = \beta_i \cdot \beta_j$. Then

$$\mathrm{Th}_L(x_{ij}) = \sum_{A'\in\mathcal{A}_n} N(L',L) \prod_{1\leq i,j\leq n} x_{ij}^{A'_{ij}}, \tag{2.3.17b}$$

where $L'$ is any lattice realising the matrix $A'$ of inner-products.

In any case, this generating function $\mathrm{Th}_L$, after making the change-of-variables $x_{ij} = e^{\pi \mathrm{i} T_{ij}}$, is a *Siegel modular form*! We return to it shortly.

Let's try to find a version of modular forms where $\mathbb{H}$ is replaced by a higher-dimensional space. Start with $\Theta_L(\tau, z)$ in equation (2.3.7), but reinterpret this as a function of the complex matrix $T := \tau A$, with entries $A_{ij} = b^{(i)} \cdot b^{(j)}$ for a basis $b^{(i)}$ of the lattice $L$. We thus get

$$\Theta(T, z) := \sum_{n\in\mathbb{Z}^n} \exp[\pi \mathrm{i} n \cdot T n + 2\pi \mathrm{i} n \cdot z]. \tag{2.3.18}$$

How far can we extend the domain $T$? We may as well restrict to symmetric matrices $T$. For which symmetric matrices $T$ does (2.3.18) converge to a holomorphic function? We know from (2.3.7) that it does whenever $T = xA + \mathrm{i}A$ for any positive-definite matrix $A$ and real number $x$, but there is no need to restrict to such $T$. Indeed, it is straightforward to obtain that (2.3.18) converges to a holomorphic function for any $z \in \mathbb{C}^n$ and any $T$ in the Siegel upper half-space $\mathbb{H}_n$ defined in Section 2.1.4.

Of course, (2.3.18) is quasi-periodic in the $z$ variable:

$$\Theta(T, z + m) = \Theta(T, z), \qquad \forall m \in \mathbb{Z}^n \tag{2.3.19a}$$

$$\Theta(T, z + Tm) = \exp[-\pi \mathrm{i} m \cdot Tm - 2\pi \mathrm{i} m \cdot z]\, \Theta(T, z), \qquad \forall m \in \mathbb{Z}^n. \tag{2.3.19b}$$

The Siegel theta function $\Theta(T, z)$ is an easy generalisation of the Jacobi theta function (2.3.7). What makes it so remarkable is its symmetries as a function of $T$:

$$\begin{aligned}\Theta((AT + B)(CT + D)^{-1}, (CT + D)^{t-1} z)&\\ = \xi_\gamma \det(CT + D)^{\frac{1}{2}} \exp[\pi \mathrm{i} z \cdot (CT + D)^{-1} z)]\, \Theta(T, z)&\end{aligned} \tag{2.3.20}$$

for all $\gamma = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathrm{Sp}_{2n}(\mathbb{Z})$ for which all diagonal entries of $A^t C$ and $B^t D$ are even. Call this subgroup $\Gamma_\theta^n$, in analogy with (2.2.5). The numbers $\xi_\gamma \in \mathbb{C}$ are certain eighth roots of unity.

We defined $\mathrm{Sp}_{2n}(\mathbb{Z})$ in Section 2.1.4. The modularity of $\Theta(T, z)$ is proved much the way that modularity of $\theta_3$ was proved. The analogue of (2.2.1a) is

$$\mathrm{Sp}_{2n}(\mathbb{Z}) = \left\langle \begin{pmatrix} I & A \\ 0 & I \end{pmatrix}, \begin{pmatrix} B & 0 \\ 0 & B^{t-1} \end{pmatrix}, \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} \mid \forall A \in M_{n\times n}(\mathbb{Z}), \right.$$
$$\left. A = A^t, \forall B \in \mathrm{GL}_n(\mathbb{Z}) \right\rangle. \tag{2.3.21}$$

If we insist the matrices $A$ in (2.3.21) have even diagonals, then we generate $\Gamma_\theta^n$. Verifying invariance of $\Theta(T, z)$ under $\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}$ and $\begin{pmatrix} B & 0 \\ 0 & B^{t-1} \end{pmatrix}$ is routine; use Poisson summation for $\begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$. The argument is given in detail in chapter 2.5 of [**439**].

Section 2.1.4 relates $\mathbb{H}_n$ to Riemann surfaces of genus $n$. As we recall, the possible period matrices $\Omega$ of a given surface form an $\mathrm{Sp}_{2n}(\mathbb{Z})$-orbit in $\mathbb{H}_n$. The Jacobian of the surface is $\mathbb{C}^n/(\mathbb{Z}^n + \Omega\mathbb{Z}^n)$. Quasi-periodicity (2.3.19) embeds these Jacobians into projective space. Most points in $\mathbb{H}_n$ (at least for $n > 2$) aren't period matrices of surfaces, and as we recall the moduli space $\mathfrak{M}_{n,0}$ can be identified with $\mathfrak{C}_n/\mathrm{Sp}_{2n}(\mathbb{Z})$ for some subset $\mathfrak{C}_n$ in $\mathbb{H}_n$.

We should thus regard $\Theta(T, z)$, $\mathrm{Sp}_{2n}(\mathbb{Z})$ and $\mathbb{H}_n$ as the genus $n$ versions of $\theta_3$, $\mathrm{SL}_2(\mathbb{Z}) \cong \mathrm{Sp}_2(\mathbb{Z})$ and $\mathbb{H}$, where $\tau$ becomes an $n \times n$ matrix. The hyperbolic geometry of $\mathbb{H}$ becomes symplectic geometry on $\mathbb{H}_n$ (see e.g. section 4 of [**395**]). As mentioned in footnote 1 of this chapter, the future will find Moonshine expanding into higher genus. The calculations will be far more complicated, and this is presumably the reason for the delay. One of the only explicit works in this direction is [**533**], which looks at the lattice ↔ theta function example of Figure 0.1 (or equivalently the bosonic string compactified on a torus) at genus 2. As expected, Siegel modular forms play a dominant role. See also [**9**] for some calculations with multi-loop heterotic strings, which heavily involve Siegel theta functions.

**Definition 2.3.3** *Let $\Gamma \subset \mathrm{Sp}_{2n}(\mathbb{Z})$ $(n > 1)$ have finite index. Then a* Siegel modular form of weight $k$ and level $\Gamma$ *is a holomorphic function $f$ on $\mathbb{H}_n$ such that*

$$f((AT + B)(CT + D)^{-1}) = \det(CT + D)^k\, f(T), \qquad \forall \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \Gamma.$$

A growth condition at the cusps (requiring holomorphicity) is automatically satisfied when $n > 1$. Another simplification of higher genus is that any subgroup $\Gamma \subset \mathrm{Sp}_{2n}(\mathbb{Z})$ of finite index includes some congruence group $\Gamma^n(N) := \{A \in \mathrm{Sp}_{2n}(\mathbb{Z}) \mid A \equiv I \pmod{N}\}$ with finite index.

For example, $\Theta(T, z)^2$ is a modular form of weight 1 and level $\Gamma^n(4)$. Eisenstein series for $\mathrm{Sp}_{2n}(\mathbb{Z})$ can be defined in the obvious way, as a sum of $\det(CT + D)^{-2k}$ over appropriately defined pairs $\{C, D\}$ of matrices (see e.g. section 14 of [**395**] for details). A final example plays the same role for $\Theta(T, z)$ that $\Theta_L(\tau)$ played for $\theta_3(\tau)$: let $L$ be any $m$-dimensional rational lattice and let $A$ be its Gram matrix, then

$$\Theta_L(T, Z) := \sum_N \exp[\pi \mathrm{i}\, \mathrm{tr}(N^t T N A) + 2\pi \mathrm{i}\, \mathrm{tr}(N^t Z)],$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

where $T \in \mathbb{H}_n$, $Z$ is an $n \times m$ complex matrix, and the sum is over all $n \times m$ $\mathbb{Z}$-matrices. This is a specialisation of $\Theta$ for $\mathrm{Sp}_{2nm}(\mathbb{Z})$, and is a Siegel modular form of weight $m/2$ for some $\Gamma^n(M)$ (see e.g. chapter 2.6 of [**439**]). We met $\Theta_L$ in (2.3.17).

Finally, let us describe the analogue of Fourier expansion here. For convenience take $\Gamma$ to be $\mathrm{Sp}_{2n}(\mathbb{Z})$. Then a modular form $f$ for $\Gamma$ obeys the periodicity $f(T+B) = f(T)$ for all $n \times n$ $\mathbb{Z}$-matrices $B$. Together with holomorphicity, this means $f$ has an expansion

$$f(T) = \sum_{M \geq 0} a(M) \exp[2\pi \mathrm{i}\, \mathrm{tr}(TM)], \tag{2.3.22}$$

where the sum is over all positive-semidefinite symmetric $n \times n$ matrices $M$ with entries $M_{ii} \in \mathbb{Z}$ and $M_{ij} \in \frac{1}{2}\mathbb{Z}$. These numbers $a(M)$ play the role of Fourier coefficients here. For example, (2.3.17b) gives the Fourier expansion of $\Theta_L(T)$.

Question 2.3.1. Prove that the Weierstrass $\wp$ function (2.1.6a) is a Jacobi form for $\mathrm{SL}_2(\mathbb{Z})$ with weight $k = 2$ and index $m = 1$.

Question 2.3.2. Let $L, L'$ be two $n$-dimensional rational lattices in $\mathbb{R}^n$, and let $u, u' \in \mathbb{R}^n$ be vectors of finite order for $L$ and $L'$, respectively.
(a) *Prove:* If $\Theta_{L+u}(\tau, z) = \Theta_{L'+u'}(\tau, z)$ for all $\tau \in \mathbb{H}$, $z \in \mathbb{C}^n$, then $L + u = L' + u'$ *as sets*.
(b) Prove that $L$ and $L'$ are isomorphic (Section 1.2.1) iff there exists an orthogonal map $T \in \mathrm{O}_n(\mathbb{R})$ such that $\Theta_L(\tau, z) = \Theta_{L'}(\tau, Tz)$ for all $\tau \in \mathbb{H}$, $z \in \mathbb{C}^n$.

Question 2.3.3. Let $L$ be any integral lattice of dimension $n$. For each $m = 0, 1, 2, \ldots$, let $L_{(m)}$ denote all the vectors $u \in L$ with norm-squared $u \cdot u = m$. Each automorphism $\omega$ of $L$ permutes the vectors in $L_{(m)}$, so for each $m$ we get a $\|L_{(m)}\|$-dimensional representation $\alpha_{(m)}$ of $\mathrm{Aut}(L)$ by permutation matrices. Thus, for each $\omega \in \mathrm{Aut}(L)$, we can *twist* $\Theta_L$ as follows: define

$$\Theta_L^{(\omega)}(\tau) := \sum_{m=0}^{\infty} \chi_{(m)}(\omega) \exp[\pi \mathrm{i} \tau m],$$

where $\chi_{(m)}$ is the character of the representation $\alpha_{(m)}$. For example, $\Theta_L^{(id)} = \Theta_L$ and $\Theta_L^{(-id)}(\tau) = 1$. Prove that, for each $\omega \in \mathrm{Aut}(L)$, $\Theta_L^{(\omega)}$ will be a modular form for some $\Gamma(N)$ and some weight $0 \leq k \leq n/2$, and that $k = n/2$ iff $\omega = id$.

Question 2.3.4. Let $f$ be a modular form of weight $k$, for some $\Gamma(N)$.
(a) Prove that, for each choice of $r \in \mathbb{Q}$, the function $g(\tau) := f(\tau + r)$ is a modular form of level $k$, for some $\Gamma(M)$ ($M$ depending on $r$).
(b) For any field $\mathbb{F}$, prove that $\mathrm{SL}_2(\mathbb{F})$ is generated by the matrices $\begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix}$, for $r \in \mathbb{F}$, together with $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. From this, prove that if $f$ is a modular form for $\Gamma(N)$ of weight $k$, then for any $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Q})$, the function $h(\tau) := f(\frac{a\tau+b}{c\tau+d})$ will be a modular form of weight $k$ for some $\Gamma(M)$ ($M$ depending on $a, b, c, d$).

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

## 2.4 Representations and modular forms

*According to I. M. Gel'fand, mathematics of any kind is representation theory.*[8] *This section applies this beautiful strategy to modular forms.*

There are at least formal similarities between quantum theory and modular forms. Wigner taught that a particle should be identified with a unitary representation of $\mathrm{SL}_2(\mathbb{C})$ or $\mathrm{SL}_2(\mathbb{R})$, in $(3+1)$- or $(2+1)$-dimensional space-time, respectively. In this section we associate modular forms to unitary representations of $\mathrm{SL}_2(\mathbb{R})$, and the picture generalises naturally to, for example, $\mathrm{SL}_2(\mathbb{C})$. Could there be some cross-fertilisation between the methods and ideas of quantum field theory and modular forms?

In the 1962 International Congress of Mathematicians, I. M. Gel'fand remarked somewhat cryptically that there is an intriguing analogy between the scattering matrix of quantum mechanics and zeta functions. Ten years later the idea was exploited and clarified by Faddeev and Pavlov, who applied the Lax–Phillips scattering theory to the theory of automorphic forms. For example, poles of the scattering matrix (which in quantum field theory would correspond to particles) correspond to zeros of the Riemann zeta function. Their work is generalised in [**371**], where we find for instance a new proof of the Selberg Trace Formula for $\mathrm{SL}_2$. These applications are significant, and hopefully a small hint of things to come. See also [**562**].

### *2.4.1 Automorphic forms*

Definitions 0.1 and 2.2.1 of modular functions and forms for $\mathrm{SL}_2(\mathbb{Z})$ should seem very arbitrary. In mathematics we attack arbitrariness through generalisation. A good generalisation helps us to see the meaning of each feature, and puts the whole theory into a broader perspective. Of course we can generalise these definitions by replacing $\mathrm{SL}_2(\mathbb{Z})$ with other Fuchsian groups $\Gamma < \mathrm{SL}_2(\mathbb{R})$, but this is too obvious to be helpful.

Much more valuable is to understand the relation between $\mathbb{H}$ and $G = \mathrm{SL}_2(\mathbb{R})$. In particular, an easy calculation shows that our action of $G$ on $\mathbb{H}$ is *transitive*. That is, any point in $\mathbb{H}$ can get mapped to any other point in $\mathbb{H}$ by a matrix in $G$. In particular, $\gamma_{x+\mathrm{i}y} = \begin{pmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{pmatrix} \in G$ sends i to $x + \mathrm{i}y$. We call $\mathbb{H}$ a *homogeneous space* for $G$. Moreover, the subgroup of $G$ fixing $\mathrm{i} \in \mathbb{H}$, say, is $K = \mathrm{SO}_2(\mathbb{R})$. Thus

$$\mathbb{H} \cong \mathrm{SL}_2(\mathbb{R})/\mathrm{SO}_2(\mathbb{R}) = G/K. \tag{2.4.1a}$$

More precisely, we have the Iwasawa decomposition

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = y^{-1/2} \begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{2.4.1b}$$

$$x + \mathrm{i}y = \frac{a\mathrm{i} + b}{c\mathrm{i} + d}, \quad e^{\mathrm{i}\theta} = \frac{d - \mathrm{i}c}{|d - \mathrm{i}c|}. \tag{2.4.1c}$$

In fact $\mathrm{SO}_2(\mathbb{R})$ is the unique (up to conjugation) maximal compact subgroup of $G$.

[8] See the quotation on page 840 of *Proc. ICM* (American Mathematical Society, Providence 1987), edited by A. M. Gleason.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

In mathematics we try to find hidden structure, and that is the spirit in which (2.4.1a) should be read. The key here was the transitive action: an expression like (2.4.1a) arises whenever one has a homogeneous space. Note that the action $\gamma.\tau$ of $G$ on $\mathbb{H}$ now reduces to matrix multiplication: $\gamma\gamma_\tau K$.

Do modular forms respect (2.4.1a)? Can we lift modular forms $f : \mathbb{H} \to \mathbb{C}$ into functions $\phi_f : G \to \mathbb{C}$? Yes, and in fact we gain something in the process. Use (2.4.1b):

$$\phi_f \begin{pmatrix} a & b \\ c & d \end{pmatrix} = f\left(\frac{a\mathrm{i}+b}{c\mathrm{i}+d}\right)(c\mathrm{i}+d)^{-k} = f(x+\mathrm{i}y)\, y^{k/2}\, e^{\mathrm{i}\theta k}, \tag{2.4.2a}$$

where $k$ is the weight of $f$. Then for any $A \in \mathrm{SL}_2(\mathbb{Z})$ and $\alpha \in \mathbb{R}$, we get

$$\phi_f\left(A\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}\right) = \phi_f\begin{pmatrix} a & b \\ c & d \end{pmatrix} e^{-\mathrm{i}k\alpha}. \tag{2.4.2b}$$

The point of multiplication by $(c\mathrm{i}+d)^{-k}$ is now clear: it makes $\phi_f$ left-invariant with respect to $\mathrm{SL}_2(\mathbb{Z}) = \Gamma$. Thus we've sacrificed $K$-invariance and $\Gamma$-covariance, for $K$-covariance and $\Gamma$-invariance. This is significant, because compact Lie groups like $K$ are much easier to handle than infinite discrete groups like $\mathrm{SL}_2(\mathbb{Z})$.

In particular, we find that the right multiplication in (2.4.2b) defines a one-dimensional representation of $K$ on $\mathbb{C}\phi_f$. We know that the finite-dimensional irreducible $K$-representations are parametrised by a nonnegative integer, and all are one-dimensional. Thus we get an algebraic interpretation for the parameter $k$ in Definition 2.2.1: it is the highest weight of a representation of the maximal compact subgroup $\mathrm{SO}_2(\mathbb{R})$ of $\mathrm{SL}_2(\mathbb{R})$.

We also get a representation of $\mathrm{SL}_2(\mathbb{R})$ on the left side, given by $\phi_f \mapsto \phi_f \circ \gamma^{-1}$. The vector space here is the infinite-dimensional function space given by the $\mathbb{C}$-span of the $\mathrm{SL}_2(\mathbb{R})$-orbit of $\phi_f$. The result is an irreducible representation of $\mathrm{SL}_2(\mathbb{R})$, which is constant on $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. This representation is unitary – in fact it is a subrepresentation of the regular representation of $G$ on the Hilbert space $L^2(\Gamma\backslash G)$.

As an aside, note that everything generalises very naturally to Siegel modular forms. There, $G$ is $\mathrm{Sp}_{2n}(\mathbb{R})$, $\Gamma$ is $\mathrm{Sp}_{2n}(\mathbb{Z})$ or a similar discrete group like $\Gamma_\theta^n$, and $K = \mathrm{SO}_{2n}(\mathbb{R}) \cap \mathrm{Sp}_{2n}(\mathbb{R}) \cong U_n(\mathbb{C})$. Once again, $\mathbb{H}_n \cong G/K$. For Jacobi forms, $G$ is a semi-direct product of $\mathrm{SL}_2(\mathbb{R})$ with the Heisenberg group (it is constructed next subsection), and $K$ is $\mathrm{SO}_2(\mathbb{R}) \times S^1$: once again $G/K \cong \mathbb{H} \times \mathbb{C}$, as it should. The weight $k$ and index $m$ in Definition 2.3.2 parametrise the irreducible one-dimensional representations of $\mathrm{SO}_2(\mathbb{R})$ and $S^1$, that is to say $K$. Thus the index of a Jacobi form has a natural algebraic interpretation, as it should.

So the generalisation of modular forms and functions is starting to be clearer. We are looking for functions on the space $\Gamma\backslash G$, for discrete subgroups $\Gamma$ of real Lie groups $G$, and we should study them via the representation of $G$ they generate. The relation between modular forms and representation theory was accomplished in the 1950s by Gel'fand and Fomin. Let's make it more precise.

The unitary irreducible representations of $G = \mathrm{SL}_2(\mathbb{R})$ were classified by Bargmann [**44**]. His motivation was physics (the Lorentz group). Of course there is the one-dimensional identity representation. The remaining irreducible unitary representations are all infinite-dimensional, and fall into three series: the principal series $\mathcal{P}_s^{\pm}$ for $s \in \mathbb{R}$,

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

the complementary series $\mathcal{C}_s$ for $0 < s < 1$, and the discrete series $\mathcal{D}_n^{\pm}$ for $n = 2, 3, \ldots$ In addition, $G$ has many irreducible non-unitary representations. See, for example, chapter 1.3 of [**243**] for explicit realisations of all the unitary representations. For example, the discrete series $\mathcal{D}_n^{+}$ consists of holomorphic functions $f$ on $\mathbb{H}$, with Peterssen Hermitian form $\langle f, g\rangle = \int_{\mathbb{H}} f(\tau)\overline{g(\tau)}y^{n-1}\mathrm{d}x\,\mathrm{d}y$, and action $f \mapsto (-c\tau + a)^{-n} f\left(\frac{d\tau - b}{-c\tau + a}\right)$. Obviously our $G$-representation associated with $\Phi_f$ is isomorphic to $\mathcal{D}_k^{+}$. What $f$'s come from the other $G$-representations?

Associated with the principal series are functions such as this analogue of the Eisenstein series, called a *Maass form*:

$$E(\tau, s) = \sum_{m,n\in\mathbb{Z}}{}' \frac{y^s}{|m\tau + n|^{2s}}, \quad s \in \mathbb{C}.$$

This may look less strange when one considers the formula $\mathrm{Im}(\gamma.\tau) = y/|c\tau + d|^2$. For fixed $\tau \in \mathbb{H}$, the Maass form is absolutely convergent for $\mathrm{Re}(s) > 1$ and has a meromorphic extension to all $s \in \mathbb{C}$. For fixed $s \in \mathbb{C}$, it is invariant under $\mathrm{SL}_2(\mathbb{Z})$. It is not a holomorphic function of $\tau$, and so cannot be a modular form in the usual sense, but holomorphicity in Definitions 0.1 and 2.2.1 is a feature we must be prepared to lose, since most real Lie groups $G$ aren't complex manifolds. In fact we lost the holomorphicity of $f$ when we wrote (2.4.2a). What takes its place?

What is holomorphicity, other than the solution to differential equations (the Cauchy–Riemann equations, or the Laplacian $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ on $\mathbb{R}^2$)? The Maass forms aren't holomorphic, but they are eigenfunctions of the Laplacian on $\mathbb{H}$, namely $-y^2\,(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})$. By the Laplacian on $\mathbb{H}$ we mean a second-order differential operator that is invariant under all isometries $\mathrm{SL}_2(\mathbb{R})$.

We are thus led to the role of differential operators. These can be understood as follows. Whenever we have a Lie group representation, we also get an associated action of the Lie algebra (the derived module of Section 1.5.5). The Lie algebra will typically act as first-order differential operators; on $L^2(G)$ it acts by Lie derivatives. More precisely, to $X \in \mathfrak{sl}_2(\mathbb{R})$ we get the action $f(g) \mapsto \frac{\mathrm{d}}{\mathrm{d}t} f(ge^{tX})|_{t=0}$. For example, $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathfrak{sl}_2(\mathbb{R})$ corresponds to $\frac{\partial}{\partial\theta}$, using the parametrisation of (2.4.1a). An action of $\mathfrak{sl}_2(\mathbb{R})$ implies an action of the universal enveloping algebra $U(\mathfrak{sl}_2(\mathbb{R}))$, in our case simply by composing differential operators to get ones of higher order. As always, the centre $Z(U(\mathfrak{sl}_2(\mathbb{R})))$ naturally plays a fundamental role. Here, it is generated by the second-order operator

$$y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) - y\frac{\partial^2}{\partial x\,\partial\theta}.$$

This is how the Laplacian arises, algebraically. By definition, it commutes with all operators, so studying its eigenspaces helps decompose $L^2(\Gamma\backslash G)$ – we used a similar idea in decomposing Lie algebra modules into weight-spaces. Understanding that decomposition is essentially equivalent to understanding the space of modular forms for $\Gamma$, and can be called the harmonic analysis of automorphic forms.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

We have only scratched the surface, but this discussion and the following definition should give the reader a glimpse of the resulting theory.

**Definition 2.4.1** *Let $\Gamma$ be a discrete subgroup of a real semi-simple Lie group $G$, and let $K$ be a maximal compact subgroup of $G$. Let $\chi$ be a one-dimensional representation of $K$. We call a smooth function $f : G \to \mathbb{C}$ an* automorphic form *for $\Gamma$ if:*

(i) $f(\gamma g k) = \chi(k) f(g)$ *for all* $\gamma \in \Gamma, g \in G, k \in K$;
(ii) $f$ *is an eigenfunction of every operator in* $Z(U(\mathfrak{g}))$;
(iii) $f$ *obeys a certain growth condition.*

The term ‘automorphic form’ (going back to Klein in 1890) is much older than this definition. Here, $\mathfrak{g}$ is the Lie algebra of $G$ and $Z(U(\mathfrak{g}))$ is the centre of its universal enveloping algebra, which will be isomorphic to a polynomial algebra in $r$ variables, where $r$ is the rank of $\mathfrak{g}$. As mentioned above, the differential equations in (ii) take the place of holomorphicity. The growth condition is too technical to give here, but for $\mathrm{SL}_2(\mathbb{Z})$ it reduces to holomorphicity at the cusps. For more on the relation between automorphic forms and representations, see, for example, [**89**].

All modern material on automorphic functions uses the language of *adèles* and *idèles*,[9] which unify and simplify the theory (at the expense of making it more abstract). However, since they have no role in the remaining material of this book, we only sketch their motivation, and remain true here to the spirit of this not-completely-self-contained subsection.

*Projective* or *inverse limits* are the way algebra ‘integrates’ an infinite tower of structures into a single structure. A classic – and relevant – example is divisibility by powers of primes. We say that a given integer $n$ is divisible by $p^a$ if the canonical projection $\mathbb{Z} \to \mathbb{Z}_{p^a}$ (‘reduce mod $p^a$’) sends $n$ to 0. Now, the rings $\mathbb{Z}_{p^a}$ and $\mathbb{Z}_{p^b}$ are related by a homomorphism $\mathbb{Z}_{p^a} \to \mathbb{Z}_{p^b}$, provided $a \geq b$. So we get a tower

$$\cdots \to \mathbb{Z}/p^3\mathbb{Z} \to \mathbb{Z}/p^2\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z} \to 0.$$

The corresponding integrated structure is the *projective limit* $\lim_{\leftarrow}\mathbb{Z}_{p^a} =: \widehat{\mathbb{Z}}_p$, the $p$-adic integers, which can be realised as formal power series $\sum_{a=0}^{\infty} a_n p^n$, $a_i \in \mathbb{Z}/p\mathbb{Z}$. Doing arithmetic on them amounts to treating all $\mathbb{Z}/p^a\mathbb{Z}$ simultaneously – in this sense it is the integration of all $\mathbb{Z}_{p^a}$. For example,

$$\sqrt{2} = 3 + 1 \cdot 7 + 2 \cdot 7^2 + 6 \cdot 7^3 + \cdots$$

in $\widehat{\mathbb{Z}}_7$. The $p$-adic rationals $\widehat{\mathbb{Q}}_p$ are the field of fractions of $\widehat{\mathbb{Z}}_p$, or equivalently the formal Laurent series $\sum_{i=-N}^{\infty} a_i p^i$, $p_i \in \mathbb{Z}/p\mathbb{Z}$. They are to the ordinary rationals much as $\mathbb{R} =: \widehat{\mathbb{Q}}_\infty$ is: a completion, on which calculus can be defined. For a readable introduction to the $p$-adics, see [**257**]. Projective limits play a huge role in Section 6.3.3.

The more intuitive notion of limit, namely the *injective* or *direct limit*, arises when all arrows are reversed (i.e. when we have a sequence of embeddings rather than projections),

[9] Idèles were introduced by Chevalley in 1935 to remove some of the analysis being used with L-functions, etc. The word comes from ‘ideal’. Adèles were introduced in 1945 as an additive version of idèles.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

and is the algebraic analogue of taking derivatives. The prototypical example is the space of smooth functions $F_M(U)$ on an open patch of a manifold $M$: the direct limit $\lim_{\rightarrow} F_M(U)$, as $U \rightarrow \{p\}$, is isomorphic to the space of germs at $p$.

The modern theory of automorphic forms collects together the $\widehat{\mathbb{Q}}_p$ into the additive group of adèles $\mathbb{A}$ and multiplicative group of idèles $\mathbb{A}^\times$. The adèles are defined to be the group of all sequences $(x_\infty, x_2, x_3, x_5, \ldots, x_p, \ldots)$, where $x_\infty \in \mathbb{R}$, $x_p \in \widehat{\mathbb{Q}}_p$, and for all but finitely many $p$, $x_p \in \widehat{\mathbb{Z}}_p$. The idèles are defined similarly, and we obtain

$$\mathbb{A}^\times \cong \mathbb{Q}^\times \times \mathbb{R}_>^\times \times \prod \widehat{\mathbb{Z}}_p^\times,$$

where $\sum_{i=0}^{\infty} a_i p^i \in \mathbb{Z}_p^\times$ if $a_0 \neq 0$. The rationals $\mathbb{Q}$ embed in each $\widehat{\mathbb{Q}}_p$, and so embed diagonally in $\mathbb{A}$ ($r \mapsto \widehat{\mathbb{Z}}_p$ for any prime $p$ not dividing the denominator of $r$). There are many generalisations of $\mathbb{A}$ and $\mathbb{A}^\times$, for example we can replace $\mathbb{Q}$ by other number fields. But what good are they? What have they to do with modular forms?

There are many situations where the level of a modular form is variable. For example, any $A \in \mathrm{SL}_2(\mathbb{Q})$ takes a modular form for $\Gamma(N)$ to one for some other $\Gamma(N')$ (see Question 2.3.4). We have natural maps from the surface $\Gamma(n)\backslash\mathbb{H}$ to any $\Gamma(d)\backslash\mathbb{H}$, when $d$ divides $n$. Collecting together this tower of surfaces $\Gamma(n)\backslash\mathbb{H}$ into a single structure amounts to taking the limit space $\widehat{\mathbb{H}} := \lim_{\leftarrow} \Gamma(n)\backslash\mathbb{H}$. Functions on $\widehat{\mathbb{H}}$ include ratios $f/g$ of modular forms of the same weight but different levels. Much as

$$\lim_{\leftarrow} \mathbb{R}/n\mathbb{Z} \cong \mathbb{A}/\mathbb{Q}$$

as topological groups, we get

$$\widehat{\mathbb{H}} \cong \mathrm{SL}_2(\mathbb{Q})\backslash\mathrm{SL}_2(\mathbb{A})/K_\infty, \tag{2.4.3}$$

where $K_\infty$ consists of all sequences of matrices $(A, I_2, I_2, \ldots)$ where $A \in \mathrm{SO}_2(\mathbb{R}) \subset \mathrm{SL}_2(\mathbb{R})$ and the $I_2$'s are the identity matrices in each $\mathrm{SL}_2(\widehat{Q}_p)$. In Section 4.3.3 we discover $\widehat{\mathbb{H}}$ naturally in nonperturbative string theory.

Similarly, a Dirichlet character (see Section 2.3.3) can be thought of as a continuous one-dimensional representation on $\mathbb{Q}^\times\backslash\mathbb{A}^\times$, and the Galois group of a finite abelian extension of $\mathbb{Q}$ can be thought of as a subgroup of $\mathbb{Q}^\times\backslash\mathbb{A}^\times$.

The Langlands conjectures suggest that the $n$-dimensional representations of the absolute Galois group $\mathrm{Gal}(\overline{\mathbb{K}}/\mathbb{K})$ of a field $\mathbb{K}$ (such as $\mathbb{Q}$) correspond to 'automorphic representations' of $\mathrm{GL}_n(\mathbb{A})$, where $\mathbb{A}$ here is the group of adèles of $\mathbb{K}$. This correspondence can be seen through the corresponding L-functions. For $\mathrm{GL}_1$ and $\mathbb{K} = \mathbb{Q}$, this correspondence involves the Kronecker–Weber Theorem and Dirichlet characters. For $\mathrm{GL}_2$ this relates two-dimensional representations of Galois groups to modular forms. A recent accessible introduction to the Langlands Programme is [**90**]. Although there are hints of some sort of relation between the Langlands conjectures and Moonshine in its more general sense, these are still too speculative to go into here. However, Section 6.3.3 may whet one's appetite.

### 2.4.2 *Theta functions as matrix entries*

The relationship between representation theory and modular forms discussed last section is quite democratic in the sense that it exists at the level of the vector space of modular forms. Democracy is all well and good, but we are not equally interested in all modular forms – some have names!

The Jacobi theta function $\theta_3(\tau, z)$ is the unique quasi-periodic entire function, in the sense that any entire function $f : \mathbb{C} \to \mathbb{C}$ obeying $f(z+1) = f(z)$ and $f(z+\tau) = a\,e^{-2\pi \mathrm{i} z}$ for some constants $\tau \in \mathbb{H}$ and $a \in \mathbb{C}$ is a constant multiple of the function

$$f(z) = 1 + \sum_{n=-\infty}^{\infty} e^{\pi \mathrm{i}(n^2-n)\tau} a^{-n} e^{2\pi \mathrm{i} n z}.$$

For an elementary analytic proof see section 1.1 of [**439**]. From this uniqueness, all properties of $\theta_3$ can be quickly derived. In this section we sketch a striking algebraic version of this argument.

Starting in the 1960s, theta functions were interpreted as matrix entries in a representation of the Heisenberg group. The motivation was pure Moonshine:

> A force d'habitude, le fait que les séries thêta définissent des fonctions modulaires a presque cessé de nous étonner. Mais l'apparition du groupe symplectique comme un *deus ex machina* dans les célèbres travaux de Siegel sur les formes quadratiques n'a rien perdu encore de son caractère mystérieux. Le but de ce mémoire, et de ceux qui lui feront suite, n'est pas, bien entendu, d'élucider définitivement la question, mais de jeter un peu de lumière sur certains aspects de cette théorie qui étaient restés dans l'ombre jusqu'à présent. [**555a**][10]

The resulting explanation of the transformation $\theta_3(-1/\tau) = \sqrt{\frac{\tau}{\mathrm{i}}}\,\theta_3(\tau)$ can be extended to many other functions arising in Moonshine. First let us sketch the basic idea, before giving details and generalisations.

The starting point is the thought of realising special functions as matrix entries of Lie group representations. An elementary example of this involves the representation of $S^1 = U_1(\mathbb{R})$ as rotations in $\mathbb{R}^2$:

$$\theta \mapsto \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \tag{2.4.4}$$

The basic properties of $\sin(\theta)$ and $\cos(\theta)$ (e.g. angle-sum formulae, or even-oddness) can quickly be derived from this. We want to do something similar with $\theta_3$.

Begin by recalling the full variable dependence of $\theta_3(\tau, z, u)$, given in (2.3.4). For fixed $u$ we get a Jacobi form, and for fixed $\tau$ and $u$ we get an elliptic function for the

[10] 'By force of habit, the fact that theta series define modular forms has nearly ceased to amaze us. But the appearance of the symplectic group as a *deus ex machina* in the famous work of Siegel on quadratic forms has still lost none of its mysterious character. The goal of this paper, and of those which follow it, is not of course to clarify definitively the question, but rather to shed a little light on certain aspects of this theory which have remained in the dark up to now.'

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

torus $\mathbb{C}/(\mathbb{Z}+\mathbb{Z}\tau)$. This leads us to consider two translation operators on the space of (say) entire functions $f : \mathbb{C} \to \mathbb{C}$, as follows. Fix $\tau \in \mathbb{H}$ and define

$$(S_b f)(z) = f(z+b), \tag{2.4.5a}$$

$$(T_a f)(z) = \exp[\pi \mathrm{i} a^2 \tau + 2\pi \mathrm{i} a z]\, f(z + a\tau), \tag{2.4.5b}$$

for any $a, b \in \mathbb{R}$. In this way, for each fixed $\tau \in \mathbb{H}$, $\mathbb{R}^2$ acts on the space of entire functions – the role of $\tau$ being primarily to parametrise different isomorphisms between the additive groups $\mathbb{R}^2$ and $\mathbb{C}$. However, an easy calculation shows that $T_a$ and $S_b$ don't commute, rather $S_b \circ T_a = \exp[2\pi \mathrm{i}\, ab]\, T_a \circ S_b$. So the group $\langle T_a, S_b \rangle$ generated by all $T_a$'s and $S_b$'s is the semi-direct product of $S^1$ with $\mathbb{R}^2$, consisting of all pairs $[\lambda, x]$ for $\lambda \in \mathbb{C}$, $|\lambda| = 1$ and $x = (x_1, x_2) \in \mathbb{R}^2$, and operation

$$[\lambda, x] \cdot [\mu, y] = [\lambda\mu \exp[2\pi \mathrm{i} x_2 y_1], x + y].$$

This group is called the *Heisenberg group $H$*. Then (2.4.5) says that $\theta_3$ is a vector in a space carrying a representation of $H$. Now, it turns out that all irreducible representations $(\pi, \mathcal{H})$ of $H$ are essentially isomorphic. A more natural and useful way to see $\theta_3$ in any such representation $(\pi, \mathcal{H})$ is by defining a vector $f_\tau \in \mathcal{H}$ and distribution $\mu_\mathbb{Z}$ such that the Hermitian product

$$\langle \pi_{[1,x]} f_\tau, \mu_\mathbb{Z} \rangle = c\, e^{\pi \mathrm{i} x_1 (\tau x_1 + x_2)}\, \theta_3(\tau, x_1\tau + x_2) \tag{2.4.6}$$

for some nonzero constant $c$. The exponential factor on the right side of (2.4.6) simplifies the quasi-periodicity of the right side.

We will see that $\mathrm{SL}_2(\mathbb{R})$ acts as automorphisms on the Heisenberg group $H$. Hence for any $\gamma \in \mathrm{SL}_2(\mathbb{R})$, we get a new representation $\pi_\gamma$ of $H$ by $[\lambda, x] \mapsto \pi_{\gamma.[\lambda,x]}$. This representation must be isomorphic to $\pi$, so there is a (unitary) operator $R_\gamma$ on $\mathcal{H}$ such that $\pi_{\gamma.[\lambda,x]} = R_\gamma \circ \pi_{[\lambda,x]} \circ R_\gamma^{-1}$. The assignment $\gamma \mapsto R_\gamma$ defines a projective representation of $\mathrm{SL}_2(\mathbb{R})$ on $\mathcal{H}$. Modularity of $\theta_3$ now follows from the calculation

$$\langle \pi_{[1,x]} f_\tau, \mu_\mathbb{Z} \rangle = \langle R_\gamma \pi_{[1,x]} f_\tau, R_\gamma \mu_\mathbb{Z} \rangle = \langle \pi_{\gamma.[1,x]} R_\gamma f_\tau, R_\gamma \mu_\mathbb{Z} \rangle, \tag{2.4.7}$$

together with the computation of $R_\gamma f_\tau$ and $R_\gamma \mu_\mathbb{Z}$ for the $\gamma \in \Gamma_\theta < \mathrm{SL}_2(\mathbb{Z})$. Let us now fill in the details.

For reasons that will be clear shortly, it is preferable to work instead of $[\lambda, x]$ with the realisation of the group $H$ given by all pairs $(\lambda, x)$ with operation

$$(\lambda, x) \cdot (\mu, y) = (\lambda\mu \exp[\pi \mathrm{i}(x_1 y_2 - x_2 y_1)], x + y).$$

The isomorphism between these realisations of $H$ is given by the correspondence

$$(\lambda, x) \longleftrightarrow [\lambda^{-1} \exp[\pi \mathrm{i} x_1 x_2], x].$$

This group $H$ is a three-dimensional real Lie group corresponding to the Heisenberg Lie algebra $\mathfrak{Heis}$ defined in (1.4.3). It is a quotient by $\mathbb{Z} \cong \left\langle \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\rangle$ of the group $\widetilde{H}$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

of upper-triangular matrices

$$\begin{pmatrix} 1 & a & c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \in \mathrm{SL}_3(\mathbb{R}).$$

$\widetilde{H}$ is the (unique) simply-connected Lie group with Lie algebra $\mathfrak{Heis}$; it isn't important that we're focusing on $H$ rather than its universal cover $\widetilde{H}$. The group $H$ and its $(2n+1)$-dimensional versions (the obvious extension of $\mathbb{R}^{2n}$ by $S^1$) were studied originally in the context of quantum mechanics, hence their name.

The representation theory of these groups was established around 1930. Let $\pi$ be a unitary irreducible representation of $H$, in a Hilbert space $\mathcal{H}$. Recall from Section 1.5.5 that this means $\pi$ is a homomorphism from $H$ into the group of unitary operators of $\mathcal{H}$; moreover, for each $f \in \mathcal{H}$, the map from $H$ to $\mathcal{H}$ given by $(\lambda, x) \mapsto \pi_{(\lambda,x)} f$ is continuous. First note that by Schur's Lemma (the analogue here of Lemma 1.1.3), the central element $(\lambda, 0) \in H$ will act in $\mathcal{H}$ by a scalar multiple $\lambda^n$ for some $n \in \mathbb{Z}$.

**Theorem 2.4.2 (Stone–von Neumann)** *Let $\pi$ be a unitary irreducible representation of $H$, obeying $\pi_{(\lambda,0)}(f) = \lambda^n f$.*

(i) *If $n \neq 0$, then $\pi$ is infinite-dimensional and any other unitary irreducible representation $\pi'$ of $H$ obeying $\pi'_{(\lambda,0)}(f) = \lambda^n f$ will be unitarily equivalent to $\pi$.*

(ii) *If $n = 0$, then $\pi$ is one-dimensional and unitarily equivalent to $(\lambda, x) \mapsto e^{\mathrm{i}\, a\cdot x} \in \mathbb{C}$ for some vector $a \in \mathbb{R}^2$.*

We're interested in the case $n = 1$; see, for example, theorem 1.2 in [**440**] for a proof of this special case. There are many different realisations for this unique irreducible representation. The simplest (sometimes called the Schrödinger representation) uses the Hilbert space $\mathcal{H} = L^2(\mathbb{R})$. The action of $(\lambda, x) \in H$ on $f \in L^2(\mathbb{R})$ is given by the unitary operator $U_{(\lambda,x)}$ defined by

$$(U_{(\lambda,x)} f)(y) = \lambda \exp\left[\pi \mathrm{i}\ (2yx_2 + x_1 x_2)\right]\ f(y + x_1).$$

This is (essentially) the exponential of the defining representation (4.2.5) of $\mathfrak{Heis}$. Incidentally, the action of $S_b, T_a$ in (2.4.5) on entire functions extends to an $n = -1$ representation of $H$; this representation is anti-linearly equivalent to the Schrödinger representation.

We want to recover the theta function naturally from the $n = 1$ representation. As always, 'natural' means free of arbitrary choices, such as a specific realisation of the $n = 1$ representation, or a specific basis of the underlying Hilbert space. Begin with any realisation $(\pi, \mathcal{H})$ of the $n = 1$ representation of $H$.

As we see in Section 1.5.5, a unitary representation $U$ of a Lie group $G$ on a space $\mathcal{H}$ induces a representation $\delta U$ (the derived module) of the corresponding Lie algebra $\mathfrak{g}$ on a dense subspace $\mathcal{H}_\infty$ of $\mathcal{H}$ by anti-Hermitian operators. For example, the representation (2.4.4) of $U_1(\mathbb{R})$ acts on the Hilbert space $\mathcal{H} = L^2(S^1) \oplus L^2(S^1)$ of all pairs $\binom{f(\theta)}{g(\theta)}$. To see how the corresponding Lie algebra $u_1(\mathbb{R}) = \mathbb{R}$ acts, decompose (2.4.4) into irreducibles

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

(i.e. diagonalise):

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{pmatrix}^{-1} \begin{pmatrix} e^{\mathrm{i}\theta} & 0 \\ 0 & e^{-\mathrm{i}\theta} \end{pmatrix} \begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{pmatrix}.$$

Thus the Lie algebra $\mathfrak{u}_1(\mathbb{R})$ acts as

$$x \mapsto \begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{i}x\frac{\mathrm{d}}{\mathrm{d}\theta} & 0 \\ 0 & -x\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\theta} \end{pmatrix} \begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{pmatrix} = \begin{pmatrix} 0 & -x\frac{\mathrm{d}}{\mathrm{d}\theta} \\ x\frac{\mathrm{d}}{\mathrm{d}\theta} & 0 \end{pmatrix}.$$

The domain of these operators isn't the whole of the Hilbert space $\mathcal{H}$, but it does contain the dense subspace consisting of the infinitely differentiable functions.

Similarly, our representation $\pi$ of $H$ on $\mathcal{H}$ induces an anti-Hermitian representation $\delta\pi$ of $\mathfrak{Heis}$ on a dense subspace $\mathcal{H}_\infty$ of $\mathcal{H}$. If we write $e^{x_1A} = (1, (x_1, 0))$, $e^{x_2B} = (1, (0, x_2))$ and $e^{tC} = (e^{2\pi\mathrm{i}t}, 0)$, then using the Baker–Campbell–Hausdorff formula (1.4.6), these generators obey $[A, B] = C$, $[A, C] = [B, C] = 0$. As an example, in the Schrödinger $n = 1$ representation on space $\mathcal{H} = L^2(\mathbb{R})$, these become the 'momentum operator' $\delta U_A f = \frac{\mathrm{d}f}{\mathrm{d}x}$, the 'position operator' $(\delta U_B f)(x) = 2\pi\mathrm{i}xf(x)$ and the central term $\delta U_C f = 2\pi\mathrm{i}f$. In this example, the dense subspace $\mathcal{H}_\infty$ is the Schwartz space $S(\mathbb{R})$ (Section 1.3.1) consisting of infinitely differentiable, rapidly decreasing functions.

We are now ready to define the two vectors $f_\tau, e_\mathbb{Z}$ in (2.4.6). Consider the subspace $W_\tau$ consisting of all $f \in \mathcal{H}$ for which $(\delta\pi_A - \tau\delta\pi_B)f$ is defined, and equals 0. This can be thought of as a holomorphicity condition $\frac{\partial}{\partial\bar{z}}f = 0$ (recall $\tau$ corresponds to $\sqrt{-1}$). We know that $W_\tau$ will be one-dimensional for our choice of $\pi$, since it manifestly is for the Schrödinger representation $U$: there, $W_\tau = \mathbb{C}e^{\pi\mathrm{i}\tau y^2}$. Choose any nonzero $f_\tau \in W_\tau$.

The map $\sigma(n) := ((-1)^{n_1n_2}, n)$ defines a homomorphism $\mathbb{Z}^2 \to H$, and obeys $(\rho \circ \sigma)(n) = n$ for the obvious projection $\rho : H \to \mathbb{R}^2$ – we say $\rho$ 'splits over $\mathbb{Z}^2$'. Define $V$ to be the common 1-eigenspace of all $U_{\sigma(n)}$. More precisely, let $V$ consist of all (tempered) distributions $\mu \in \mathcal{H}_0^*$ with the property that, for all $n \in \mathbb{Z}^2$ and all $f \in \mathcal{H}_\infty$, $\langle \pi_{\sigma(n)} f, \mu\rangle = \langle f, \mu\rangle$. For example, in the Schrödinger representation, we must have $e^{2\pi\mathrm{i}n_2y}\mu(y + n_1) = \mu(y)$ for all $n \in \mathbb{Z}^2$. Note that $\mu(y) = \Sigma_{n\in\mathbb{Z}}\delta(y + n)$ satisfies that, and using test functions $f(y) = e^{2\pi\mathrm{i}my}$ it quickly follows that this $\mu$ is unique up to scalar multiplication. Therefore, for our representation $\pi$, $V$ will also be one-dimensional. Choose any nonzero $\mu_\mathbb{Z} \in V$. It encodes quasi-periodicity.

Thus we obtain, in the Schrödinger representation,

$$\langle U_{(1,x)} f_\tau, \mu_\mathbb{Z}\rangle = \left\langle e^{\pi\mathrm{i}\,(2yx_2 + x_1x_2)} e^{\pi\mathrm{i}\tau\,(y+x_1)^2}, \sum_n \delta(y + n) \right\rangle,$$

which simplifies to the right side of (2.4.6) with $c = 1$. Therefore, by uniqueness of $\pi$ and basis independence of the Hermitian product $\langle\, ,\, \rangle$, we get that (2.4.6) holds regardless of the realisation $(\pi, \mathcal{H})$ and vectors $f_\tau, \mu_\mathbb{Z}$ we choose.

The reader can verify that quasi-periodicity is automatic (Question 2.4.3). The modularity is of course more difficult (and more interesting). To do this, we need to describe the action of $\mathrm{SL}_2(\mathbb{R})$ on the space $\mathcal{H}$ (which we can take to be $L^2(\mathbb{R})$).

Any $\gamma \in \mathrm{SL}_2(\mathbb{R})$ defines an automorphism of $H$ by $(\lambda, x) \mapsto (\lambda, \gamma.x)$, by

$$\gamma.x = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} dx_1 - cx_2 \\ -bx_1 + ax_2 \end{pmatrix}. \qquad (2.4.8)$$

The precise form of this action is chosen so that (2.4.10a) below will involve the usual Möbius action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$. We can twist by $\gamma$ and thus get a new representation $(\pi', \mathcal{H})$ of $H$, defined by $\pi'_{(\lambda,x)} f = \pi_{(\lambda,\gamma.x)} f$. Obviously $\pi'$ is also irreducible and has central parameter $n = 1$, so by the Stone–von Neumann Theorem must be unitarily equivalent to $\pi$. That is, there exists a unitary operator $R_\gamma$ on the Hilbert space $\mathcal{H}$ that intertwines $\pi$ and $\pi'$: $R_\gamma \pi = \pi' R_\gamma$. The assignment $\gamma \mapsto R_\gamma$ is only defined up to a constant, and so we get a projective representation of $\mathrm{SL}_2(\mathbb{R})$ on $\mathcal{H}$. As we learn in Section 3.1.1, projective representations become true representations when we centrally extend. In particular, we get a true representation when we replace $\mathrm{SL}_2(\mathbb{R})$ with a double-cover called the *metaplectic group* $\mathrm{Mp}_2(\mathbb{R})$.

The metaplectic group is the unique connected double-cover of $\mathrm{SL}_2(\mathbb{R})$. It can be thought of as a way of keeping track of which branch of the square-root we're on in equations like (2.3.5b), and this provides its easiest realisation. Define $\mathrm{Mp}_2(\mathbb{R})$ to be the set of all pairs $(\gamma, s)$, where $\gamma \in \mathrm{SL}_2(\mathbb{R})$ and $s = s(\tau)$ is a choice of holomorphic square-root of $c\tau + d$. Since there are two choices for $s$ (differing by a sign), this is indeed a double-cover. The group operation is

$$(\gamma, s(\tau))(\gamma', s'(\tau)) = (\gamma\gamma', s(\gamma'.\tau)\, s'(\tau)), \qquad (2.4.9)$$

as can be seen by calculating from (2.3.6) with $k = 1/2$.

Returning to the $\gamma$-twist $\pi'$ of the representation $\pi$ of $H$, it is possible to choose unitary operators $R_{(\gamma,s)}$, for each $(\gamma, s) \in \mathrm{Mp}_2(\mathbb{R})$, such that $R_{(\gamma,s)}\pi = \pi' R_{(\gamma,s)}$ and $(\gamma, s) \mapsto R_{(\gamma,s)}$ defines a representation of the metaplectic group $\mathrm{Mp}_2(\mathbb{R})$.

Recalling the definition of $f_\tau$ and $\mu_{\mathbb{Z}}$ as eigenvectors, it isn't difficult to see that

$$R_{(\gamma,s)} f_\tau = s(\tau)^{-1} f_{\gamma.\tau}, \qquad \forall (\gamma, s) \in \mathrm{Mp}_2(\mathbb{R}), \qquad (2.4.10a)$$

$$R_{(\gamma,s)} \mu_{\mathbb{Z}} = \mu_{(\gamma,s)} e_{\mathbb{Z}}, \qquad \forall (\gamma, s) \in \widetilde{\Gamma}_\theta = \{(\gamma, s) \in \mathrm{Mp}_2(\mathbb{R}) \,|\, \gamma \in \Gamma_\theta\}, \qquad (2.4.10b)$$

where $\gamma.\tau$ is the usual action (2.1.4a) and where $\mu : \widetilde{\Gamma}_\theta \to \mathbb{C}^*$ is some one-dimensional representation (with values in eighth roots of unity). See chapter 8 of [**440**] for the detailed calculation. We now immediately obtain from (2.4.6) and (2.4.7) that

$$c\, e^{\pi \mathrm{i} x_1 (\tau x_1 + x_2)} \theta_3(\tau, x_1\tau + x_2) = \langle \pi_{(1,\gamma.x)} R_\gamma f_\tau, R_\gamma \mu_{\mathbb{Z}} \rangle = s(\tau)^{-1} \langle \pi_{(1,\gamma.x)} f_{\gamma.\tau}, \mu_{(\gamma,s)} e_{\mathbb{Z}} \rangle$$
$$= c\, s(\tau)^{-1} \mu_{(\gamma,s)} \exp\left[\pi \mathrm{i}\,(dx_1 - cx_2)\left(\frac{a\tau + b}{c\tau + d}(dx_1 - cx_2) + (-bx_1 + ax_2)\right)\right]$$
$$\times\, \theta_3\left(\frac{a\tau + b}{c\tau + d}, (dx_1 - cx_2)\frac{a\tau + b}{c\tau + d} + (-bx_1 + ax_2)\right), \qquad (2.4.11)$$

for all $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_\theta$, which simplifies down to the desired modularity (2.3.5b).

Last subsection we learned that $\mathrm{SL}_2(\mathbb{R})$ acts transitively on $\mathbb{H}$. Using this and (2.4.10a), we can refine (2.4.6) and write $\theta_3$ as a matrix entry of a unitary representation of the

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

obvious semi-direct product of $\mathrm{Mp}_2(\mathbb{R})$ with $H$. We obtain

$$c\, e^{\pi \mathrm{i} x_1(\tau x_1+x_2)}\,\theta_3(\tau, x_1\tau + x_2) = \sqrt{c\mathrm{i}+d}\,\langle \pi_{(1,x)} R_{(\gamma,s)} f_\tau, \mu_{\mathbb{Z}}\rangle, \tag{2.4.12}$$

where $\tau = \frac{bd+ac+\mathrm{i}}{c^2+d^2}$, for $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$.

This argument is far longer and more technically difficult than the other proofs of theta function modularity given in this chapter, and it is easy to get lost in the details. But it is a remarkable argument, and much more conceptual than, for example, Poisson summation. The modular group $\mathrm{SL}_2(\mathbb{Z})$ (or rather its subgroup $\Gamma_\theta$) arises here as a group of automorphisms of $H$ transforming in a controlled way the vectors $f_\tau$ and $\mu_{\mathbb{Z}}$. The intrinsically algebraic nature of the argument means it generalises easily, and with little extra effort we could have given the proof for Siegel theta functions. (Nonholomorphic) Eisenstein series can also be constructed and studied in a similar way (by first lifting to $\mathrm{SL}_2(\mathbb{R})$). But as with the previous modularity proofs, new ideas would be needed to generalise it beyond these classical functions into a general device providing uniform proofs of modularity for Moonshine functions. In the next subsection though we explain why it might after all have something to do with Moonshine.

### *2.4.3 Braided #2: from the trefoil to Dedekind*

The decomposition (2.4.1b) says that $\mathrm{SL}_2(\mathbb{R})$ is topologically homeomorphic to $\mathbb{R}^2 \times S^1$, i.e. the interior of a solid torus (or if one prefers, the complement of $S^1$ in $\mathbb{R}^3$). In remarkable work in the context of computing $k_2(\mathbb{Z})$ (see Section 2.5.1), Quillen showed that the space $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$ is naturally diffeomorphic to the complement of the trefoil knot in the sphere $S^3$ (see pages 84–5 of [**419**] for the elementary argument). Namely, the Eisenstein series $a = G_4, b = G_6$ in (0.1.5) identify the space $\mathrm{GL}_2(\mathbb{Z})\backslash\mathrm{GL}_2(\mathbb{R})$ of two-dimensional lattices with the complement of the complex curve $20a^3 - 49b^2 = 0$ (which corresponds to degenerate lattices); the intersection of $20a^3 - 49b^2 = 0$ with the sphere $|a|^2 + |b|^2 = 1$ in $\mathbb{C}^2$ (to get instead $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$) is then identified with the trefoil (the (2,1)-torus knot, drawn in Figure 1.10). Now, in Section 2.4.1 we lift modular forms for $\mathrm{SL}_2(\mathbb{Z})$ to the space $L^2(\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R}))$: thus, for example, the $j$-function is a complex-valued function on the complement of the trefoil. More generally, as we will see later, the characters of an affine algebra, or vertex operator algebra, or rational conformal field theory, are vector-valued functions on the complement of the trefoil. The cusps of $\mathbb{H}$ can be interpreted as rational points on the trefoil. Can modular forms and functions somehow see this topological trefoil? The answer is yes!

First, the fundamental group of the complement of the trefoil is easy to compute using the Wirtinger presentation (Section 6.2.5), and is naturally isomorphic to the braid group $\mathcal{B}_3$. This suggests the following picture. Write $G$ for $\mathrm{SL}_2(\mathbb{R})$, $\widetilde{G}$ for its universal cover and $\Gamma$ for $\mathrm{SL}_2(\mathbb{Z})$. Then

$$\widetilde{G} \xrightarrow{\pi} G \xrightarrow{q} \Gamma\backslash G. \tag{2.4.13}$$

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Of course $\pi$ is surjective and has kernel $\pi_1(G) \cong \mathbb{Z}$. $\widetilde{G}$ is also the universal cover of the trefoil-complement $\Gamma\backslash G$, and the kernel of this surjective map $q \circ \pi$ is the central extension $\pi_1(\Gamma\backslash G) \cong \mathcal{B}_3$ of the modular group $\mathrm{SL}_2(\mathbb{Z})$. The map $\mathcal{B}_3 \to \mathrm{SL}_2(\mathbb{Z})$ is simply the reduced Burau representation (1.1.11b) specialised to $w = -1$ (recall (1.1.10a)).

So what does this mean for modular forms? Recall from Section 2.2.1 that modular forms for $\mathrm{SL}_2(\mathbb{Z})$ have multiplier $\mu$ that carries a *projective* representation of $\mathrm{SL}_2(\mathbb{Z})$ – it will be a true representation only when the weight $k$ is an integer. As we emphasise in Section 3.1.1, *projective* representations become *true* representations when one centrally extends. Especially when the weight is fractional, the role of $\mathrm{SL}_2(\mathbb{R})$ really should be played by the more fundamental Lie group $\widetilde{\mathrm{SL}_2(\mathbb{R})}$, and likewise the modular group $\mathrm{SL}_2(\mathbb{Z})$ should be replaced by its central extension $\mathcal{B}_3$.

For a good example, recall the Dedekind eta function $\eta(\tau)$ of (2.2.6b). As we see in (2.2.8), it is a modular form for $\mathrm{SL}_2(\mathbb{Z})$ of weight $\frac{1}{2}$, whose multiplier $\mu$ is quite complicated as a function on $\mathrm{SL}_2(\mathbb{Z})$. But $\mathcal{B}_3$ is the more fundamental transformation group underlying $\eta(\tau)$. Indeed, in terms of $\mathcal{B}_3$, the multiplier is trivial to describe:

$$\mu(\beta) = \exp\left[\frac{2\pi \mathrm{i}}{24}\deg\beta\right], \tag{2.4.14}$$

where the degree of a braid is the length of its word in $\sigma_1, \sigma_2$ (Section 1.1.4). More generally, the multiplier for any modular form for $\mathrm{SL}_2(\mathbb{Z})$ will be similar, with '24' replaced by some other rational. Surely this algebraic interpretation of Dedekind sums in terms of $\mathcal{B}_3$ is related to the topological interpretation of Dedekind sums reviewed and explored in [**24**]; see also [**23**], [**43**].

Of course the multiplier of $\eta$ is almost as trivial if we write $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ as a monomial in the generators $S, T$, but finding that monomial isn't easy. On the other hand, finding '$\deg\beta$' by looking at the braid $\beta$ is easy: just count the crossings in $\beta$, with signs. The multiplier, as a function of $\beta$, is far simpler than as a function of $a, b, c, d$. Our topological considerations have been rewarded!

Likewise, the multiplier in the vector-valued Jacobi form (2.3.8) (again of weight $\frac{1}{2}$) defines a four-dimensional projectivere presentation of $\mathrm{SL}_2(\mathbb{Z})$, given by the tensor product of the one-dimensional representation $\exp[2\pi \mathrm{i}\deg\beta/8]$ of $\mathcal{B}_3$, with a true four-dimensional representation of $\mathrm{SL}_2(\mathbb{Z})$.

Of course the metaplectic group was introduced last subsection for essentially the same reason ($\mathrm{Mp}_2(\mathbb{R})$ is also a quotient of $\widetilde{\mathrm{SL}_2(\mathbb{R})}$). Indeed, since most modular forms arising in the literature have weight in $\frac{1}{2}\mathbb{Z}$, the metaplectic group is a large enough central extension, and $\widetilde{\mathrm{SL}_2(\mathbb{R})}$ may seem like overkill. But modular forms with fractional weight exist in abundance for arbitrarily large denominator (see e.g. [**303**] for examples). The important 'one-point functions on a torus' (Section 4.3.2) in conformal field theory (CFT), to which family the Moonshine functions naturally belong, can form vector-valued modular forms of arbitrary rational weight. We will see in Section 7.2.4 how nicely the CFT machinery accommodates this universal $\mathcal{B}_3$ action, and also how other

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

considerations in (Monstrous) Moonshine are trying to focus our attention on the relation of $\mathcal{B}_3$ to modular functions.

The braid group $\mathcal{B}_3$ is at least as relevant for the nonholomorphic automorphic forms of $\mathrm{SL}_2(\mathbb{Z})$, alluded to in Section 2.4.1. For a simple example, [**379**] studies the Maass cusp forms $u(\tau)$ (with weight 0), identifying them with 'period functions' $\psi(z)$; the exact symmetry $u(-1/\tau) = u(\tau)$ becomes $\psi(1/z) = z^{2s}\psi(z)$, where $s$ is the 'spectral parameter' of $u$. This transformation of the $\psi$'s, with the factor $z^{2s}$, is what one would expect from the braid group (compare (7.2.4)).

We should regard $\mathcal{B}_3$ as the universal symmetry of (not necessarily holomorphic) modular forms for $\mathrm{SL}_2(\mathbb{Z})$. If instead we have modular forms for some subgroup $\Gamma$ of $\mathrm{SL}_2(\mathbb{Z})$, then the role of $\mathcal{B}_3$ is replaced by its subgroup that projects (via the reduced Burau representation (1.1.11b) specialised to $w = -1$) to $\Gamma$. For instance, the principal congruence subgroup $\Gamma(2)$ corresponds to the pure braid group $\mathcal{P}_3$. It would be interesting to find the topological interpretation of $\Gamma_0(p)+$ in (7.1.5) and the other modular groups appearing in Monstrous Moonshine.

The lesson of Section 2.4.1 is that, whenever we have some sort of modularity for, for example, $\mathrm{SL}_2(\mathbb{Z})$, we should lift the domain to that of the relevant Lie group (e.g. $\mathrm{SL}_2(\mathbb{R})$). This should be especially valuable for providing perspective and clarity when we are investigating a new modular-like phenomenon. To give one example among many, [**519**] introduces nonholomorphic deformations of familiar modular forms relevant to strings on a pp-wave background (a 1-parameter deformation of flat space-time). Of more direct relevance to us is the question: *Is it natural to regard the modular functions (characters) of RCFT, VOAs and Moonshine as functions on $SL_2(\mathbb{R})$?*

The lesson of this subsection is that an $\mathrm{SL}_2(\mathbb{Z})$-action may become simpler when lifted to its central extension $\mathcal{B}_3$. The braid group provides a clean universal formulation especially appropriate when metaplectic groups or other central extensions of $\mathrm{SL}_2(\mathbb{Z})$ arise. Mathematics thrives on having alternate interpretations for the same phenomemon: here we replace the matrix group $\mathrm{SL}_2(\mathbb{Z})$ (or its subgroups) with the topologically defined $\mathcal{B}_3$ (or its subgroups). Some things will be easier in one formalism, and presumably other things in the other (e.g. the multipliers $\mu$ are much easier for $\mathcal{B}_3$). It is tempting to apply this to the so-called S-duality of superstrings (Section 3.2.5). *Are there other ways modular forms for $SL_2(\mathbb{Z})$ see the trefoil?*

The modularity argument of Section 2.4.2 has never been applied to Monstrous Moonshine, to this author's knowledge. But one hint that it might be the shadow of such a device is that the braid group lurks here. In particular, there is an action of $\mathcal{B}_3$ on $G \times G$, for any group $G$ (Question 2.4.4); the action (2.4.8) of $\mathrm{SL}_2(\mathbb{Z})$ on $H$ is really this action of $\mathcal{B}_3$ on $\mathbb{R}^2$ – it factors through to $\mathrm{SL}_2(\mathbb{Z})$ because $\mathbb{R}^2$ is abelian. In Section 7.3.3 we use this same action, this time applied to $\mathbb{M} \times \mathbb{M}$, to identify the group-theoretic property of the Monster $\mathbb{M}$ that could be responsible for the genus-0 properties of the McKay–Thompson series $T_g$.

Another hint, perhaps more substantial, of its relevance to Moonshine-like phenomena is the repeated appearance of Maslov indices in the study of gluing anomalies in three-dimensional topological field theory (see chapter IV of [**534**]). This suggested to Turaev

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

an intimate relation of topological field theory with the Segal–Shale–Weil representations of the metaplectic groups. These representations also appear in the context of braids and subfactors [**252**] – metaplectic representations arise naturally there when constructing knot invariants from braids. Much of the mathematical background is developed in [**387**], where we also learn that the universal cover $\widetilde{\mathrm{SL}_2(\mathbb{R})}$ can easily be expressed using Maslov indices.

Question 2.4.1. Use the decomposition (2.4.1b) to find a (noncanonical) group structure on $\mathbb{H}$, inherited from that of $\mathrm{SL}_2(\mathbb{R})$.

Question 2.4.2. Show that uniqueness of the representation in Theorem 2.4.2 fails if $H$ is replaced with infinitely many coupled Heisenberg groups. (This is a major complication for quantum field theory, as we see in Section 4.2.2 in the context of Haag's Theorem.)

Question 2.4.3. Verify that any function of the form $F(x) = \langle \pi_{(1,x)} f, \mu_{\mathbb{Z}} \rangle$, for any $f$ for which $F$ is defined, necessarily obeys $F(x+n) = (-1)^{n_1 n_2} e^{\pi \mathrm{i}(n_1 x_2 - n_2 x_1)} F(x)$. Hence $\mu_{\mathbb{Z}}$ is responsible for the quasi-periodicity (2.3.5a) of $\theta_3$.

Question 2.4.4. (a) Let $G$ be a finite group. Verify that we obtain a right braid group $\mathcal{B}_3$ action on the Cartesian product $G \times G \times G$, by defining

$$(g, h, k).\sigma_1 = (ghg^{-1}, g, k), \ \ (g, h, k).\sigma_2 = (g, hkh^{-1}, h), \tag{2.4.15a}$$

where $\sigma_i$ are the usual generators of $\mathcal{B}_3$ (recall (1.1.9)). Also, verify that there is a right $\mathcal{B}_3$-action on $G \times G$, generated by

$$(g, h).\sigma_1 = (g, gh), \ \ (g, h).\sigma_2 = (gh^{-1}, h). \tag{2.4.15b}$$

(b) Let $C \subseteq G \times G$ consist of all pairs $(g, h)$ where $gh = hg$. Show that this $\mathcal{B}_3$ action takes $C$ to itself, and that its restriction to $C$ actually defines an action of $\mathrm{SL}_2(\mathbb{Z})$ on $C$.
(c) Extend the $\mathcal{B}_3$ actions of (a) to $\mathcal{B}_n$ actions on $G^n$ and $G^{n-1}$.

Question 2.4.5. (a) Show that $\mathrm{SL}_2(\mathbb{R})$ is isomorphic to the group

$$\mathrm{SU}_{1,1}(\mathbb{C}) := \left\{ \begin{pmatrix} \alpha & \overline{\beta} \\ \beta & \overline{\alpha} \end{pmatrix} \in \mathrm{SL}_2(\mathbb{C}) \right\},$$

by showing they are conjugate in $\mathrm{SL}_2(\mathbb{C})$.
(b) Verify that $\mathrm{SU}_{1,1}(\mathbb{C})$ is isomorphic to the set of all pairs $(\gamma, \theta)$, where $|\gamma| < 1$ and $-1 < \theta \le 1$, with group operation $(\gamma, \theta)(\gamma', \theta') = (\theta'', \theta'')$ where

$$\gamma'' = \frac{\gamma + \gamma' e^{-2\pi \mathrm{i}\theta}}{1 + \overline{\gamma}\gamma' e^{-2\pi \mathrm{i}\theta}}, \qquad \theta'' = \theta + \theta' + \frac{1}{2\pi \mathrm{i}} \log \frac{\gamma + \gamma' e^{-2\pi \mathrm{i}\theta}}{1 + \overline{\gamma}\gamma' e^{-2\pi \mathrm{i}\theta}} \ (\mathrm{mod}\ 2).$$

(c) Using (b), realise the universal cover $\widetilde{\mathrm{SL}_2(\mathbb{R})}$ of $\mathrm{SL}_2(\mathbb{R})$.
(d) Realise $\mathcal{B}_3$ as a subgroup of $\widetilde{\mathrm{SL}_2(\mathbb{R})}$.

## 2.5 Meta-patterns in mathematics

### *2.5.1 Twenty-four*

There are lots of 'meta-patterns' in mathematics, i.e. collections of seemingly different problems that have similar answers, or structures that appear more often than we would have expected. Once one of these meta-patterns is identified it is always helpful to understand what is responsible for it, to see what simple structure or basic lemma underlies it. Why are groups so important in mathematics and science? Because they are the devices through which we 'act' on sets, spaces, etc. Mathematics is not above metaphysics; like any area it grows by asking questions, and changing one's perspective – even to a metaphysical one – should suggest new questions.

To give a trivial example, years ago while the author was writing up his PhD thesis he noticed in several places the numbers 1, 2, 3, 4 and 6. For instance $\cos(2\pi r) \in \mathbb{Q}$ for $r \in \mathbb{Q}$ iff the denominator of $r$ is 1, 2, 3, 4 or 6. Likewise, the theta function $\Theta_{\mathbb{Z}+r}(\tau)$ for $r \in \mathbb{Q}$ can be written as $\sum a_i \theta_3(b_i \tau)$ for some $a_i, b_i \in \mathbb{R}$ iff the denominator of $r$ is 1, 2, 3, 4 or 6. This pattern is easy to explain: they are precisely those positive integers $n$ with Euler totient $\phi(n) \leq 2$, that is there are at most two positive numbers less than $n$ coprime to $n$. The various incidences of these numbers can usually be reduced to this $\phi(n) \leq 2$ property. For example, the number field $\mathbb{Q}[\cos(2\pi \frac{a}{b})]$ (see Section 1.7.1), considered as a vector space over $\mathbb{Q}$, has dimension $\phi(b)/2$.

A more interesting meta-pattern involves the number 24 and its divisors (especially 8). One sees 24 wherever modular forms naturally appear. For instance, we see it in the critical dimensions in string theory: the bosonic string lives in a background spacetime of dimension $24 + 2$, while the fermionic string lives in $8 + 2$ dimensions. Another example: the dimensions of even self-dual positive-definite lattices must be a multiple of 8 (e.g. the $E_8$ root lattice has dimension 8, while the Leech lattice has dimension 24). The meta-pattern 24 is also easy to understand: the fundamental problem for which it is the answer is the following one. Fix $n$, and consider the congruence $x^2 \equiv 1 \pmod{n}$. Certainly in order to have a chance of satisfying this, $x$ and $n$ must be coprime. The extreme situation[11] is when *every* number $x$ coprime to $n$ satisfies this congruence: that is,

$$\gcd(x, n) = 1 \qquad \Longleftrightarrow \qquad x^2 \equiv 1 \pmod{n}. \tag{2.5.1}$$

The reader can try to verify the following simple fact: $n$ obeys this extreme situation (2.5.1) iff $n$ divides 24. What does this congruence property have to do with these other occurrences of 24? The elementary argument for even self-dual positive-definite lattices involves the construction $L\{T\}$ of Section 2.3.3 and is sketched in Question 2.5.1.

The '24' appearing in the $q^{1/24}$ of $\eta$ is the same as the 24 in $c/24$ appearing in, for example, (3.1.10); in both cases they come from $\zeta(-1) = -1/12$ or equivalently

[11] This is a standard trick in mathematics: when some sort of bound is established, look at the extremal cases that realise that bound. If your bound is a good one, it should be possible to say something about those extremal cases, and having something to say is always of paramount importance. This strategy is used, for instance, in the definition of normal subgroup in Section 1.1.1 and of simple-currents in Section 6.1.1.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

$\zeta(2) = \pi^2/6$. Are these the same as the 24 in (2.5.1)? Note from the right side of (2.2.6b) that

$$\eta(\tau) = \Theta_{\mathbb{Z}+\frac{1}{12}}(12\tau) - \Theta_{\mathbb{Z}+\frac{5}{12}}(12\tau).$$

Using this identity, the fact that $\eta(\tau+1)$ is a constant multiple of $\eta(\tau)$ is indeed related to (2.5.1). Moreover, this '1/24' is directly related to the abelianisation

$$\mathrm{SL}_2(\mathbb{Z})/[\mathrm{SL}_2(\mathbb{Z}), \mathrm{SL}_2(\mathbb{Z})] \cong \mathbb{Z}_{12} \tag{2.5.2}$$

of $\mathrm{SL}_2(\mathbb{Z})$: writing $\eta(-1/\tau)^2 = a\tau\eta(\tau)^2$ and $\eta(\tau+1)^2 = b\eta(\tau)^2$, the multiplier $s \mapsto a, t \mapsto b$ must define a one-dimensional representation of $\mathrm{SL}_2(\mathbb{Z})$, since $\eta^2$ has weight 1 (recall Question 2.2.3); for any group $G$, and in particular $\mathrm{SL}_2(\mathbb{Z})$, the abelianisation $G/[GG]$ is isomorphic to the group of all one-dimensional representations of $G$. This argument forces $b$ to be some12th root of unity, and $a$ to be $b^3$.

Perhaps the most intriguing '24' occurs as a K-theoretic invariant of the integers. *K-theory* is a generalised (co)homology theory, and as such associates a sequence of abelian groups $K_i(X)$ to the object $X$, which can capture some subtle aspects of $X$. When $X$ is a ring, the definition of these invariants $K_i(X)$ is quite involved, and their calculation is very difficult (see e.g. [**419**] – for example, for $X = \mathbb{Z}$ the groups are known only for $0 \le i \le 5$, where they equal $\mathbb{Z}, \mathbb{Z}_2, \mathbb{Z}_2, \mathbb{Z}_{48}, 0, \mathbb{Z}$, respectively. $K_0(\mathbb{Z}) \cong \mathbb{Z}$ says that the projective $\mathbb{Z}$-modules are the free $\mathbb{Z}$-modules $\mathbb{Z}^n$, while $K_1(\mathbb{Z}) \cong \mathbb{Z}_2$ tells us that the Euclidean domain $\mathbb{Z}$ has only two units (namely, $\pm 1$). The first interesting group in this list is $\mathbb{Z}_{48}$, which arises naturally here as an extension of $\mathbb{Z}_{24}$. Thus 24 (or 48) is a number intimately associated with $\mathbb{Z}$. This author knows no direct connection with our definition (2.5.1) of 24, but there is a conjectural relation of $\|K_{4n-2}(\mathbb{Z})_{torsion}\|/\|K_{4n-1}(\mathbb{Z})_{torsion}\|$ with values $\zeta(1-2n)$ of the Riemann zeta function (see e.g. [**230a**]). In particular, $K_3(\mathbb{Z}) \cong \mathbb{Z}_{48}$ is related to $\zeta(-1) = -\frac{1}{12}$, which in turn is related to our 24.

### *2.5.2 A–D–E*

A much deeper and still not-completely-understood meta-pattern is called *A–D–E* (see [**16**] for a discussion and examples). The name comes from the *simply-laced Lie algebras*, i.e. the simple finite-dimensional Lie algebras whose Coxeter–Dynkin diagrams – see Figure 1.17 – contain only single edges (i.e. no arrows). These are the $A_\star$- and $D_\star$-series, along with the $E_6$, $E_7$ and $E_8$ exceptionals. The observation is that many other problems, which don't have anything directly in common with simple Lie algebras, have a solution that falls into this *A–D–E* pattern. Of course, for an object to be meaningfully labelled $X_\ell$ at least some of the data associated with the algebra $X_\ell$ should reappear in some form in that object. Let's look at some examples.

Consider any even positive-definite integral lattice $L$ (Section 1.2.1). The smallest possible nonzero length-squareds in $L$ will be 2, and the vectors of length-squared 2 are special and are called *roots* (Question 1.2.5). It is important in lattice theory to know the lattices that are spanned by their roots; it turns out these are precisely the orthogonal direct sums of lattices called $A_n$, $D_n$ and $E_6$, $E_7$ and $E_8$ (Theorem 1.2.2). They carry

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

those names for a number of reasons. For example, the lattice called $X_n$ has a basis $\{\alpha_1, \ldots, \alpha_n\}$ with the property that the Gram matrix $A_{ij} := \alpha_i \cdot \alpha_j$ is the Cartan matrix (see Section 1.4.5) for the Lie algebra $X_n$. Also, the group generated by reflections in the roots of the lattice $X_n$ is naturally isomorphic to the Weyl group of the Lie algebra $X_n$. Moreover, to any simple Lie algebra there is canonically associated a lattice called the root lattice; for the simply-laced algebras, these are isomorphic to the lattice of the same name. Incidentally, the root lattices for the non-simply-laced simple Lie algebras are (up to rescalings) orthogonal direct sums of the simply-laced root lattices.

A famous $A$–$D$–$E$ example is due to McKay.[12] Consider any finite subgroup $G$ of the Lie group $\mathrm{SU}_2(\mathbb{C})$ (i.e. the $2 \times 2$ unitary matrices with determinant 1). For example, there is the cyclic group $\mathbb{Z}_n$ of $n$ elements generated by the matrix

$$M_n = \begin{pmatrix} \exp[2\pi \mathrm{i}/n] & 0 \\ 0 & \exp[-2\pi \mathrm{i}/n] \end{pmatrix}.$$

There are also the (doubles of) dihedral groups $\mathcal{D}_n$, and the binary tetrahedral, binary octahedral and binary icosahedral groups of orders 24, 48 and 120, respectively. Let $R_i$ be the irreducible representations of $G$. For instance, for $\mathbb{Z}_n$, there are precisely $n$ of these, all one-dimensional, given by sending the generator $M_n$ to $\exp[2\pi \mathrm{i}k/n]$ for each $k = 1, 2, \ldots, n$. Now consider the tensor product $G \otimes R_i$, where we interpret $G \subset \mathrm{SU}_2(\mathbb{C})$ here as a two-dimensional representation. By Theorem 1.1.2 we can decompose that product into a direct sum $\oplus_j m_{ij} R_j$ of irreducibles (the $m_{ij}$ here are multiplicities). Now create a graph with one node for each $R_i$, and with the $i$th and $j$th nodes ($i \neq j$) connected with precisely $m_{ij}$ directed edges $i \to j$. If $m_{ij} = m_{ji}$, we agree to erase the double arrows from the $m_{ij}$ edges. Then McKay [**411**] observed that this graph, for any of these finite $G < \mathrm{SU}_2(\mathbb{C})$, is a distinct extended Coxeter–Dynkin diagram of $A$–$D$–$E$ type (these are all listed in Figure 3.2). For instance, the cyclic group with $n$ elements yields the extended graph of $A_{n-1}$.

How was McKay led to his remarkable correspondence? He knew that the sum of the labels $a_i = 1, 2, 3, 4, 5, 6, 4, 2, 3$ associated with each node of the extended $E_8$ diagram (Figure 3.2) equals 30, the Coxeter number of $E_8$. So what do their *squares* add to? 120, which he recognised as the cardinality of one of the exceptional finite subgroups of $\mathrm{SU}_2(\mathbb{C})$, and that got him thinking . . .

A deep example of $A$–$D$–$E$, due to Arnol'd, are the *simple singularities*. A *singularity* or *critical point* of a smooth function $f : \mathbb{C}^n \to \mathbb{C}$ is a point $z \in \mathbb{C}^n$ where all first partial derivatives $\partial_i f$ vanish. For example, $f(z) = z^{k+1}$ has a singularity at $z = 0$ for any integer $k \geq 1$. We identify singularities if locally they merely differ by a change-of-coordinates – see, for example, [**19**] for details. For example, any singularity of $f : \mathbb{C} \to \mathbb{C}$ is equivalent to one of the form $f(z) = z^{k+1}$. A simple singularity is an isolated singularity and behaves like the poles $f(z) = z^{-n}$ of usual complex analysis – again see [**19**] for the precise definition. For example, $z_1^2 + z_2^{k+1}$ is simple but $z_1^4 + 3z_1^2 z_2^2 + z_2^4$ is not (the coefficient '3' can be deformed, yielding a continuum of inequivalent singularities).

[12] He is the same John McKay we celebrated in Chapter 0.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

Table 2.2. *The simple singularities in* $\mathbb{C}^2$

| Name | $A_k$ | $D_k$ | $E_6$ | $E_7$ | $E_8$ |
|---|---|---|---|---|---|
| Representative | $x^2+y^{k+1}$ | $x^2y+y^{k-1}$ | $x^3+y^4$ | $x^3+xy^3$ | $x^3+y^5$ |

Table 2.2 lists the simple singularities in $\mathbb{C}^2$ up to equivalence. In higher dimensions we get the same list, with the extra variables coming in as $z_3^2+\cdots+z_n^2$. These singularities can be related to McKay's $A$–$D$–$E$ as follows. The group $\mathrm{SU}_2(\mathbb{C})$ acts on $\mathbb{C}^2$ in the obvious way (matrix multiplication). If $G$ is a discrete subgroup of $\mathrm{SU}_2(\mathbb{C})$, then consider the ring of polynomials in two variables $w_1, w_2$ invariant under $G$. It turns out it will have three generators $x(w_1,w_2)$, $y(w_1,w_2)$, $z(w_1,w_2)$, which are connected by one polynomial relation (syzygy). For instance, take $G$ to be the cyclic group $\mathbb{Z}_n$, then we're interested in polynomials $p(w_1,w_2)$ invariant under $w_1\mapsto\exp[2\pi\mathrm{i}/n]w_1$, $w_2\mapsto\exp[-2\pi\mathrm{i}/n]w_2$. Any such invariant $p(w_1,w_2)$ is clearly generated by (i.e. can be written as a polynomial in) $w_1w_2$, $w_1^n$ and $w_2^n$. Choosing instead the generators $x=\frac{w_1^n-w_2^n}{2}$, $y=w_1w_2$, $z=\mathrm{i}\,\frac{w_1^n+w_2^n}{2}$, we get the syzygy $y^n=-(x^2+z^2)$. For any $G$, generators $x,y,z$ can always be found so that the syzygy will be one of the polynomials in Table 2.2 (with '$+z^2$'appended), and this will give the equation of the algebraic surface $\mathbb{C}^2/G$ as a two-dimensional complex surface in $\mathbb{C}^3$. For example, the complex surfaces $\mathbb{C}^2/\mathbb{Z}_n$ and $\{(x,y,z)\in\mathbb{C}^3\,|\,x^2+y^2+z^n=0\}$ are equivalent.

There are other ways these singularities can be associated with $A$–$D$–$E$. Given a surface $\Sigma\subset\mathbb{C}^3$ with a single singularity, a *resolution* $\widetilde{\Sigma}$ is a smooth surface without singularities that agrees with $\Sigma$ away from the singularity (again see [**19**] for details). A *minimal resolution* is one through which any other resolution must factor. The minimal resolution exists and is unique. For example, the $A_1$ singularity $x^2+y^2+z^2=0$ has the resolution

$$\widetilde{\Sigma}=\{(x,y,z,(a,b))\in\mathbb{C}^3\times\mathbb{P}^1(\mathbb{C})\,|\,x^2+y^2+z^2=0,\ xb=ya\}.$$

For $(x,y)\neq(0,0)$, $xb=ya$ uniquely determines the homogeneous coordinates $(a,b)$, but the singularity $(x,y)=(0,0)$ is blown up into the sphere $\mathbb{P}^1(\mathbb{C})$; the points on the sphere parametrise the different (complex) directions in which the singularity can be approached.

More generally, given a minimal resolution $\pi:\widetilde{\Sigma}\to\Sigma$ of a simple singularity, $\pi^{-1}(0)$ will be a union of $r$ spheres $\cup C_i$. duVal [**165**] noticed that these classes $[C_i]$ form a basis of the homology group $H_2(\widetilde{\Sigma},\mathbb{Z})$, on which there is defined a $\mathbb{Z}$-valued intersection form; this form makes $H_2(\widetilde{\Sigma},\mathbb{Z})$ into a negative-definite lattice isomorphic (up to a factor of $\sqrt{-1}$) to the root lattice of $X_r$, where $[C_i]$ map to a basis of simple roots. The Weyl group of $X_r$ is isomorphic to the so-called monodromy group of the singularity (see [**19**] for details).

Incidentally, the *McKay correspondence* refers to the strategy of describing the geometry of the resolution of the orbifold singularities $\mathbb{C}^n/G$ for finite subgroups $G$ of $\mathrm{SL}_n(\mathbb{C})$,

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

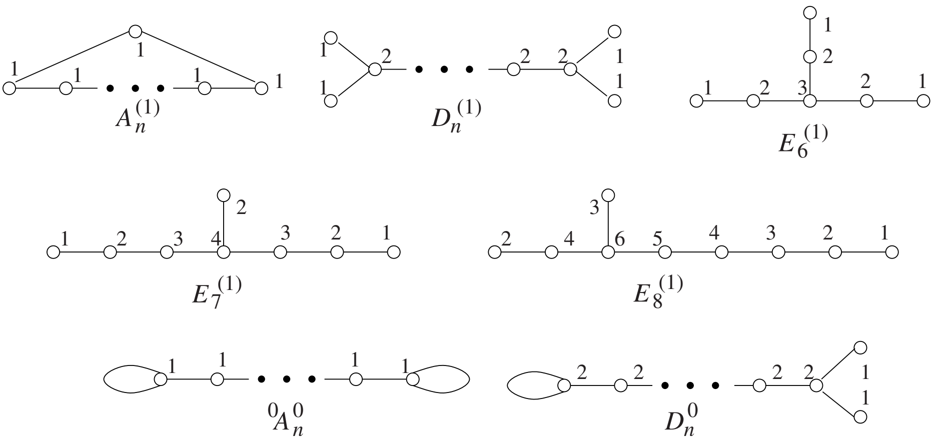

Fig. 2.9 The connected multigraphs with largest eigenvalue 2.

through the representation theory of $G$. See [**254**] for the $n = 2$ story (i.e. for the simple singularities) and [**471**] for fascinating speculations on what happens in dimension $n > 2$.

Arguably the first $A$–$D$–$E$ classification goes back to Theaetetus, who classified the regular solids in 400 B.C. For instance, the tetrahedron can be associated with $E_6$ while the cube is matched with $E_7$. This $A$–$D$–$E$ is only partial, as there are no regular solids assigned to the $A$-series, and to get the $D$-series one must look at 'degenerate regular solids', that is the regular polygons.

The closest we have to an explanation of the $A$–$D$–$E$ meta-pattern would seem to be graphs of small eigenvalues. Consider any multigraph $\mathcal{G}$ – that is, we allow multiple edges (there can be more than one edge connecting two vertices) and loops (an edge running from a node to itself), but all edges are undirected. We can also assume without loss of generality that $\mathcal{G}$ is connected. Assign a positive number $a_i$ to each node. If this assignment has the property that for each $i$, $2a_i = \sum a_j$ where the sum is over all nodes $j$ adjacent to $i$ (counting multiplicities of edges), then we call it 'PF2'. The column vector $(a_1, \ldots, a_n)^t$ will be a strictly positive eigenvector (called the Perron–Frobenius eigenvector) of the adjacency matrix of $\mathcal{G}$, with eigenvalue 2. A multigraph has a PF2 assignment iff the eigenvalue $\lambda$ of its adjacency matrix with largest absolute value $|\lambda|$ is $\lambda = 2$ (see Theorem 2.5.1 below). For instance, for the multigraph ○=○, corresponding to adjacency matrix $\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$, the assignment $a_1 = 1 = a_2$ is PF2 but the assignment $a_1 = 1, a_2 = 2$ is not. The question is, which multigraphs have a PF2 assignment? The answer is given in Figure 2.9. The names $A_n{}^{(1)}$ to $E_6{}^{(1)}$ there come from Figure 3.2; the names ${}^0A_n^0$ and $D_n^0$ are invented. We see that the PF2 multigraphs without loops are precisely the extended Coxeter–Dynkin diagrams of $A$–$D$–$E$ type, and their PF2 assignments are unique (up to constant proportionality) and are given by the *labels* $a_i$ of the corresponding affine algebra (i.e. the numbers attached to the graphs in Figures 2.9 and 3.2).

The unextended diagrams have a similar depiction. For them, we assign positive numbers $a_i$ to each node so that $2a_i \geq \sum_j a_j$, where as before we sum over all adjacent $j$. We also require that for at least one vertex $i$, we don't get an equality. Call this a PF$2^-$ assignment. A multigraph $\mathcal{G}$ has a PF$2^-$ assignment iff the absolute value $|\lambda|$ of each eigenvalue $\lambda$ of its adjacency matrix is $< 2$. In Figure 1.4 we list all multigraphs for which there is a PF$2^-$ assignment.

*Perron–Frobenius theory* studies the eigenvectors/eigenvalues of nonnegative matrices. We revisit this theory elsewhere in the book. The basic result is:

**Theorem 2.5.1 (Perron–Frobenius)** *Let $A$ be an $n \times n$ matrix with real nonnegative entries $A_{ij} \geq 0$ $(1 \leq i, j \leq n)$.*

(a) *Let $\rho(A) := \max_\lambda |\lambda|$ be the maximum of the absolute values of the eigenvalues of $A$. Then $\rho(A)$ is itself an eigenvalue of $A$, called the 'Perron–Frobenius eigenvalue', and it has an eigenvector $(a_1, \ldots, a_n)^t \geq 0$ (i.e. each $a_i \geq 0$), called a 'Perron–Frobenius eigenvector'.*

(b) *If it is not possible to simultaneously permute the rows and columns of $A$ so that $A$ takes the form*

$$A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$$

*for submatrices $B, C, D$ (such a matrix $A$ is called 'irreducible'), then the Perron–Frobenius eigenvector is strictly positive and is unique up to scalar multiples.*

(c) *Suppose $A$ is irreducible in the sense of (b), and $B$ is an $n \times n$ matrix obeying $0 \leq B_{ij} \leq A_{ij}$ $\forall i, j$. Then $\rho(B) \leq \rho(A)$, with equality iff $B = A$.*

See, for example, [**420**] for a proof and further results of this kind. In our case $A$ is the adjacency matrix of a connected multigraph and so, being symmetric, is irreducible in the sense of (b). The classification of all PF2 and PF$2^-$ multigraphs follows by repeatedly applying Theorem 2.5.1(c) (see Question 2.5.2).

What do eigenvalues have to do with the other $A$–$D$–$E$ classifications? Consider a finite subgroup $G$ of $\mathrm{SU}_2(\mathbb{C})$. Take the dimension of the equation $G \otimes R_i = \oplus_j m_{ij} R_j$: we get $2d_i = \sum_j m_{ij} d_j$, where $d_j = \dim(R_j)$. Hence the dimensions of the irreducible representations define a PF2 assignment for each of McKay's graphs, and hence those graphs must be of $A$–$D$–$E$ type (provided we know $m_{ij} = m_{ji}$ and $m_{ii} = 0$).

Or consider lattices: let $\alpha_i$ be a basis of a positive-definite lattice, with all norm-squareds $\alpha_i \cdot \alpha_i = 2$. Then by the Cauchy–Schwarz inequality, $\alpha_i \cdot \alpha_j \in \{0, \pm 1\}$ for $i \neq j$. For $i < j$, if $\alpha_i \cdot \alpha_j = +1$ then replace $\alpha_j$ with $\alpha_j - \alpha_i$. What this means is that we can assume that each $\alpha_i \cdot \alpha_j \in \{0, -1\}$ for $i \neq j$. Put $A_{ij} = \alpha_i \cdot \alpha_j$ and $B = 2I - A$. Then $B$ is a symmetric $\mathbb{N}$-matrix with zeros down the diagonal, and is easily seen to have Perron–Frobenius eigenvalue $< 2$. Thus $B$ falls into the $A$–$D$–$E$ pattern.

**Suggestion** *There are two different, though related, fundamental $A$–$D$–$E$ patterns: namely, the* PF2 *and* PF$2^-$ *multigraph classifications. Any other instance of an $A$–$D$–$E$ pattern reduces to one or the other of these.*

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

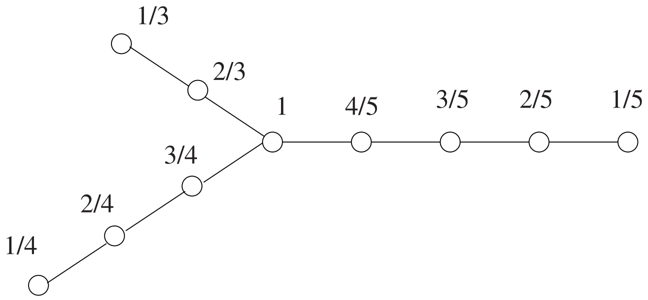

Fig. 2.10 The tree corresponding to $p = 3, q = 4, r = 5$.

This suggestion should be treated with some caution – as simple singularities illustrate, the same area may realise both types of $A$–$D$–$E$ patterns, depending on the specific questions asked. In particular, duVal corresponds to Figure 1.4 and McKay to Figure 2.9. What relates these is that one of the nodes in the McKay graph (namely, that corresponding to the identity) is distinguished, and when it is deleted duVal's graph is recovered. We return to singularities in Section 3.2.5.

We encounter other $A$–$D$–$E$'s later in this book. One of these (Theorem 6.2.2) is the only instance of $A$–$D$–$E$ known to this author that hasn't yet been related to PF2 or PF$2^-$.

Incidentally, it is commonly suggested that a possible explanation for $A$–$D$–$E$ may be the set of all triples $p, q, r \in \mathbb{N}$ for which

$$\frac{1}{p} + \frac{1}{q} + \frac{1}{r} > 1. \tag{2.5.3}$$

Then $(1, q, r)$, $(2, 2, r)$ and $(2, 3, 3), (2, 3, 4), (2, 3, 5)$ (corresponding to $A_{q+r-1}$, $D_{r+2}$, $E_{6,7,8}$, respectively) exhausts all solutions except for $p = 1, q \neq r$. However, this is not as fundamental as the graph explanation suggested above. In particular, given any triple obeying (2.5.3), construct the tree consisting of three strings leaving a common central vertex, of lengths $p - 1, q - 1, r - 1$, respectively (see Figure 2.10). Give this graph the assignment indicated in the figure – that is, label the $i$th vertex from the end of the first (respectively second, third) string $\frac{i}{p}$ (respectively $\frac{i}{q}$, $\frac{i}{r}$). Then inequality (2.5.3) is precisely the statement that this assignment is PF$2^-$, and thus that the graph will be of (unextended) $A$–$D$–$E$ type. The reverse implication, showing that any PF$2^-$ graph $\mathcal{G}$ will necessarily correspond to a triple obeying (2.5.3), is much less elementary.

What comes after $A$–$D$–$E$? Natural candidates should be the graphs with largest eigenvalue $\rho = 3$, say. For the same reason that those with $\rho = 2$ arise in so many contexts, those with $\rho = 3$ surely will too. The difference is that the number and variety of graphs grows dramatically with the largest eigenvalue $\rho$. The list of graphs with $\rho = 2$ has such a simple and tight structure that different situations will automatically share a family resemblance, provided only that they depend critically on graphs with $\rho = 2$. For instance, the eigenvalues of any graph with $\rho = n$ must be character values of an $n$-dimensional representation of SU$_n$, if the graph is to have a chance at being the

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press

McKay graph of a finite subgroup of $\mathrm{SU}_n$; although this is automatic for $n = 2$, it is a severe constraint for $n \geq 3$. A different $\rho = 3$ situation can carry with it its own severe constraints, which would thus overwhelm the presence of the $\rho = 3$ graphs. We could say that $\rho = 2$ is a dominant gene, while $\rho = 3$ is recessive; this is why $A$–$D$–$E$ is so ubiquitous, and why there seems to be no effective successor meta-pattern to $A$–$D$–$E$. (But see Section 6.3.2.)

For a final meta-pattern, consider ‘modular functions’. After all, they appear in many places and disguises. Maybe we shouldn’t regard their ubiquity as fortuitous. Instead, perhaps there’s a deeper common ‘situation’ that is the source for that ubiquity. Two-dimensional lattices, perhaps? Riemann surfaces? The braid group $\mathcal{B}_3$?

Question 2.5.1. Let $L \subset \mathbb{R}^n$ be an even self-dual $n$-dimensional lattice. Assume there exists an orthonormal basis $e_i$ of $\mathbb{R}^n$ and a number $k$ such that the orthogonal lattice $2^k(\mathbb{Z}e_1 \oplus \cdots \oplus \mathbb{Z}e_n)$ is a sublattice of $L$ (this is true for any self-dual $L$ – see theorem 3.15 of [**238**]).

(a) Let $L'$ be the orthonormal lattice $\mathbb{Z}e_1 \oplus \cdots \oplus \mathbb{Z}e_n$. Then the abelian group $L/(L \cap L')$ must be isomorphic to $\mathbb{Z}_{2^{k_1}} \times \cdots \times \mathbb{Z}_{2^{k_m}}$ for $0 < k_m \leq \cdots \leq k_1 \leq k$. Generators $\omega_1, \ldots, \omega_m \in L$ for it can be chosen so that $\omega_i = \frac{1}{2^{k_i}} \sum_j \omega_{ij} e_j$, where $\omega_{ij} \in \mathbb{Z}$, such that $\sum_i c_i \omega_i \in L'$ for $c_i \in \mathbb{Z}$ iff $2^{k_i}$ divides $c_i$ for each $i$. Prove that there exist vectors $r_1, \ldots, r_j \in L'$ such that $r_i \cdot \omega_j \equiv \frac{1}{2^{k_i}} \delta_{ij}$ (mod 1).

(b) Let $x = \sum_i 2^{k_i - k_m} r_i^2 \omega_i = \frac{1}{2^{k_m}} \sum_j x_j e_j$, so $x \in L$ and $x_j \in \mathbb{Z}$. Prove that each $x_j$ is odd (*Hint*: consider $\sum_i \omega_{ij} r_j - e_i$).

(c) Conclude from (2.5.1) that 8 must divide the dimension $n$.

Question 2.5.2. Using Theorem 2.5.1(c), prove that the multigraphs in Figures 1.4 and 2.9 exhaust all connected multigraphs whose eigenvalues $\lambda$ all obey $|\lambda| \leq 2$.

Question 2.5.3. Why are there no loops in the McKay graph corresponding to any finite subgroup of $\mathrm{SL}_2(\mathbb{C})$? Why don’t these McKay graphs have directed edges?

Question 2.5.4. The classifications in Figures 1.4 and 2.9 depend on the requirement that the matrices be symmetric, i.e. that the multigraphs have no arrows. Find all $2 \times 2$ nonnegative integer matrices whose eigenvalues $\lambda$ all obey $|\lambda| \leq 2$.

https://doi.org/10.1017/9781009401548.003 Published online by Cambridge University Press