1
Classical algebra

In this chapter we sketch the basic material – primarily algebra – needed in later chapters. As mentioned in the Introduction, the aspiration of this book isn't to 'Textbookhood'. There are plenty of good textbooks on the material of this chapter (e.g. [**162**]). What is harder to find are books that describe the ideas beneath and the context behind the various definitions, theorems and proofs. This book, and this chapter, aspire to that. What we lose in depth and detail, we hope to gain in breadth and conceptual content. The range of readers in mind is diverse, from mathematicians expert in other areas to physicists, and the chosen topics, examples and explanations try to reflect this range.

Finite groups (Section 1.1) and lattices (Section 1.2.1) appear as elementary examples throughout the book. Lie algebras (Section 1.4), more than their nonlinear partners the Lie groups, are fundamental to us, especially through their representations (Section 1.5). Functional analysis (Section 1.3), category theory (Section 1.6) and algebraic number theory (Section 1.7) play only secondary roles. Section 1.2 provides some background geometry, but for proper treatments consult [**113**], [**104**], [**527**], [**59**], [**478**].

Note the remarkable unity of algebra. Algebraists look at mathematics and science and see structure; they study *form* rather than *content*. The foundations of a new theory are laid by running through a fixed list of questions; only later, as the personality quirks of the new structure become clearer, does the theory become more individual. For instance, among the first questions asked are: What does 'finite' mean here? and What plays the role of a prime number? Mathematics (like any subject) evolves by asking questions, and though a good original question thunders like lightning at night, it is as rare as genius itself. See the beautiful book [**504**] for more of algebra presented in this style.

1.1 Discrete groups and their representations

The notion of a group originated essentially in the nineteenth century with Galois, who also introduced normal subgroups and their quotients G/N, all in the context of what we now call *Galois theory* (Section 1.7.2). According to Poincaré, when all of mathematics is stripped of its contents and reduced to pure form, the result is group theory.[1] Groups are the devices that act, which explains their fundamental role in mathematics. In physics like much of Moonshine, groups arise through their representations. Standard references for representation theory are [**308**], [**219**]; gentle introductions to various aspects of group theory are [**162**], [**421**] (the latter is especially appropriate for physicists).

[1] See page 499 of J.-P. Serre, *Notices Amer. Math. Soc.* (May 2004).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

1.1.1 Basic definitions

A *group* is a set G with an associative product gg' and an identity e, such that each element $g \in G$ has an inverse g^{-1}. The number of elements $\|G\|$ of a group is called its *order*, and is commonly denoted $|G|$.

If we're interested in groups, then we're interested in *comparing* groups, that is we're interested in functions $\varphi : G \to H$ that respect group structure. What this means is φ takes products in G to products in H, the identity e_G in G to the identity e_H in H, and the inverse in G to the inverse in H (the last two conditions are redundant). Such φ are called *group homomorphisms*.

Two groups G, H are considered equivalent or *isomorphic*, written $G \cong H$, if as far as the essential group properties are concerned (think 'form' and not 'content'), the two groups are indistinguishable. That is, there is a group homomorphism $\varphi : G \to H$ that is a bijection (so φ^{-1} exists) and $\varphi^{-1} : H \to G$ is itself a group homomorphism (this last condition is redundant). An *automorphism* (or symmetry) of G is an isomorphism $G \to G$; the set $\operatorname{Aut} G$ of all automorphisms of G forms a group.

For example, consider the cyclic group $\mathbb{Z}_n = \{[0], [1], \ldots, [n-1]\}$ consisting of the integers taken mod n, with group operation addition. Write $\mathrm{U}_1(\mathbb{C})$ for the group of complex numbers with modulus 1, with group operation multiplication. Then $\varphi([a]) = e^{2\pi \mathrm{i} a/n}$ defines a homomorphism between $\mathbb{Z}_n$ and $\mathrm{U}_1(\mathbb{C})$. The group of positive real numbers under multiplication is isomorphic to the group of real numbers under addition, the isomorphism being given by logarithm – as far as their group structure is concerned, they are identical. $\operatorname{Aut}\mathbb{Z} \cong \mathbb{Z}_2$, corresponding to multiplying the integers by ± 1, while $\operatorname{Aut}\mathbb{Z}_n$ is the multiplicative group $\mathbb{Z}_n^\times$, consisting of all numbers $1 \le \ell \le n$ coprime to n (i.e. $\gcd(\ell, n) = 1$), with the operation being multiplication mod n.

Field is an algebraic abstraction of the concept of number: in one we can add, subtract, multiply and divide, and all the usual properties like commutativity and distributivity are obeyed. Fields were also invented by Galois. $\mathbb{C}$, $\mathbb{R}$ and $\mathbb{Q}$ are fields, while $\mathbb{Z}$ is not (you can't always divide an integer by, for example, 3 and remain in $\mathbb{Z}$). The integers mod n, i.e. $\mathbb{Z}_n$, are a field iff n is prime (e.g. in $\mathbb{Z}_4$, it is not possible to divide by the element [2] even though $[2] \neq [0]$ there). $\mathbb{C}$ and $\mathbb{R}$ are examples of fields of characteristic 0 – this means that 0 is the only integer k with the property that $kx = 0$ for all x in the field. We say $\mathbb{Z}_p$ has characteristic p since multiplying by the integer p has the same effect as multiplying by 0. There is a finite field with q elements iff q is a power of a prime, in which case the field is unique and is called $\mathbb{F}_q$. Strange fields have important applications in, for example, coding theory and, ironically, in number theory itself – see Sections 1.7.1 and 2.4.1.

The *index* of a subgroup H in G is the number of 'cosets' gH; for finite groups it equals $\|G\|/\|H\|$. A *normal* subgroup N of a group is one obeying $gNg^{-1} = N$ for all $g \in G$. Its importance arises because the set G/H of cosets gH has a natural group structure precisely when H is normal. If H is a normal subgroup of G we write $H \lhd G$; if H is merely a subgroup of G we write $H < G$. The kernel $\ker(\varphi) = \varphi^{-1}(e_H)$ of a homomorphism $\varphi : G \to H$ is always normal in G, and $\operatorname{Im}\varphi \cong G/\ker\varphi$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

By the *free group $\mathcal{F}_n$ with generators* $\{x_1, \ldots, x_n\}$ we mean the set of all possible words in the 'alphabet' $x_1, x_1^{-1}, \ldots, x_n, x_n^{-1}$, with group operation given by concatenation. The identity e is the empty word. The only identities obeyed here are the trivial ones coming from $x_i x_i^{-1} = x_i^{-1} x_i = e$. For example, $\mathcal{F}_1 \cong \mathbb{Z}$. The group $\mathcal{F}_2$ is already maximally complicated, in that all the other $\mathcal{F}_n$ arise as subgroups.

We call a group G *finitely generated* if there are finitely many elements $g_1, \ldots, g_n \in G$ such that $G = \langle g_1, \ldots, g_n \rangle$, that is any $g \in G$ can be written as some finite word in the alphabet $g_1^{\pm 1}, \ldots, g_n^{\pm 1}$. For example, any finite group is finitely generated, while the additive group $\mathbb{R}$ is not. Any finitely generated group G is the homomorphic image $\varphi(\mathcal{F}_n)$ of some free group $\mathcal{F}_n$, i.e. $G \cong \mathcal{F}_n / \ker(\varphi)$ (why?). This leads to the idea of *presentation*: $G \cong \langle X \mid \mathcal{R} \rangle$ where X is a set of generators of G and $\mathcal{R}$ is a set of relations, that is words that equal the identity e in G. Enough words must be chosen so that $\ker \varphi$ equals the smallest normal subgroup of $\mathcal{F}_n$ containing all of $\mathcal{R}$. For example, here is a presentation for the dihedral group $\mathcal{D}_n$ (the symmetries of the regular n-sided polygon):

$$\mathcal{D}_n = \langle a, b \mid a^n = b^2 = abab = e \rangle. \tag{1.1.1}$$

For two interesting presentations of the trivial group $G = \{e\}$, see [**416**]. To define a homomorphism $\varphi : G \to H$ it is enough to give the value $\varphi(g_i)$ of each generator of G, and verify that φ sends all relations of G to identities in H.

We say G equals the (internal) direct product $N \times H$ of subgroups if every element $g \in G$ can be written uniquely as a product nh, for every $n \in N, h \in H$, and where N, H are both normal subgroups of G and $N \cap H = \{e\}$. Equivalently, the (external) direct product $N \times H$ of two groups is defined to be all ordered pairs (n, h), with operations given by $(n, h)(n', h') = (nn', hh')$; G will be the internal direct product of its subgroups N, H iff it is isomorphic to their external direct product. Of course, $N \cong G/H$ and $H \cong G/N$. Direct product is also called 'homogeneous extension' in the physics literature.

More generally, G is an (internal) semi-direct product $N \rtimes H$ of subgroups if all conditions of the internal direct product are satisfied, except that H need not be normal in G (but as before, $N \lhd G$). Equivalently, the (external) semi-direct product $N \rtimes_\theta H$ of two groups is defined to be all ordered pairs (n, h) with operation given by

$$(n, h)(n', h') = (n\, \theta_h(n'), hh'),$$

where $h \mapsto \theta_h \in \operatorname{Aut} N$ can be any group homomorphism. It's a good exercise to verify that $N \rtimes_\theta H$ is a group for any such θ, and to relate the internal and external semi-direct products. Note that $N \cong \{(n, e_H)\}$, $H \cong \{(e_N, h)\} \cong G/N$. Also, choosing the trivial homomorphism $\theta_h = id$. recovers the (external) direct product. The semi-direct product is also called the 'inhomogeneous extension'.

For example, the dihedral group is a semi-direct product of $\mathbb{Z}_n$ with $\mathbb{Z}_2$. The group of isometries (distance-preserving maps) in 3-space is $\mathbb{R}^3 \rtimes (\{\pm I\} \times \mathrm{SO}_3)$, where $\mathbb{R}^3$ denotes the additive subgroup of translations, $-I$ denotes the reflection $x \mapsto -x$ through the origin, and SO_3 is the group of rotations. This continuous group is an example of a

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Lie group (Section 1.4.2). Closely related is the *Poincaré group*, which is the semi-direct product of translations $\mathbb{R}^4$ with the Lorentz group $\mathrm{SO}_{3,1}$.

Finally and most generally, if N is a normal subgroup of G then we say that G is an (internal) extension of N by the quotient group G/N. Equivalently, we say a group G is an (external) extension of N by H if each element g in G can be identified with a pair (n, h), for $n \in N$ and $h \in H$, and where the group operation is

$$(n, h)(n', h') = (\text{stuff}, hh'),$$

provided only that $(n, e_H)(n', e_H) = (nn', e_H)$.

That irritating *carry* in base 10 addition, which causes so many children so much grief, is the price we pay for building up our number system by repeatedly extending by the group $\mathbb{Z}_{10}$ (one for each digit) (see Question 1.1.8(c)).

A group G is *abelian* if $gh = hg$ for all $g, h \in G$. So $\mathbb{Z}_n$ is abelian, but the symmetric group $\mathcal{S}_n$ for $n > 2$ is not. A group is *cyclic* if it has only one generator. The only cyclic groups are the abelian groups $\mathbb{Z}_n$ and $\mathbb{Z}$. The *centre* $Z(G)$ of a group is defined to be all elements $g \in G$ commuting with all other $h \in G$; it is always a normal abelian subgroup.

Theorem 1.1.1 (Fundamental theorem of finitely generated abelian groups) *Let G be a finitely generated abelian group. Then*

$$G \cong \mathbb{Z}^r \times \mathbb{Z}_{m_1} \times \cdots \times \mathbb{Z}_{m_h}$$

where $\mathbb{Z}^r = \mathbb{Z} \times \cdots \times \mathbb{Z}$ (r times), and m_1 divides m_2 which divides ... which divides m_h. The numbers r, m_i, h are unique. The group G is finite iff $r = 0$.

The proof isn't difficult – for example, see page 43 of [**504**]. Theorem 1.1.1 is closely related to other classical decompositions, such as that of the Jordan canonical form for matrices.

1.1.2 Finite simple groups

Theorem 1.1.1 gives among other things the classification of all finite abelian groups. In particular, the number of abelian groups G of order $\|G\| = n = \prod_p p^{a_p}$ is $\prod_p P(a_p)$, where $P(m)$ is the partition number of m (the number of ways of writing m as a sum $m = \sum_i m_i, m_1 \geq m_2 \geq \cdots \geq 0$).

What can we say about the classification of arbitrary finite groups? This is almost certainly hopeless. All groups of order p or p^2 (for p prime) are necessarily abelian. The smallest non-abelian group is the symmetric group $\mathcal{S}_3$ (order 6); next are the dihedral group $\mathcal{D}_4$ and the quaternion group $\mathcal{Q}_4 = \{\pm 1, \pm i, \pm j, \pm k\}$ (both order 8). Table 1.1 summarises the situation up to order 50 – for orders up to 100, see [**418**]. This can't be pushed that much further, for example the groups of order 128 (there are 2328 of them) were classified only in 1990. One way to make progress is to restrict the class of groups considered.

Every group has two trivial normal subgroups: itself and $\{e\}$. If these are the only normal subgroups, the group is called *simple*. It is conventional to regard the trivial

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Table 1.1. *The numbers of non-abelian groups of order* < 50

$\|G\|$	6	8	10	12	14	16	18	20	21	22	24	26	27
#	1	2	1	3	1	9	3	3	1	1	12	1	2

$\|G\|$	28	30	32	34	36	38	39	40	42	44	46	48
#	2	3	44	1	10	1	1	11	5	2	1	47

group $\{e\}$ as not simple (just as '1' is conventionally regarded as not prime). An alternate definition of a simple group G is that if $\varphi : G \to H$ is any homomorphism, then either φ is constant (i.e. $\varphi(G) = \{e\}$) or φ is one-to-one.

The importance of simple groups is provided by the *Jordan–Hölder Theorem*. By a 'composition series' for a group G, we mean a nested sequence

$$G = H_0 > H_1 > H_2 > \cdots > H_k > H_{k+1} = \{e\} \tag{1.1.2}$$

of groups such that H_i is normal in H_{i-1} (though not necessarily normal in H_{i-2}), and the quotient H_{i-1}/H_i (called a 'composition factor') is simple. An easy induction shows that any finite group G has at least one composition series. If $H'_0 > \cdots > H'_{\ell+1} = \{e\}$ is a second composition series for G, then the Jordan–Hölder Theorem says that $k = \ell$ and, up to a reordering π, the simple groups H_{i-1}/H_i and $H'_{\pi j-1}/H'_{\pi j}$ are isomorphic.

The cyclic group $\mathbb{Z}_n$ is simple iff n is prime. Two composition series of $\mathbb{Z}_{12} = \langle 1\rangle$ are

$$\mathbb{Z}_{12} > \langle 2\rangle > \langle 4\rangle > \langle 12\rangle,$$
$$\mathbb{Z}_{12} > \langle 3\rangle > \langle 6\rangle > \langle 12\rangle,$$

corresponding to composition factors $\mathbb{Z}_2, \mathbb{Z}_2, \mathbb{Z}_3$ and $\mathbb{Z}_3, \mathbb{Z}_2, \mathbb{Z}_2$. This is reminiscent of $2 \cdot 2 \cdot 3 = 3 \cdot 2 \cdot 2$ both being prime factorisations of 12. When all composition factors of a group are cyclic, the group is called *solvable*. The deep *Feit–Thompson Theorem* tells us that any group of odd order is solvable, as are all abelian groups and any group of order < 60 (Question 1.1.2). The name 'solvable' comes from Galois theory (Section 1.7.2).

Finite groups are a massive generalisation of the notion of number. The number n can be identified with the cyclic group $\mathbb{Z}_n$. The divisor of a number corresponds to a normal subgroup, so a prime number corresponds to a simple group. The Jordan–Hölder Theorem generalises the uniqueness of prime factorisations. Building up any number by multiplying primes becomes building up a group by (semi-)direct products and, more generally, by group extensions. Note however that $\mathbb{Z}_6 \times \mathbb{Z}_2$ and $\mathcal{S}_3 \times \mathbb{Z}_2$ – both different from $\mathbb{Z}_{12}$ – also have $\mathbb{Z}_2, \mathbb{Z}_2, \mathbb{Z}_3$ as composition factors. The lesson: unlike for numbers, 'multiplication' here does not give a unique answer. The semi-direct product $\mathbb{Z}_3 \rtimes \mathbb{Z}_2$ can equal either $\mathbb{Z}_6$ or $\mathcal{S}_3$, depending on how the product is taken.

The composition series (1.1.2) tells us that the finite group G is obtained inductively from the trivial group $\{e\}$ by extending $\{e\}$ by the simple group H_k/H_{k+1} to get H_k, then extending H_k by the simple group H_{k-1}/K_k to get H_{k-1}, etc. In other words, any

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

finite group G can be obtained from the trivial group by extending inductively by simple groups; those simple groups are its 'prime factors' = composition factors.

Thus simple groups have an importance for group theory approximating what primes have for number theory. One of the greatest accomplishments of twentieth-century mathematics is the classification of the finite simple groups. Of course we would have preferred the complete finite group classification, but the simple groups are a decent compromise! This work, completed in the early 1980s (although gaps in the arguments are continually being discovered and filled [**22**]), runs to approximately 15 000 journal pages, spread over 500 individual papers, and is the work of a whole generation of group theorists (see [**256**], [**512**] for historical remarks and some ideas of the proof). A modern revision is currently underway to simplify the proof and find and fill all gaps, but the final proof is still expected to be around 4000 pages long. The resulting list, probably complete, is:

- the cyclic groups $\mathbb{Z}_p$ (p a prime);
- the alternating groups $\mathcal{A}_n$ for $n \geq 5$;
- 16 families of Lie type;
- 26 sporadic groups.

The alternating group $\mathcal{A}_n$ consists of the even permutations in the symmetric group $\mathcal{S}_n$, and so has order(=size) $\frac{1}{2}n!$. The groups of Lie type are essentially Lie groups (Section 1.4.2) defined over the finite fields $\mathbb{F}_q$, sometimes 'twisted'. See, for example, chapter I.4 of [**92**] for an elementary treatment. The simplest example is $\mathrm{PSL}_n(\mathbb{F}_q)$, which consists of the $n \times n$ matrices with entries in $\mathbb{F}_q$, with determinant 1, quotiented out by the centre of $\mathrm{SL}_n(\mathbb{F}_q)$ (namely the scalar matrices $\mathrm{diag}(a, a, \ldots, a)$ with $a^n = 1$) ($\mathrm{PSL}_2(\mathbb{Z}_2)$ and $\mathrm{PSL}_2(\mathbb{Z}_3)$ aren't simple so should be excluded). The 'P' here stands for 'projective' and refers to this quotient, while the 'S' stands for 'special' and means determinant 1.

The determinant $\det(\rho(g))$ of any representation ρ (Section 1.1.3) of a noncyclic simple group must be identically 1, and the centre of any noncyclic simple group must be trivial (why?). Hence in the list of simple groups of Lie type are found lots of P's and S's.

The smallest noncyclic simple group is $\mathcal{A}_5$, with order 60. It is isomorphic to $\mathrm{PSL}_2(\mathbb{Z}_5)$ and $\mathrm{PSL}_2(\mathbb{F}_4)$, and can also be interpreted as the group of all rotational symmetries of a regular icosahedron (reflections have determinant -1 and so cannot belong to any simple group $\ncong \mathbb{Z}_2$). The simplicity of $\mathcal{A}_5$ is ultimately responsible for the fact that the zeros of a general quintic polynomial cannot be solved by radicals (see Section 1.7.2).

The smallest sporadic group is the Mathieu group M_{11}, order 7920, discovered in 1861.[2] The largest is the Monster $\mathbb{M}$,[3] conjectured independently by Fischer and Griess

[2] ... although his arguments apparently weren't very convincing. In fact some people, including the Camille Jordan of Jordan–Hölder fame, argued in later papers that the largest of Mathieu's groups, M_{24}, couldn't exist. We now know it does, for example an elegant realisation is as the automorphisms of Steiner system S(5,8,24).

[3] Griess also came up with the symbol for the Monster; Conway came up with the name. It's a little unfortunate (but perhaps inevitable) that the Monster is not named after its codiscoverers, Berndt Fischer and Robert Griess; the name 'Friendly Giant' was proposed in [**263**] as a compromise, but 'Monster' stuck.

in 1973 and finally proved to exist by Griess [**263**] in 1980. Its order is given in (0.2.2). 20 of the 26 sporadic groups are involved in (i.e. are quotients of subgroups of) the Monster, and play some role in Moonshine, as we see throughout Section 7.3. We study the Monster in more detail in Section 7.1.1. Some relations among $\mathbb{M}$, the Leech lattice Λ and the largest Mathieu group M_{24} are given in chapters 10 and 29 of [**113**]. We collect together some of the data of the sporadics in Table 7.1.

This work reduces the construction and classification of all finite groups to understanding the possible extensions by simple groups. Unfortunately, group extensions turn out to be technically quite difficult and lead one into group cohomology.

There are many classifications in mathematics. Most of them look like phone books, and their value is purely pragmatic: for example, as a list of potential counterexamples, and as a way to prove some theorems by exhaustion. And of course obtaining them requires *at least* one paper, and with it some breathing space before those scoundrels on the grant evaluation boards. But when the classification has structure, it can resemble in ways a tourist guide, hinting at new sites to explore. The 18 infinite families in the finite simple group classification are well known and generic, much like the chain of MacDonald's restaurants, useful and interesting in their own ways. But the eye skims over them, and is drawn instead to the 26 sporadic groups and in particular to the largest: the Monster.

1.1.3 Representations

Groups typically arise as 'things that act'. This is their *raison d'être*. For instance, the symmetries of a square form the dihedral group $\mathcal{D}_4$ – that is, the elements of $\mathcal{D}_4$ act on the vertices by permuting them. When a group acts on a structure, you generally want it to preserve the essential features of the structure. In the case of our square, we want adjacent vertices to remain adjacent after being permuted.

So when a group G acts on a vector space V (over $\mathbb{C}$, say), we want it to act 'linearly'. The action $g.v$ of G on V gives V the structure of a *G-module*. In completely equivalent language, it is a *representation* ρ of G on $V \cong \mathbb{C}^n$, that is as a group homomorphism from G to the invertible matrices $\mathrm{GL}_n(\mathbb{C})$. So a representation ρ is a realisation of the group G by matrices, where multiplication in G corresponds to matrix multiplication:

$$\rho(gh) = \rho(g)\,\rho(h).$$

The identification of V with $\mathbb{C}^n$ is achieved by choosing a basis of V, so the module language is 'cleaner' in the sense that it is basis-independent, but this also tends to make it less conducive for practical calculations. The module action $g.v$ is now written $\rho(g)\mathbf{v}$, where $\mathbf{v}$ is the column vector consisting of the components of $v \in V$ with respect to the given basis. If $\rho(g)$ are $n \times n$ matrices, we say ρ is an n-dimensional representation.

For a practise example, consider the symmetric group

$$\mathcal{S}_3 = \{(1), (12), (23), (13), (123), (132)\}. \tag{1.1.3}$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

These cycles multiply as $(13)(123) = (12)$. One representation of $\mathcal{S}_3$ is one-dimensional, and sends all six elements of $\mathcal{S}_3$ to the 1×1 identity matrix:

$$\rho_1(\sigma) = (1), \qquad \forall \sigma \in \mathcal{S}_3.$$

Obviously (1.1.3) is satisfied, and so this defines a representation. But it's trivial, projecting away all structure in the group $\mathcal{S}_3$. Much more interesting is the defining representation ρ_3, which assigns to each $\sigma \in \mathcal{S}_3$ a 3×3 permutation matrix by using σ to permute the rows of the identity matrix I. For example

$$(12) \mapsto \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (13) \mapsto \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad (123) \mapsto \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

This representation is faithful, that is different permutations σ are assigned different matrices $\rho_3(\sigma)$. From this defining representation ρ_3, we get a second one-dimensional one – called the sign representation ρ_s – by taking determinants. For example, $(1) \mapsto (+1)$, $(12) \mapsto (-1)$, $(13) \mapsto (-1)$ and $(123) \mapsto (+1)$.

The most important representation associated with a group G is the *regular representation* given by the group algebra $\mathbb{C}G$. That is, consider the $\|G\|$-dimensional vector space (over $\mathbb{C}$, say) consisting of all formal linear combinations $\sum_{h\in G} \alpha_h h$, where $\alpha_h \in \mathbb{C}$. This has a natural structure of a G-module, given by $g.(\sum \alpha_h h) = \sum_h \alpha_h gh$.

When G is infinite, there will be convergence issues and hence analysis since infinite sums $\sum \alpha_h h$ are involved. The most interesting possibility is to interpret $h \mapsto \alpha_h$ as a $\mathbb{C}$-valued function $\alpha(h)$ on G. Suppose we have a G-invariant measure $\mathrm{d}\mu$ on this space of functions $\alpha : G \to \mathbb{C}$ – this means that the integral $\int_{gU} \alpha(h)\, \mathrm{d}\mu(h)$ will exist and equal $\int_U \alpha(gh)\, \mathrm{d}\mu(h)$ whenever the latter exists. For example, if G is discrete, define '$\int_G \alpha\, \mathrm{d}\mu$' to be $\sum_{h\in G} \alpha(h)$, while if G is the additive group $\mathbb{R}$, $\mathrm{d}\mu(x)$ is the Lebesgue measure (see Section 1.3.1). Looking at the g-coefficient of the product $(\sum_h \alpha_h h)(\sum_k \beta_k k)$, we get the formula $g \mapsto \sum \alpha_h \beta_{h^{-1}g}$, which we recognise as the convolution product (recall $(\alpha * \beta)(x) = \int \alpha(x)\, \beta(x - y)\, \mathrm{d}y$) in, for example, Fourier analysis. In this context, the regular representation of G becomes the Hilbert space $L^2(G)$ of square-integrable functions (i.e. $\int |\alpha|^2 \mathrm{d}\mu < \infty$); the convolution product defines an action of $L^2(G)$ on itself. Note however that the $L^2(\mathbb{R})$-module $L^2(\mathbb{R})$, for a typical example, doesn't restrict to an $\mathbb{R}$-module: the action of $\mathbb{R}$ on $\alpha \in L^2(\mathbb{R})$ by $(x.\alpha)(y) = \alpha(x + y)$ corresponds to the convolution product of α with the 'Dirac delta' distribution δ centred at x. We return to $L^2(G)$ in Section 1.5.5.

Two representations ρ, ρ' are called *equivalent* if they differ merely by a change of coordinate axes (basis) in the ambient space $\mathbb{C}^n$, that is if there exists a matrix U such that $\rho'(g) = U\rho(g)U^{-1}$ for all g. The *direct sum* $\rho' \oplus \rho''$ of representations is given by

$$(\rho' \oplus \rho'')(g) = \begin{pmatrix} \rho'(g) & 0 \\ 0 & \rho''(g) \end{pmatrix}. \tag{1.1.4a}$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

The tensor product $\rho' \otimes \rho''$ of representations is given by $(\rho' \otimes \rho'')(g) = \rho'(g) \otimes \rho''(g)$, where the Kronecker product $A \otimes B$ of matrices is defined by the following block form:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots \\ a_{21}B & a_{22}B & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{1.1.4b}$$

The *contragredient* or *dual* ρ^* of a representation is given by the formula

$$\rho^*(g) = (\rho(g^{-1}))^t, \tag{1.1.4c}$$

so called because it's the natural representation on the space V^* dual to the space V on which ρ is defined. For any finite group representation defined over a subfield of $\mathbb{C}$, the dual ρ^* is equivalent to the complex conjugate representation $g \mapsto \overline{\rho(g)}$.

Returning to our $\mathcal{S}_3$ example, the given matrices for ρ_3 were obtained by having $\mathcal{S}_3$ act on coordinates with respect to the standard basis $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. If instead we choose the basis $\{(1, 1, 1), (1, -1, 0), (0, 1, -1)\}$, these matrices become

$$(12) \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad (13) \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \quad (123) \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}.$$

It is manifest here that ρ_3 is a direct sum of ρ_1 (the upper-left 1×1 block) with a two-dimensional representation ρ_2 (the lower-right 2×2 block) given by

$$(12) \mapsto \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}, \quad (13) \mapsto \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \quad (123) \mapsto \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

An *irreducible* or *simple* module is a module that contains no nontrivial submodule. 'Submodule' plays the role of divisor here, and 'irreducible' the role of prime number. A module is called *completely reducible* if it is the direct sum of finitely many irreducible modules. For example, the $\mathcal{S}_3$ representations ρ_1, ρ_s and ρ_2 are irreducible, while $\rho_3 \cong \rho_1 \oplus \rho_2$ is completely reducible.

A representation is called *unitary* if it is equivalent to one whose matrices $\rho(g)$ are all unitary (i.e. their inverses equal their complex conjugate transposes). A more basis-independent definition is that a G-module V is unitary if there exists a Hermitian form $\langle u, v\rangle \in \mathbb{C}$ on V such that

$$\langle g.u, g.v\rangle = \langle u, v\rangle.$$

By definition, a *Hermitian form* $\langle u, v\rangle : V \times V \to \mathbb{C}$ is linear in v and anti-linear in u, i.e.

$$\langle au + a'u', bv + b'v'\rangle = \overline{a}b\langle u, v\rangle + \overline{a}b'\langle u, v'\rangle + \overline{a'}b\langle u', v\rangle + \overline{a'}b'\langle u', v'\rangle,$$

for all $a, a', b, b' \in \mathbb{C}$, $u, u', v, v' \in V$, and finally $\langle u, u\rangle > 0$ for all nonzero $u \in V$. When V is finite-dimensional, a basis can always be found in which its Hermitian form looks like $\langle x, y\rangle = \sum_i \overline{x}_i y_i$. Most representations of interest in quantum physics are unitary. Unitary representations are much better behaved than non-unitary ones.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

For instance, an easy argument shows that finite-dimensional unitary representation is completely reducible.

An *indecomposable* module is one that isn't the direct sum of smaller ones. An indecomposable module may be *reducible*: its matrices could be put into the form

$$\rho(g) = \begin{pmatrix} A(g) & B(g) \\ 0 & D(g) \end{pmatrix},$$

where for some g the submatrix $B(g)$ isn't the 0-matrix (otherwise we would recover (1.1.4a)). Then $A(g)$ is a subrepresentation, but $D(g)$ isn't. For finite groups, however, irreducible=indecomposable:

Theorem 1.1.2 (Burnside, 1904) *Let G be finite and the field be $\mathbb{C}$. Any G-module is unitary and will be completely reducible if it is finite-dimensional. There are only finitely many irreducible G-modules; their number equals the number of conjugacy classes of G.*

The *conjugacy classes* are the sets $K_g = \{h^{-1}gh \mid h \in G\}$. This fundamental result fails for infinite groups. For example, take G to be the additive group $\mathbb{Z}$ of integers. Then there are uncountably many one-dimensional representations of G, and there are representations that are reducible but indecomposable (see Question 1.1.6(a)). Theorem 1.1.2 is proved using a projection defined by certain averaging over G, as well as:

Lemma 1.1.3 (Schur's Lemma) *Let G be finite and ρ, ρ' be representations.*

(a) *ρ is irreducible iff the only matrices A commuting with all matrices $\rho(g)$, $g \in G$ – that is $A\rho(g) = \rho(g)A$ – are of the form $A = aI$ for $a \in \mathbb{C}$, where I is the identity matrix.*

(b) *Suppose both ρ and ρ' are irreducible. Then ρ and ρ' are isomorphic iff there is a nonzero matrix A such that $A\rho(g) = \rho'(g)A$ for all $g \in G$.*

Schur's Lemma is an elementary observation central to representation theory. It's proved by noting that the kernel (nullspace) and range (column space) of A are G-invariant.

The character[4] ch_ρ of a representation ρ is the map $G \to \mathbb{C}$ given by the trace:

$$\mathrm{ch}_\rho(g) = \mathrm{tr}\,(\rho(g)). \tag{1.1.5}$$

We see that equivalent representations have the same character, because of the fundamental identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Remarkably, for finite groups (and $\mathbb{C}$), the converse is also true: inequivalent representations have different character. That trace identity also tells us that the character is a 'class function', i.e. $\mathrm{ch}_\rho(hgh^{-1}) = \mathrm{tr}(\rho(h)\,\rho(g)\,\rho(h)^{-1}) = \mathrm{ch}_\rho(g)$ so ch_ρ is constant on each conjugacy class K_g. Group characters are enormously simpler than representations: for example, the smallest nontrivial representation of the Monster

[4] Surprisingly, characters were invented before group representations, by Frobenius in 1868. He defined characters indirectly, by writing the 'class sums' C_j in terms of the idempotents of the centre of the group algebra. It took him a year to realise they could be reinterpreted as the traces of matrices.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Table 1.2. *The character table of* $\mathcal{S}_3$

$\mathrm{ch}\backslash\sigma$	(1)	(12)	(123)
ch_1	1	1	1
ch_s	1	−1	1
ch_2	2	0	−1

$\mathbb{M}$ consists of about 10^{54} matrices, each of size $196\,883 \times 196\,883$, while its character consists of 194 complex numbers. The reason is that the representation matrices have a lot of redundant, basis-dependent information, to which the character is happily oblivious.

The Thompson trick mentioned in Section 0.3 tells us: *A dimension can (and should) be twisted; that twist is called a character.* Indeed, $\mathrm{ch}_\rho(e) = \dim(\rho)$, where the dimension of ρ is defined to be the dimension of the underlying vector space V, or the size n of the $n \times n$ matrices $\rho(g)$. When we see a positive integer, we should try to interpret it as a dimension of a vector space; if there is a symmetry present, then it probably acts on the space, in which case we should see what significance the other character values may have.

Algebra searches for structure. What can we say about the set of characters? First, note directly from (1.1.4) that we can add and multiply characters:

$$\mathrm{ch}_{\rho\oplus\rho'}(g) = \mathrm{ch}_\rho(g) + \mathrm{ch}_{\rho'}(g), \tag{1.1.6a}$$

$$\mathrm{ch}_{\rho\otimes\rho'}(g) = \mathrm{ch}_\rho(g)\,\mathrm{ch}_{\rho'}(g), \tag{1.1.6b}$$

$$\mathrm{ch}_{\rho^*}(g) = \overline{\mathrm{ch}_\rho(g)}. \tag{1.1.6c}$$

Therefore the complex span of the characters forms a (commutative associative) *algebra*. For G finite (and the field algebraically closed), each matrix $\rho(g)$ is separately diagonalisable, with eigenvalues that are roots of 1 (why?). This means that each character value $\mathrm{ch}_\rho(g)$ is a sum of roots of 1.

By the *character table* of a group G we mean the array with rows indexed by the characters ch_ρ of irreducible representations, and the columns by conjugacy classes K_g, and with entries $\mathrm{ch}_\rho(g)$. An example is given in Table 1.2. Different groups can have identical character tables: for instance, for any n, the dihedral group $\mathcal{D}_{4n}$ has the same character table as the quaternionic group $\mathcal{Q}_{4n}$ defined by the presentation

$$\mathcal{Q}_{4n} = \langle a, b \mid a^2 = b^{2n},\ abab = e \rangle. \tag{1.1.7}$$

In spite of this, the characters of a group G tell us much about G – for example, its order, all of its normal subgroups, whether or not it's simple, whether or not it's solvable ... In fact, the character table of a finite simple group determines the group uniquely [**100**] (its order alone usually distinguishes it from other simple groups). This suggests:

Problem *Suppose G and H have identical character tables (up to appropriate permutations of rows and columns). Must they have the same composition factors?*

After all, the answer is certainly yes for solvable G (why?).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

It may seem that 'trace' is a fairly arbitrary operation to perform on the matrices $\rho(g)$ – certainly there are other invariants we can attach to a representation ρ so that equivalent representations are assigned equal numbers. For example, how about $g \mapsto \det\rho(g)$? This is too limited, because it is a group homomorphism (e.g. what happens when G is simple?). But more generally, choose an independent variable x_g for each element $g \in G$, and for any representation ρ of G define the group determinant of ρ

$$\Theta_\rho = \det\left(\sum_{g\in G} x_g\, \rho(g)\right).$$

This is a multivariable polynomial Θ_ρ, homogeneous of degree $n = \dim(\rho)$. The character $\mathrm{ch}_\rho(g)$ can be obtained from the group determinant Θ_ρ: it is the coefficient of the $x_g x_e^{n-1}$ term. In fact, the group G is uniquely determined by the group determinant of the regular representation $\mathbb{C}G$. See the review article [**315**].

One use of characters is to identify representations. For this purpose the orthogonality relations are crucial: given any characters ch, ch′ of G, define the Hermitian form

$$\langle \mathrm{ch}, \mathrm{ch}'\rangle = \frac{1}{\|G\|}\sum_{g\in G} \mathrm{ch}(g)\overline{\mathrm{ch}'(g)}. \tag{1.1.8a}$$

Write ρ_i for the irreducible representations, and ch_i for the corresponding traces. Then

$$\langle \mathrm{ch}_i, \mathrm{ch}_j\rangle = \delta_{ij}, \tag{1.1.8b}$$

that is the irreducible characters ch_i are an orthonormal basis with respect to (1.1.8a). If the rows of a matrix are orthonormal, so are the columns. Hence (1.1.8b) implies

$$\sum_i \mathrm{ch}_i(g)\,\overline{\mathrm{ch}_i(h)} = \frac{\|G\|}{\|K_g\|}\,\delta_{K_g,K_h}. \tag{1.1.8c}$$

The decomposition of $\mathbb{C}G$ into irreducibles is now immediate:

$$\mathbb{C}G \cong \bigoplus_i (\dim\rho_i)\,\rho_i,$$

that is each irreducible representation appears with multiplicity given by its dimension. Taking the dimension of both sides, we obtain the useful identity

$$\|G\| = \sum_i (\dim\rho_i)^2.$$

The notion of vector space and representation can be defined over any field $\mathbb{K}$. One thing that makes representations over, for example, the finite field $\mathbb{K} = \mathbb{Z}_p$ much more difficult is that characters no longer distinguish inequivalent representations. For instance, take $G = \{e\}$ and consider the representations

$$\rho(1) = (1) \qquad \text{and} \qquad \rho'(1) = \begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}.$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

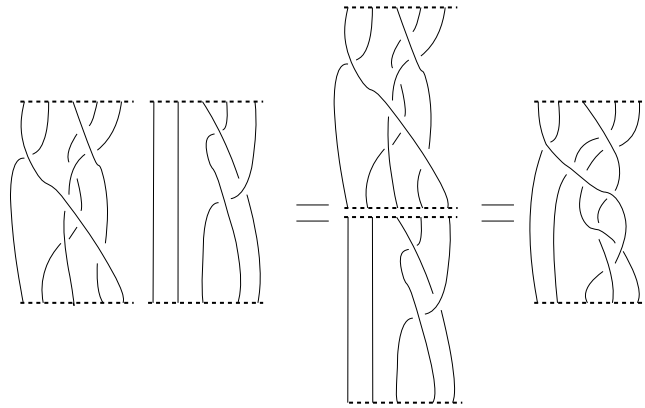

Fig. 1.1 Multiplication in the braid group $\mathcal{B}_5$.

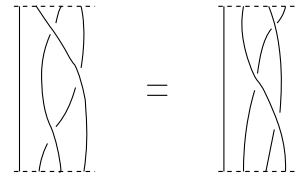

Fig. 1.2 The relation $\sigma_2\sigma_3\sigma_2 = \sigma_3\sigma_2\sigma_3$ in $\mathcal{B}_4$.

These are certainly different representations – their dimensions are different. But over the field $\mathbb{Z}_2$, their characters ch and ch′ are identical. Theorem 1.1.2 also breaks down here. Unless otherwise stated, in this book we restrict to characteristic 0 (but see *modular Moonshine* in Section 7.3.5).

1.1.4 Braided #1: the braid groups

Fundamental to us are the braid groups, especially $\mathcal{B}_3$. By an *n-braid* we mean n non-intersecting strands as in Figure 1.1. We are interested here in how the strands interweave, and not how they knot, and so we won't allow the strands to double-back on themselves. We regard two n-braids as equivalent if they can be deformed continuously into each other – we make this notion more precise in Section 1.2.3. The set of equivalence classes of n-braids forms a group, called the *braid group* $\mathcal{B}_n$, with multiplication given by vertical concatenation, as in Figure 1.1.

Artin (1925) gives a very useful presentation of $\mathcal{B}_n$:

$$\mathcal{B}_n = \langle \sigma_1, \ldots, \sigma_{n-1} \,|\, \sigma_i\sigma_j = \sigma_j\sigma_i,\ \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1},\ \text{whenever } |i-j| \geq 2 \rangle. \tag{1.1.9}$$

Here σ_i denotes the braid obtained from the identity braid by interchanging the ith and $(i+1)$th strands, with the ith strand on top. See Figure 1.2 for an illustration.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Of course $\mathcal{B}_1$ is trivial and $\mathcal{B}_2 \cong \mathbb{Z}$, but the other $\mathcal{B}_n$ are quite interesting. Any non-trivial element in $\mathcal{B}_n$ has infinite order. Let $\sigma = \sigma_1\sigma_2\cdots\sigma_{n-1}$; then $\sigma\sigma_i = \sigma_{i+1}\sigma$, so the generators are all conjugate and the braid $Z = \sigma^n$ lies in the centre of $\mathcal{B}_n$. In fact, for $n \geq 2$ the centre $Z(\mathcal{B}_n) \cong \mathbb{Z}$, and is generated by that braid Z. We're most interested in $\mathcal{B}_3$: then $Z = (\sigma_1\sigma_2)^3$ generates the centre, and we will see shortly that

$$\mathcal{B}_3/\langle Z^2\rangle \cong \mathrm{SL}_2(\mathbb{Z}), \tag{1.1.10a}$$

$$\mathcal{B}_3/\langle Z\rangle \cong \mathrm{PSL}_2(\mathbb{Z}). \tag{1.1.10b}$$

There is a surjective homomorphism $\phi : \mathcal{B}_n \to \mathcal{S}_n$ taking a braid α to the permutation $\phi(\alpha) \in \mathcal{S}_n$, where the strand of α starting at position i on the top ends on the bottom at position $\phi(\alpha)(i)$. For example, $\phi(\sigma_i)$ is the transposition $(i, i+1)$. The kernel of ϕ is called the *pure braid group* $\mathcal{P}_n$. A presentation for $\mathcal{P}_n$ is given in lemma 1.8.2 of [**59**]. We find that $\mathcal{P}_2 = \langle\sigma_1^2\rangle \cong \mathbb{Z}$ and

$$\mathcal{P}_3 = \left\langle \sigma_1^2, \sigma_2^2, Z \right\rangle \cong \mathcal{F}_2 \times \mathbb{Z}. \tag{1.1.10c}$$

Another obvious homomorphism is the degree map $\deg : \mathcal{B}_n \to \mathbb{Z}$, defined by $\deg(\sigma_i^{\pm 1}) = \pm 1$. It is easy to show using (1.1.9) that 'deg' is well defined and is the number of signed crossings in the braid. Its kernel is the commutator subgroup $[\mathcal{B}_n\mathcal{B}_n]$ (see Question 1.1.7(a)).

The most important realisation of the braid group is as a fundamental group (see (1.2.6)). It is directly through this that most appearances of $\mathcal{B}_n$ in Moonshine-like phenomena arise (e.g. Jones' braid group representations from subfactors, or Kohno's from the monodromy of the Knizhnik–Zamolodchikov equation).

The relation of $\mathcal{B}_3$ to modularity in Moonshine, however, seems more directly to involve the faithful action of $\mathcal{B}_n$ on the free group $\mathcal{F}_n = \langle x_1, \ldots, x_n\rangle$ (see Question 6.3.5). This action allows us to regard $\mathcal{B}_n$ as a subgroup of $\mathrm{Aut}\,\mathcal{F}_n$.

As is typical for infinite discrete groups, $\mathcal{B}_n$ has continua of representations. For instance, there is a different one-dimensional for every choice of nonzero complex number $w \neq 0$, namely $\alpha \mapsto w^{\deg\alpha}$. It seems reasonable to collect these together and regard them as different specialisations of a single one-dimensional $\mathbb{C}[w^{\pm 1}]$-representation, which we could call $w^{\deg}$, where $\mathbb{C}[w^{\pm 1}]$ is the (Laurent) polynomial algebra in w and w^{-1}.

The *Burau representation* (Burau, 1936) of $\mathcal{B}_n$ is an n-dimensional representation with entries in the Laurent polynomials $\mathbb{C}[w^{\pm 1}]$, and is generated by the matrices

$$\sigma_i \mapsto I_{i-1} \oplus \begin{pmatrix} 1-w & w \\ 1 & 0 \end{pmatrix} \oplus I_{n-i-1}, \tag{1.1.11a}$$

where I_k here denotes the $k \times k$ identity matrix. $\mathbb{C}[w^{\pm 1}]$ isn't a field, but checking determinants confirms that all matrices $\rho(\sigma_i)$ are invertible over it. The Burau representation is reducible – in particular the column vector $v = (1, 1, \ldots, 1)^t$ is an eigenvector with eigenvalue 1, for all the matrices in (1.1.11a), and hence $\mathcal{B}_n$ acts trivially on the subspace $\mathbb{C}v$. The remaining $(n-1)$-dimensional representation is the *reduced Burau*

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

representation. For example, for $\mathcal{B}_3$ it is

$$\sigma_1 \mapsto \begin{pmatrix} -w & 1 \\ 0 & 1 \end{pmatrix}, \qquad \sigma_2 \mapsto \begin{pmatrix} 1 & 0 \\ w & -w \end{pmatrix}, \tag{1.1.11b}$$

and so the centre-generator Z maps to the scalar matrix $w^3 I$. Note that the specialisation $w = -1$ has image $\mathrm{SL}_2(\mathbb{Z})$ – in fact it gives the isomorphism (1.1.10a) – while $w = 1$ has image $\mathcal{S}_3$ and is the representation ρ_2.

There are many natural ways to obtain the representation (1.1.11a). The simplest uses derivatives $\frac{\partial}{\partial x_i}$ acting in the obvious way on the group algebra $\mathbb{C}\mathcal{F}_n$. To any n-braid $\alpha \in \mathcal{B}_n$ define the $n \times n$ matrix whose (i, j)-entry is given by

$$w^{\deg} \frac{\partial}{\partial x_j}(\alpha . x_i),$$

where $\alpha . x_i$ denotes the action of $\mathcal{B}_n$ on $\mathcal{F}_n$ and where $w^{\deg}$ is the obvious representation of $\mathcal{F}_n$, extended linearly to $\mathbb{C}\mathcal{F}_n$. Then this recovers (1.1.11a).

All irreducible representations of $\mathcal{B}_3$ in dimension ≤ 5 are found in [**531**]. Most are non-unitary. For example, any two-dimensional irreducible representation is of the form

$$\sigma_1 \mapsto \begin{pmatrix} \lambda_1 & \lambda_2 \\ 0 & \lambda_2 \end{pmatrix}, \qquad \sigma_2 \mapsto \begin{pmatrix} \lambda_2 & 0 \\ -\lambda_1 & \lambda_1 \end{pmatrix},$$

for some nonzero complex numbers λ_1, λ_2 (compare (1.1.11b)). This representation will be unitary iff both $|\lambda_1| = |\lambda_2| = 1$ and $\lambda_1/\lambda_2 = e^{it}$ for $\pi/3 < t < 5\pi/3$. Not all representations of $\mathcal{B}_3$ are completely reducible, however (Question 1.1.9).

Question 1.1.1. Identify the group $\mathrm{PSL}_2(\mathbb{Z}_2)$ and confirm that it isn't simple.

Question 1.1.2. If G and H are any two groups with $\|G\| = \|H\| < 60$, explain why they will have the same composition factors.

Question 1.1.3. Verify that the dihedral group $\mathcal{D}_n$ (1.1.1) has order $2n$. Find its composition factors. Construct $\mathcal{D}_n$ as a semi-direct product of $\mathbb{Z}_2$ and $\mathbb{Z}_n$.

Question 1.1.4. (a) Using the methods and results given in Section 1.1.3, compute the character table of the symmetric group $\mathcal{S}_4$.
(b) Compute the tensor product coefficients of $\mathcal{S}_4$. That is, if $\rho_1, \rho_2, \ldots$ are the irreducible representations of $\mathcal{S}_4$, compute the multiplicities T_{ij}^k defined by

$$\rho_i \otimes \rho_j \cong \oplus_k T_{ij}^k \rho_k.$$

Question 1.1.5. Prove that $\mathrm{ch}(g^{-1}) = \overline{\mathrm{ch}(g)}$. Can you say anything about the relation of $\mathrm{ch}(g^\ell)$ and $\mathrm{ch}(g)$, for other integers ℓ?

Question 1.1.6. (a) Find a representation over the field $\mathbb{C}$ of the additive group $G = \mathbb{Z}$, which is indecomposable but not irreducible. Hence show that inequivalent (complex) finite-dimensional representations of $\mathbb{Z}$ can have identical characters.
(b) Let p be any prime dividing some $n \in \mathbb{N}$. Find a representation of the cyclic group $G = \mathbb{Z}_n$ over the field $\mathbb{K} = \mathbb{Z}_p$, which is indecomposable but not irreducible.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Question 1.1.7. (a) Let G be any group, and define the *commutator subgroup* $[G, G]$ to be the subgroup generated by the elements $ghg^{-1}h^{-1}$, for all $g, h \in G$. Prove that $[G, G]$ is a normal subgroup of G, and that $G/[G, G]$ is abelian. (In fact, $G/[G, G]$ is isomorphic to the group of all one-dimensional representations of G.)
(b) Show that the free groups $\mathcal{F}_n \cong \mathcal{F}_m$ iff $n = m$, by using Theorem 1.1.2.

Question 1.1.8. (a) Explicitly show how the semi-direct product $\mathbb{Z}_3 \rtimes_\theta \mathbb{Z}_2$ can equal $\mathbb{Z}_6$ or $\mathcal{S}_3$, depending on the choice of θ.
(b) Show that $\mathbb{Z}_2 \rtimes_\theta H \cong \mathbb{Z}_2 \times H$, for any group H and homomorphism θ.
(c) Hence $\mathbb{Z}_4$ can't be written as a semi-direct product of $\mathbb{Z}_2$ with $\mathbb{Z}_2$. Explicitly construct it as an external group extension of $\mathbb{Z}_2$ by $\mathbb{Z}_2$.

Question 1.1.9. Find a two-dimensional representation of the braid group $\mathcal{B}_3$ that is not completely reducible.

1.2 Elementary geometry

Geometry and algebra are opposites. We inherited from our mammalian ancestors our subconscious facility with geometry; to us geometry is intuitive and has implicit meaning, but because of this it's harder to generalise beyond straightforward extensions of our visual experience, and rigour tends to be more elusive than with algebra. The power and clarity of algebra comes from the conceptual simplifications that arise when content is stripped away. But this is equally responsible for algebra's blindness. Although recently physics has inspired some spectacular developments in algebra, traditionally geometry has been the most reliable star algebraists have been guided by. We touch on geometry throughout this book, though for us it adds more colour than essential substance.

1.2.1 Lattices

Many words in mathematics have multiple meanings. For example, there are vector *fields* and number *fields*, and *modular* forms and *modular* representations. 'Lattice' is another of these words: it can mean a 'partially ordered set', but to us a lattice is a discrete maximally periodic set – a toy model for everything that follows.

Consider the real vector space $\mathbb{R}^{m,n}$: its vectors look like $x = (x_+; x_-)$, where x_+ and x_- are m- and n-component vectors, respectively, and inner-products are given by $x \cdot y = x_+ \cdot y_+ - x_- \cdot y_-$. The inner-products $x_\pm \cdot y_\pm$ are given by the usual $\sum_i (x_\pm)_i (y_\pm)_i$. For example, the familiar Euclidean (positive-definite) space is $\mathbb{R}^n = \mathbb{R}^{n,0}$, while the Minkowski space-time of special relativity is $\mathbb{R}^{3,1}$.

Now choose any basis $\beta = \{b^{(1)}, \ldots, b^{(m+n)}\}$ in $\mathbb{R}^{m,n}$. If we consider all possible linear combinations $\sum_i a_i b^{(i)}$ over the real numbers $\mathbb{R}$, then we recover $\mathbb{R}^{m,n}$; if instead we consider linear combinations over the integers only, we get a *lattice*.

Definition 1.2.1 *Let V be any n-dimensional inner-product space, and let $\{b^{(1)}, \ldots, b^{(n)}\}$ be any basis. Then $L(\beta) := \mathbb{Z}b^{(1)} + \cdots + \mathbb{Z}b^{(n)}$ is called a* lattice.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

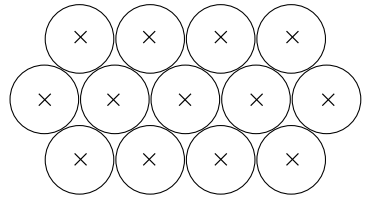

Fig. 1.3 Part of the A_2 disc packing.

A lattice is discrete and closed under sums and integer multiples. For example, $\mathbb{Z}^{m,n}$ is a lattice (take the standard basis in $\mathbb{R}^{m,n}$). A more interesting lattice is the hexagonal lattice (also called A_2), given by the basis $\beta = \{(\frac{\sqrt{2}}{2}, \frac{\sqrt{6}}{2}), (\sqrt{2}, 0)\}$ of $\mathbb{R}^2$ – try to plot several points. If you wanted to slide a bunch of identical coins on a table together as tightly as possible, their centres would form the hexagonal lattice (Figure 1.3). Another important lattice is $II_{1,1} \subset \mathbb{R}^{1,1}$, given by $\beta = \{(\frac{1}{\sqrt{2}}; \frac{1}{\sqrt{2}}), (\frac{1}{\sqrt{2}}; \frac{-1}{\sqrt{2}})\}$; equivalently, it can be thought of as the set of all pairs $(a, b) \in \mathbb{Z}^2$ with inner-product

$$(a, b) \cdot (c, d) = ad + bc. \tag{1.2.1}$$

Different bases may or may not result in a different lattice. For a trivial example, consider $\beta = \{1\}$ and $\beta' = \{-1\}$ in $\mathbb{R} = \mathbb{R}^{1,0}$: they both give the lattice $\mathbb{Z} = \mathbb{Z}^{1,0}$. Two lattices $L(\beta) \subset V$ and $L(\beta') \subset V'$ are called *equivalent* or *isomorphic* if there is an orthogonal transformation $T : V \to V'$ such that the lattices $T(L(\beta))$ and $L(\beta')$ are identical as sets, or equivalently if $b'_i = T \sum_j c_{ij} b_j$, for some integer matrix $C = (c_{ij}) \in \mathrm{GL}_n(\mathbb{Z})$ with determinant ± 1.

This notion of lattice equivalence is important in that it emphasises the essential properties of a lattice and washes away the unpleasant basis-dependence of Definition 1.2.1. In particular, the ambient space V in which the lattice lives, and the basis β, are non-essential. The transformation T tells us we can change V, and C is a change-of-basis matrix for which both C and C^{-1} are defined over $\mathbb{Z}$.

For example, $\beta = \{(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})\}$ in $\mathbb{R}^2$ yields a lattice equivalent to $\mathbb{Z}^2$. The basis $\beta' = \{(-1, 1, 0), (0, -1, 1)\}$ for the plane $a + b + c = 0$ in $\mathbb{R}^3$ yields the lattice $L(\beta') = \{(a, b, c) \in \mathbb{Z}^3 \mid a + b + c = 0\}$, equivalent to the hexagonal lattice A_2.

The *dimension* of the lattice is the dimension $\dim(V)$ of the ambient vector space. The lattice is called *positive-definite* if it lies in some $\mathbb{R}^m$ (i.e. $n = 0$), and *integral* if all inner-products $x \cdot y$ are integers, for $x, y \in L$. A lattice L is called *even* if it is integral and in addition all norm-squareds $x \cdot x$ are *even* integers. For example, $\mathbb{Z}^{m,n}$ is integral but not even, while A_2 and $II_{1,1}$ are even. The *dual* L^* of a lattice L consists of all vectors $x \in V$ such that $x \cdot L \subset \mathbb{Z}$. A natural basis for the dual $L(\beta)^*$ is the *dual basis* β^*, consisting of the vectors $c_j \in V$ obeying $b_i \cdot c_j = \delta_{ij}$ for all i, j. A lattice is integral iff $L \subseteq L^*$. A lattice is called *self-dual* if $L = L^*$. The lattices $\mathbb{Z}^{m,n}$ and $II_{1,1}$ are self-dual, but A_2 is not. We are most interested in even positive-definite lattices.

To any n-dimensional lattice $L(\beta)$, define an $n \times n$ matrix A (called a *Gram matrix*) by $A_{ij} = b_i \cdot b_j$. Two lattices with identical Gram matrices are necessarily equivalent, but the converse is not true. Note that the Gram matrix of $L(\beta^*)$ is the inverse of the Gram

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

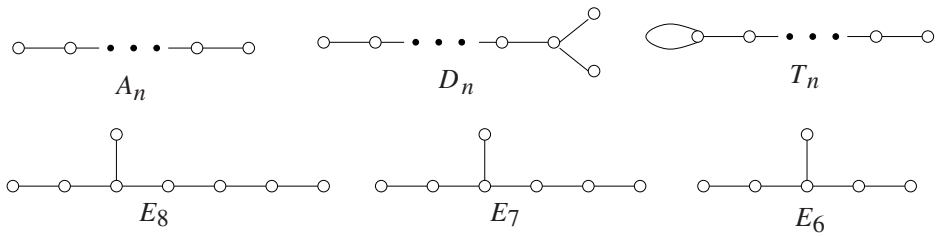

Fig. 1.4 The graphs with largest eigenvalue < 2.

matrix for $L(\beta)$. The *determinant* $|L|$ of a lattice is the determinant of the Gram matrix; geometrically, it is the volume-squared of the fundamental parallelepiped of L defined by the basis. This will always be positive if L is positive-definite. The determinant of a lattice is independent of the specific basis β chosen; equivalent lattices have equal determinant, though the converse isn't true. An integral lattice L is self-dual iff $|L| = \pm 1$. If $L' \subseteq L$ are of equal dimension, then the quotient L/L' is a finite abelian group of order

$$\|L/L'\| = \sqrt{|L'|/|L|} \in \mathbb{N}. \tag{1.2.2}$$

Given two lattices L, L', their (orthogonal) *direct sum* $L \oplus L'$ is defined to consist of all pairs (x, x'), for $x \in L, x' \in L'$, with inner-product defined by $(x, x') \cdot (y, y') = x \cdot y + x' \cdot y'$. The dimension of $L \oplus L'$ is the sum of the dimensions of L and L'. The direct sum $L \oplus L'$ will be integral (respectively self-dual) iff both L and L' are integral (respectively self-dual).

An important class of lattices are the so-called *root lattices* A_n, D_n, E_6, E_7, E_8 associated with simple Lie algebras (Section 1.5.2). They can be defined from the graph ('Coxeter–Dynkin diagram') in Figure 1.4 (but ignore the 'tadpole' T_n for now): label the nodes of such a graph from 1 to n, put $A_{ii} = 2$ and put $A_{ij} = -1$ if nodes i and j are connected by an edge. Then this matrix A (the *Cartan matrix* of Definition 1.4.5) is the Gram matrix of a positive-definite integral lattice. Realisations of some of these are given shortly; bases can be found in table VII of [**214**], or planches I–VII of [**84**]. Of these, E_8 is the most interesting as it is the even self-dual positive-definite lattice of smallest dimension.

The following theorem characterises norm-squared 1,2 vectors.

Theorem 1.2.2 *Let L be an n-dimensional positive-definite integral lattice.*
(a) *Then L is equivalent to the direct sum $\mathbb{Z}^m \oplus L'$, where L has precisely $2m$ unit vectors and L' has none.*
(b) *If L is spanned by its norm-squared 2 vectors, then L is a direct sum of root lattices.*

Theorem 1.2.2(b) gives the point-of-contact of Lie theory and lattices. The densest packing of circles in the plane (Figure 1.3) is A_2, in the sense that the centres of these circles are the points of A_2. The obvious pyramidal way to pack oranges is also the densest, and likewise gives the A_3 root lattice. The densest known sphere packings in dimensions 4, 5, 6, 7, 8 are the root lattices D_4, D_5, E_6, E_7, E_8, respectively.

The Leech lattice Λ is one of the most distinguished lattices, and like E_8 is directly related to Moonshine. It can be constructed using 'laminated lattices'

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

([**113**], chapter 6). Start with the zero-dimensional lattice $L_0 = \{0\}$, consisting of just one point. Use it to construct a one-dimensional lattice L_1, with minimal (nonzero) norm 2, built out of infinitely many copies of L_0 laid side by side. The result of course is simply the even integers $2\mathbb{Z}$. Now construct a two-dimensional lattice L_2, of minimal norm 2, built out of infinitely many copies of L_1 stacked next to each other. There are lots of ways to do this, but choose the densest lattice possible. The result is the hexagonal lattice A_2 rescaled by a factor of $\sqrt{2}$. Continue in this way: L_3, L_4, L_5, L_6, L_7 and L_8 are the root lattices A_3, D_4, D_5, E_6, E_7 and E_8, respectively, all rescaled by $\sqrt{2}$.

The 24th repetition of this construction yields uniquely the Leech lattice $\Lambda = L_{24}$. It is the unique 24-dimensional even self-dual lattice with no norm-squared 2-vectors, and provides among other things the densest known packing of 23-dimensional spheres S^{23} in $\mathbb{R}^{24}$. It is studied throughout [**113**]. After dimension 24, chaos reigns in lamination (23 different 25-dimensional lattices have an equal right to be called L_{25}, and over 75 000 are expected for L_{26}). So lamination provides us with a sort of no-input construction of the Leech lattice. Like the Mandelbrot set, the Leech lattice is a subtle structure with an elegant construction – a good example of the mathematical meaning of 'natural'.

Question 1.2.1 asks you to come up with a definition for the automorphism group of a lattice. An automorphism is a symmetry, mapping the lattice to itself, preserving all essential lattice properties. It is how group theory impinges on lattice theory.

Most (positive-definite) lattices have trivial automorphism groups, consisting only of the identity and the reflection $x \mapsto -x$ through the origin. But the more interesting lattices tend to have quite large groups. The reflection through the hyperplane orthogonal to a norm-squared 2-vector in an integral lattice defines an automorphism; together, these automorphisms form what Lie theory calls a Weyl group.

Typically the Weyl group has small index in the full automorphism group, though a famous counterexample is the Leech lattice (which, as we know, has trivial Weyl group). Its automorphism group is denoted Co_0 and has approximately 8×10^{18} elements. The automorphism $x \mapsto -x$ lies in its centre; if we quotient by this 2-element centre we get a sporadic simple group Co_1. Define Co_2 and Co_3 to be the subgroups of Co_0 consisting of all $g \in Co_0$ fixing some norm-squared 4-vector and some norm-squared 6-vector, respectively. These three groups Co_1, Co_2, Co_3 are all simple. In fact, a total of 12 sporadic finite simple groups appear as subquotients in Co_0, and can best be studied geometrically in this context. Gorenstein [**256**] wrote:

> ... if Conway had studied the Leech lattice some 5 years earlier, he would have discovered a total of 7 new simple groups! Unfortunately he had to settle for 3. However, as consolation, his paper on .0[=Co_0] will stand as one of the most elegant achievements of mathematics.

1.2.2 Manifolds

On what structures do lattices act naturally? An obvious place is on their ambient space ($\mathbb{R}^n$, say). They act by addition. Quotient out by this action. Topologically, we have

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

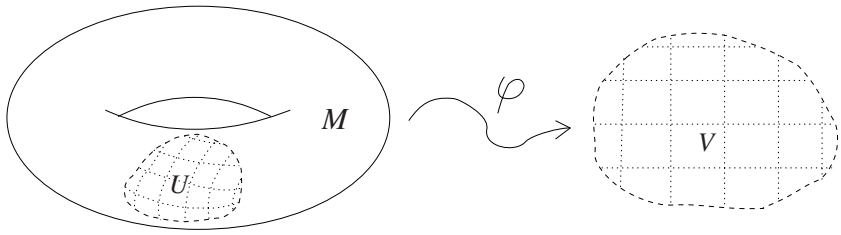

Fig. 1.5 A coordinate patch.

created a *manifold* (to be defined shortly); to each point on this manifold corresponds an orbit in $\mathbb{R}^n$ of our lattice action.

Consider first the simplest case. The number $n \in \mathbb{Z}$ acts on $\mathbb{R}$ by sending $x \in \mathbb{R}$ to $x + n$. The orbits are the equivalence classes of the reals mod 1. We can take as representatives of these equivalence classes, that is as points of $\mathbb{R}/\mathbb{Z}$, the half-open interval $[0, 1)$. This orbit space inherits a *topology* (i.e. a qualitative notion of points being close; for basic point-set topology see e.g. [**481**], [**104**]), from that of $\mathbb{R}$, and this is almost captured by the interval $[0, 1)$. The only problem is that the orbit of $0.999 \equiv -0.0001$ is pretty close to that of 0, even though they are at opposite ends of the interval. What we should do is identify the two ends, i.e. glue together 0 and 1. The result is a circle.

We say that $\mathbb{R}/\mathbb{Z}$ is topologically the circle S^1. The same argument applies to $\mathbb{R}^n/L$, and we get the n-torus $S^1 \times \cdots \times S^1$ (see Question 1.2.2).

The central structure in geometry is a manifold – geometries where calculus is possible. Locally, a manifold looks like a piece of $\mathbb{R}^n$ (or $\mathbb{C}^n$), but these pieces can be bent and stitched together to create more interesting shapes. For instance, the n-torus is an n-dimensional manifold. The definition of manifold, due to Poincaré at the turn of the century, is a mathematical gem; it explains how flat patches can be sewn together to form smooth and globally interesting shapes.

Definition 1.2.3 *A C^∞ manifold M is a topological space with a choice of open sets $U_\alpha \subset M$, $V_\alpha \subset \mathbb{R}^n$ and homeomorphisms $\varphi_\alpha : U_\alpha \to V_\alpha$, as in Figure 1.5, such that the U_α cover M (i.e. $M = \cup_\alpha U_\alpha$) and whenever $U_\alpha \cap U_\beta$, the map $\varphi_\alpha \circ \varphi_\beta^{-1}$ is a C^∞ map from some open subset (namely $\varphi_\beta(U_\alpha \cap U_\beta)$) of V_β to some open subset (namely $\varphi_\alpha(U_\alpha \cap U_\beta)$) of V_α.*

A homeomorphism means an invertible continuous map whose inverse is also continuous. By a C^∞ map f between open subsets of $\mathbb{R}^n$, we mean that

$$f(x) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))$$

is continuous, and all partial derivatives $\frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}} f_j$ exist and are also continuous.

This is the definition of a *real* manifold; a *complex* manifold is similar. An n-dimensional complex manifold is a $2n$-dimensional real one. A one-dimensional manifold is called a *curve*, and a two-dimensional one a *surface*. 'Smooth' is often used for C^∞.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Using φ_α, each 'patch' $U_\alpha \subset M$ inherits the structure of $V_\alpha \subset \mathbb{R}^n$. For instance, we can coordinatise V_α and do calculus on it, and hence we get coordinates for, and can do calculus on, U_α. The overlap condition for $\varphi_\alpha \circ \varphi_\beta^{-1}$ guarantees compatibility. For example, the familiar latitude/longitude coordinate system comes from covering the Earth with two coordinate patches V_i – one centred on the North pole and the other on the South, and both stretching to the Equator – with polar coordinates chosen on each V_i.

More (or less) structure can be placed on the manifold, by constraining the overlap functions $\varphi_\alpha \circ \varphi_\beta^{-1}$ more or less. For example, a 'topological manifold' drops the C^∞ constraint; the result is that we can no longer do calculus on the manifold, but we can still speak of continuous functions, etc. A *conformal manifold* requires that the overlap functions preserve angles in $\mathbb{R}^n$ – the angle between intersecting curves in $\mathbb{R}^n$ is defined to be the angle between the tangents to the curves at the point of intersection. Conformal manifolds inherit the notion of angle from $\mathbb{R}^n$. Stronger is the notion of *Riemannian manifold*, which also enables us to speak of length.

It is now easy to compare structures on different manifolds. For instance, given two manifolds M, M', a function $f : M \to M'$ is 'C^∞' if each composition $\varphi'_\beta \circ f \circ \varphi_\alpha^{-1}$ is a C^∞ map from some open subset of V_α to V'_β; M and M' are C^∞-*diffeomorphic* if there is an invertible C^∞-function $f : M \to M'$ whose inverse is defined and is also C^∞.

Note that our definition doesn't assume the manifold M is embedded in some ambient space $\mathbb{R}^m$. Although it is true (Whitney) that any n-dimensional real manifold M can be embedded in Euclidean space $\mathbb{R}^{2n}$, this embedding may not be natural. For example, we are told that we live in a 'curved' four-dimensional manifold called space-time, but its embedding in $\mathbb{R}^8$ presumably has no physical significance.

Much effort in differential geometry has been devoted to questions such as: Given some topological manifold M, how many inequivalent differential structures (compatible with the topological structure) can be placed on M? It turns out that for any topological manifold of dimension ≤ 3, this differential structure exists and is unique. Moreover, $\mathbb{R}^n$ has a unique differential structure as well in dimensions ≥ 5. Remarkably, in four (and only four) dimensions it has uncountably many different differential structures (see [**195**])! Could this have anything to do with the appearance of macroscopic space-time being $\mathbb{R}^4$? Half a century before that discovery, the physicist Dirac prophesied [**139**]:

> ... as time goes on it becomes increasingly evident that the rules which the mathematician finds interesting are the same as those which Nature has chosen ... only four-dimensional space is of importance in physics, while spaces with other dimensions are of about equal interest in mathematics. It may well be, however, that this discrepancy is due to the incompleteness of present-day knowledge, and that future developments will show four-dimensional space to be of far greater mathematical interest than all the others.

Given any open set U in a manifold M, write $C^\infty(U)$ for the space of C^∞-functions $f : U \to \mathbb{R}$. When $U \subseteq U_\alpha$, we can use local coordinates and write $f(x^1, \ldots, x^n)$ (local coordinates are often written with superscripts). A fundamental lesson of geometry (perhaps learned from physics) is that one studies the manifold M through the (local) smooth

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

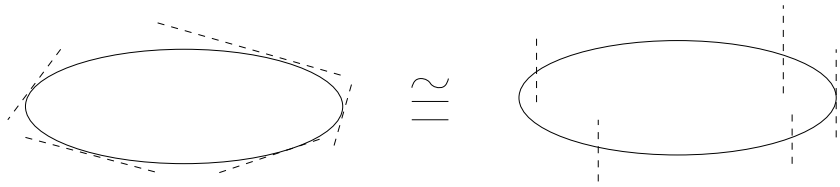

Fig. 1.6 The tangent bundle of S^1.

functions $f \in C^\infty(U)$ that live on it. This approach to geometry has been axiomatised into the notion of *sheaf* (see e.g. [**537**]), to which we return in Section 5.4.2.

For example, identifying S^1 with $\mathbb{R}/\mathbb{Z}$, the space $C^\infty(S^1)$ consists of the smooth period-1 functions $f : \mathbb{R} \to \mathbb{R}$, i.e. $f(\theta + 1) = f(\theta)$. Or we can identify S^1 with the locus $x^2 + y^2 = 1$, in which case $C^\infty(S^1)$ can be identified with the algebra $C^\infty(\mathbb{R}^2)$ of smooth functions in two variables, quotiented by the subalgebra (in fact ideal) consisting of all smooth functions $g(x, y)$ vanishing on all points satisfying $x^2 + y^2 = 1$; when $f(x, y)$, $g(x, y)$ are polynomials, then they are identical functions in $C^\infty(S^1)$ iff their difference $f(x, y) - g(x, y)$ is a polynomial multiple of $x^2 + y^2 - 1$.

Fix a point $p \in M$ and an open set U containing p. In Section 1.4.2 we need the notion of tangent vectors to a manifold M. An intuitive approach starts from the set $S(U, p)$ of curves passing through p, i.e. $\sigma : (-\epsilon, \epsilon) \to U_\alpha$ is smooth and $\sigma(0) = p$. Call curves $\sigma_1, \sigma_2 \in S(U, p)$ equivalent if they touch each other at p, that is

$$\sigma_1 \approx_p \sigma_2 \quad \text{iff} \quad \frac{\mathrm{d}}{\mathrm{d}t} f(\sigma_1(t))|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} f(\sigma_2(t))|_{t=0}, \qquad \forall f \in C^\infty(U, p). \tag{1.2.3a}$$

This defines an equivalence relation; the equivalence class $\langle\sigma\rangle_p$ consisting of all curves equivalent to σ is an infinitesimal curve at p. Equivalently, define a tangent vector to be a linear map $\xi : C^\infty(M) \to \mathbb{R}$ that satisfies the Leibniz rule

$$\xi(fg) = \xi(f)\, g(p) + f(p)\, \xi(g). \tag{1.2.3b}$$

In local coordinates $\xi = \sum_{i=1}^n \alpha_i \frac{\partial}{\partial x^i}|_{x=p}$, where the α_i are arbitrary real numbers. The bijection between these two definitions associates with any infinitesimal curve $v = \langle\sigma\rangle_p$ the tangent vector called the *directional derivative* $D_v : C^\infty(M) \to \mathbb{R}$, given by

$$D_v(f) = \frac{\mathrm{d}}{\mathrm{d}t} f(\sigma(t))|_{t=0}. \tag{1.2.3c}$$

The *tangent space* $T_p(M)$ *at* p is the set of all tangent vectors. Equation (1.2.3b) shows that $T_p(M)$ has a natural vector space structure; its dimension equals that of M. These tangent spaces can be glued together into a $2n$-dimensional manifold called the *tangent bundle* TM. Figure 1.6 shows why TS^1 is the cylinder $S^1 \times \mathbb{R}$. However, this is exceptional: although *locally* the tangent bundle TM of any manifold is trivial – that is, each TU_α is diffeomorphic to the direct product $U_\alpha \times \mathbb{R}^n$ – *globally* most tangent bundles TM are different from $M \times \mathbb{R}^n$.

A *vector field* is an assignment of a tangent vector to each point on the manifold, a smooth map $X : M \to TM$ such that $X(p) \in T_p(M)$. Equivalently, we can regard it

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

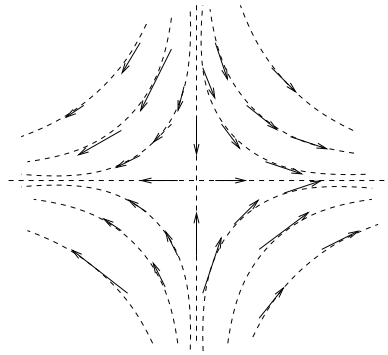

Fig. 1.7 The flow of a vector field.

as a *derivation* $X : C^\infty(M) \to C^\infty(M)$, i.e. a first-order differential operator acting on functions $f : M \to \mathbb{R}$ and obeying $X(fg) = X(f)\,g + f\,X(g)$. For example, the vector fields on the circle consist of the operators $g(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}$ for any smooth period-1 function $g(\theta)$.

Let $\mathrm{Vect}(M)$ denote the set of all vector fields on a manifold M. Of course this is an infinite-dimensional vector space, but we see in Section 1.4.1 that it has a much richer algebraic structure: it is a Lie algebra. $\mathrm{Vect}(S^1)$ is central to Moonshine, and in Section 3.1.2 we start exploring its properties.

A vector field X on M can be interpreted as being the instantaneous velocity of a fluid confined to M. We can 'integrate' this, by solving a first-order ordinary differential equation, thus covering M with a family of non-intersecting curves. Each curve describes the motion, or *flow*, of a small particle dropped into the fluid at the given point $p \in M$. The tangent vector to the curve at the given point p equals $X(p)$ – see Figure 1.7. Equivalently, corresponding to a vector field X is a continuous family $\varphi_t : M \to M$ of diffeomorphisms of M, one for each 'time' t, obeying $\varphi_t \circ \varphi_s = \varphi_{t+s}$, where $\varphi_t(p)$ is defined to be the position on M where the point p flows to after t seconds.

So it is natural to ask, what can we do with a diffeomorphism α of M? Clearly, α gives rise to an automorphism of the algebra $C^\infty(M)$, defined by $f \mapsto f^\alpha = f \circ \alpha$. Using this, we get an automorphism of $\mathrm{Vect}(M)$, $X \mapsto X^\alpha$, given by $X^\alpha(f) = (X(f^\alpha))^{\alpha^{-1}}$, or more explicitly, $X^\alpha(f)(p) = X(f \circ \alpha)(\alpha^{-1}(p))$. We return to this in Section 1.4.2.

One thing you can do with a *continuous* family of diffeomorphisms is construct a derivative for the algebras $C^\infty(M)$, $\mathrm{Vect}(M)$, etc. Defining a derivative of, say, a vector field X requires that we compare tangent vectors $X(p)$, $X(p')$ at neighbouring points on the manifold. This can't be done directly, since $X(p) \in T_pM$ and $X(p') \in T_{p'}M$ lie in different spaces. Given a vector field X, and corresponding flow φ_t, define the Lie derivative $\mathcal{L}_X(Y) \in \mathrm{Vect}(M)$ of any vector field $Y \in \mathrm{Vect}(M)$ by

$$\mathcal{L}_X(Y)(p) = \lim_{t\to 0} \frac{Y^{\varphi_t}(p) - Y(p)}{t} \in T_pM.$$

The Lie derivative $\mathcal{L}_X(f)$ of a function $f \in C^\infty(M)$ is defined similarly, and equals $X(f)$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Dual to the tangent vectors are the *differential 1-forms*. Just as the tangent spaces $T_p(M)$ together form the $2n$-dimensional tangent bundle TM, so their duals $T_p^*(M)$ form the $2n$-dimensional *cotangent bundle* T^*M. At least for finite-dimensional manifolds, the vector spaces $T_p^*(M)$ and $T_p(M)$, as well as the manifolds T^*M and TM, are homeomorphic, but without additional structure on M this homeomorphism is not canonical (it is basis- or coordinate-dependent). If $x = (x^1, \ldots, x^n) \mapsto M$ is a coordinate chart for manifold M, then $\partial_i := \frac{\partial}{\partial x^i}|_p$ is a basis for the tangent space T_pM, and its dual basis is written $\mathrm{d}x^i \in T_p^*(M)$: by definition they obey $\mathrm{d}x^i(\partial_j) = \delta_{ij}$.

Changing local coordinates from x to $y = y(x)$, the chain rule tells us

$$\frac{\partial}{\partial y^i} = \sum_{j=1}^{n} \frac{\partial x^j}{\partial y^i} \frac{\partial}{\partial x^j}, \tag{1.2.4a}$$

and hence the 1-form basis changes by the inverse formula:

$$\mathrm{d}y^i = \sum_{j=1}^{n} \frac{\partial y^i}{\partial x^j} \mathrm{d}x^j. \tag{1.2.4b}$$

The main purpose of differential forms is integration (hence their notation). If we regard the integrand of a line-integral as a 1-form field (i.e. a choice of 1-form for each point $p \in M$), we make manifest the choice of measure. Rather than saying the ambiguous 'integrate the constant function "$f(p) = 1$" along the manifold S^1', we say the unambiguous 'integrate the 1-form "$\omega_p = \mathrm{d}\theta$" along the manifold S^1'. Likewise, the integrands of double-, triple-, etc. integrals are 2-forms, 3-forms, etc., dual to tensor products of tangent spaces. We can evaluate these integrals by introducing coordinate patches and thus reducing them to usual $\mathbb{R}^n$ integrals over components of the differential form. The spirit of manifolds is to have a coordinate-free formalism; changing local coordinates (e.g. when moving from one coordinate patch to an overlapping one) changes those components as in (1.2.4b) in such a way that the value of the integral won't change.

A standard example of a 1-form field is the *gradient* $\mathrm{d}f$ of a function $f \in C^\infty(M)$, defined at each point $p \in M$ by the rule: given any tangent vector $D_v \in T_p(M)$, define the number $(\mathrm{d}f)(D_v)$ to be the value of the directional derivative $D_v(f)(p)$ at p.

A familiar example of a *2-form* $g_p \in T_p^*(M) \otimes T_p^*(M)$ is the *metric tensor* on $T_p(M)$. Given two vectors $u, v \in T_p$, the number $g_p(u, v)$ is to be thought of as their inner-product. A *Riemannian manifold* is a manifold M together with a 2-form field g, which is symmetric and nondegenerate (usually positive-definite).[5] Given a local coordinate about $p \in M$, a basis for the tangent space T_pM is $\frac{\partial}{\partial x^i}$ and we can describe the metric tensor g_p using an $n \times n$ matrix whose ij-entry is $g_{ij}(p) := g_p(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j})$, or in infinitesimal language as $\mathrm{d}s^2 = \sum_{i,j=1}^{n} g_{ij}\mathrm{d}x^i\mathrm{d}x^j$, a form more familiar to most physicists.

[5] Whitney's aforementioned embedding of M into Euclidean space implies that any manifold can be given a Riemannian structure, since a submanifold of a Riemannian manifold naturally inherits the Riemannian structure. The *Beautiful Mind* of John Nash proved that any Riemannian structure on a given n-dimensional manifold M can likewise be inherited from its embedding into some sufficiently large-dimensional Euclidean space.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Much structure comes with this metric tensor field g. Most important, of course, we can define lengths of curves and the angles with which they intersect. In particular, the arc-length of the curve $\gamma : [0, 1] \to M$ is the integral

$$\int_0^1 \sqrt{g_{\gamma(t)}\left(\frac{\mathrm{d}\gamma}{\mathrm{d}t}, \frac{\mathrm{d}\gamma}{\mathrm{d}t}\right)}\,\mathrm{d}t,$$

a quantity independent of the specific parametrisation $t \mapsto \gamma(t)$ chosen (verify this).

Also, we can use the metric to identify each T_p^* with T_p, just as the standard inner-product in $\mathbb{R}^n$ permits us to identify a column vector $u \in \mathbb{R}^n$ with its transpose $u^t \in \mathbb{R}^{n*}$. Moreover, given any curve $\sigma : [0, 1] \to M$ connecting $\sigma(0) = p$ to $\sigma(1) = q$, we can identify the tangent spaces $T_p(M)$ and $T_q(M)$ by *parallel-transport*. Using this, we can define a derivative (the so-called 'covariant derivative') that respects the metric, and a notion of geodesic (a curve that parallel-transports its own tangent vector, and which plays the role of 'straight line' here). In short, on a Riemannian manifold geometry in its fullest sense is possible. See, for example, [**104**] for more details.

Many manifolds locally look like a Cartesian product $A \times B$. A *fibre bundle* $p : E \to B$ locally (i.e. on small open sets U of E) looks like $F \times V$, where $F \cong p^{-1}(b)$ (for any $b \in B$) is called the *fibre*, and V is an open set in the *base* B. For example, the (open) cylinder and Möbius stip are both fibre bundles with base S^1 and fibre $(0, 1) \subset \mathbb{R}$. A *section* $s : B \to E$ obeys $p \circ s = id.$, that is for each small open set V of B it is a function $V \to F$. A *vector bundle* is a fibre bundle with fibre a vector space $F = V$, for example the tangent bundle TM is a vector bundle with base M and fibre $\cong T_pM$. We write $\Gamma(E)$ for the space of sections of a vector bundle E. A *line bundle* is a vector bundle with one-dimensional fibre (so the sections of a line bundle locally look like complex- or real-valued functions on the base). A *connection* on a vector bundle $E \to B$ is a way to differentiate sections (the *covariant derivative*). An example is a Riemannian structure on the tangent bundle $E = TB$. See, for example, [**104**] for details and examples.

Felix Klein's *Erlangen Programm* (so called because he announced it there) is a strategy relating groups and geometry. Geometry, it says, consists of a manifold (the space of points) and a group of automorphisms (transformations) of the manifold preserving the relevant geometric structures (e.g. length, angle, lines, etc.). Conversely, given a manifold and a group of automorphisms, we should determine the invariants relative to the group. Several different geometries are possible on the same manifold, distinguished by their preferred transformations.

For example, Euclidean geometry in its strongest sense (i.e. with lengths, angles, lines, etc.) has the group of symmetries generated by rotations, reflections and translations – that is any transformation of the form $x \mapsto xA + a$, where $x, a \in \mathbb{R}^n$ (regarded as row vectors, say) and A is an orthogonal $n \times n$ matrix. If our context is scale-independent (e.g. when studying congruent triangles), we can allow A to obey $AA^t = \lambda I$ for any $\lambda \in \mathbb{R}$.

More interesting is *projective geometry*. Here, angles and lengths are no longer invariants, but lines are. Projective geometry arose from the theory of perspective in art. The transformations of projective n-geometry come from projections $\mathbb{R}^{n+1} \to \mathbb{R}^n$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

More precisely, consider real projective n-space $\mathbb{P}^n(\mathbb{R})$. We coordinatise it using *homogeneous coordinates*: $\mathbb{P}^n(\mathbb{R}) = \mathbb{R}^{n+1\prime}/\sim$ consists of $(n+1)$-tuples of real numbers, where we identify points with their multiples. The origin $(0, 0, 0, \ldots, 0)$ in $(n+1)$-space is excluded from projective space (hence the prime), as it belongs to all such lines. A projective 'point' consists of points on the same line through the origin; a projective 'line' consists of planes through the origin; etc. By convention, any equation in homogeneous coordinates is required to be homogeneous (so that a point satisfies an equation iff its whole line does). Complex projective space $\mathbb{P}^n(\mathbb{C})$ is defined similarly.

To see what projective geometry is like, consider first the projective line $\mathbb{P}^1(\mathbb{R})$. Take any point in $(x, y) \in \mathbb{P}^1(\mathbb{R})$. If $y \neq 0$ we may divide by it, and we get points of the form $(x', 1)$. These are in one-to-one correspondence with the points in the real line. If, on the other hand, $y = 0$, then we know $x \neq 0$ and so we should divide by x: what we get is the point $(1, 0)$, which we can think of as the infinite point $(\frac{1}{0}, 1)$. Thus the real projective line $\mathbb{P}^1(\mathbb{R})$ consists of the real line, together with a point 'at infinity'. Similarly, the complex projective line consists of the complex plane $\mathbb{C}$ together with a point at infinity; topologically, this is a sphere named after Riemann.

More generally, $\mathbb{P}^n(\mathbb{R})$ consists of the real space $\mathbb{R}^n$, together with a copy of $\mathbb{P}^{n-1}(\mathbb{R})$ as the hyperplane of infinite points. These points at infinity are where parallel lines meet. Intuitively, projective geometry allows us to put 'finite' and 'infinite' points on an equal footing; we can see explicitly how, for example, curves look at infinity.

For example, the 'parallel' lines $x = 0$ and $x = 1$ in $\mathbb{P}^2(\mathbb{R})$ correspond to the homogeneous equations $x = 0$ and $x = z$, and so to the points with homogeneous coordinates $(0, y, z)$ and (x, y, x). They intersect at the 'infinite' point $(0, y, 0) \sim (0, 1, 0)$. The parabola $y = x^2$ has only one infinite point (namely (0,1,0)), the hyperbola $xy = 1$ has two infinite points ((1,0,0) and (0,1,0)), while the circle $x^2 + y^2 = 1$ doesn't have any. Intuitively, the parabola is an ellipse tangent to the line (really, circle) at infinity, while the hyperbola is an ellipse intersecting it transversely.

Klein's group of transformations here is the projective linear group $\mathrm{PGL}_{n+1}(\mathbb{R})$, that is all invertible $(n+1) \times (n+1)$ matrices A where we identify A with λA for any nonzero number λ. It acts on the homogeneous coordinates in the usual way: $x \mapsto xA$. This group mixes thoroughly the so-called infinite points with the finite ones, and emphasises that infinite points in projective geometry are completely on a par with finite ones. For example, the transformation $A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ maps the parabola $y = x^2$ to the hyperbola $xy = 1$, indicating that these are projectively identical curves.

Projective geometry is central to modern geometry. The projective plane can be axiomatised, for example one axiom says that any two lines intersect in exactly one point. A remarkable property of projective geometry is that any theorem remains a theorem if the words 'line' and 'point' are interchanged.

In summary, there are many different geometries. Which geometry to use (e.g. Euclidean, projective, conformal) in a given context depends on the largest possible group of transformations that respect the basic quantities.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

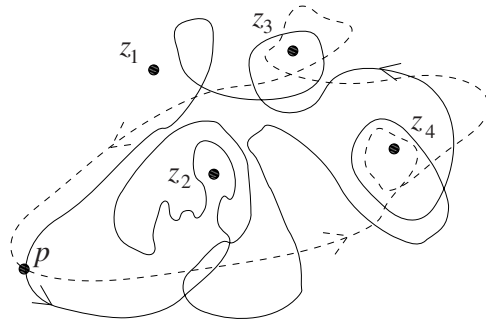

Fig. 1.8 Two homotopic loops in $\pi_1 \cong \mathcal{F}_4$.

1.2.3 Loops

The last subsection used curves to probe the infinitesimal neighbourhood of any point $p \in M$. We can also use curves to probe global features of manifolds.

Let M be any manifold, and put $I = [0, 1]$. A *loop* at $p \in M$ is any continuous curve $\sigma : I \to M$ with $\sigma(0) = \sigma(1) = p$. So σ starts and ends at the point p. Let $\Omega(M, p)$ be the set of all such loops. Loops $\sigma_0, \sigma_1 \in \Omega(M, p)$ are *homotopic* if σ_0 can be continuously deformed into σ_1, that is if there is a continuous map $F : I \times I \to M$ with $\sigma_i(\star) := F(\star, i) \in \Omega(M, p)$, for $i = 0, 1$. This defines an equivalence relation on $\Omega(M, p)$. For instance all loops in $M = \mathbb{R}^n$ are homotopic, while the homotopy equivalence classes for the circle $M = S^1$ are parametrised by their winding number $n \in \mathbb{Z}$, that is by the contour integral $\frac{1}{2\pi i} \int_{\sigma(I)} \frac{\mathrm{d}z}{z}$.

Let $\pi_1(M, p)$ denote the set of all homotopy equivalence classes for $\Omega(M, p)$. It has a natural group structure: $\sigma\sigma'$ is the curve that first goes from p to p following σ, and then from p to p following σ'. More precisely,

$$(\sigma\sigma')(t) = \begin{cases} \sigma(2t) & \text{if } 0 \le t \le \frac{1}{2} \\ \sigma'(2t-1) & \text{if } \frac{1}{2} \le t \le 1 \end{cases} . \tag{1.2.5}$$

For instance, the inverse σ^{-1} is given by the curve traversed in the opposite direction: $t \mapsto \sigma(1-t)$. The identity is the constant curve $\sigma(t) = p$. With this operation $\pi_1(M, p)$ is called the *fundamental group* of M (the subscript '1' reminds us that a loop is a map from S^1; likewise π_k considers maps from the k-sphere S^k to M). As long as any two points in M can be connected with a path, then all $\pi_1(M, p)$ will be isomorphic and we can drop the dependence on 'p'. When $\pi_1(M) = \{e\}$, we say M is *simply connected*.

For example, $\pi_1(\mathbb{R}^n) \cong 1$ and $\pi_1(S^1) \cong \mathbb{Z}$. The complex plane $\mathbb{C}$ with n points removed has fundamental group $\pi_1(\mathbb{C}\backslash\{z_1, \ldots, z_n\}) \cong \mathcal{F}_n$, the free group – Figure 1.8 gives two paths homotopic to $x_4 x_3^{-1} \in F_4$. The torus $S^1 \times S^1$ has $\pi_1 \cong \mathbb{Z} \oplus \mathbb{Z}$.

The braid group (1.1.9), as with any group, also has a realisation as a fundamental group. Let $\mathfrak{C}_n$ be $\mathbb{C}^n$ with all diagonals removed:

$$\mathfrak{C}_n = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_i \neq z_j \text{ whenever } i \neq j\}. \tag{1.2.6}$$

Then it is easy to see that the pure braid group $\mathcal{P}_n$ is isomorphic to $\pi_1(\mathfrak{C}_n)$ – indeed, given any braid $\alpha \in \mathcal{B}_n$, the value of the ith coordinate $\sigma(t)_i$ of the corresponding loop

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

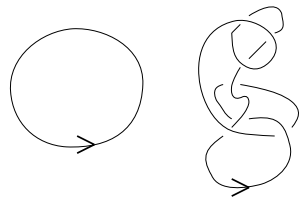

Fig. 1.9 Some trivial knots.

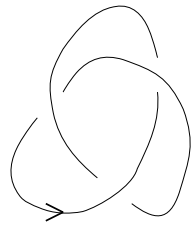

Fig. 1.10 The trefoil.

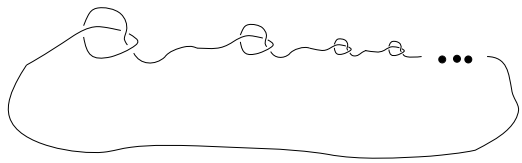

Fig. 1.11 A wild knot.

$\sigma \in \pi_1(\mathfrak{C}_n)$ will be the position of the ith strand when we take a slice at t through our braid ($t = 0$ is the top of the braid, $t = 1$ the bottom). Now, the symmetric group $\mathcal{S}_n$ acts freely (i.e. without fixed points) on $\mathfrak{C}_n$ by permuting the coordinates: $\pi.z = (z_{\pi 1}, \ldots, z_{\pi n})$. The space $\mathfrak{C}_n/\mathcal{S}_n$ of orbits under this action has fundamental group $\pi_1(\mathfrak{C}_n/\mathcal{S}_n) \cong \mathcal{B}_n$.

Note that if $f : M' \to M$ is a homeomorphism, then it induces a group homomorphism $f_* : \pi_1(M') \to \pi_1(M)$. We return to this in Section 1.7.2.

By a *link* we mean a diffeomorphic image of $S_1 \cup \cdots \cup S^1$ into $\mathbb{R}^3$. A *knot* is a link with one strand – see Figures 1.9 and 1.10. Since S^1 comes with an orientation, so does each strand of a link. The reason for requiring the embedding $f : S^1 \cup \cdots \cup S^1 \to \mathbb{R}^3$ to be *differentiable* is that we want to avoid 'wild knots' (see Figure 1.11); almost every *homeo*morphic image of $S^1 \cup \cdots \cup S^1$ will be wild at almost every point.

Two links are equivalent, i.e. *ambient isotopic*, if continuously deforming one link yields the other. The word 'ambient' is used because the isotopy is applied to the ambient space $\mathbb{R}^3$. This is the intuitive notion of equivalent knots in a string, except that we glue the two ends of the string together (we can trivially untie any knotted open string by slipping the knot off an end). By a trivial knot or the *unknot* we mean any knot homotopic to (say) the unit circle in the xy-plane in $\mathbb{R}^3$.

We choose $\mathbb{R}^3$ for the ambient space because any link in $\mathbb{R}^n$, for $n \geq 4$, is trivial, and the Jordan Curve Theorem tells us that there are only two different 'knots' in $\mathbb{R}^2$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

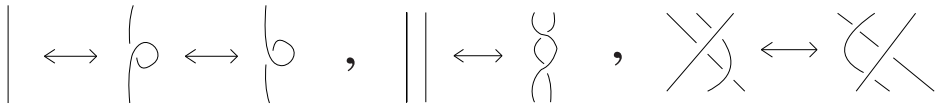

Fig. 1.12 The Reidemeister moves I, II, III, respectively.

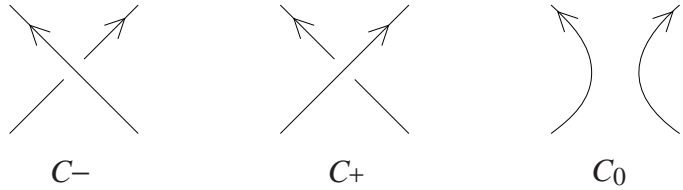

Fig. 1.13 The possible (non)crossings.

(distinguished by their orientation). More generally, knotted k-spheres S^k in $\mathbb{R}^n$ are nontrivial only when $n = k + 2$ [**478**].

It isn't difficult to show [**478**] that two links are ambient isotopic iff their diagrams can be related by making a finite sequence of moves of the form given in Figure 1.12. The Reidemeister moves are useless at deciding directly whether two knots are equivalent, or even whether a given knot is trivial. Indeed, this seems difficult no matter which method is used, although a finite algorithm (by Häken and Hemion [**283**]) apparently exists. A very fruitful approach has been to assign to a link a quantity (called a *link invariant*), usually a polynomial, in such a way that ambient isotopic links get the same quantity. One of these is the *Jones polynomial* J_L, which can be defined recursively by a *skein relation*. Start with any (oriented) link diagram and choose any crossing; up to a rotation it will either look like the crossing C_+ or C_- in Figure 1.13. There are two things we can do to this crossing: we can pass the strings through each other (so the crossing of type $C_\pm$ becomes one of type $C_\mp$); or we can erase the crossing as in C_0. In this way we obtain three links: the original one (which we could call $L_\pm$ depending on the orientation of the chosen crossing) and the two modified ones ($L_\mp$ and L_0). The skein relation is

$$t^{-1}\, J_{L_+}(t) - t\, J_{L_-}(t) + \left(t^{-\frac{1}{2}} - t^{\frac{1}{2}}\right) J_{L_0}(t) = 0. \tag{1.2.7}$$

We also define the polynomial $J(t)$ of the unknot to be identically 1.

For a link with an odd number of components, $J_L(t) \in \mathbb{Z}[t^{\pm 1}]$ is a Laurent polynomial in t, while for an even number $J_L(t) \in \sqrt{t}\mathbb{Z}[t^{\pm 1}]$. For example, applying (1.2.7) twice, we get that the Jones polynomial of the trefoil in Figure 1.10 is $J(t) = -t^4 + t^3 + t$.

Are the trefoil and its mirror image ambient isotopic? The easiest argument uses the Jones polynomial: taking the mirror image corresponds to replacing t with t^{-1}, and we see that the Jones polynomial of the trefoil is not invariant under this transformation.[6]

[6] More generally, a knot with odd crossing number will be inequivalent with its mirror image (the crossing number is the minimum number of crossings needed in a diagram of the knot).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

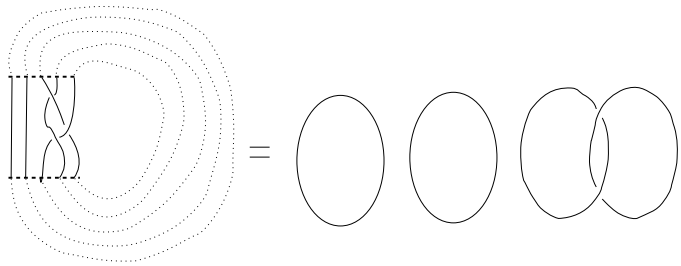

Fig. 1.14 The link associated with a braid.

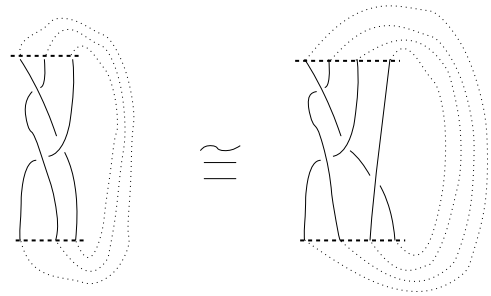

Fig. 1.15 A Markov move of type II.

The Reidemeister moves quickly prove $J_L(t)$ is a knot invariant, i.e. equivalent knots have the same polynomial, although inequivalent knots can also have the same one. But it was the first new knot polynomial in 56 years. It triggered discoveries of several other invariants while making unexpected connections elsewhere (Section 6.2.6), and secured for Jones a Fields medal. The problem then became that there were too many link invariants. We explain how we now organise them in Section 1.6.2.

Braids and links are directly related by theorems of Alexander (1923) and Markov (1935). Given any braid α we can define a link by connecting the ith spot on the bottom of the braid with the ith spot on the top, as in Figure 1.14. Alexander's theorem tells us that all links come from a braid in this way. Certainly though, different braids can correspond to the same link – for example, take any $\alpha, \beta \in \mathcal{B}_n$, then the links of α and $\beta\alpha\beta^{-1}$ are the same (slide the braid β^{-1} counterclockwise around the link until it is directly above, and hence cancels, β). This is called a Markov move of type I. A Markov move of type II changes the number of strands in a braid by ± 1, in a simple way – see Figure 1.15. Markov's theorem [**59**] says that two braids $\alpha \in \mathcal{B}_n$, $\beta \in \mathcal{B}_m$ correspond to equivalent links iff they are related by a finite sequence of Markov[7] moves. In Section 6.2.5, we explain how to use these two theorems to construct link invariants.

Question 1.2.1. Come up with a reasonable definition for the automorphism group of a lattice. Prove that the automorphism group of a positive-definite lattice is always finite.

[7] His father is the Markov of Markov chains.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Question 1.2.2. Let $x = (x_1, x_2)$ be any vector with nonzero coordinate x_2. Write $L(x)$ here for the lattice $\mathbb{Z}(1,0) + \mathbb{Z}x$, and $T(x)$ for the torus $\mathbb{R}^2/L(x)$. Which x's give pointwise identical lattices (i.e. given x, find all y such that $L(x) = L(y)$)? Verify that all tori are diffeomorphic. Which tori $T(x)$ are obviously conformally equivalent?

Question 1.2.3. If we drop the requirement in Definition 1.2.1 that the $x^{(i)}$ be a basis, does anything really bad happen?

Question 1.2.4. Prove Theorem 1.2.2.

Question 1.2.5. Let L be an integral lattice. What is special about the reflection r_α through a vector $\alpha \in L$ with norm-squared $\alpha \cdot \alpha = 2$? (The formula for the reflection r_α is $r_\alpha(x) = x - \frac{2x\cdot\alpha}{\alpha\cdot\alpha}\,\alpha$.)

Question 1.2.6. Prove from (1.2.7) that the Jones polynomial for a link and its mirror image can be obtained from each other by the switch $t \leftrightarrow t^{-1}$. Prove that the Jones polynomial of a link is unchanged if the orientation of any component (i.e. the arrow on any strand) is reversed.

Question 1.2.7. Find the Jones polynomial of the disjoint union of n circles.

1.3 Elementary functional analysis

Moonshine concerns the occurrence of modular forms in algebra and physics, and care is taken to avoid analytic complications as much as possible. But spaces here are unavoidably infinite-dimensional, and through this arise subtle but significant points of contact with analysis. For example, the $q^{1/24}$ prefactor in the Dedekind eta (2.2.6b), and the central extension of loop algebras (3.2.2a), are analytic fingerprints. Lie group representations usually involve functional analysis (see e.g. Section 2.4.2 where we relate the Heisenberg group to theta functions). Much of functional analysis was developed to address mathematical concerns in quantum theory, and perhaps all of the rich subtleties of quantum field theory can be interpreted as functional analytic technicalities. For example, *anomalies* (which for instance permit derivations of the Atiyah–Singer Index Theorem from super Yang–Mills calculations) can be explained through a careful study of domains of operators [**172**]. Moreover, the natural culmination of the Jones knot polynomial is a deep relation between subfactors and conformal field theories (Section 6.2.6). The necessary background for all this is supplied in this section.

In any mature science such as mathematics, the division into branches is a convenient lie. In this spirit, analysis can be distinguished from, say, algebra by the central role played in the former by numerical inequalities. For instance, inequalities appear in the definition of derivatives and integrals as limits. Functional analysis begins with the reinterpretation of derivatives and integrals as linear operators on vector spaces. These spaces, which consist of appropriately restricted functions, are infinite-dimensional. The complexity and richness of the theory comes from this infinite-dimensionality.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Section 1.3.1 assumes familiarity with elementary point-set topology, as well as the definition of Lebesgue measure. All the necessary background is contained in standard textbooks such as [**481**].

1.3.1 Hilbert spaces

By a vector space $\mathcal{V}$, we mean something closed under *finite* linear combinations $\sum_{i=1}^{n} a_i v^{(i)}$. Here we are primarily interested in infinite-dimensional spaces over the complex numbers (i.e. the scalars a_i are taken from $\mathbb{C}$), and the vectors v are typically functions f. By a (complex) *pre-Hilbert space* we mean a vector space $\mathcal{V}$ with a Hermitian form $\langle f, g\rangle \in \mathbb{C}$ ('Hermitian form' is defined in Section 1.1.3). All complex n-dimensional pre-Hilbert spaces are isomorphic to $\mathbb{C}^n$ with Hermitian form

$$\langle u, v\rangle = \overline{u_1} v_1 + \cdots + \overline{u_n} v_n.$$

The analogue of $\mathbb{C}^n$ in countably many dimensions is $\ell^2(\infty)$, which consists of all sequences $u = (u_1, u_2, \ldots)$ with finite sum $\sum_{i=1}^{\infty} |u_i|^2 < \infty$. The reader can verify that it is closed under sums and thus forms a pre-Hilbert space. Another example consists of the C^∞-functions $f : \mathbb{R}^n \to \mathbb{C}$, say, with 'compact support' (that means that the set of all $x \in \mathbb{R}^n$ for which $f(x) \neq 0$ is bounded). The Hermitian form here is

$$\langle f, g\rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \overline{f(x)}\, g(x)\, \mathrm{d}^n x\ ; \tag{1.3.1}$$

this pre-Hilbert space is denoted $\mathcal{C}^\infty_{cs}(\mathbb{R}^n)$. For instance, the function defined by

$$f(x) = \begin{cases} \exp[\frac{1}{x^2-1}] & \text{for } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

lies in $\mathcal{C}^\infty_{cs}(\mathbb{R})$. A larger space, arising for instance in quantum mechanics, is denoted $\mathcal{S}(\mathbb{R}^n)$ and consists of all functions $f \in C^\infty(\mathbb{R}^n)$ that, together with their derivatives, decrease to 0 faster than any power of $|x|^{-1}$, as $|x| \to \infty$. The space $\mathcal{S}$ is a pre-Hilbert space, again using (1.3.1). It contains functions such as $\mathrm{poly}(x_1, \ldots, x_n)\, e^{-x_1^2 - \cdots - x_n^2}$.

A pre-Hilbert space has a notion of distance, or *norm* $\|f\|$, given by $\|f\|^2 = \langle f, f\rangle$. Using this we can define limits, Cauchy sequences, etc. in the usual way [**481**]. We call a subset X of $\mathcal{V}$ *dense* in $\mathcal{V}$ if for any $f \in \mathcal{V}$ there is a sequence $f_n \in X$ that converges to f. For instance, the rationals $\mathbb{Q}$ are dense in the reals $\mathbb{R}$, but the integers aren't. Any convergent sequence is automatically Cauchy; a pre-Hilbert space $\mathcal{V}$ is called *complete* if conversely all Cauchy sequences in it converge.

Definition 1.3.1 *A* Hilbert space $\mathcal{H}$ *is a complete pre-Hilbert space.*

For example, each $\mathbb{C}^n$ is Hilbert, as is $\ell^2(\infty)$. Most pre-Hilbert spaces aren't Hilbert, for example neither $\mathcal{C}^\infty_{cs}(\mathbb{R}^n)$ nor $\mathcal{S}(\mathbb{R}^n)$ are. However, given any pre-Hilbert space $\mathcal{V}$, there is a Hilbert space $\mathcal{H}$ that contains $\mathcal{V}$ as a dense subspace. This Hilbert space $\mathcal{H}$ is called the completion $\overline{\mathcal{V}}$ of $\mathcal{V}$, and is unique up to isomorphism. The construction of $\mathcal{H}$ from $\mathcal{V}$ is analogous to the construction of $\mathbb{R}$ from $\mathbb{Q}$, obtained by defining an equivalence relation on the Cauchy sequences.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

The Hilbert space completion $\overline{\mathcal{C}^\infty_{cs}(\mathbb{R}^n)} = \overline{\mathcal{S}(\mathbb{R}^n)}$ is defined using the 'Lebesgue measure' μ, which is an extension of the usual notion of length to a much more general class of subsets $X \subset \mathbb{R}$ than the intervals, and the 'Lebesgue integral' $\int f(x)\,\mathrm{d}\mu(x)$, which is an extension of the usual Riemann integral to a much more general class of functions than the piecewise continuous ones. For example, what is the length of the set X consisting of all rational numbers between 0 and 1? This isn't defined, but its Lebesgue measure is easily seen to be 0. We won't define Lebesgue measures and integrals here, because we don't really need them; a standard account is [**481**]. The completion of $\mathcal{C}^\infty_{cs}(\mathbb{R}^n)$ is the Hilbert space $L^2(\mathbb{R}^n)$ consisting of all square-integrable functions $f : \mathbb{R}^n \to \mathbb{C} \cup \{\infty\}$. The Hermitian form is given by $\langle f, g\rangle = \int_{\mathbb{R}^n} \overline{f(x)}\, g(x)\,\mathrm{d}\mu(x)$. By f 'square-integrable' we mean that f is 'measurable' (e.g. any piecewise continuous function is measurable) and $\langle f, f\rangle < \infty$. We must identify two functions f, g if they agree *almost everywhere*, that is the set X of all $x \in \mathbb{R}^n$ at which $f(x) \neq g(x)$ has Lebesgue measure 0. This is because any two such functions have the property that $\langle f, h\rangle = \langle g, h\rangle$ for all h.

All Hilbert spaces we will consider, such as $L^2(\mathbb{R}^n)$, are *separable*. This means that there is a countable orthonormal set X of vectors $e_n \in \mathcal{H}$ (so $\langle e_n, e_m\rangle = \delta_{nm}$) such that the pre-Hilbert space span(X) consisting of all *finite* linear combinations $\sum a_m e_m$ is dense in $\mathcal{H}$. That is, given any $f \in \mathcal{H}$, $f = \lim_{n\to\infty} \sum_{i=1}^n \langle e_i, f\rangle\, e_i$ – we say that the *topological span* of X is $\mathcal{H}$. *All infinite-dimensional separable Hilbert spaces are isomorphic to* $\ell^2(\infty)$. The easy proof sends $f \in \mathcal{H}$ to the sequence $(\langle e_1, f\rangle, \langle e_2, f\rangle, \ldots) \in \ell^2(\infty)$.

We are interested in linear maps. The first surprise is that continuity is not automatic. In fact, let $T : \mathcal{V}_1 \to \mathcal{V}_2$ be a linear map between pre-Hilbert spaces. Then T is continuous at one point iff it's continuous at all points, iff it is *bounded* – that is, iff there exists a constant C such that $\|Tf\| \leq C\,\|f\|$, for all $f \in \mathcal{V}_1$. If $\mathcal{V}_1$ is finite-dimensional, then it is easy to show that any linear T is bounded. But in quantum mechanics, for example, most operators of interest are unbounded.

Another complication of infinite-dimensionality is that in practise we're often interested in linear operators whose domain is only a (dense) subspace of $\mathcal{H}$. For example, the domains of the operators $f(x) \mapsto xf(x)$ or $f(x) \mapsto \frac{\mathrm{d}}{\mathrm{d}x} f(x)$ (the 'position' and 'momentum' operators of quantum mechanics – see Section 4.2.1) are proper subspaces of $L^2(\mathbb{R})$. Those operators are well-defined though on $\mathcal{S}(\mathbb{R})$ (indeed, this is precisely why the space $\mathcal{S}$ is so natural for quantum mechanics). Once again bounded operators are simpler: if T is a bounded linear operator on some dense subspace $\mathcal{V}$ of a Hilbert space $\mathcal{H}$, then there is one and only one way to continuously extend the domain of T to all of $\mathcal{H}$.

The *dual* (or *adjoint*) $\mathcal{V}^*$ of a pre-Hilbert space $\mathcal{V}$ is defined as the space of all continuous linear maps (*functionals*) $\mathcal{V} \to \mathbb{C}$. In general, $\mathcal{V}$ can be regarded as a subspace of $\mathcal{V}^*$, with $f \in \mathcal{V}$ being identified with the functional $g \mapsto \langle \overline{f}, g\rangle$; when $\mathcal{V}$ is a Hilbert space $\mathcal{H}$, this identification defines an isomorphism $\mathcal{H}^* \cong \mathcal{H}$.

The functionals for C^∞_{cs} are called *distributions*, while those for $\mathcal{S}$ are *tempered distributions*. For example, the Dirac delta '$\delta(x - a)$' is defined as the element of $\mathcal{S}(\mathbb{R})^*$ sending functions $\varphi \in \mathcal{S}(\mathbb{R})$ to the number $\varphi(a) \in \mathbb{C}$. (Tempered) distributions F can all be realised (non-uniquely) as follows: given $a \in \mathbb{N}$ and a continuous function $f(x)$ of

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

polynomial growth, we get a functional $F \in \mathcal{S}(\mathbb{R})^*$ by

$$F(\varphi) = \int_{\mathbb{R}} f(x) \frac{\mathrm{d}^a \varphi}{\mathrm{d}x^a} \, \mathrm{d}x. \tag{1.3.2}$$

A similar realisation holds for the spaces $\mathcal{S}(\mathbb{R}^n)$ and $C^\infty_{cs}(\mathbb{R}^n)$. Of course distributions are not functions, and we cannot rewrite (1.3.2) as $\int g(x)\, \varphi(x)\, \mathrm{d}^n x$ for some function g. Note that the Dirac delta is not well-defined on the completion $L^2(\mathbb{R})$ of $\mathcal{S}$, since the elements $f \in L^2(\mathbb{R})$ are equivalence classes of functions and hence have ambiguous function values $f(a)$. This beautiful interpretation of distributions like δ as linear functionals is due to Sobolev and was developed by Schwartz, the 1950 Fields medalist. Another interpretation, using formal power series, is given in Section 5.1.2.

Distributions can be differentiated arbitrary numbers of times, and their partial derivatives commute (something not true of all differentiable functions). However, they usually cannot be multiplied together and thus form only a vector space, not an algebra. For more on distributions, see chapter 2 of [**67**] or chapter I of [**244**].

We're most interested in unitary and self-adjoint operators. First, let's define the adjoint. Let $T : \mathcal{V} \to \mathcal{H}$ be linear, where $\mathcal{V}$ is a subspace of $\mathcal{H}$. Let $\mathcal{U}$ be the set of all $g \in \mathcal{H}$ for which there is a unique vector $g^* \in \mathcal{H}$ such that for all $f \in \mathcal{V}$, $\langle g^*, f\rangle = \langle g, Tf\rangle$. Define the map (*adjoint*) $T^* : \mathcal{U} \to \mathcal{H}$ by $T^*(g) = g^*$. The adjoint T^* exists (i.e. its domain $\mathcal{U}$ is non-empty) iff $\mathcal{V}$ is dense in $\mathcal{H}$. In particular, T^{**} need not equal T. When $\mathcal{V}$ is dense in $\mathcal{H}$, $\mathcal{U}$ is a vector space and T^* is linear. When T is bounded, so is T^*, and its domain $\mathcal{U}$ is all of $\mathcal{H}$. Note that $\langle g, Tf\rangle = \langle T^* g, f\rangle$ for all $f \in \mathcal{V}$, $g \in \mathcal{U}$, but that relation doesn't uniquely specify T^*.

We call T *self-adjoint* if $T = T^*$ (so in particular this implies that their domains $\mathcal{V}, \mathcal{U}$ are equal). This implies $\langle Tf, g\rangle = \langle f, Tg\rangle$, but as before the converse can fail. If T is self-adjoint and unbounded, then its domain cannot be all of $\mathcal{H}$.

A linear map $T : \mathcal{H}_1 \to \mathcal{H}_2$ between Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ is *unitary* if it is both onto and obeys $\langle Tf, Tg\rangle = \langle f, g\rangle$. Equivalently, $T^*T = TT^* = 1$. The surjectivity assumption is not redundant in infinite dimensions (Question 1.3.2). A unitary map is necessarily bounded. A famous example of a unitary operator is the Fourier transform $f \mapsto \widehat{f}$, which, as usually defined, maps $\mathcal{S}(\mathbb{R}^n)$ onto itself; it extends to a unitary operator on $L^2(\mathbb{R}^n)$.

To define limits, etc., one needs only a topology. This need not come from a norm, and in general many different topologies can naturally be placed on a space. For an artificial example, consider the real line $\mathbb{R}$ endowed with the discrete topology (in which any subset of $\mathbb{R}$ is open): then any function $f : \mathbb{R} \to \mathbb{R}$ will be continuous, a sequence $x_n \in \mathbb{R}$ will converge iff there is some N such that $x_N = x_{N+1} = x_{N+2} = \cdots$, and $\mathbb{R}$ with this topology is again complete. In the topology coming from the Hermitian form (1.3.1), $\mathcal{S}(\mathbb{R}^n)$ is incomplete, however it is common to refine that topology somewhat. In this new topology, a sequence $f_m \in \mathcal{S}(\mathbb{R})$ converges to 0 iff for every $a, b \in \mathbb{N}$ we have

$$\lim_{m\to\infty} \sup_{x\in\mathbb{R}} |x|^b \left| \frac{\mathrm{d}^a f_m(x)}{\mathrm{d}x^a} \right| = 0.$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

This topology comes from interpreting $\mathcal{S}$ as the intersection of countably many Hilbert spaces; with it, $\mathcal{S}$ is complete. When we speak of $\mathcal{S}(\mathbb{R}^n)$ elsewhere in this book, we always take its topology to be this one (or its higher-dimensional analogue). Similar comments can be made for $C^\infty_{cs}(\mathbb{R}^n)$ – see chapter I of [**244**] for details. With these new topologies, both $\mathcal{S}(\mathbb{R}^n)$ and $C^\infty_{cs}(\mathbb{R}^n)$ are examples of *nuclear spaces*;[8] although they are not themselves Hilbert spaces (the completeness in Definition 1.3.1 must be in terms of the norm topology), they behave in a more finite-dimensional way, as is indicated by the Spectral Theorem given below. See, for example, [**244**] for more on nuclear spaces.

The Spectral Theorem tells us in which sense we can diagonalise self-adjoint and unitary operators. To state it precisely, we need a small generalisation of the construction of $\ell^2(\infty)$. Consider any measure space (e.g. $X = \mathbb{R}$ or S^1 with Lebesgue measure μ). Fix $n = 1, 2, \ldots, \infty$, and suppose that for each $x \in X$ there is associated a copy $\mathcal{H}_n$ of $\mathbb{C}^n$ or (if $n = \infty$) $\ell^2(\infty)$. We want to define the (orthogonal) *direct integral* over $x \in X$ of these $\mathcal{H}_n$'s. Consider all functions $h : X \to \mathcal{H}_n, x \mapsto h_x$ that aren't too wild and that obey the finiteness condition $\int_X \|h_x\|^2 \, \mathrm{d}\mu < \infty$. As usual, we identify two such functions h, g if they agree everywhere except on a subset of X of μ-measure 0. Defining a Hermitian form by $\langle h, g\rangle = \int_X \langle h_x, g_x\rangle \, \mathrm{d}\mu$, the set of all such (equivalence classes of) h constitutes a Hilbert space denoted $\int_X \mathcal{H}_n \, \mathrm{d}\mu$ (completeness is proved as for $\ell^2(\infty)$). It is trivial to drop the requirement that the separable space $\mathcal{H}_n$ be fixed – see, for example, chapter 2 of [**67**] for details of the direct integral $\int_X \mathcal{H}(x) \, \mathrm{d}\mu$.

In finite dimensions any self-adjoint operator is diagonalisable. This fails in infinite dimensions, for example both the 'momentum operator' $\mathrm{i}\frac{\partial}{\partial x}$ and the 'position operator' $f(x) \mapsto x\, f(x)$ are self-adjoint on the dense subspace $\mathcal{S}(\mathbb{R})$ of $L^2(\mathbb{R})$, but neither have any eigenvectors anywhere in $L^2(\mathbb{R})$. So we need to generalise eigen-theory.

The statement of the Spectral Theorem simplifies when our operators act on $\mathcal{S}$. So let $T : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$ be linear. Diagonalising T would mean finding a basis for $\mathcal{S}$ consisting of eigenvectors of T. We can't do that, but we get something almost as good. By a *generalised eigenvector* corresponding to the *generalised eigenvalue* $\lambda \in \mathbb{C}$, we mean a tempered distribution $F \in \mathcal{S}^*$ such that $F(T\varphi) = \lambda\, F(\varphi)$ for all $\varphi \in \mathcal{S}$. For each λ, let $E_\lambda \subset \mathcal{S}^*$ be the generalised eigenspace consisting of all such F. We say that the set of all generalised eigenvectors $\cup_\lambda E_\lambda$ is *complete* if they distinguish all vectors in $\mathcal{S}$, i.e. if, for any $\varphi, \varphi' \in \mathcal{S}$, we have $F(\varphi) = F(\varphi')$ for all generalised eigenvectors $F \in \cup_\lambda E_\lambda$ iff $\varphi = \varphi'$.

Theorem 1.3.2 (Spectral Theorem)
(a) *Let $U : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$ be unitary. Then U extends uniquely to a unitary operator on all of $L^2(\mathbb{R}^n)$. All generalised eigenvalues λ lie on the unit circle $|\lambda| = 1$. We can express $L^2(\mathbb{R}^n)$ as a direct integral $\int_{|\lambda|=1} \mathcal{H}(\lambda) \, \mathrm{d}\mu(\lambda)$ of Hilbert spaces $\mathcal{H}(\lambda) \subseteq E_\lambda$, so*

[8] Nuclear spaces were first formulated by Grothendieck, who began his mathematical life as a functional analyst before revolutionising algebraic geometry. The term 'nuclear' comes from 'noyau' (French for both 'nucleus' and 'kernel'), since the *Kernel Theorem* is a fundamental result holding for them. The 'L' in both ℓ^2 and L^2 is in honour of Lebesgue, and the symbol $\mathcal{S}$ honours Schwartz.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

that U sends the function $h \in L^2(\mathbb{R}^n)$ to the function Uh with λ-component $(Uh)_\lambda = \lambda h_\lambda \in \mathcal{H}(\lambda)$. Moreover, the generalised eigenvectors are complete.
(b) *Suppose $A : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$ is self-adjoint. Then all generalised eigenvalues λ lie on the real line $\mathbb{R}$. We can express $L^2(\mathbb{R}^n)$ as a direct integral $\int_{-\infty}^{\infty} \mathcal{H}(\lambda)\, \mathrm{d}\mu(\lambda)$ of Hilbert spaces $\mathcal{H}(\lambda) \subseteq E_\lambda$, so that for each $h \in \mathcal{S}(\mathbb{R}^n)$, Ah has λ-component $(Ah)_\lambda = \lambda h_\lambda$. Moreover, the generalised eigenvectors are complete.*

For a simple example, consider the linear map $U : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ acting by translation: $(Uf)(x) = f(x+1)$. This is unitary, but it has no true eigenvectors in L^2. On the other hand, each point $\lambda = e^{\mathrm{i}y}$ on the unit circle is a generalised eigenvalue, corresponding to generalised eigenvector F_λ given by $F_\lambda(\varphi) = \int_{-\infty}^{\infty} e^{-\mathrm{i}yx} \varphi(x)\, \mathrm{d}x$. The direct integral interpretation of L^2 corresponds to the association of any $f(x) \in L^2$ with its Fourier transform $f_\lambda = \widehat{f}(y) = \int_{-\infty}^{\infty} e^{\mathrm{i}yx} f(x)\, \mathrm{d}x$. The completeness of the generalised eigenvectors is implied by the Plancherel identity

$$\int |f(x)|^2 \mathrm{d}x = \frac{1}{2\pi} \int |\widehat{f}(y)|^2 \mathrm{d}y. \tag{1.3.3}$$

The Spectral Theorem as formulated also holds for C^∞_{cs} in place of $\mathcal{S}$, and more generally for any *rigged* (or *equipped*) *Hilbert space* $\mathcal{V} \subset \mathcal{H} \subset \mathcal{V}^*$, where $\mathcal{H}$ is separable and $\mathcal{V}$ is nuclear (chapter I of [**244**]). They help provide a mathematically elegant formulation of quantum theories.

1.3.2 Factors

von Neumann algebras (see e.g. [**319**], [**177**]) can be thought of as symmetries of a (generally infinite) group. Their building blocks are called *factors*. Vaughn Jones initiated the combinatorial study of *subfactors* N of M (i.e. inclusions $N \subseteq M$ where M, N are factors), relating it to, for example, knots, and for this won a Fields medal in 1990. In Section 6.2.6 we describe Jones's work and the subsequent developments; this subsection provides the necessary background. Our emphasis is on accessibility.

Let $\mathcal{H}$ be a (separable complex) Hilbert space. By $\mathcal{L}(\mathcal{H})$ we mean the algebra of all bounded operators on $\mathcal{H}$ (we write '1' for the identity). For example, $\mathcal{L}(\mathbb{C}^n)$ is the space $M_n(\mathbb{C})$ of all $n \times n$ complex matrices. Let '$*$' be the adjoint (defined in the last subsection). Given a set S of bounded operators, denote by S' its *commutant*, that is the set of all bounded operators $x \in \mathcal{L}(\mathcal{H})$ that commute with all $y \in S$: $xy = yx$. We write $S'' := (S')'$ for the commutant of the commutant – clearly, $S \subseteq S''$.

Definition 1.3.3 *A* von Neumann algebra *M is a subalgebra of $\mathcal{L}(\mathcal{H})$ containing the identity* 1*, which obeys $M = M^*$ and $M = M''$.*

This is like defining a group by a representation. A von Neumann algebra can also be defined abstractly, which is equivalent except that (as we will see shortly) the natural notions of isomorphism are different in the concrete and abstract settings (just as the same group can have non-isomorphic representations).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Of course $\mathcal{L}(\mathcal{H})$ is a von Neumann algebra. Given any subset $S \subset \mathcal{L}(\mathcal{H})$ with $S^* = S$, the double-commutant S'' is a von Neumann algebra, namely the smallest one containing S. The space $L^\infty(\mathbb{R})$ of bounded functions $f : \mathbb{R} \to \mathbb{C}$ forms an abelian von Neumann algebra on the Hilbert space $\mathcal{H} = L^2(\mathbb{R})$ by pointwise multiplication. More generally (replacing $\mathbb{R}$ with any other measure space X and allowing multiple copies of the Hilbert space $L^2(X)$), all abelian von Neumann algebras are of that form.

The centre $Z(M) = M \cap M'$ of a von Neumann algebra M is an abelian one. Using the above characterisation $Z(M) = L^\infty(X)$, we can write M as a direct integral $\int_X M(\lambda)\,\mathrm{d}\lambda$ of von Neumann algebras $M(\lambda)$ with trivial centre: $Z(M(\lambda)) = \mathbb{C}1$. The direct integral, discussed last subsection, is a continuous analogue of direct sum.

Definition 1.3.4 *A* factor M *is a von Neumann algebra with centre* $Z(M) = \mathbb{C}I$.

Thus the study of von Neumann algebras is reduced to that of factors – the simple building blocks of any von Neumann algebra. $\mathcal{L}(\mathcal{H})$ is a factor. In finite dimensions, any (concrete) factor is of the form $M_n(\mathbb{C}) \otimes \mathbb{C}I_m$ acting in the Hilbert space $\mathbb{C}^n \otimes \mathbb{C}^m$ ('I_m' is the $m \times m$ identity matrix). Whenever the factor is (abstract) isomorphic to some $\mathcal{L}(\mathcal{H})$, its concrete realisation will have a similar tensor product structure, which is the source of the name 'factor'. In quantum field theory, where von Neumann algebras arise as algebras of operators (Section 4.2.4), a factor means there is no observable that can be measured simultaneously (with infinite precision) with all others.

The richness of the theory is because there are other factors besides $\mathcal{L}(\mathcal{H})$. In particular, factors fall into different families:

Type I_n: the factors (abstract) isomorphic to $\mathcal{L}(\mathcal{H})$ ($n = \dim \mathcal{H}$).
Type II_1: infinite-dimensional but it has a *trace* (i.e. a linear functional $\mathrm{tr} : M \to \mathbb{C}$ such that $\mathrm{tr}(xy) = \mathrm{tr}(yx)$).
Type II_∞: the factors isomorphic to $\mathrm{II}_1 \otimes \mathcal{L}(\mathcal{H})$.
Type III: everything else.

Choosing the normalisation $\mathrm{tr}(1) = 1$, the type II_1 trace will be unique. This is a very coarse-grained breakdown, and in fact the complete classification of factors is not known. There are uncountably many inequivalent type II_1 factors. Type III is further subdivided into families III_λ for all $0 \leq \lambda \leq 1$. von Neumann regarded the type III factors as pathological, but this was unfair (see Section 6.2.6). Almost every factor is isomorphic to type III_1 (i.e. perturbing an infinite-dimensional factor typically gives you one of type III_1). *Hyperfinite* factors are limits in some sense of finite-dimensional factors. There is a unique (abstract) hyperfinite factor of type II_1, II_∞ and III_λ for $0 < \lambda \leq 1$; we are interested in the hyperfinite II_1 and III_1 factors. Incidentally, the von Neumann algebras arising in quantum field theory are always of type III_1.

Discrete groups impinge on the theory through the *crossed-product construction* of factors. Start with any von Neumann algebra $M \subset \mathcal{L}(\mathcal{H})$, and let G be a discrete group acting on M (so $g.(xy) = (g.x)(g.y)$ and $g.x^* = (g.x)^*$). Let $\mathcal{H}_G = \mathcal{H} \otimes \ell^2(G)$ be the Hilbert space consisting of all column vectors $\zeta = (\zeta_g)_{g \in G}$ with entries $\zeta_g \in \mathcal{H}$ and

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

obeying $\sum_{g\in G}\|\zeta_g\|^2 < \infty$. M acts on $\mathcal{H}_G$ by $\zeta \mapsto \pi(x)(\zeta)$, where π is defined by

$$(\pi(x)(\zeta))_g := (g^{-1}.x)\zeta_g. \tag{1.3.4a}$$

In (1.3.4a), $g^{-1}.x$ is the action of G on M, and $g^{-1}.x \in M \subset \mathcal{L}(\mathcal{H})$ acts on $\zeta_g \in \mathcal{H}$ by definition. Likewise, G acts on $\mathcal{H}_G$ by $\zeta \mapsto \lambda(h)(\zeta)$, where λ is defined by

$$(\lambda(h)(\zeta))_g := \zeta_{h^{-1}g}. \tag{1.3.4b}$$

We can regard π and λ as embedding M and G in $\mathcal{L}(\mathcal{H}_G)$. The crossed-product is simply the smallest von Neumann algebra containing both these images:

$$M \rtimes G := (\pi(M) \cup \lambda(G))''. \tag{1.3.4c}$$

More explicitly (using the obvious orthonormal basis), any bounded operator $\tilde{y} \in \mathcal{L}(\mathcal{H}_G)$ is a matrix $\tilde{y} = (\tilde{y}_{g,h})$ with entries $\tilde{y}_{g,h} \in \mathcal{L}(\mathcal{H})$ for $g, h \in G$, and where $(\tilde{y}\zeta)_g = \sum_{h\in G} \tilde{y}_{g,h}\,\zeta_h$ (defining the infinite sum on the right appropriately [**319**]). Then for all $x \in M$ and $g, h, k \in G$, we get the matrix entries

$$\pi(x)_{g,h} = \delta_{g,h}\, h^{-1}.x,$$
$$\lambda(k)_{g,h} = \delta_{g,kh}\, 1.$$

The crossed-product is now a space of functions $y : G \to M$:

$$M \rtimes G \cong \{y : G \to M \mid \exists \tilde{y} \in \mathcal{L}(\mathcal{H}_G) \text{ such that } \tilde{y}_{g,h} = h^{-1}.(y_{gh^{-1}})\ \forall g, h \in G\} \tag{1.3.4d}$$

(see lemma 1.3.1 of [**319**]). In this notation the algebra structure of $M \rtimes G$ is given by

$$(xy)(g) = \sum_{h\in G} h^{-1}.(x_{gh^{-1}}\, y_h), \tag{1.3.4e}$$

$$(y^*)_g = g^{-1}.(y_{g^{-1}})^*. \tag{1.3.4f}$$

Crossed-products allow for elegant constructions of factors. For example, the (von Neumann) group algebra $\mathbb{C} \rtimes G$ is type II_1, for any discrete group G acting trivially on $\mathbb{C}$ and with the property that all of its conjugacy classes (apart from $\{e\}$) are infinite (examples of such G are the free groups $\mathcal{F}_n$ or $\mathrm{PSL}_2(\mathbb{Z})$). Also, any type III_1 factor is of the form $M \rtimes \mathbb{R}$, where M is type II_∞ and the $\mathbb{R}$ action scales the trace.

A proper treatment of factors (which this subsection is not) would involve *projections* onto closed subspaces, that is elements $p \in M$ satisfying $p = p^* = p^2$. These span (in the appropriate sense) the full von Neumann algebra. In the case of $M = M_n(\mathbb{C})$, the projections are precisely the orthogonal projections onto subspaces of $\mathbb{C}^n$, and thus have a well-defined dimension (namely the dimension of that subspace, so some integer between 0 and n). Remarkably, the same applies to any projection in any factor. For type II_1 this 'dimension' $\dim(p)$ is the trace $\tau(p)$, which we can normalise so that $\tau(1) = 1$. Then we get that $\dim(p)$ continuously fills out the interval $[0, 1]$. For type II_∞, the dimensions fill out $[0, \infty]$. For type III, every nonzero projection is equivalent (in a certain sense) to the identity and so the (normalised) dimensions are either 0 or 1.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Finally, one can ask for the relation between the abstract and concrete definitions of M – in other words, given a factor M, what are the different representations (= modules) of M, that is realisations of M as bounded operators on a Hilbert space $\mathcal{H}$. For example, for M of type I_n, these are of the form $M \otimes \mathbb{C}^M = M \oplus \cdots \oplus M$ (m times) for m finite, as well as $M \otimes \ell^\infty$. We see the type I_n modules are in one-to-one correspondence with the 'multiplicity' $m \in \{0, 1, \ldots, \infty\}$, which we can denote $\dim_M(\mathcal{H})$ and think of as $\dim(\mathcal{H})/\dim(M)$, at least when M is finite-dimensional. There is a similar result for type II: for each choice $d \in [0, \infty]$ there is a unique module $\mathcal{H}_d$, and any module $\mathcal{H}$ is equivalent to a unique $\mathcal{H}_d$. Finally, any two nontrivial representations of a type III factor will be equivalent. For a general definition of $\dim_M(\mathcal{H})$ and a proof of this representation theory, see theorem 2.1.6 in [**319**].

For type II_1, this parameter $d =: \dim_M(\mathcal{H})$ is sometimes called by von Neumann's unenlightening name 'coupling constant'. Incidentally, $\mathcal{H}_1$ is constructed in Question 1.3.6.

Question 1.3.1. (a) Verify explicitly that the position $f(x) \mapsto xf(x)$ and momentum $\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}x}$ operators are neither bounded nor continuous, for the Hilbert space $L^2(\mathbb{R})$.
(b) Verify explicitly that the position operator of (a) is not defined everywhere.

Question 1.3.2. Consider the shift operator $S(x_1, x_2, \ldots) = (0, x_1, x_2, \ldots)$ in $\ell^2(\infty)$. Verify that $S^*S = 1$ but $SS^* \neq 1$.

Question 1.3.3. Apply the Spectral Theorem to the momentum operator $\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}x}$.

Question 1.3.4. Let $\mathcal{V} = \{f \in C^\infty(S^1) \mid f(0) = 0\}$.
(a) Verify that $\mathcal{V}$ is dense in $\mathcal{H} = L^2(S^1)$.
(b) Verify that $D = \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}\theta}$ obeys $\langle Df, g\rangle = \langle f, Dg\rangle$ for all $f, g \in \mathcal{V}$.
(c) Construct the adjoint D^* of $D : \mathcal{V} \to \mathcal{H}$. Is D self-adjoint?
(d) For each $\lambda \in \mathbb{C}$, define $\mathcal{V}_\lambda$ to be the extension of $\mathcal{V}$ consisting of all functions smooth on the interval $[0, 2\pi]$ and with $f(0) = \lambda f(2\pi)$. Extend D in the obvious way to $\mathcal{V}_\lambda$. For which λ is D now self-adjoint?

Question 1.3.5. Let the free group $\mathcal{F}_2$ act trivially on $\mathbb{C}$. Find a trace for $\mathbb{C} \rtimes \mathcal{F}_2$. What is the centre of $\mathbb{C} \rtimes \mathcal{F}_2$?

Question 1.3.6. Let M be type II_1. Prove M is a pre-Hilbert space by defining $\langle x, y\rangle$ appropriately (*Hint*: use the trace). Let $L^2(M)$ be its completion. Show that $L^2(M)$ is a module over M.

1.4 Lie groups and Lie algebras

Undergraduates are often disturbed (indeed, reluctant) to learn that the vector-product $u \times v$ really only works in three dimensions. Of course, there are several generalisations to other dimensions: for example an antisymmetric $(N-1)$-ary product (a determinant) in N dimensions, or the wedge product of k-forms in $2k+1$ dimensions. Arguably the most fruitful generalisation is that of a Lie algebra, defined below. They are the tangent spaces of those differential manifolds whose points can be 'multiplied' together.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

As we know, much of algebra is developed by analogy with elementary properties of integers. For a finite-dimensional Lie algebra, a divisor is called an ideal; a prime is called simple; and multiplying corresponds to semi-direct sum (Lie algebras behave simpler than groups but not as simple as numbers). In particular, simple Lie algebras are important for similar reasons that simple groups are, and can also be classified (with *much* less effort). One non-obvious discovery is that they are rigid: the best way to capture the structure of a simple Lie algebra is through a graph. We push this thought further in Section 3.3. For an elementary introduction to Lie theory, [**92**] is highly recommended.

1.4.1 Definition and examples of Lie algebras

An *algebra* is a vector space with a way to multiply vectors that is compatible with the vector space structure (i.e. the vector-valued product is required to be *bilinear*: $(au + a'u') \times (bv + b'v') = ab\,u \times v + ab'\,u \times v' + a'b\,u' \times v + a'b'\,u' \times v'$). For example, the complex numbers $\mathbb{C}$ form a two-dimensional algebra over $\mathbb{R}$ (a basis is 1 and $\mathrm{i} = \sqrt{-1}$; the *scalars* here are real numbers and the *vectors* are complex numbers). The quaternions are four-dimensional over $\mathbb{R}$ and the octonions are eight-dimensional over $\mathbb{R}$. Incidentally, these are the only finite-dimensional normed algebras over $\mathbb{R}$ that obey the cancellation law: $u \neq 0$ and $u \times v = 0$ implies $v = 0$ (does the vector-product of $\mathbb{R}^3$ fail the cancellation law?). This important little fact makes several unexpected appearances [**29**]. For instance, imagine a ball (i.e. S^2) covered in hair. No matter how you comb it, there will be a part in the hair, or at least a point where the hair leaves in all directions, or some such problem. More precisely, there is no continuous nowhere-zero vector field on S^2. On the other hand, it is trivial to comb the hair on the circle S^1 without singularity: just comb it clockwise, for example. More generally, the even spheres S^{2k} can never be combed. Now try something more difficult: place k wigs on S^k, and try to comb all k of them so that at each point on S^k the k hairs are linearly independent. This is equivalent to saying that the tangent bundle $T S^k$ equals $S^k \times \mathbb{R}^k$. The only k-spheres S^k that can be 'k-combed' in this way (i.e. for which there exist k linearly independent continuous vector fields) are for $k = 1$, 3 and 7. This is intimately connected with the existence of $\mathbb{C}$, the quaternions and octonions (namely, S^1, S^3 and S^7 are the length 1 complex numbers, quaternions and octonions, respectively) [**104**].

Definition 1.4.1 *A* Lie algebra $\mathfrak{g}$ *is an algebra with product (usually called a 'bracket' and written* $[xy]$*) that is both 'anti-commutative' and 'anti-associative':*

$$[xy] + [yx] = 0; \tag{1.4.1a}$$

$$[x[yz]] + [y[zx]] + [z[xy]] = 0. \tag{1.4.1b}$$

Like most other identities in mathematics, (1.4.1b) is named after Jacobi (although he died years before Lie theory was created). Usually we consider Lie algebras over $\mathbb{C}$, but sometimes over $\mathbb{R}$. Note that (1.4.1a) is equivalent to demanding $[xx] = 0$ (except for fields of characteristic 2).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

A homomorphism $\varphi : \mathfrak{g}_1 \to \mathfrak{g}_2$ between Lie algebras must preserve the linear structure as well as the bracket – i.e. φ is linear and $\varphi[xy] = [\varphi(x)\varphi(y)]$ for all $x, y \in \mathfrak{g}_1$. If φ is in addition invertible, we call $\mathfrak{g}_1, \mathfrak{g}_2$ *isomorphic*.

One important consequence of bilinearity is that it is enough to know the values of all the brackets $[x^{(i)}x^{(j)}]$ for $i < j$, for any basis $\{x^{(1)}, x^{(2)}, \ldots\}$ of the vector space $\mathfrak{g}$. (The reader should convince himself of this before proceeding.)

A trivial example of a Lie algebra is a vector space $\mathfrak{g}$ with a bracket identically 0: $[xy] = 0$ for all $x, y \in \mathfrak{g}$. Any such Lie algebra is called *abelian*, because in any representation (i.e. realisation by matrices) its matrices will commute. Abelian Lie algebras of equal dimension are isomorphic.

In fact, the only one-dimensional Lie algebra (for any choice of field $\mathbb{F}$) is the abelian one $\mathfrak{g} = \mathbb{F}$. It is straightforward to find all two- and three-dimensional Lie algebras (over $\mathbb{C}$) up to isomorphism: there are precisely two and six of them, respectively (though one of the six depends on a complex parameter). Over $\mathbb{R}$, there are two and nine (with two of the latter depending on real parameters). This exercise cannot be continued much further – for example, not all seven-dimensional Lie algebras (over $\mathbb{C}$ say) are known. Nor is it obvious that this would be a valuable exercise. We should suspect that our definition of Lie algebra is probably too general for anything obeying it to be automatically interesting. Most commonly, a classification yields a stale and useless list – a phone book more than a tourist guide.

Two of the three-dimensional Lie algebras are important in what follows. One of them is well known to the reader: the vector-product in $\mathbb{C}^3$. Taking the standard basis $\{e_1, e_2, e_3\}$ of $\mathbb{C}^3$, the bracket can be defined by the relations

$$[e_1e_2] = e_3, \qquad [e_1e_3] = -e_2, \qquad [e_2e_3] = e_1. \tag{1.4.2a}$$

This algebra, denoted A_1 or $\mathfrak{sl}_2(\mathbb{C})$, deserves the name 'mother of all Lie algebras' (Section 1.4.3). Its more familiar realisation uses a basis $\{e, f, h\}$ with relations

$$[ef] = h, \qquad [he] = 2e, \qquad [hf] = -2f. \tag{1.4.2b}$$

The reader can find the change-of-basis (valid over $\mathbb{C}$ but not $\mathbb{R}$) showing that equations (1.4.2) define isomorphic *complex* (though not *real*) Lie algebras.

Another important three-dimensional Lie algebra is the Heisenberg algebra[9] $\mathfrak{Heis}$, the algebra of the canonical commutation relations in quantum mechanics, defined by

$$[xp] = h, \qquad [xh] = [ph] = 0. \tag{1.4.3}$$

The most basic source of Lie algebras are the $n \times n$ matrices with *commutator*:

$$[AB] = [A, B] := AB - BA \tag{1.4.4}$$

(the reader can verify that the commutator always obeys (1.4.1)). Let $\mathfrak{gl}_n(\mathbb{R})$ (respectively $\mathfrak{gl}_n(\mathbb{C})$) denote the Lie algebra of all $n \times n$ matrices with coefficients in $\mathbb{R}$ (respectively

[9] There actually is a family of 'Heisenberg algebras', with (1.4.3) being the one of least dimension.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

$\mathbb{C}$), with Lie bracket given by (1.4.4). More generally, if $\mathcal{A}$ is any associative algebra, then $\mathcal{A}$ becomes a Lie algebra by defining the bracket $[xy] = xy - yx$.

Another general construction of Lie algebras starts with any (not necessarily associative or commutative) algebra $\mathcal{A}$. By a *derivation* of $\mathcal{A}$, we mean any linear map $\delta : \mathcal{A} \to \mathcal{A}$ obeying the Leibniz rule $\delta(ab) = \delta(a)\, b + a\, \delta(b)$. We can compose derivations, but in general the result $\delta_1 \circ \delta_2$ won't be a derivation. However, an easy calculation verifies that the commutator $[\delta_1 \delta_2] = \delta_1 \circ \delta_2 - \delta_2 \circ \delta_1$ of derivations is also a derivation. Hence the vector space of derivations is naturally a Lie algebra. If $\mathcal{A}$ is finite-dimensional, so will be its Lie algebra of derivations.

In particular, vector fields $X \in \mathrm{Vect}(M)$ are derivations. We can compose them $X \circ Y$, but this results in a second-order differential operator. Instead, the natural 'product' is their commutator $[X, Y] = X \circ Y - Y \circ X$, as it always results in a vector field. $\mathrm{Vect}(M)$ with this bracket is an infinite-dimensional Lie algebra. For example, recall $\mathrm{Vect}(S^1)$ from Section 1.2.2 and compare

$$\left(f(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\right) \circ \left(g(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\right) = f(\theta)\, g(\theta)\, \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} + f(\theta)\, g'(\theta)\, \frac{\mathrm{d}}{\mathrm{d}\theta},$$

$$\left[f(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}, g(\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\right] = (f(\theta)\, g'(\theta) - f'(\theta)\, g(\theta))\frac{\mathrm{d}}{\mathrm{d}\theta}.$$

Incidentally, another natural way to multiply vector fields X, Y of vector fields, the Lie derivative $\mathcal{L}_X(Y)$ defined in Section 1.2.2, equals the commutator $[X, Y]$ and so gives the same Lie algebra structure on $\mathrm{Vect}(M)$.

1.4.2 Their motivation: Lie groups

From Definition 1.4.1 it is far from clear that Lie algebras, as a class, should be natural and worth studying. After all, there are infinitely many possible axiomatic systems: why should this one be anything special *a priori*? Perhaps the answer could have been anticipated by the following line of reasoning.

Axiom *Groups are important and interesting.*
Axiom *Manifolds are important and interesting.*

Definition 1.4.2 *A* Lie group *G is a manifold with a compatible group structure.*

This means that 'multiplication' $\mu : G \times G \to G$ (which sends the pair (a, b) to ab) and 'inverse' $\iota : G \to G$ (which sends a to a^{-1}) are both differentiable maps. The manifold structure (Definition 1.2.3) of G can be chosen as follows: fix any open set U_e about the identity $e \in G$; then the open set $U_g := gU_e$ will contain $g \in G$. The real line $\mathbb{R}$ is a Lie group under addition: obviously, μ and ι defined by $\mu(a, b) = a + b$ and $\iota(a) = -a$ are both differentiable. A circle is also a Lie group: parametrise the points with the angle θ defined mod 2π; the 'product' of the point at angle θ_1 with the point at angle θ_2 is the point at angle $\theta_1 + \theta_2$. Surprisingly, the only other k-sphere that is a Lie group is S^3

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

(the product can be defined using quaternions of unit length,[10] or by identifying S^3 with the matrix group $\mathrm{SU}_2(\mathbb{C})$). This is because it is always possible to 'n-comb the hair' on an n-dimensional Lie group (Section 1.4.1) – more precisely, the tangent bundle TG of any Lie group is trivial $G \times \mathbb{R}^n$, something easy to see using the charts U_g.

A *complex* Lie group G is a complex manifold with a compatible group structure. For example, the only one-dimensional compact real Lie group is S^1, whereas there are infinitely many compact one-dimensional complex Lie groups, namely the tori or 'elliptic curves' $\mathbb{C}/L$, for any two-dimensional lattice L in the plane $\mathbb{C}$. Thought of as real Lie groups (i.e. forgetting their complex structure), elliptic curves all are real-diffeomorphic to $S^1 \times S^1$; they differ in their complex-differential structure. We largely ignore the complex Lie groups; unless otherwise stated, by 'Lie group' we mean 'real Lie group'.[11]

Many but not all Lie groups can be expressed as matrix groups whose operation is matrix multiplication. The most important are GL_n (invertible $n \times n$ matrices) and SL_n (ones with determinant 1).

Incidentally, Hilbert's 5th problem[12] asked how important the differentiability hypothesis is here. It turns out it isn't (see [**569**] for a review): if a group G is a topological manifold, and μ and ι are merely continuous, then it is possible to endow G with a differentiable structure in one and only one way so that μ and ι are differentiable.

In any case, a consequence of the above axioms is surely:

Corollary *Lie groups should be important and interesting.*

Indeed, Lie groups appear throughout mathematics and physics, as we will see again and again. For example, the Lie groups of relativistic physics (Section 4.1.2) come from the group $\mathrm{O}_{3,1}(\mathbb{R})$ consisting of all 4×4 matrices Λ obeying $\Lambda G \Lambda^t = G$, where $G = \mathrm{diag}(1, 1, 1, -1)$ is the Minkowski metric. Any such Λ must have determinant ± 1, and has $|\Lambda_{44}| \geq 1$; these 2×2 possibilities define the four connected components of $\mathrm{O}_{3,1}(\mathbb{R})$. The (restricted) *Lorentz group* $\mathrm{SO}^+_{3,1}(\mathbb{R})$ consists of the determinant 1 matrices Λ in $\mathrm{O}_{3,1}(\mathbb{R})$ with $\Lambda_{44} \geq 1$. It describes rotations in 3-space, as well as 'boosts' (changes of velocity). $\mathrm{SO}^+_{3,1}(\mathbb{R})$ has a double-cover (i.e. an extension by $\mathbb{Z}_2$) isomorphic to $\mathrm{SL}_2(\mathbb{C})$, which is more fundamental. Finally, the *Poincaré group* is the semi-direct product of $\mathrm{SO}^+_{3,1}(\mathbb{R})$ with $\mathbb{R}^4$, corresponding to adjoining to $\mathrm{SO}^+_{3,1}(\mathbb{R})$ the translations in space-time $\mathbb{R}^4$. The Lorentz group is six-dimensional, while the Poincaré group is 10-dimensional.

As said in Section 1.2.2, the tangent spaces of manifolds are vector spaces of dimension equal to that of the manifold. The space structure is easy to see for Lie groups: choose any infinitesimal curves $u = \langle g(t) \rangle_e$, $v = \langle h(t) \rangle_e \in T_e G$, so $g(0) = h(0) = e$, and let $a, b \in \mathbb{R}$. Then $au + bv$ corresponds to the curve $t \mapsto g(at)\, h(bt)$.

Not surprisingly, G acts on the tangent vectors: let $u \in T_h G$ correspond to curve $h(t)$, with $h(0) = h$, and define gu for any $g \in G$ to be the vector in $T_{gh} G$ corresponding

[10] Similarly, the 7-sphere inherits from the octonions a *non-associative* (hence nongroup) product, compatible with its manifold structure.

[11] Our vector spaces (e.g. Lie algebras) are usually complex; our manifolds (e.g. Lie groups) are usually real.

[12] In the International Congress of Mathematicians in 1899, David Hilbert announced several problems chosen to anticipate (and direct) major areas of study. His list was deeply influential.

to the curve $t \mapsto g(h(t))$. This means that conjugating gug^{-1} for any element $u \in T_eG$ gives another element of T_eG, that is T_eG carries a representation of the group G called the adjoint representation.

This is all fine. However, we have a rich structure on our manifold – namely the group structure – and it would be deathly disappointing if this adjoint representation were the high-point of the theory. Fortunately we can go *much* further. Consider any $u, v \in T_eG$, where $v = \langle g(t) \rangle_e$. Then $g(t)\, u\, g(t)^{-1}$ lies in the vector space T_eG for all t, and hence so will the derivative. It turns out that the quantity

$$[uv] := \frac{\mathrm{d}}{\mathrm{d}t}(g(t)\, u\, g(t)^{-1})|_{t=0} \tag{1.4.5}$$

depends only on u and v (hence the notation). A little work shows that it is bilinear, anti-symmetric, and anti-associative. That is, T_eG is a Lie algebra!

In the last subsection we indicated that $\mathrm{Vect}(M)$ carries a Lie algebra structure, for any manifold M. It is tempting to ask: when M is a Lie group G, what is the relation between the infinite-dimensional Lie algebra $\mathrm{Vect}(G)$, and the finite-dimensional Lie algebra T_eG? Note that G acts on the space $\mathrm{Vect}(M)$ by 'left-translation', that is if X is a vector field, which we can think of as a derivation of the algebra $C^\infty(G)$ of real-valued functions on G, and $g \in G$, then $g.X$ is the vector field given by $(g.X)(f)(h) = X(f)(gh)$. Then the Lie algebra T_eG is isomorphic to the subalgebra of $\mathrm{Vect}(G)$ consisting of the 'left-invariant vector fields', that is those X obeying $g.X = X$. Given any manifold M, the Lie algebra $\mathrm{Vect}(M)$ corresponds to the infinite-dimensional Lie group $\mathrm{Diff}^+(M)$ of orientation-preserving diffeomorphisms of M; when M is itself a Lie group, the left-invariant vector fields correspond in $\mathrm{Diff}^+(M)$ to a copy of M given by left-multiplication.

Fact *The tangent space of a Lie group is a Lie algebra. Conversely, any (finite-dimensional real or complex) Lie algebra is the tangent space T_eG to some Lie group.*

For example, consider the Lie group $G = \mathrm{SL}_n(\mathbb{R})$. Let $A(t) = \big(A_{ij}(t)\big)$ be any curve in G with $A(0) = I_n$. We see that only one term in the expansion of $\det A(t)$ can contribute to its derivative at $t = 0$, namely the diagonal term $A_{11}(t) \cdots A_{nn}(t)$, so differentiating $\det(A(t)) = 1$ at $t = 0$ tells us that $A'_{11}(0) + \cdots + A'_{nn}(0) = 0$. Thus the tangent space $T_{I_n}G$ consists of all trace-zero $n \times n$ matrices, since the algebra like the group must be $(n^2 - 1)$-dimensional. We write it $\mathfrak{sl}_n(\mathbb{R})$. Now choose any matrices $U, V \in \mathfrak{sl}_n(\mathbb{R})$, and let $A(t)$ be the curve in $\mathrm{SL}_n(\mathbb{R})$ corresponding to V. Differentiating $A(t)\, A(t)^{-1} = I_n$, we see that $(A^{-1})' = -A^{-1}A'A^{-1}$ and thus (1.4.5) becomes

$$[V\, U] = A'(0)\, U\, I_n^{-1} + I_n\, U\, \left(-I_n^{-1}\, A'(0)\, I_n^{-1}\right).$$

In other words, the bracket in $\mathfrak{sl}_n(\mathbb{R})$ – as with any other matrix algebra – is given by the commutator (1.4.4).

Given the above fact, a safe guess would be:

Conjecture *Lie algebras are important and interesting.*

From this line of reasoning, it should be expected that historically Lie groups arose first. Indeed that is the case: Sophus Lie introduced them in 1873 to try to develop a Galois

theory for ordinary differential equations. Galois theory can be used for instance to show that not all fifth degree (or higher) polynomials can be explicitly 'solved' using radicals (Section 1.7.2). Lie wanted to study the explicit solvability (integrability) of differential equations, and this led him to develop what we now call Lie theory. The importance of Lie groups, however, has grown well beyond this initial motivation.

A Lie algebra, being a linearised Lie group, is much simpler and easier to handle. The algebra preserves the local properties of the group, though it loses global topological properties (like compactness). A Lie group has a single Lie algebra, but a Lie algebra corresponds to many different Lie groups. The Lie algebra corresponding to both $\mathbb{R}$ and S^1 is $\mathfrak{g} = \mathbb{R}$ with trivial bracket. The Lie algebra corresponding to both $S^3 = \mathrm{SU}_2(\mathbb{C})$ and $\mathrm{SO}_3(\mathbb{R})$ is the vector-product algebra (1.4.2a) (usually called $\mathfrak{so}_3(\mathbb{R})$).

We saw earlier that many (but not all) examples of Lie groups are matrix groups, that is subgroups of $\mathrm{GL}_n(\mathbb{R})$ or $\mathrm{GL}_n(\mathbb{C})$. The *Ado–Iwasawa Theorem* (see e.g. chapter VI of [**314**]) says that all finite-dimensional Lie algebras (over any field) are realisable as Lie subalgebras of $\mathfrak{gl}_n(\mathbb{R})$ or $\mathfrak{gl}_n(\mathbb{C})$. This is analogous to Cayley's Theorem, which says any finite group is a subgroup of some symmetric group $\mathcal{S}_n$. Now, choose any Lie algebra $\mathfrak{g} \subseteq \mathfrak{gl}_n(\mathbb{C})$. Let G be the topological closure of the subgroup of $\mathrm{GL}_n(\mathbb{C})$ generated by all matrices e^A for $A \in \mathfrak{g}$, where e^A is defined by the Taylor expansion

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

Then the Lie group G has Lie algebra $\mathfrak{g}$. Remarkably, the group operation on G (at least close to the identity) can be deduced from the bracket: the first few terms of the *Baker–Campbell–Hausdorff formula* read

$$\exp(X)\,\exp(Y) = \exp\left(X + Y + \frac{1}{2}[XY] + \frac{1}{12}[[XY]X] + \frac{1}{12}[[XY]Y] + \cdots\right). \tag{1.4.6}$$

See, for example, [**475**] for the complete formula and some of its applications.

We saw earlier that the condition 'determinant = 1' for matrix groups translates to the Lie algebra condition 'trace = 0'. This also follows from the identity $\det(e^A) = e^{\mathrm{tr}\,A}$, which follows quickly from the Jordan canonical form of A.

Of course all undergraduates are familiar, at least implicitly, with exponentiating operators. Taylor's Theorem tells us that for any analytic function f and any real number a, the operator $e^{a\frac{\mathrm{d}}{\mathrm{d}x}}$ sends $f(x)$ to $f(x+a)$. Curiously, the operator $\log(\frac{\mathrm{d}}{\mathrm{d}x})$ also has a meaning, in the context of, for example, affine Kac–Moody algebras [**344**].

The definition of a Lie algebra makes sense over any field $\mathbb{K}$. However, the definition of Lie groups is much more restrictive, because they are analytic rather than merely linear and hence require fields like $\mathbb{C}$, $\mathbb{R}$ or the p-adic rationals $\widehat{\mathbb{Q}}_p$. A good question is: which Lie-like group structures do Lie algebras correspond to, for the other fields? A good answer is: algebraic groups, which are to algebraic geometry what Lie groups are to differential geometry. See, for example, part III of [**92**] for an introduction.

The main relationship between real Lie groups and algebras is summarised by:

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Theorem 1.4.3 *To any finite-dimensional real Lie algebra $\mathfrak{g}$, there is a unique connected simply-connected Lie group $\widetilde{G}$, called the universal cover group. If G is any other connected Lie group with Lie algebra $\mathfrak{g}$, then there exists a discrete subgroup H of the centre of $\widetilde{G}$, such that $G \cong \widetilde{G}/H$ and $H \cong \pi_1(G)$, the fundamental group of G.*

The definitions of simply-connected and π_1 are given in Section 1.2.3. The universal cover $\widetilde{\mathbb{R}}$ of the Lie algebra $\mathbb{R}$ is the additive group $\mathbb{R}$; the circle $G = S^1$ has the same Lie algebra and can be written as $S^1 \cong \mathbb{R}/\mathbb{Z}$. The real Lie groups $\mathrm{SU}_2(\mathbb{C})$ and $\mathrm{SO}_3(\mathbb{R})$ both have Lie algebra $\mathfrak{so}_3(\mathbb{R})$; $\mathrm{SU}_2(\mathbb{C}) \cong S^3$ is the universal cover, and $\mathrm{SO}_3(\mathbb{R}) \cong \mathrm{SU}_2(\mathbb{C})/\{\pm I_2\}$ is the 3-sphere with antipodal points identified. $\pi_1(\mathrm{SL}_2(\mathbb{R})) \cong S^1$, and its universal cover (see Question 2.4.4) is an example of a Lie group that is not a matrix group.

So the classification of (connected) Lie groups reduces to the much simpler classification of Lie algebras, together with the classification of discrete groups in the centre of the corresponding $\widetilde{G}$. The condition that G be connected is clearly necessary, as the direct product of a Lie group with any discrete group leaves the Lie algebra unchanged.

Lie group structure theory is merely a major generalisation of linear algebra. The basic constructions familiar to undergraduates have important analogues valid in many Lie groups. For instance, in our youth we were taught to solve linear equations and invert matrices by reducing a matrix to row-echelon form using row operations. This says that any matrix $A \in \mathrm{GL}_n(\mathbb{C})$ can be factorised $A = BPN$, where N is upper-triangular with 1's on the diagonal, P is a permutation matrix and B is an upper-triangular matrix. This is essentially the Bruhat decomposition of the Lie group $\mathrm{GL}_n(\mathbb{C})$. More generally (where it applies to any 'reductive' Lie group G), P will be an element of the so-called Weyl group of G, and B will be in a 'Borel subgroup'. For another example, everyone knows that any nonzero real number x can be written uniquely as $x = (\pm 1) \cdot |x|$, and many of us remember that any invertible matrix $A \in \mathrm{GL}_n(\mathbb{R})$ can be uniquely written as a product $A = OP$, where O is orthogonal and P is positive-definite. More generally, this is called the Cartan decomposition for a real semi-simple Lie group. This encourages us to interpret a linear algebra theorem as a special case of a Lie group theorem . . . a squirrel.

1.4.3 Simple Lie algebras

The reader already weary of such algebraic tedium won't be surprised to read that the typical algebraic definitions can be imposed on Lie theory. The analogue of direct product of groups here is direct sum $\mathfrak{g}_1 \oplus \mathfrak{g}_2$, with bracket $[(x_1, x_2), (y_1, y_2)] = ([x_1 y_1]_1, [x_2 y_2]_2)$. Semi-direct sum is defined as usual. The analogue of normal subgroup here is called an *ideal*: a subspace $\mathfrak{h}$ of $\mathfrak{g}$ such that $[\mathfrak{g}\mathfrak{h}] := \mathrm{span}\{[xy] \mid x \in \mathfrak{g}, y \in \mathfrak{h}\}$ is contained in $\mathfrak{h}$. A Lie group N is a normal subgroup of Lie group G iff the Lie algebra of N is an ideal of that of G. Given an ideal $\mathfrak{h}$ of a Lie algebra $\mathfrak{g}$, the quotient space $\mathfrak{g}/\mathfrak{h}$ has a natural Lie algebra structure; if $\varphi : \mathfrak{g}_1 \to \mathfrak{g}_2$ is a Lie algebra homomorphism, then the kernel $\ker(\varphi)$ is an ideal of $\mathfrak{g}_1$ and the image $\varphi(\mathfrak{g}_1)$ is a subalgebra of $\mathfrak{g}_2$ isomorphic to $\mathfrak{g}_1/\ker(\varphi)$. The name 'ideal' comes from number theory (Section 1.7.1). The centre $Z(\mathfrak{g}) := \{x \in \mathfrak{g} \mid [x\mathfrak{g}] = 0\}$ of $\mathfrak{g}$ always forms an ideal, as does $[\mathfrak{g}\mathfrak{g}]$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

A *simple* Lie algebra is one with no proper ideals. It is standard though to exclude the one-dimensional Lie algebras, much like is often done with the cyclic groups $\mathbb{Z}_p$. A *semi-simple* Lie algebra is defined as any $\mathfrak{g}$ for which $[\mathfrak{g}\mathfrak{g}] = \mathfrak{g}$; it turns out that $\mathfrak{g}$ is semi-simple iff $\mathfrak{g}$ is the (Lie algebra) direct sum $\oplus_i \mathfrak{g}_i$ of simple Lie algebras $\mathfrak{g}_i$. A *reductive* Lie algebra $\mathfrak{g}$ is defined by the relation $[\mathfrak{g}\mathfrak{g}] \oplus Z(\mathfrak{g}) = \mathfrak{g}$; $\mathfrak{g}$ is reductive iff $\mathfrak{g}$ is the direct sum of a semi-simple Lie algebra with an abelian one. Of course simple Lie algebras are more important, but semi-simple and reductive ones often behave similarly.

The finite-dimensional simple Lie algebras constitute an important class of Lie algebras. Although it is doubtful the reader has leapt out of his chair with surprise at this pronouncement, it is good to see explicit indications of this importance.

Simple Lie algebras serve as building blocks for all other finite-dimensional Lie algebras, in the following sense (called *Levi decomposition* – see, for example, chapter III.9 of [**314**] for a proof): any finite-dimensional Lie algebra $\mathfrak{g}$ over $\mathbb{C}$ or $\mathbb{R}$ can be written *as a vector space* in the form $\mathfrak{g} = \mathfrak{r} \oplus \mathfrak{h}$, where $\mathfrak{h}$ is the largest semi-simple Lie subalgebra of $\mathfrak{g}$, and $\mathfrak{r}$ is called the *radical* of $\mathfrak{g}$ and is by definition the maximal 'solvable' ideal of $\mathfrak{g}$. This means $\mathfrak{g}$ is the semi-direct sum of $\mathfrak{r}$ with $\mathfrak{h} \cong \mathfrak{g}/\mathfrak{r}$. A *solvable* Lie algebra is the repeated semi-direct sum by one-dimensional Lie algebras; more concretely, it is isomorphic to a subalgebra of the upper-triangular matrices in some $\mathfrak{gl}_n$. Levi decomposition is the Lie theoretic analogue of the Jordan–Hölder Theorem of Section 1.1.2.

It is reassuring that we can also see the importance of simple Lie algebras geometrically: given any finite-dimensional real Lie group that is 'compact' as a manifold (i.e. bounded and contains all its limit points), its Lie algebra is reductive. Conversely, any reductive real Lie algebra is the Lie algebra of a compact Lie group.

In our struggle to understand a structure, it is healthy to find new ways to capture old information. Let us begin with a canonical way to associate linear endomorphisms (which the basis-hungry of us can regard as square matrices) to elements of the Lie algebra $\mathfrak{g}$. Define the 'adjoint operator' $\operatorname{ad} x : \mathfrak{g} \to \mathfrak{g}$ to be the linear map given by $(\operatorname{ad} x)(y) = [xy]$. In this language, anti-associativity of the bracket translates to the facts that: (i) for each $x \in \mathfrak{g}$, $\operatorname{ad} x$ is a derivation of $\mathfrak{g}$; and (ii) the assignment $x \mapsto \operatorname{ad} x$ defines a 'representation' of $\mathfrak{g}$, called the *adjoint representation* (more on this next section).

The point is that there are basis-independent ways to get numbers out of matrices. The *Killing form* $\kappa : \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}$ of a (complex) Lie algebra $\mathfrak{g}$ is defined by

$$\kappa(x|y) := \operatorname{tr}(\operatorname{ad} x \circ \operatorname{ad} y), \qquad \forall x, y \in \mathfrak{g}. \tag{1.4.7a}$$

By 'trace' we mean to choose a basis, get matrices, and take the trace in the usual way; the answer is independent of the basis chosen. The Killing form is symmetric, respects the linear structure of $\mathfrak{g}$ (i.e. is bilinear) and respects the bracket in the sense that

$$\kappa([xy]|z) = \kappa(x|[yz]), \qquad \forall x, y, z \in \mathfrak{g}. \tag{1.4.7b}$$

This property of κ is called *invariance* (Question 1.4.6(b)).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Table 1.3. *Freudenthal's Magic Square: the Lie algebra* $\mathfrak{g}(\mathcal{A}_1, \mathcal{A}_2)$

$\mathcal{A}_i$	$\mathbb{R}$	$\mathbb{C}$	quat	oct
$\mathbb{R}$	$\mathfrak{so}_3(\mathbb{R})$	$\mathfrak{su}_3(\mathbb{R})$	$\mathfrak{sp}_3(\mathbb{R})$	F_4
$\mathbb{C}$	$\mathfrak{su}_3(\mathbb{R})$	$\mathfrak{su}_3(\mathbb{R}) \oplus \mathfrak{su}_3(\mathbb{R})$	$\mathfrak{su}_6(\mathbb{R})$	E_6
quat	$\mathfrak{sp}_3(\mathbb{R})$	$\mathfrak{su}_6(\mathbb{R})$	$\mathfrak{so}_{12}(\mathbb{R})$	E_7
oct	F_4	E_6	E_7	E_8

Let A, B be two $n \times n$ real matrices; then

$$\mathrm{tr}(AB) = \sum_{i=1}^{n} A_{ii} B_{ii} + \sum_{1 \le i < j \le n} (A_{ij} B_{ji} + A_{ji} B_{ij}),$$

which can be interpreted as an indefinite inner-product on $\mathbb{R}^{n^2}$. Thus the Killing form $\kappa(x|y)$ should be thought of as an inner-product on the vector space $\mathfrak{g}$. It arose historically by expanding the characteristic polynomial $\det(\mathrm{ad}\, x - \lambda I)$ (Question 1.4.6(c)).

An inner-product on a complex space V has only one invariant: the dimension of the subspace of null vectors. More precisely, define the radical of the Killing form to be

$$\mathfrak{s}(\kappa) := \{x \in \mathfrak{g} \mid \kappa(x|y) = 0 \ \forall y \in \mathfrak{g}\}.$$

By invariance of κ, $\mathfrak{s}$ is an ideal. It is always solvable.

Theorem 1.4.4 (Cartan's criterion) *Let* $\mathfrak{g}$ *be a (complex or real) finite-dimensional Lie algebra. Then* $\mathfrak{g}$ *is semi-simple iff* κ *is nondegenerate, i.e.* $\mathfrak{s}(\kappa) = 0$.

Moreover, $\mathfrak{g}$ is solvable iff $[\mathfrak{g}\mathfrak{g}] \subseteq \mathfrak{s}(\kappa)$. The nondegeneracy of the Killing form plays a crucial role in the theory of semi-simple $\mathfrak{g}$. For instance, it is an easy orthogonality argument that a semi-simple Lie algebra is the direct sum of its simple ideals.

The classification of simple finite-dimensional Lie algebras over $\mathbb{C}$ was accomplished at the turn of the century by Killing and Cartan. There are four infinite families A_r ($r \ge 1$), B_r ($r \ge 3$), C_r ($r \ge 2$) and D_r ($r \ge 4$), and five exceptionals E_6, E_7, E_8, F_4 and G_2. A_r can be thought of as $\mathfrak{sl}_{r+1}(\mathbb{C})$, the $(r+1) \times (r+1)$ matrices with trace 0. The orthogonal algebras B_r and D_r can be identified with $\mathfrak{so}_{2r+1}(\mathbb{C})$ and $\mathfrak{so}_{2r}(\mathbb{C})$, respectively, where $\mathfrak{so}_n(\mathbb{C})$ is all $n \times n$ anti-symmetric matrices $A^t = -A$. The symplectic algebra C_r is $\mathfrak{sp}_{2r}(\mathbb{C})$, i.e. all $2r \times 2r$ matrices A obeying $A\Omega = -\Omega A^t$, where $\Omega = \begin{pmatrix} 0 & I_r \\ -I_r & 0 \end{pmatrix}$ and I_r is the identity. In all these cases the bracket is the commutator (1.4.4). The exceptional algebras can be constructed using, for example, the octonions. For instance, G_2 is the algebra of derivations of octonions. In fact, given any pair $\mathcal{A}_1, \mathcal{A}_2$ of normed division rings (so $\mathcal{A}_i$ are $\mathbb{R}$, $\mathbb{C}$, the quaternions or the octonions), there is a general construction of a simple Lie algebra $\mathfrak{g}(\mathcal{A}_1, \mathcal{A}_2)$ (over $\mathbb{R}$) – see, for example, section 4 of [**29**]. The results are summarised in *Freudenthal's Magic Square* (Table 1.3). The interesting thing here is the uniform construction of four of the five exceptional Lie algebras. In Sections 1.5.2 and 1.6.2 we give further reasons for thinking of the exceptional Lie algebras as fitting into a sequence – a nice paradigm whenever multiple exceptional structures are present.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

To verify that (1.4.2b) truly is $\mathfrak{sl}_2(\mathbb{C})$, put

$$e = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad f = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{1.4.8}$$

The names A, B, C, D have no significance: since the four series start at $r = 1, 2, 3, 4$, they were called A, B, C, D, respectively. Unfortunately, misfortune struck: at random $B_2 \cong C_2$ was called orthogonal, although the affine Coxeter–Dynkin diagrams (Figure 3.2) reveal that it is actually symplectic and only accidentally looks orthogonal. In hindsight the names of the B- and C-series really should have been switched.

For reasons we explain in Section 1.5.2, all semi-simple finite-dimensional Lie algebras over $\mathbb{C}$ have a presentation of the following form.

Definition 1.4.5 (a) *A* Cartan$_{ss}$ matrix A *is an* $n \times n$ *matrix with integer entries* a_{ij}*, such that:*

C1. *each diagonal entry* $a_{ii} = 2$*;*
C2. *each off-diagonal entry* a_{ij}*,* $i \neq j$*, is a nonpositive integer;*
C3. *the zeros in* A *are symmetric about the main diagonal (i.e.* $a_{ij} = 0$ *iff* $a_{ji} = 0$*); and*
C4. *there exists a positive diagonal matrix* D *such that the product* AD *is positive-definite (i.e.* $(AD)^t = AD$ *and* $x^t ADx > 0$ *for any real column vector* $x \neq 0$*).*

(b) *Given any Cartan*$_{ss}$ *matrix* A*, define a Lie algebra* $\mathfrak{g}(A)$ *by the following presentation. It has* $3n$ *generators* e_i*,* f_i*,* h_i*, for* $i = 1, \ldots, n$*, and obeys the relations*

R1. $[e_i f_j] = \delta_{ij} h_i$, $[h_i e_j] = a_{ij} e_j$, $[h_i f_j] = -a_{ij} f_j$*, and* $[h_i h_j] = 0$*, for all* i, j*; and*
R2. $(\operatorname{ad} e_i)^{1-a_{ij}} e_j = (\operatorname{ad} f_i)^{1-a_{ij}} f_j = 0$ *whenever* $i \neq j$*.*

'ss' stands for 'semi-simple'; it is standard to call these matrices A 'Cartan matrices', but this can lead to terminology complications when in Section 3.3.2 we doubly generalise Definition 1.4.5(a). As always, $\operatorname{ad} e : \mathfrak{g} \to \mathfrak{g}$ is defined by $(\operatorname{ad} e) f = [ef]$, so if $a_{ij} = 0$ then $[e_i e_j] = 0$, while if $a_{ij} = -1$ then $[e_i [e_i e_j]] = 0$. It is a theorem of Serre (1966) that $\mathfrak{g}(A)$ is finite-dimensional semi-simple, and any complex finite-dimensional semi-simple Lie algebra $\mathfrak{g}$ equals $\mathfrak{g}(A)$ for some Cartan$_{ss}$ matrix A.

The terms 'generators' and 'basis' are sometimes confused. Both build up the whole algebra; the difference lies in which operations you are permitted to use. For a basis, you are only allowed to use linear combinations (i.e. addition of vectors and multiplication by numbers), while for generators you are also permitted multiplication of vectors (the bracket here).'Dimension' refers to basis, while 'rank' usually refers to generators. For instance, the (commutative associative) algebra of polynomials in one variable x is infinite-dimensional, but the single polynomial x is enough to generate it (so its rank is 1). Although $\mathfrak{g}(A)$ has $3r$ generators, its dimension will usually be far greater.

The entries of Cartan$_{ss}$ matrices are mostly zeros, so it is more transparent to realise them with a graph, called the Coxeter–Dynkin diagram.[13] The diagram corresponding

[13] The more common name 'Dynkin diagram' is historically inaccurate. Coxeter was the first to introduce these graphs, originally in the context of reflection groups, but in 1934 he applied them also to Lie algebras. Dynkin's involvement with them occurred over a decade later.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

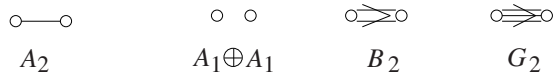

Fig. 1.16 The rank 2 Coxeter–Dynkin diagrams.

to matrix A has r nodes; the ith and jth nodes are connected with $a_{ij}a_{ji}$ edges, and if $a_{ij} \neq a_{ji}$, we put an arrow over those edges pointing to i if $a_{ij} < a_{ji}$.

For example, the 2×2 Cartan$_{ss}$ matrices are

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & -1 \\ -3 & 2 \end{pmatrix}.$$

The third and fourth matrices can be replaced by their transposes, which correspond to isomorphic algebras. Their Coxeter–Dynkin diagrams are given in Figure 1.16.

To get a better feeling for relations R1, R2, consider a fixed i. The generators $e = e_i$, $f = f_i$, $h = h_i$ obey (1.4.2b). In other words, every node in the Coxeter–Dynkin diagram corresponds to a copy of the A_1 Lie algebra. The lines connecting these nodes tell how these r copies of A_1 intertwine. For instance, the first Cartan matrix given above corresponds to the Lie algebra $A_2 = \mathfrak{sl}_3(\mathbb{C})$. The two A_1 subalgebras that generate it (one for each node) can be chosen to be the trace-zero matrices of the form

$$\begin{pmatrix} \star & \star & 0 \\ \star & \star & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 & 0 \\ 0 & \star & \star \\ 0 & \star & \star \end{pmatrix}.$$

The Lie algebra corresponding to a disjoint union $\cup_i \mathcal{D}_i$ of diagrams is the direct sum $\oplus_i \mathfrak{g}_i(\mathcal{D}_i)$ of algebras. Thus we may require the matrix A to be *indecomposable*, or equivalently that the Coxeter–Dynkin diagram be connected, in which case the Lie algebra $\mathfrak{g}(\mathcal{D})$ will be simple. Of the four in Figure 1.16, only the second is decomposable.

Theorem 1.4.6 (a) *The complete list of indecomposable Cartan$_{ss}$ matrices, or equivalently the connected Coxeter–Dynkin diagrams, is given in Figure 1.17. The series A_r, B_r, C_r, D_r are defined for $r \geq 1$, $r \geq 3$, $r \geq 2$, $r \geq 4$, respectively.*
(b) *The complete list of finite-dimensional simple Lie algebras over $\mathbb{C}$ are $\mathfrak{g}(\mathcal{D})$ for each of the Coxeter–Dynkin diagrams in Figure 1.17.*

This classification changes if the field – the choice of scalars – is changed. As always, $\mathbb{C}$ is better behaved than $\mathbb{R}$ because every polynomial can be factorised completely over $\mathbb{C}$ (we say $\mathbb{C}$ is *algebraically closed*). This implies every matrix has an eigenvector over $\mathbb{C}$, something not true over $\mathbb{R}$. Over $\mathbb{C}$, each simple algebra has its own symbol $X_r \in \{A_r, \ldots, G_2\}$; over $\mathbb{R}$, each symbol corresponds to a number of inequivalent algebras. See section VI.10 of [**348**] or chapter 8 of [**214**] for details. For example, 'A_1' corresponds to three different real simple Lie algebras, namely the matrix algebras $\mathfrak{sl}_2(\mathbb{R})$, $\mathfrak{sl}_2(\mathbb{C})$ (interpreted as a *real* vector space) and $\mathfrak{su}_2(\mathbb{C}) \cong \mathfrak{so}_3(\mathbb{R})$. The simple Lie algebra classification is known in any characteristic $p > 7$ (see e.g. [**559**]). Smaller primes usually behave poorly, and the classification for characteristic 2 is probably hopeless.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

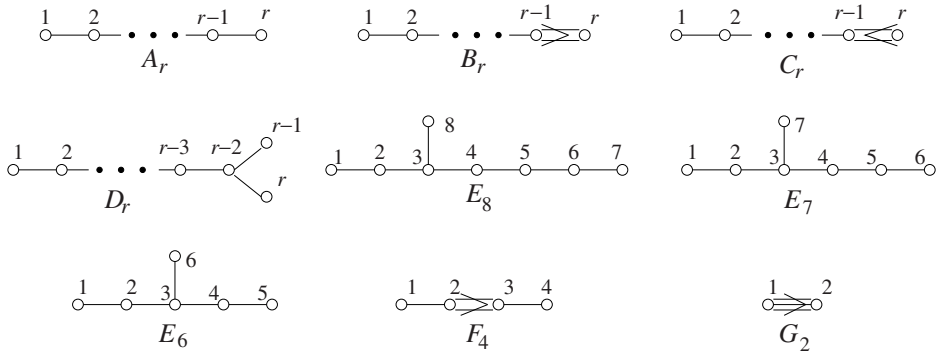

Fig. 1.17 The Coxeter–Dynkin diagrams of the simple Lie algebras.

Simple Lie algebras need not be finite-dimensional. An example is the *Witt algebra* $\mathfrak{Witt}$, defined (over $\mathbb{C}$) by the basis[14] ℓ_n, $n \in \mathbb{Z}$, and relations

$$[\ell_m \ell_n] = (m-n)\ell_{m+n}. \tag{1.4.9}$$

Using the realisation $\ell_n = -\mathrm{i}e^{-\mathrm{i}n\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}$, $\mathfrak{Witt}$ is seen to be the polynomial subalgebra of the complexification $\mathbb{C} \otimes \mathrm{Vect}(S^1)$ (i.e. the scalar field of $\mathrm{Vect}(S^1)$ is changed from $\mathbb{R}$ to $\mathbb{C}$). Incidentally, infinite-dimensional Lie algebras need not have a Lie group: for example, the real algebra $\mathrm{Vect}(S^1)$ has the Lie group $\mathrm{Diff}(S^1)$ of diffeomorphisms $S^1 \to S^1$, but its complexification $\mathbb{C} \otimes \mathrm{Vect}(S^1)$ has no Lie group (Section 3.1.2). The Witt algebra is fundamental to Moonshine. We study it in Section 3.1.2.

Question 1.4.1. Let G be a finite group, and $\mathbb{C}G$ be its group algebra (i.e. all formal linear combinations $\sum_g a_g g$ over $\mathbb{C}$). Verify that $\mathbb{C}G$ becomes a Lie algebra when given the bracket $[g, h] = gh - hg$ (extend linearly to all of $\mathbb{C}G$). Identify this Lie algebra.

Question 1.4.2. Let $\mathbb{K}$ be any field. Find all two-dimensional Lie algebras over $\mathbb{K}$, up to (Lie algebra) isomorphism.

Question 1.4.3. Prove the Witt algebra (1.4.9) is simple.

Question 1.4.4. Prove the Lie algebraic analogue of the statement that any homomorphism $f : G \to H$ between simple groups is either constant or a group isomorphism.

Question 1.4.5. The nonzero quaternions $a1 + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$, for $a, b, c, d \in \mathbb{R}$, form a Lie group by multiplication (recall that $\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1$, $\mathrm{ij} = -\mathrm{ji} = \mathrm{k}$, $\mathrm{jk} = -\mathrm{kj} = \mathrm{i}$ and $\mathrm{ki} = -\mathrm{ik} = \mathrm{j}$). Find the Lie algebra.

Question 1.4.6. (a) Verify that $\mathrm{ad}\,[xy] = \mathrm{ad}\,x \circ \mathrm{ad}\,y - \mathrm{ad}\,y \circ \mathrm{ad}\,x$, for any elements x, y in a Lie algebra $\mathfrak{g}$.
(b) Verify that the Killing form is invariant (i.e. obeys (1.4.7b)) for any Lie algebra.

[14] In order to avoid convergence complications, only finite linear combinations of basis vectors are typically permitted in algebra. Infinite linear combinations would require taking some completion.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

(c) Let $\mathfrak{g}$ be n-dimensional and semi-simple. Choose any $x \in \mathfrak{g}$. Verify that the coefficient of λ^{n-2} in the characteristic polynomial $\det(\mathrm{ad}\, x - \lambda I)$ is proportional to $\kappa(x|x)$.

Question 1.4.7. Consider the complex Lie algebra $\mathfrak{g}(A)$, for $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, defined in Definition 1.4.5(b).
(a) Prove that a basis for $\mathfrak{g}$ is $\{e_i, f_i, h_i, [e_1 e_2], [f_1 f_2]\}$ and thus that $\mathfrak{g}$ is eight-dimensional. Prove from first principles that $\mathfrak{g}$ is simple.
(b) Verify that the following generates a Lie algebra isomorphism of $\mathfrak{g}$ with $\mathfrak{sl}_3(\mathbb{C})$:

$$e_1 \mapsto \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad f_1 \mapsto \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad h_1 \mapsto \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$
$$e_2 \mapsto \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad f_2 \mapsto \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad h_2 \mapsto \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Question 1.4.8. Show that property C3 can be safely dropped. That is, given a $\mathbb{Z}$-matrix A obeying C1, C2 and C4, show that there is a Cartan matrix A' such that the Lie algebra $\mathfrak{g}(A)$ (defined as in Definition 1.4.5(b)) is isomorphic to $\mathfrak{g}(A')$.

Question 1.4.9. Are $\mathrm{Vect}(\mathbb{R})$ and $\mathrm{Vect}(S^1)$ isomorphic as Lie algebras?

1.5 Representations of simple Lie algebras

The representation theory of the simple Lie algebras can be regarded as an enormous generalisation of trigonometry. For instance, the facts that $\frac{\sin(nx)}{\sin(x)}$ can be written as a polynomial in $\cos(x)$ for any $n \in \mathbb{Z}$, and that

$$\frac{\sin(mx)\,\sin(nx)}{\sin(x)} = \sin((m+n)x) + \sin((m+n-2)x) + \cdots + \sin((m-n)x)$$

for any $m, n \in \mathbb{N}$ are both easy special cases of the theory. Representation theory is vital to the classification and structure of simple Lie algebras, and leads to the beautiful geometry and combinatorics of root systems. The relevance of Lie algebras to Moonshine and conformal field theory – which is considerable – is through their representations. The book [**219**] is a standard treatment of Lie representation theory; it is presented with more of a conformal field theoretic flavour in [**214**].

1.5.1 Definitions and examples

Although we have learned over the past couple of centuries that commutativity can be dropped without losing depth and usefulness, most interesting algebraic structures obey some form of associativity. In fact, true associativity (as opposed to, for example, anti-associativity) really simplifies the arithmetic. Given the happy accident that the commutator $[x, y] := xy - yx$ in any associative algebra obeys anti-associativity, it is tempting to seek ways in which associative algebras $\mathfrak{A}$ can 'model' or *represent* a given

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Lie algebra. That is, we would like a map $\rho : \mathfrak{g} \to \mathfrak{A}$ that preserves the linear structure (i.e. ρ is linear) and sends the bracket $[xy]$ in $\mathfrak{g}$ to the commutator $[\rho(x), \rho(y)]$ in $\mathfrak{A}$.

In practise groups often appear as symmetries, and algebras as their infinitesimal generators. These symmetries often act linearly. In other words, the preferred associative algebras are usually matrix algebras, and so we are interested in Lie algebra homomorphisms $\rho : \mathfrak{g} \to \mathfrak{gl}_n$. The *dimension* of this representation is the number n.

Completely equivalent to a representation is the notion of '$\mathfrak{g}$-module M', as is the case for finite groups (Section 1.1.3). A $\mathfrak{g}$-*module* is a vector space M on which $\mathfrak{g}$ acts (on the left) by product $x.v$, for $x \in \mathfrak{g}$, $v \in M$. This product must be bilinear, and must obey $[xy].v = x.(y.v) - y.(x.v)$. We use 'module' and 'representation' interchangeably.

Lie algebra modules behave much like finite group modules. Let $\rho_i : \mathfrak{g} \to \mathfrak{gl}(V_i)$ be two representations of $\mathfrak{g}$. We define their *direct sum* $\rho_1 \oplus \rho_2 : \mathfrak{g} \to \mathfrak{gl}(V_1 \oplus V_2)$ as usual by

$$(\rho_1 \oplus \rho_2)(x)(v_1, v_2) = (\rho_1(x)(v_1), \rho_2(x)(v_2)), \qquad \forall x \in \mathfrak{g},\ v_i \in V_i. \tag{1.5.1a}$$

Lie algebras are special in that (like groups) we can *multiply* their representations: define the *tensor product representation* $\rho_1 \otimes \rho_2 : \mathfrak{g} \to \mathfrak{gl}(V_1 \otimes V_2)$ through

$$(\rho_1 \otimes \rho_2)(x)(v_1 \otimes v_2) = (\rho_1(x)v_1) \otimes v_2 + v_1 \otimes (\rho_2(x)v_2), \qquad \forall x \in \mathfrak{g}_1,\ v_i \in V_i. \tag{1.5.1b}$$

Recall that the vector space $V_1 \otimes V_2$ is defined to be the span of all $v_1 \otimes v_2$, so the value $(\rho_1 \otimes \rho_2)(x)(v)$ on generic vectors $v \in V_1 \otimes V_2$ requires (1.5.1b) to be extended linearly. It is easy to verify that (1.5.1b) defines a representation of $\mathfrak{g}$; the obvious but incorrect attempt $(\rho_1(x)v_1) \otimes (\rho_2(x)v_2)$ would lose linear dependence on x. As usual, the dimension of $\rho_1 \oplus \rho_2$ is $\dim(\rho_1) + \dim(\rho_2)$, while $\dim(\rho_1 \otimes \rho_2)$ is $\dim(\rho_1)\dim(\rho_2)$.

A rich representation theory requires in addition a notion of *dual* or *contragredient*. Recall that the dual space V^* is the space of all linear functionals $v^* : V \to \mathbb{C}$. Given a $\mathfrak{g}$-module V, the natural module structure on V^* is the contragredient, defined by

$$(x.v^*)(u) = -v^*(x.u), \qquad \forall x \in \mathfrak{g},\ v^* \in V^*,\ u \in V. \tag{1.5.1c}$$

This defines $\rho^*(x)v^* \in V^*$ by its value at each $u \in V$. In terms of matrices, (1.5.1c) amounts to choosing $\rho^*(x)$ to be $-\rho(x)^t$, the negative of the transpose of $\rho(x)$. The negative sign is needed for the Lie brackets to be preserved.

The definition of unitary representation ρ for finite groups says each $\rho(g)$ should be a unitary matrix. Since the exponential of a Lie algebra representation should be a Lie group representation, we would like to say that a unitary representation ρ of a Lie algebra should obey $\rho(x)^\dagger = -\rho(x)$ for any $x \in \mathfrak{g}$, where '$\dagger$' is the adjoint (complex conjugate-transpose), that is to say all matrices $\rho(x)$ should be anti-self-adjoint. This works for real Lie algebras, but not for complex ones: if $\rho(x)$ is anti-self-adjoint, then $\rho(\mathrm{i}x) = \mathrm{i}\rho(x)$ will be self-adjoint!

The correct notion of unitary representation $\rho : \mathfrak{g} \to \mathfrak{gl}(V)$ for complex Lie algebras is that there is an anti-linear map $\omega : \mathfrak{g} \to \mathfrak{g}$ obeying $\omega[xy] = -[\omega x, \omega y]$, such that $\rho(x)^\dagger = \rho(\omega x)$. 'Anti-linear' means $\omega(ax + y) = \overline{a}\omega(x) + \omega(y)$. Equivalently, ρ is

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

unitary if the complex vector space V has a Hermitian form $\langle u, v\rangle \in \mathbb{C}$ on it, such that

$$\langle u, \rho(x)v\rangle = \langle \rho(\omega x)u, v\rangle. \tag{1.5.2}$$

For the case of real Lie algebras, $\omega x = -x$ works. For the complex semi-simple Lie algebra $\mathfrak{g}(A)$ of Definition 1.4.5, the most common choice is $\omega e_i = f_i, \omega f_i = e_i, \omega h_i = h_i$ (this is the negative of the so-called *Chevalley involution*).

A *submodule* of a $\mathfrak{g}$-module V is a subspace $U \subseteq V$ obeying $\mathfrak{g}.U \subseteq U$. The obvious submodules are $\{0\}$ and V; an *irreducible module* is one whose only submodules are those trivial ones. Schur's Lemma (Lemma 1.1.3) holds verbatim, provided G is replaced with a finite-dimensional Lie algebra $\mathfrak{g}$, and ρ, ρ' are also finite-dimensional.

Finding all possible modules, even for the simple Lie algebras, is probably hopeless. For example, all simple Lie algebras have uncountably many irreducible ones. However, it is possible to find all of their *finite-dimensional* modules.

Theorem 1.5.1 *Let $\mathfrak{g}$ be a complex finite-dimensional semi-simple Lie algebra of rank r. Then any finite-dimensional $\mathfrak{g}$-module is completely reducible into a direct sum of irreducible modules. Moreover, there is a unique unitary irreducible module $L(\lambda)$ for each r-tuple $\lambda = (\lambda_1, \ldots, \lambda_r)$ of nonnegative integers, and all irreducible ones are of that form.*

Let $P_+ = P_+(\mathfrak{g})$ denote the set of all r-tuples λ of nonnegative integers; $\lambda \in P_+$ are called *dominant integral weights*. The module $L(\lambda)$ is called the irreducible module with *highest weight* λ. We explain how to prove Theorem 1.5.1 and construct $L(\lambda)$ in Section 1.5.3, but to get an idea of what $L(\lambda)$ looks like, consider A_1 from (1.4.2b). For any $\lambda \in \mathbb{C}$, define $x_0 \neq 0$ to formally obey $h.x_0 = \lambda x_0$ and $e.x_0 = 0$. Define inductively $x_{i+1} := f.x_i$ for $i = 0, 1, \ldots$ The span of all x_i, call it $M(\lambda)$, is an infinite-dimensional A_1-module: the calculations $h.x_{i+1} = h.(f.x_i) = ([hf] + fh).x_i = (-2f + fh).x_i$ and $e.x_{i+1} = e.(f.x_i) = ([ef] + fe).x_i = (h + fe).x_i$ show inductively that $h.x_m = (\lambda - 2m)x_m$ and $e.x_m = (\lambda - m + 1)m, x_{m-1}$. The linear independence of the x_i follow from these. $M(\lambda)$ is called a *Verma module* with highest weight λ, and x_0 its highest-weight vector.

Is $M(\lambda)$ unitary? Here, ω interchanges e and f, and fixes h. The calculation

$$\langle x_i, x_i\rangle = \langle f.x_{i-1}, x_i\rangle = \langle x_{i-1}, e.x_i\rangle = (\lambda - i + 1)\langle x_{i-1}, x_{i-1}\rangle \tag{1.5.3}$$

tells us that the norm-squares $\langle x_i, x_i\rangle$ and $\langle x_{i-1}, x_{i-1}\rangle$ can't both be positive, if i is sufficiently large. Thus no Verma module $M(\lambda)$ is unitary.

Now specialise to $\lambda = n \in \mathbb{N} := \{0, 1, 2, \ldots\}$. Since $e.x_{n+1} = 0$ and $h.x_{n+1} = (-n - 2)x_{n+1}$, $M(n)$ contains a *submodule* with highest-weight vector x_{n+1}, isomorphic to $M(-n-2)$. x_{n+1} is called a *singular* or *null vector*, because by (1.5.3) it has norm-squared $\langle x_{n+1}, x_{n+1}\rangle = 0$. In other words, we could set $x_{n+1} := 0$ and still have an A_1-module – a *finite-dimensional* module $L(n) := M(n)/M(-n-2)$ with basis $\{x_0, x_1, \ldots, x_n\}$ and dimension $n + 1$. This basis is orthogonal and $L(n)$ is unitary.

For example, the basis $\{x_0, x_1\}$ of $L(1)$ recovers the representation $\mathfrak{sl}_2(\mathbb{C})$ of (1.4.8). The *adjoint representation* of Section 1.5.2 is $L(2)$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

The situation for the other simple Lie algebras X_r is similar (Section 1.5.3). On the other hand, non-semi-simple Lie algebras have a much more complicated representation theory. They have finite-dimensional modules that aren't completely reducible. For example, given any finite-dimensional representation $\rho : \mathfrak{g} \to \mathfrak{gl}(V)$ of any *solvable* Lie algebra $\mathfrak{g}$, a basis can be found for V such that every matrix $\rho(x)$ will be upper-triangular (i.e. the entries $\rho(x)_{ij}$ will equal 0 when $i > j$) – see Lie's Theorem in section 4.1 of [**300**]. This implies that any finite-dimensional irreducible module of a solvable $\mathfrak{g}$ is one-dimensional, and thus a finite-dimensional representation ρ will be completely reducible iff all matrices $\rho(x)$ are simultaneously diagonalisable. See Question 1.5.2.

1.5.2 The structure of simple Lie algebras

Representation theory is important in the structure theory of the Lie algebra itself, and as such is central to the classification of simple Lie algebras. In particular, any Lie algebra $\mathfrak{g}$ is itself a $\mathfrak{g}$-module with action $x.y := (\operatorname{ad} x)(y) = [xy]$ – the so-called adjoint representation. In this subsection we use this representation to associate a Cartan matrix to each semi-simple $\mathfrak{g}$.

Consider for concreteness the $\mathfrak{g} = \mathfrak{sl}_n(\mathbb{C})$, the Lie algebra of all trace-0 $n \times n$ matrices, for $n \geq 2$. Let $\mathfrak{h}$ be the set of all diagonal trace-0 matrices. Then the matrices in $\mathfrak{h}$ commute with themselves, so $\mathfrak{h}$ is an abelian Lie subalgebra of $\mathfrak{g}$. Restricting the adjoint representation of $\mathfrak{g}$, we can regard $\mathfrak{g}$ as an $(n^2 - 1)$-dimensional $\mathfrak{h}$-module. Unlike most $\mathfrak{h}$-modules, this one is completely reducible.

In particular, let $E_{(ab)}$ be the $n \times n$ matrix with entries $(E_{(ab)})_{ij} = \delta_{ai}\delta_{bj}$, that is with 0's everywhere except for a '1' in the ab entry. Since $E_{(ab)}E_{(cd)} = \delta_{bc}E_{(ad)}$, we get

$$[E_{(ab)}, E_{(cd)}] = \delta_{bc}E_{(ad)} - \delta_{ad}E_{(cb)}. \tag{1.5.4a}$$

Now, a basis for $\mathfrak{h}$ is $A_a = E_{(a,a)} - E_{(a+1,a+1)}$ for $a = 1, \ldots, n-1$. Thus

$$[A_a, E_{(cd)}] = (\delta_{ad} + \delta_{a+1,c} - \delta_{ac} - \delta_{a+1,d})E_{(cd)} \tag{1.5.4b}$$

and the basis $\{E_{(cd)}\}_{1\leq c\neq d\leq n} \cup \{A_a\}_{1\leq a<n}$ of $\mathfrak{g}$ simultaneously diagonalises all endomorphisms $\operatorname{ad} A_a$. In other words, this representation $\operatorname{ad}\mathfrak{h}$ decomposes into a direct sum of one-dimensional $\mathfrak{h}$-modules. Define functionals $\alpha_{(cd)} \in \mathfrak{h}^*$ by

$$\alpha_{(cd)}(A_a) = \delta_{ad} + \delta_{a+1,c} - \delta_{ac} - \delta_{a+1,d}.$$

Then we can write

$$\mathfrak{g} = \oplus_{1\leq c\neq d\leq n}\mathbb{C}E_{(cd)} \oplus \operatorname{span}\{A_a\}_{1\leq a<n} = \oplus_{\alpha\in\Phi}\mathfrak{g}_\alpha \oplus \mathfrak{h}, \tag{1.5.4c}$$

where $\Phi = \{\alpha_{(cd)}\}_{1\leq c\neq d\leq n}$ and $\mathfrak{g}_{\alpha_{(cd)}} = \mathbb{C}E_{(cd)}$. The functional $\alpha = \alpha_{(cd)}$ is called a *root* because $\alpha(A)$ is the eigenvalue of the operator $\operatorname{ad} A$ on the eigenspace $\mathbb{C}E_{(cd)}$ and thus is a *root* of the characteristic polynomial of $\operatorname{ad} A$. We avoid calling 0 (the functional for $\mathfrak{h}$) a root because it behaves differently, for example $\mathfrak{g}_0 = \mathfrak{h}$ has dimension $n-1$ but all other $\mathfrak{g}_\alpha$ have dimension 1. In Section 3.3.1 we identify 0 though as a precursor to the so-called imaginary roots of Kac–Moody algebras.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

From the identity

$$(\mathrm{ad}\, A)[xy] = [(\mathrm{ad}\, A)x,\, y] + [x,\, (\mathrm{ad}\, A)y]$$

(which holds in any Lie algebra), or more concretely from (1.5.4a), we see that the decomposition (1.5.4c) defines a grading $[\mathfrak{g}_\alpha,\, \mathfrak{g}_\beta] \subseteq \mathfrak{g}_{\alpha+\beta}$, for any roots $\alpha,\, \beta \in \Phi$, where we put $\mathfrak{g}_{\alpha+\beta} = \{0\}$ if $\alpha + \beta \notin \Phi$. In fact, a little more care verifies that equality always holds:

$$[\mathfrak{g}_\alpha,\, \mathfrak{g}_\beta] = \mathfrak{g}_{\alpha+\beta}, \qquad \forall \alpha,\, \beta \in \Phi. \tag{1.5.4d}$$

In Question 1.5.3 you compute the Killing form (1.4.7a). We find that $\kappa(E_{(ab)}|E_{(cd)}) = 0$ unless $(d,\, c) = (a,\, b)$, and that κ is positive-definite when restricted to the real $(n-1)$-dimensional space $\mathfrak{h}_\mathbb{R}$ spanned over $\mathbb{R}$ by $A_1, \ldots, A_{n-1}$.

The roots $\alpha_1 = \alpha_{(1,2)}, \ldots, \alpha_{n-1} = \alpha_{(n-1,n)}$ form a basis Δ for the dual space $\mathfrak{h}^*$, and are called *simple roots*. Explicitly, the root $\alpha_{(cd)} \in \Phi$ is

$$\alpha_{(cd)} = \begin{cases} \alpha_c + \alpha_{c+1} + \cdots + \alpha_{d-1} & \text{if } c < d \\ -\alpha_d - \alpha_{d+1} - \cdots - \alpha_{c-1} & \text{if } c > d \end{cases}.$$

Note that for each root $\alpha = \alpha_{(ab)}$, the elements $e_\alpha := E_{(ab)}$, $f_\alpha := E_{(ba)}$, $h_\alpha := A_a - A_b$ span a copy of $\mathfrak{sl}_2$. In particular, the $\mathfrak{sl}_2$'s coming from the simple roots α_i generate all of $\mathfrak{sl}_n(\mathbb{C})$, thanks to the grading (1.5.4d). For each $\alpha_i, \alpha_j \in \Delta$, let

$$a_{ij} = \alpha_i(h_{\alpha_j}) = \begin{cases} 2 & \text{if } i = j \\ -1 & \text{if } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}.$$

This defines a Cartan matrix A. To verify that $\mathfrak{g}(A)$ is $\mathfrak{sl}_n(\mathbb{C})$, do calculations such as

$$[e_{\alpha_i}[e_{\alpha_i} e_{\alpha_{i\pm1}}]] \in \mathfrak{g}_{2\alpha_i + \alpha_{i\pm1}} = \{0\}.$$

This analysis continues to hold for any semi-simple $\mathfrak{g}$. The space $\mathfrak{h}$ of diagonal matrices becomes any subalgebra of $\mathfrak{g}$, all of whose elements x have diagonalisable operator $\mathrm{ad}\, x$. Any maximal such Lie subalgebra is called a *Cartan subalgebra*. Since almost every polynomial has distinct roots, almost every matrix is diagonalisable; for semi-simple $\mathfrak{g}$, almost every $\mathrm{ad}\, x$ is diagonalisable. A Cartan subalgebra is necessarily abelian.

Given a Cartan subalgebra $\mathfrak{h}$, we get a *root-space decomposition*

$$\mathfrak{g} = \oplus_{\alpha\in\Phi} \mathfrak{g}_\alpha \oplus \mathfrak{h} \tag{1.5.5a}$$

as in (1.5.4c), by simultaneously diagonalising all $\mathrm{ad}\, \mathfrak{h}$. The $\alpha \in \Phi \subset \mathfrak{h}^*$ are called roots as before; the *root spaces* $\mathfrak{g}_\alpha$ are defined to be the simultaneous eigenspaces

$$\mathfrak{g}_\alpha := \{x \in \mathfrak{g} \,|\, [hx] = \alpha(h)x\}. \tag{1.5.5b}$$

The $\mathfrak{g}_\alpha$ are always one-dimensional and define a grading as in (1.5.4d). The Killing form κ is a nondegenerate inner-product, with $\kappa(\mathfrak{g}_\alpha|\mathfrak{g}_\beta) = 0$ unless $\beta = -\alpha$. The finite set Φ of roots is called the *root system*; the full algebra $\mathfrak{g}$ can be reconstructed directly from Φ.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Each $\mathfrak{g}$ has uncountably many possible Cartan subalgebras. They are related by automorphisms of $\mathfrak{g}$ – in fact 'inner automorphisms' $\exp(\operatorname{ad} x)$ (Section 1.5.4) – so they yield equivalent root systems Φ. Let $N(\mathfrak{h})$ denote the set of all inner automorphisms that map the space $\mathfrak{h}$ onto itself, and let $C(\mathfrak{h}) = \exp(\operatorname{ad}\mathfrak{h})$ denote the set of all inner automorphisms that fix $\mathfrak{h}$ pointwise. Then $C(\mathfrak{h})$ is a normal subgroup of $N(\mathfrak{h})$, and the quotient $N(\mathfrak{h})/C(\mathfrak{h})$ of these continuous groups is a finite group called the *Weyl group* W. It is a symmetry of the data of $\mathfrak{g}$, as we will see.

The Killing form identifies $\mathfrak{h}$ and its dual (this is the raising/lowering of indices familiar to any physicist, or transpose familiar to everyone else). We thus get an inner-product on the dual space $\mathfrak{h}^*$, positive-definite on the real span of the roots. For increased readability, we write $(\beta|\beta')$ in place of $\kappa(\beta|\beta')$, for $\beta, \beta' \in \mathfrak{h}^*$. The Weyl group W acts on $\mathfrak{h}^*$; in particular it is generated by the reflections

$$r_\alpha(\beta) = \beta - 2\frac{(\beta|\alpha)}{(\alpha|\alpha)}\alpha \tag{1.5.5c}$$

through each root $\alpha \in \Phi$ (recall Question 1.2.5). The Weyl group W permutes the roots and preserves the Killing form. Each reflection r_α fixes the hyperplane orthogonal to α. Removing those hyperplanes decomposes $\mathfrak{h}^*$ into connected components, one for every element of W. Choose one at random and call it the *positive chamber* C.

The $\mathbb{Z}$-span of the roots $\alpha \in \Phi$ is called the root lattice of $\mathfrak{g}$; it is positive-definite, the orthogonal direct sum of copies of $\mathbb{Z}$ and the lattices A_n, D_n, E_6, E_7, E_8 of Section 1.2.1, all appropriately scaled. The Weyl group is a group of automorphisms of the root lattice, normal and of small index in the full automorphism group.

Let $\alpha_1, \ldots, \alpha_r$ be the roots orthogonal to the walls of the positive chamber C, with the sign of each α_i chosen so that $(\alpha_i|C)$ is positive. Then those α_i form a basis Δ for $\mathfrak{h}^*$, called a *base*; the α_i are called *simple roots*. Moreover, given any root $\alpha \in \Phi$, either α or $-\alpha$ lies in $\mathbb{N}\alpha_1 + \cdots + \mathbb{N}\alpha_r$ – we say α is *positive* or *negative*, respectively. The root-space decomposition (1.5.5a) can be written in the form

$$\mathfrak{g} = \eta_+ \oplus \mathfrak{h} \oplus \eta_-, \tag{1.5.5d}$$

called a *triangular decomposition*, where $\eta_\pm$ is the sum of the positive (negative) root spaces. The grading implies $[\mathfrak{h}\mathfrak{h}] = 0$, $[\eta_\pm\eta_\pm] \subseteq \eta_\pm$, $[\mathfrak{h}\eta_\pm] \subseteq \eta_\pm$. Any Lie algebra with a triangular decomposition has Verma modules, as we will see [**432**].

Once we have a base Δ, we get a Cartan matrix A (and hence a Coxeter–Dynkin diagram) through the formula

$$a_{ij} = 2\frac{(\alpha_i|\alpha_j)}{(\alpha_j|\alpha_j)}.$$

For each simple root $\alpha_i \in \Delta$, we get elements $e_i \in \mathfrak{g}_{\alpha_i}$, $f_i \in \mathfrak{g}_{-\alpha_i}$, $h_i \in \mathfrak{h}$ that span a copy of $\mathfrak{sl}_2(\mathbb{C})$, and together these $3r$ elements generate all of $\mathfrak{g}$. In fact, these are the elements referred to in Definition 1.4.5(b), and $\mathfrak{g}$ is isomorphic to that Lie algebra $\mathfrak{g}(A)$. The cardinality r of any base is called the *rank* of $\mathfrak{g}$. Incidentally, an arrow between vertices i, j in a diagram always points towards the simple root of smaller norm.

Thus we get a Coxeter–Dynkin diagram from $\mathfrak{g}$ by making two arbitrary choices: a Cartan subalgebra $\mathfrak{h}$ and a positive chamber C. Different choices are related by symmetries

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Table 1.4. *The simple roots and fundamental weights for the classical algebras*

Algebra	Simple root α_i	Fundamental weight ω_i
A_r	$e_i - e_{i+1},\ 1 \le i \le r$	$\sum_{j=1}^{i} e_j - \frac{i}{r+1}\sum_{j=1}^{r+1} e_j$
B_r	$e_i - e_{i+1},\ 1 \le i < r$	$e_1 + \cdots + e_i,\ 1 \le i < r$
	$2e_r$	$\frac{1}{2}(e_1 + \cdots + e_r)$
C_r	$\sqrt{2}(e_i - e_{i+1}),\ 1 \le i < r$	$\frac{1}{\sqrt{2}}(e_1 + \cdots + e_i),\ 1 \le i \le r$
	$\sqrt{2}e_r$	
D_r	$e_i - e_{i+1},\ 1 \le i < r$	$e_1 + \cdots + e_i,\ 1 \le i < r-1$
	$e_{r-1} + e_r$	$\frac{1}{2}(e_1 + e_2 + \cdots + e_{r-2} + e_{r-1} - e_r),\ i = r-1$
		$\frac{1}{2}(e_1 + e_2 + \cdots + e_r),\ i = r$

(inner automorphisms) of $\mathfrak{g}$, and the resulting diagram is uniquely determined. This is a powerful paradigm: to understand and classify a rigid structure, find and study a combinatorial characterisation. Later we apply this strategy to conformal field theories.

These choices though should disturb the mathematician in us. Perhaps the presence of the Weyl group in the following is a hint that we are doing Lie theory badly. Just as the vector space 'symmetry' GL_n is the artificial consequence of choosing a basis, so is the Weyl group the bad karma caused by selecting one positive chamber over all others. Probably an approach based on Vogel's universal Lie algebra (Section 1.6.2) will ultimately be preferable.

In any case, we are most interested in the Killing form and Weyl group restricted to $\mathfrak{h}^*$. Given simple roots α_i, define *fundamental weights* $\omega_i \in \mathfrak{h}^*$ to be the dual basis $(\omega_i|\alpha_j) = \delta_{ij}$. They lie on the edges of the chamber C. Their $\mathbb{Z}$-span is the lattice dual to the root lattice, called the *weight lattice*. Denote by P_+ the intersection of the weight lattice with C, so $\lambda \in P_+$ if and only if $\lambda = \sum_{i=1}^{r} \lambda_i \omega_i$ where each *Dynkin label* λ_i lies in $\mathbb{N}$. These $\lambda \in \mathbb{N}$, called *dominant integral weights*, are the r-tuples of Theorem 1.5.1.

Table 1.4 gives the α_i and ω_i for the classical algebras, using an orthonormal basis of $\mathbb{R}^r$ ($\mathbb{R}^{r+1}$ for A_r). Nodes are labelled as in Figure 1.17 – this is the labelling used in, for example, [**328**] but not by all other authors. The table makes manifest the Killing form on $\mathfrak{h}^*$, and is useful in the study of affine Kac–Moody algebras (Section 3.2). More data for the simple Lie algebras, including the exceptional ones (avoided here for reasons of brevity), can be found in section 6.7 of [**328**], chapter 7 of [**214**], and especially pages 265–90 of [**84**].

The Weyl group of $\mathfrak{g} = \mathfrak{sl}_n(\mathbb{C})$ is the symmetric group $\mathcal{S}_n$ and acts on $\mathfrak{h}^*$ by permuting the subscripts: $\sigma \sum_i h_i \omega_i = \sum_i h_i \omega_{\sigma i}$. Figure 1.18 gives the root systems of the semi-simple Lie algebras of rank 2. A choice of simple roots is indicated by the numerals '1' and '2'. In Figure 1.19 a portion of the weight lattices of $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{C})$ and $\mathfrak{g} = \mathfrak{sl}_3(\mathbb{C})$ are displayed, along with simple roots and fundamental weights, and the Weyl reflections $r_i = r_{\alpha_i}$ through the simple roots. Note the $\mathcal{S}_2 \cong \{\pm 1\}$ symmetry of the A_1 weight lattice, and the $\mathcal{S}_3$ symmetry of the A_2 weight lattice.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

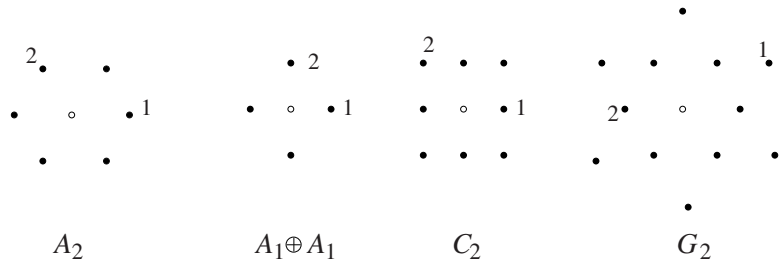

Fig. 1.18 The root systems of the rank 2 algebras.

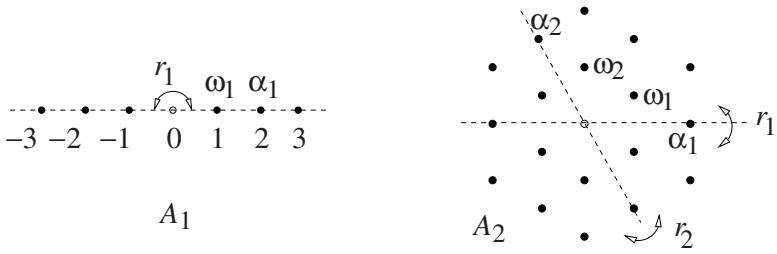

Fig. 1.19 Some of the weights of A_1 and A_2.

							$A_1, 2$
						$u_1, 1$	$A_2, 3$
					$0, 1$	$A_1, 3$	$G_2, 7$
				$0, 1$	$2u_1, 2$	$3A_1, 4$	$D_4, 8$
			$0, 2$	$A_1, 5$	$A_2, 8$	$C_3, 14$	$F_4, 26$
		$0, 1$	$2u_1, 3$	$A_2, 2$	$A_2, 9$	$A_5, 15$	$E_6, 27$
	$u_1, 2$	$A_1, 4$	$3A_1, 8$	$C_3, 14$	$A_5, 20$	$D_6, 32$	$E_7, 56$
$A_1, 3$	$A_2, 8$	$G_2, 14$	$D_4, 28$	$F_4, 52$	$E_6, 78$	$E_7, 133$	$E_8, 248$

Fig. 1.20 Cvitanović's Magic Triangle.

The first hint that the exceptional Lie algebras are not especially exceptional (i.e. that they fall into a common series) is Freudenthal's Magic Square (Table 1.3). A second is *Cvitanović's Magic Triangle* [**126**], [**129**] (Figure 1.20). The clearest example of a family of Lie algebras is A_n, where in fact the representation rings of smaller A_n embed in those of the larger (the characters are the Schur polynomials in infinitely many variables, appropriately restricted). For example, the formulae $L(\omega_1) \otimes L(\omega_k) = L(\omega_1 + \omega_k) \oplus L(\omega_{k+1})$ and $\dim L(\omega_k) = \binom{n+1}{k}$ hold for all k and A_n, although, for example, $L(\omega_2) = L(0)$ and $L(\omega_3) = 0$ for A_1. Something similar (though more complicated) happens for the 'exceptional series', i.e. the Lie algebras in the bottom row of the Magic Triangle. For instance, the decomposition of various powers $\mathfrak{g}^{\otimes_k}$ of the adjoint modules

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

into irreducibles take the same form (e.g. $\mathfrak{g} \otimes \mathfrak{g} = L(0) \oplus Y_2 \oplus Y_2^* \oplus \mathfrak{g} \oplus X_2$, where for, for example, $\mathfrak{g} = G_2, F_4, E_8$, respectively we have $Y_2 = L(2\omega_1), L(2\omega_1), L(2\omega_7), Y_2^* = L(2\omega_2), L(2\omega_4), L(\omega_1), \mathfrak{g} = L(\omega_1), L(\omega_1), L(\omega_7)$ and $X_2 = L(3\omega_2), L(\omega_2), L(\omega_6)$), and the dimension of the adjoint representation is given by the uniform equation $\dim \mathfrak{g} = 2(5h^\vee - 6)(h^\vee + 1)/(h^\vee + 6)$, where $h^\vee$ is the dual Coxeter number of Section 3.2.3. For more examples, see [**126**], [**129**] and references therein.

Note that the exceptional series is nested:

$$A_1 \subset A_2 \subset G_2 \subset D_4 \subset F_4 \subset E_6 \subset E_7 \subset E_8.$$

Taking any pair $\mathfrak{h} \subset \mathfrak{g}$, the corresponding entry in the Magic Triangle is the centraliser $\mathfrak{c}$ of $\mathfrak{h}$ in $\mathfrak{g}$, and the number there is the dimension of an irreducible module of $\mathfrak{c}$, unique up to outer automorphism, defined by the decomposition of $\mathfrak{g}$ as a $\mathfrak{c} \oplus \mathfrak{h}$-module. For simplicity Figure 1.20 is watered-down by using Lie algebras in place of Lie groups (e.g. the 0's along the top diagonal are really finite groups) – see [**129**] for details. This exceptional series is explained by Vogel's universal Lie algebra (Section 1.6.2).

1.5.3 Weyl characters

Let $\mathfrak{g}$ be any complex finite-dimensional semi-simple Lie algebra. The analysis of the last subsection on the adjoint representation can be generalised to the other finite-dimensional $\mathfrak{g}$-modules. Recall the notation introduced last subsection. Let Φ^+ be the positive roots. For each $\alpha \in \Phi^+$, choose $e_\alpha \in \mathfrak{g}_\alpha$, $f_\alpha \in \mathfrak{g}_{-\alpha}$ and $h_\alpha \in \mathfrak{h}$ as before, and write e_i, f_i, h_i for these corresponding to the simple root $\alpha_i \in \Delta$. Let ω_i be the fundamental weights, as before.

For all representations $\rho : \mathfrak{g} \to \mathfrak{gl}(V)$ of interest to us, in particular all of the finite-dimensional ones, the matrices $\rho(h)$ for $h \in \mathfrak{h}$ will be simultaneously diagonalisable. The analogue of (1.5.4c) is the *weight-space decomposition*

$$V = \oplus_{\beta \in \Omega(\rho)} V_\beta, \tag{1.5.6a}$$

where these functionals $\beta \in \Omega(\rho) \subset \mathfrak{h}^*$ are called the *weights* of ρ. For example, the non-zero weights of the adjoint representation ad $\mathfrak{g}$ are the roots. For any finite-dimensional ρ, the β all lie in the weight lattice $\mathbb{Z}\omega_1 + \cdots + \mathbb{Z}\omega_r$. These *weight spaces*

$$V_\beta := \{v \in V \mid h.v = \beta(h)v \ \forall h \in \mathfrak{h}\} \tag{1.5.6b}$$

will no longer be one-dimensional in general – the dimension $\dim V_\beta$ is called the multiplicity of β in ρ. The grading (1.5.4d) now becomes

$$f_\alpha V_\beta \subseteq V_{\beta+\alpha}, \qquad e_\alpha V_\beta \subseteq V_{\beta-\alpha}. \tag{1.5.6c}$$

The weight-space decomposition, or equivalently the weights $\beta \in \Omega(\rho)$ and their multiplicities, uniquely determines any finite-dimensional module (up to equivalence). The Weyl group W acts on weights via (1.5.5c), and preserves multiplicities:

$$\dim V_\beta = \dim V_{w\beta}, \qquad \forall w \in W, \ \beta \in \Omega(\rho). \tag{1.5.6d}$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

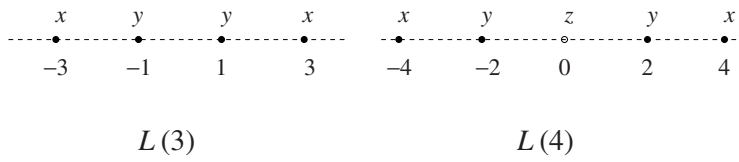

Fig. 1.21 The weights of representations of A_1.

In Section 3.2.3 we learn that this innocent symmetry (1.5.6d) is a key to the appearance of modularity in affine Kac–Moody algebras.

For an example, recall the Verma module $M(\lambda)$ for $\mathfrak{sl}_2(\mathbb{C})$ constructed in Section 1.5.1. Strictly speaking we should write $\lambda\omega_1$ for the highest weight λ. This representation has weights $(\lambda - 2j)\omega_1$ for $j = 0, 1, 2, \ldots$, all with multiplicity 1. Moreover, the unitary module $L(n) = L(n\omega_1)$ has weights $(n - 2j)\omega_1$ for $j = 0, 1, \ldots, n$, again all with multiplicity 1. The weight-spaces $L(n)_m$ are $\mathbb{C}\, x_{(n-m)/2}$. The Weyl group $W \cong \mathbb{Z}_2$ acts here by sending $i\omega_1$ to $-i\omega_1$. See Figure 1.21 for the weights of A_1-representations $L(3\omega_1)$ and $L(4\omega_1)$. We label weights in the same Weyl orbit with the same letter.

Given any functional $\lambda = \sum_{i=1}^{r} \lambda_i \omega_i \in \mathfrak{h}^*$, a *highest-weight module M with highest weight* λ is a $\mathfrak{g}$-module generated by a nonzero vector $v \in M$ obeying

$$e_\alpha . v = 0, \quad \forall \alpha \in \Phi^+, \tag{1.5.7a}$$

$$h_i . v = \lambda_i v, \quad 1 \le i \le r. \tag{1.5.7b}$$

Of course by linearity (1.5.7b) implies that $h_\alpha . v = (\lambda|\alpha)v$ for all positive roots α (not just the simple ones), and more generally $h.v = \lambda(h)v \ \forall h \in \mathfrak{h}$. The module M is generated by v in the sense that M is the span of all vectors of the form

$$x_{(1)} \cdots x_{(m)} . v := x_{(1)} . (\cdots (x_{(m)} . v) \cdots),$$

as the vectors $x_{(j)}$ range over all of $\mathfrak{g}$. This v is called the *highest-weight vector*. The $\mathfrak{g}$-modules of greatest interest to us are the highest-weight ones. The name comes from the fact that for all $\mu \in \Omega(\lambda)$ except $\mu = \lambda$, $\lambda - \mu$ lies in the positive chamber C.

By the *Verma module* $M(\lambda)$ we mean the largest or universal or free $\mathfrak{g}$-module with highest weight λ. Any other $\mathfrak{g}$-module with highest weight λ can be constructed from this. To make this more precise, we first define the analogue here of the group algebra $\mathbb{C}G$.

As we know, a basis for $\mathfrak{g}$ is e_α, f_α for all positive roots $\alpha \in \Phi^+$, together with the elements h_i. The *universal enveloping algebra* $U(\mathfrak{g})$ is the largest associative algebra generated by those $\|\Phi\| + \|\Delta\|$ symbols e_α, f_α, h_i, which obey all identities of the form $xy - yx = [xy]$ for all $x, y \in \mathfrak{g}$. More precisely, $U(\mathfrak{g})$ is the quotient of the free associative algebra on those $\|\Phi\| + \|\Delta\|$ symbols, with the ideal generated by all elements $xy - yx - [xy]$. The starting point for the theory of $U(\mathfrak{g})$ is:

Theorem 1.5.2 (Poincaré – Birkhoff–Witt) *A basis for $U(\mathfrak{g})$ is the set of monomials*

$$\left(\prod_\alpha f_\alpha^{m_\alpha}\right)\left(\prod_\alpha e_\alpha^{n_\alpha}\right)\left(\prod_{i=1}^{r} h_i^{p_i}\right),$$

for all choices of integers $m_\alpha \ge 0$, $n_\alpha \ge 0$, $p_i \ge 0$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

The basis element corresponding to $m_\alpha = n_\alpha = p_i = 0$ is denoted 1. The associative algebra $U(\mathfrak{g})$ is not commutative, so to define the products $\prod_\alpha$ we must make some arbitrary ordering of the positive roots Φ^+ – it doesn't matter how we do this. The Poincaré–Birkhoff–Witt Theorem holds for any Lie algebra (not necessarily semi-simple). See the proof and discussion in chapter III of [**348**]. That those monomials span $U(\mathfrak{g})$ is clear; more difficult is to show that they are linearly independent.

In Section 6.2.3 we use $U(\mathfrak{g})$ to construct *quantum groups*. Here what is significant is that its representation theory is identical to that of $\mathfrak{g}$. This isn't deep: the matrices $\rho(x)$, for $x \in \mathfrak{g}$, generate an associative (matrix) algebra. Thus we have replaced the task of finding modules of the *non-associative* algebra $\mathfrak{g}$ with the simpler but equivalent task of finding modules of the *associative* (though infinite-dimensional) algebra $U(\mathfrak{g})$. The relation between $\mathfrak{g}$ and $U(\mathfrak{g})$ is quite analogous to that between G and $\mathbb{C}G$, except that $\mathbb{C}G$ is somewhat simpler due to G already having an associative product.

Let $J(\lambda)$ be the left-ideal of $U(\mathfrak{g})$ generated by all e_α and all $h_i - \lambda_i 1$. This means

$$J(\lambda) = \left\{ \sum_\alpha x_\alpha e_\alpha + \sum_i y_i \, (h_i - \lambda_i 1) \,|\, x_\alpha, \, y_i \in U(\mathfrak{g}) \right\}.$$

The Verma module $M(\lambda)$ can now be defined to be the quotient of $U(\mathfrak{g})$ by $J(\lambda)$. It is a (left) $U(\mathfrak{g})$-module, and hence a $\mathfrak{g}$-module. By the Poincaré–Birkhoff–Witt Theorem, the infinite set of elements of the form

$$v_{\{m\}} := \left(\prod_\alpha f_\alpha^{m_\alpha} \right) v, \tag{1.5.8a}$$

for all integers $m_\alpha \geq 0$, forms a basis for $M(\lambda)$. The action of $e_\alpha, f_\alpha, h_i \in \mathfrak{g}$ on these vectors $v_{\{m\}}$ is obtained using the commutation relations of $\mathfrak{g}$ together with (1.5.7). In particular, we find that $v_{\{m\}}$ is an eigenvector for all operators h_i, and corresponds to weight $\lambda - \sum_\alpha m_\alpha \alpha$. Thus the weight-space decomposition of the Verma module $M(\lambda)$ is

$$M(\lambda) = \bigoplus_{\alpha' \in \mathbb{N}\alpha_1 + \cdots + \mathbb{N}\alpha_r} M(\lambda)_{\lambda - \alpha'}, \tag{1.5.8b}$$

where $M(\lambda)_{\lambda-\alpha'}$ has basis consisting of all $v_{\{m\}}$ with $\alpha' = \sum_{\alpha \in \Delta^+} m_\alpha \alpha$.

The Verma module $M(\lambda)$ is indecomposable but may or may not be reducible (see Question 1.5.5). The general way to handle modules that aren't completely reducible is to use quotients, exactly as we did with the composition series for finite groups. In particular, $M(\lambda)$ always has a unique maximal submodule $K(\lambda) \neq M(\lambda)$, and for it the quotient $L(\lambda) := M(\lambda)/K(\lambda)$ is irreducible. More generally, every $U(\mathfrak{g})$-module with highest weight λ can be obtained by quotienting $M(\lambda)$ by some submodule; the quotient $L(\lambda)$ can thus be regarded as the smallest $U(\mathfrak{g})$-module with highest weight λ, and is the module in which we are primarily interested. In particular, the finite-dimensional irreducible modules named in Theorem 1.5.1 are precisely these quotients $L(\lambda)$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

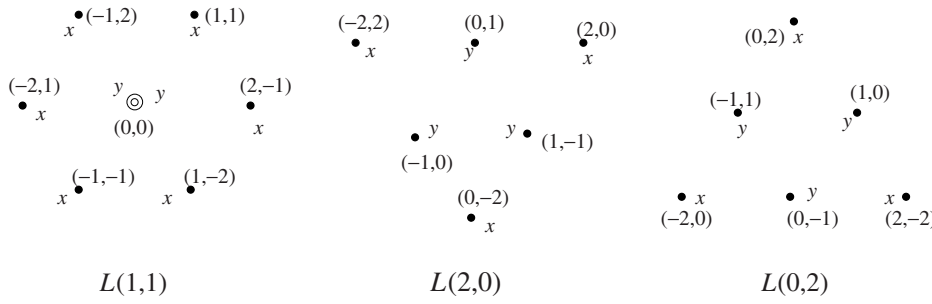

Fig. 1.22 The weights of modules of A_2.

This maximal submodule $K(\lambda)$ is the space of all null-vectors. For dominant integral weights $\lambda \in P_+$, it is the span of all vectors of the form

$$\left(\prod_\alpha f_\alpha^{c_\alpha}\right)(f_i)^{\lambda_i+1}v,$$

for any choice of integers $c_\alpha \in \mathbb{N}$, and any i.

Figure 1.22 gives the weights for A_2-modules $L(\omega_1+\omega_2)$, $L(2\omega_1)$ and $L(2\omega_2)$. We denote a weight $\beta = \sum_i \beta_i\omega_i$ by its *Dynkin labels* $\beta_i \in \mathbb{Z}$. All multiplicities in Figures 1.21 and 1.22 are 1 except for $L(\omega_1+\omega_2)_{(0,0)}$, which has multiplicity 2. Incidentally, $L(\omega_1+\omega_2)$ is the adjoint representation, while $L(2\omega_1)$ and $L(2\omega_2)$ are contragredient.

As usual, it is hard to compare modules directly: ρ and ρ' could be equivalent (i.e. differ merely by a change-of-basis) but look very different. Or given some module, we may wish to decompose it into the direct sum of irreducible modules $L(\lambda^{(i)})$. For finite groups, we use characters to clarify their representation theory, projecting away the extraneous basis-dependent details; Weyl showed that something similar works here.

The *character* of a $\mathfrak{g}$-module V, with weight-space decomposition (1.5.6a), is

$$\mathrm{ch}_V(z) := \sum_{\beta\in\Omega(V)} \dim V_\beta\, e^{\beta(z)}, \tag{1.5.9a}$$

for any $z \in \mathfrak{h}$. If we coordinatise $\mathfrak{h}$ and $\mathfrak{h}^*$ by $z = \sum_i z_i h_i$ and $\beta = \sum_i \beta_i\omega_i$, we can use

$$\beta(z) = \sum_{i=1}^{r} \beta_i z_i \frac{2}{(\alpha_i|\alpha_i)}. \tag{1.5.9b}$$

For example, for the A_1-module $L(n\omega_1)$ we find

$$\mathrm{ch}_{L(n\omega_1)}(zh) = \sum_{i=0}^{n} e^{(n-2i)z} = \frac{e^{(n+1)z} - e^{-(n+1)z}}{e^z - e^{-z}}, \tag{1.5.10a}$$

where we obtained the formula on the right by summing the geometric series. Note that its numerator and denominator are alternating sums over the Weyl group $\mathcal{S}_2$ of A_1. By comparison, the character for the Verma module $M(\lambda\omega_1)$ is

$$\mathrm{ch}_{M(\lambda\omega_1)}(zh) = \sum_{i=0}^{\infty} e^{(\lambda-2i)z} = \frac{e^{\lambda z}}{1-e^{-2z}}. \tag{1.5.10b}$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

More generally, the Verma module $M(\lambda)$ for $\mathfrak{g}$ has character

$$\mathrm{ch}_{M(\lambda)}(z) = \frac{e^{\lambda(z)}}{\prod_{\alpha\in\Delta^+}(1-e^{-\alpha(z)})}. \tag{1.5.10c}$$

The *Weyl character formula* expresses the character of any finite-dimensional irreducible module $L(\lambda)$ for any semi-simple $\mathfrak{g}$ as a fraction: the numerator is an alternating sum over the Weyl group W, and the denominator is a product over positive roots $\alpha \in \Delta^+$. More precisely,

$$\mathrm{ch}_\lambda(z) := \mathrm{ch}_{L(\lambda)}(z) = e^{-\rho\cdot z}\,\frac{\sum_{w\in W}\det(w)\,e^{w(\lambda+\rho)\cdot z}}{\prod_{\alpha\in\Delta^+}(1-e^{-\alpha(z)})}, \tag{1.5.11}$$

where $\rho = \sum_{i=1}^r \omega_i$ here is the *Weyl vector*. For a proof see, for example, chapter 14 of [**214**]. This formula and its generalisations have profound consequences (see Section 3.4.2).

Finite groups have only finitely many irreducible modules, while Lie algebras have infinitely many. Otherwise their theory is quite analogous, and in particular Lie algebra characters work as effectively as finite group characters.

Theorem 1.5.3 *Let $\mathfrak{g}$ be a finite-dimensional semi-simple Lie algebra, and M, N two finite-dimensional modules. Then $\mathrm{ch}_M(z) = \mathrm{ch}_N(z)$ for all $z \in \mathfrak{h}$ iff M and N are equivalent as $\mathfrak{g}$-modules. Moreover, $\mathrm{ch}_{M\oplus N}(z) = \mathrm{ch}_M(z) + \mathrm{ch}_N(z)$, $\mathrm{ch}_{M\otimes N}(z) = \mathrm{ch}_M(z)\,\mathrm{ch}_N(z)$ and $\mathrm{ch}_{M^*}(z) = \mathrm{ch}_M(\overline{z})$.*

As before, the characters are also enormously simpler than the modules themselves: for example, the smallest nontrivial representation of $\mathfrak{g} = E_8$ is a map from $\mathbb{C}^{248}$ to the space of 248×248 matrices, while its character is a function $\mathbb{C}^8 \to \mathbb{C}$. But why is Weyl's definition (1.5.9a) natural? How did he come up with it?

He used the relation with groups. Consider for concreteness $\mathfrak{g} = A_r$. Given any representation ρ, the map $e^x \mapsto e^{\rho(x)}$ is a representation of the Lie group $G = \mathrm{SL}_{r+1}(\mathbb{C})$ corresponding to $\mathfrak{g}$ (the exponential e^A of a matrix is defined by the usual power series, and always converges). The trace of the matrix $e^{\rho(x)}$ is the *group* character value at $e^x \in G$, so we define it to be the *algebra* character value at $x \in \mathfrak{g}$. Again, it suffices to restrict to representatives of each conjugacy class of G, because the character is a class function. Now, almost every matrix is diagonalisable (since almost any $n \times n$ matrix has n distinct eigenvalues), and so we shouldn't lose much by restricting $x \in \mathfrak{g}$ to *diagonalisable* matrices. Hence we may take our conjugacy class representatives to be *diagonal* matrices $x \in \mathfrak{g}$, i.e. to $x \in \mathfrak{h}$. So the Lie algebra character can be chosen to be a function of $z \in \mathfrak{h}$. Finally, the trace of the matrix $e^{\rho(x)}$ is the sum (with multiplicities) of its eigenvalues, which gives us (1.5.9a). This is the intuition behind Weyl's definition (1.5.9a) of character.

However, different diagonal matrices can be conjugate. For instance in A_1,

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

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

so e^{zh} and e^{-zh} lie in the same $G = \mathrm{SL}_2(\mathbb{C})$ conjugacy class and $\mathrm{ch}_M(z) = \mathrm{ch}_M(-z)$. This $z \mapsto -z$ symmetry is the Weyl group action on the Cartan subalgebra $\mathfrak{h} = \mathbb{C}h$. Each character $\mathrm{ch}_{L(\lambda)}$ of any semi-simple $\mathfrak{g}$ is similarly invariant under the Weyl group of $\mathfrak{g}$, thanks to (1.5.6d).

At first glance, it may seem that the Weyl character formula (1.5.9a) is not very practical, at least for large rank. For instance, the numerator of (1.5.9a) for E_8 would involve an alternating sum over the Weyl group, which has about 700 million elements! On the other hand, one alternating sum is very easy to compute: a determinant is an alternating sum over the symmetric group (the Weyl group of the A-series). Since all Weyl groups have symmetric subgroups of relatively small index, the numerators and denominators of (1.5.9a) actually can be computed quite effectively.

It is common practise in physics to use dimensions to specify irreducible modules. For example, the defining representation of $\mathfrak{sl}_3(\mathbb{C})$ is denoted **3**, and its contragredient by $\overline{\mathbf{3}}$. This is a terrible habit, as many unrelated modules can have identical dimension. For instance, $\mathfrak{sl}_5$ has six different irreducible modules with dimension 175: namely $L(\lambda)$ with $\lambda = (1, 2, 0, 0)$, $(1, 1, 0, 1)$, $(0, 3, 0, 0)$ and their contragredients (0,0,2,1), (1,0,1,1), (0,0,3,0). The practise should rather be to use highest weights, not dimensions, when labelling finite-dimensional modules.

1.5.4 Twisted #1: automorphisms and characters

A fundamental theme of this book is twisting by automorphisms. As we see later, it is central to conformal field theory and string theory, as well as vertex operator algebras, and is implicit in the definition of the McKay–Thompson series T_g. Its role in finite-dimensional Lie theory is more elementary, but can be regarded as a toy model for several of this book's most important subsections.

Let $\mathfrak{g}$ be any Lie algebra over $\mathbb{C}$. An automorphism γ of $\mathfrak{g}$ is an endomorphism (i.e. an invertible linear map $\gamma : \mathfrak{g} \to \mathfrak{g}$) that obeys

$$\gamma[x, y] = [\gamma x, \gamma y], \qquad \forall x, y \in \mathfrak{g}.$$

Write $\mathrm{Aut}(\mathfrak{g})$ for the group of automorphisms of $\mathfrak{g}$. When $\gamma \in \mathrm{Aut}(\mathfrak{g})$ has order $n < \infty$, it is diagonalisable on the space $\mathfrak{g}$ (why?). Hence we can write $\mathfrak{g}$ as a direct sum

$$\mathfrak{g} = \oplus_{k=0}^{n-1} \mathfrak{g}_k \tag{1.5.12a}$$

of eigenspaces of γ, where

$$\gamma x = \xi_n^k x, \qquad \forall x \in \mathfrak{g}_k \tag{1.5.12b}$$

(as always, ξ_n denotes the root of unity $\exp[2\pi \mathrm{i}/n]$). Because γ is an automorphism, (1.5.12a) defines a $\mathbb{Z}_n$-grading on $\mathfrak{g}$, in the sense that

$$[\mathfrak{g}_k, \mathfrak{g}_\ell] \subseteq \mathfrak{g}_{k+\ell}. \tag{1.5.12c}$$

The γ-invariant space $\mathfrak{g}_0$ is a subalgebra of $\mathfrak{g}$, and the other subspaces $\mathfrak{g}_k$ are $\mathfrak{g}_0$-modules.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

For example, let $\mathfrak{g} = \mathfrak{sl}_3(\mathbb{C})$ and choose the usual basis e_1, e_2, $e_{12} := [e_1 e_2]$, f_1, f_2, $f_{12} := [f_1 f_2]$, h_1, h_2. There is an order-2 automorphism γ of $\mathfrak{sl}_3$, corresponding to the left–right symmetry of the A_2 Coxeter–Dynkin diagram. It exchanges e_1 and e_2; therefore

$$\gamma e_{12} = [\gamma e_1, \gamma e_2] = [e_2 e_1] = -e_{12}.$$

Continuing in this way, we find that γ exchanges f_1 and f_2, as well as h_1 and h_2, and sends f_{12} to $-f_{12}$. Thus

$$\mathfrak{g}_0 = \text{span}\{e_1 + e_2,\, f_1 + f_2,\, h_1 + h_2\},$$
$$\mathfrak{g}_1 = \text{span}\{e_1 - e_2,\, e_{12},\, f_1 - f_2,\, f_{12},\, h_1 - h_2\}.$$

The reader can verify that the Lie subalgebra $\mathfrak{g}_0$ is isomorphic to $\mathfrak{sl}_2(\mathbb{C})$, while $\mathfrak{g}_1$ is the irreducible five-dimensional A_1-module.

Every Lie algebra has nontrivial automorphisms. For instance, let $x \in \mathfrak{g}$ be such that the operator $\operatorname{ad} x$ on $\mathfrak{g}$ is nilpotent, that is there is some integer k such that

$$(\operatorname{ad} x)^k y := [x[x \cdots [xy] \cdots]] = 0, \qquad \forall y \in \mathfrak{g}.$$

For instance, any $x = e_i$ or $x = f_j$ works when $\mathfrak{g}$ is semi-simple. Then $\exp(\operatorname{ad} x)$ (defined by the usual power series expansion) is a well-defined invertible operator on $\mathfrak{g}$ and is in fact an automorphism. These automorphisms $\exp(\operatorname{ad} x)$ together generate a normal subgroup of $\operatorname{Aut}(\mathfrak{g})$ called the *inner automorphisms* of $\mathfrak{g}$. The quotient of $\operatorname{Aut}(\mathfrak{g})$ by the inner automorphisms defines a group called the *outer automorphisms*.

For example, for $\mathfrak{g} = \mathfrak{sl}_n(\mathbb{C})$ the inner automorphisms form a group isomorphic to $\mathrm{PGL}_n(\mathbb{C}) \cong \mathrm{GL}_n(\mathbb{C})/\{\mathbb{C}^\times I_n\}$, and the group of outer automorphisms is $\mathbb{Z}_2$ for $n > 2$ (and $\{1\}$ for $n = 2$). The outer automorphism takes a matrix $x \in \mathfrak{sl}_n(\mathbb{C})$ to $-x^t$.

As an aside, the group of inner automorphisms of a simple Lie algebra over $\mathbb{C}$ is always a simple group (though infinite). It could be hoped that the same would be true if instead we consider Lie algebras over a finite field $\mathbb{F}_q$. Indeed this is the case (except for five small counterexamples, involving the fields $\mathbb{F}_2$ and $\mathbb{F}_3$). This gives rise to nine of the infinite families of finite simple groups of Lie type (Section 1.1.2); the seven remaining ones are various twists of these groups.

Given any two Cartan subalgebras $\mathfrak{h}_1$, $\mathfrak{h}_2$ of a simple algebra $\mathfrak{g}$, an inner automorphism can be found mapping $\mathfrak{h}_1$ to $\mathfrak{h}_2$ (we say the inner automorphisms *act transitively* on the set of Cartan subalgebras). Moreover, for any choice of Cartan subalgebra $\mathfrak{h}$, if we take the subgroup of inner automorphisms mapping $\mathfrak{h}$ to itself, and quotient it by the subgroup of inner automorphisms fixing $\mathfrak{h}$ pointwise, then we get the Weyl group of $\mathfrak{g}$. This means that (modulo an inner automorphism) an automorphism of $\mathfrak{g}$ permutes the simple roots; conversely, this permutation uniquely determines it. In other words, the outer automorphisms for semi-simple $\mathfrak{g}$ are in a natural one-to-one correspondence with the symmetries of the Coxeter–Dynkin diagram. These are the most important choices of automorphisms, for our purposes, as the fixed-point subalgebras $\mathfrak{g}_0$ are maximally large.

In particular, the fixed-point subalgebra $\mathfrak{g}_0$ for $\mathfrak{g} = \mathfrak{sl}_{2n}$, when γ is taken to be the outer automorphism permuting e_i and e_{2n-i} (the order-2 diagram symmetry), is isomorphic

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

to $\mathfrak{sp}_{2n} \cong C_n$. Likewise, taking $\mathfrak{g}$ to be (respectively) A_{2n+1}, D_n, D_4 and E_6 and taking γ to be the diagram symmetry of order 2, 2, 3 and 2, yields a fixed-point subalgebra $\mathfrak{g}_0$ isomorphic to $\mathfrak{so}_{2n+1} = B_n$, $\mathfrak{so}_{2n-1} \cong B_{n-1}$, G_2 and F_4, respectively.

The automorphism group $\mathrm{Aut}(\mathfrak{g})$ permutes the $\mathfrak{g}$-modules, through the formula

$$\rho^\gamma(x) = \rho(\gamma x).$$

Sometimes ρ^γ and ρ are isomorphic as $\mathfrak{g}$-modules. In this case there is a matrix $A \in \mathrm{GL}(V)$, where V is the underlying space of ρ and ρ^γ, such that

$$\rho(\gamma x) = A^{-1}\rho(x)A, \qquad \forall x \in \mathfrak{g}.$$

Let us assume for convenience that ρ is irreducible. Then by Schur's Lemma this matrix A will be well defined up to a scalar multiple.

In fact, for $\mathfrak{g}$ semi-simple and γ corresponding to a diagram symmetry, and ρ the module $L(\lambda)$, ρ^γ will be the module $L(\gamma\lambda)$, where γ acts on weights by permuting Dynkin labels. Thus $\rho^\gamma \cong \rho$ iff $\gamma\lambda = \lambda$. In this case there is a canonical choice of matrix A, sending weight-space $L(\lambda)_\beta$ to weight-space $L(\lambda)_{\gamma\beta}$, given by

$$e_{m_1} \cdots e_{m_k}.v_\lambda \mapsto e_{\gamma m_1} \cdots e_{\gamma m_k}.v_\lambda.$$

Recall Thompson's trick: twisting the graded dimension (0.3.2) to get the McKay–Thompson series T_g of (0.3.3). Here, this becomes the *γ-twisted* or *twining character*

$$\mathrm{ch}^\gamma_\lambda(h) := \mathrm{tr}_V A \, \exp[\rho(h)] = \sum_{\beta=\gamma\beta} \mathrm{tr}(A_\beta) \, \exp[\beta(h)], \tag{1.5.13}$$

where we can restrict the sum to all weights $\beta \in \Omega(L(\lambda))$ that are fixed by γ, and where A_β is the restriction of A to the weight-space $L(\lambda)_\beta$. The term 'twining', introduced in [**213**], is short for 'intertwining'. In terms of the basis (1.5.8a) for the weight-spaces $L(\lambda)_\beta$, A_β is a permutation matrix when β is fixed by γ (only these β survive in (1.5.13)).

For example, consider first $\mathfrak{g} = D_4$ and γ the diagram automorphism interchanging the third and fourth nodes. The dominant weight $\lambda = (1, 0, 0, 0)$ is invariant under γ. The D_4-representation $L(\lambda)$ is eight-dimensional, with all weight-spaces $L(\lambda)_\beta$ having dimension 1. It is thus easy to compute the twisted character $\mathrm{ch}^\gamma_\lambda$: it has a term with coefficient 1 for each γ-invariant weight $\beta = (\pm1, 0, 0, 0), (0, \pm1, 0, 0)$. For a more complicated example, consider D_4 again, but with the order-3 automorphism ('triality') and the invariant dominant weight $\lambda = (0, 1, 0, 0)$: this D_4-representation is 28-dimensional but only its weights $\beta = (0, \pm1, 0, 0), \pm(1, -1, 1, 1), \pm(1, -2, 1, 1), (0, 0, 0, 0)$ are triality-invariant. Of those, the weight-spaces are all one-dimensional except for $L(0, 1, 0, 0)_{(0,0,0,0)}$, which is four-dimensional. A basis for that weight-space consists of

$$f_3f_2f_4f_1f_2.v, \;\; f_4f_2f_3f_1f_2.v, \;\; f_4f_2f_3f_4f_2.v, \;\; f_2f_3f_4f_1f_2.v.$$

The map $A_{(0,0,0,0)}$ cyclically permutes the first three basis vectors, but fixes the fourth. Thus the twisted character has seven terms, each with coefficient 1. For similar calculations with small-rank algebras, the concrete bases given in [**383**] are useful.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

If we restrict to h in the Cartan subalgebra of the fixed-point subalgebra $\mathfrak{g}_0$, the result will lie in the character ring of $\mathfrak{g}_0$. Thus the twisted character $\mathrm{ch}^\gamma_\lambda$ is a *virtual character* for the fixed-point subalgebra $\mathfrak{g}_0$, that is, it is a linear combination over $\mathbb{Z}$ of true characters. However, $\mathrm{ch}^\gamma_\lambda$ itself need not be a true character of $\mathfrak{g}_0$.

For example, recall the example $\mathfrak{g} = D_4$, weight $\lambda = (1, 0, 0, 0)$, and γ interchanging nodes 3 and 4. Then the fixed-point subalgebra $\mathfrak{g}_0$ is B_3 and the twisted character $\mathrm{ch}^\gamma_\lambda$ is the virtual B_3 character $\mathrm{ch}^B_{(1,0,0)} - \mathrm{ch}^B_{(0,0,0)}$ (Question 1.5.8(a)). On the other hand, the other D_4 example has fixed-point subalgebra G_2, and the twisted character equals the true character $L(0, 1)$.

Surprisingly, $\mathrm{ch}^\gamma_\lambda$ is always a *true* character for the algebra $\mathfrak{g}_0^{\mathrm{op}}$ obtained by reversing the arrows in the diagram of $\mathfrak{g}_0$. $\mathfrak{g}^{\mathrm{op}}$ is called the *orbit Lie algebra* in [**213**].

For example, when $\mathfrak{g} = D_4$ and $\lambda = (1, 0, 0, 0)$, we find (Question 1.5.8(b)) that the twisted character $\mathrm{ch}^\gamma_{(1,0,0,0)}$ equals the character $\mathrm{ch}^C_{(1,0,0)}$ of the orbit Lie algebra $\mathfrak{g}_0^{\mathrm{op}}$.

More generally, we find:

Theorem 1.5.4 [213] *Let $\mathfrak{g}$ be semi-simple and finite-dimensional, and let γ be the automorphism of $\mathfrak{g}$ corresponding to a Coxeter–Dynkin diagram symmetry. Let $\lambda \in P_+(\mathfrak{g})$ be any dominant integral weight fixed by γ. Then the twisted character $\mathrm{ch}^\gamma_\lambda$ defined in (1.5.13), restricted to the Cartan subalgebra of the fixed-point subalgebra $\mathfrak{g}_0$, is a virtual character of $\mathfrak{g}_0$ and a true character $\chi_{\overline{\lambda}}$ of the orbit Lie algebra $\mathfrak{g}_0^{\mathrm{op}}$, for some $\overline{\lambda} \in P_+(\mathfrak{g}_0^{\mathrm{op}})$.*

A weight $\lambda \in P_+(A_{2n})$ fixed by the order-two diagram symmetry looks like $\lambda = (\lambda_1, \ldots, \lambda_n, \lambda_n, \ldots, \lambda_1)$; likewise, $\lambda \in P_+(A_{2n-1})$ fixed by the order-two diagram symmetry looks like $\lambda = (\lambda_1, \ldots, \lambda_{n-1}, \lambda_n, \lambda_{n-1}, \ldots, \lambda_1)$; while a weight $\lambda \in P_+(D_{n+1})$ fixed by the $n - 1 \leftrightarrow n$ diagram symmetry looks like $\lambda = (\lambda_1, \ldots, \lambda_n, \lambda_n)$. The orbit Lie algebra $\mathfrak{g}_0^{\mathrm{op}}$ here is C_n for A_{2n} or D_{n+1}, and B_n for A_{2n-1}. In all three cases, $\overline{\lambda}$ has Dynkin labels $(\lambda_1, \ldots, \lambda_n)$.

The proof of Theorem 1.5.4 follows that of the Weyl character formula. Although Theorem 1.5.4 is not itself important for us, the obvious generalisation holds for affine algebras (Theorem 3.4.1), and provides a striking special case of the important orbifold construction in string theory and vertex operator algebras.

In hindsight it is easy to see that $\mathfrak{g}_0^{\mathrm{op}}$ is the more natural algebra: for modules, $\mathfrak{h}^*$ is more relevant than $\mathfrak{h}$ since that is where the weights live. Consider, for example, D_4 again, with diagram symmetry $3 \leftrightarrow 4$. Then a γ-invariant weight looks like $\beta_1\omega_1 + \beta_2\omega_2 + \beta_3(\omega_3 + \omega_4)$. Using Table 1.4, we see that these vectors $\{\omega_1, \omega_2, \omega_3 + \omega_4\}$ have the same inner-products with each other that the fundamental weights of C_3 have (up to a global factor of 2, which is merely conventional).

Incidentally, some version of these remarks holds for finite groups. Let γ be an automorphism of a finite group G; then γ permutes the irreducible representations of G, $\rho \mapsto \rho \circ \gamma$, as before. Choose any irreducible representation $\rho \cong \rho \circ \gamma$ and let A be the isomorphism. The γ-twisted character of ρ is the trace $\mathrm{ch}^\gamma_\rho(g) := \mathrm{tr}\, A\, \rho(g)$. It won't be a class function of G – for example, for the inner automorphism $g \mapsto h^{-1}gh$, $\mathrm{ch}^h_\rho(g) = \mathrm{ch}_\rho(hg)$. But this calculation shows that it suffices to consider outer

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

automorphisms. In particular, diagram automorphisms of finite reductive groups should be interesting in this context.

1.5.5 Representations of Lie groups

We are more interested in (complex) Lie algebras, but (real) Lie groups do occasionally arise. Once again, it is their representation theory that is of greatest interest to us.

Let G be a real finite-dimensional Lie group, and let $\mathcal{H}$ be a complex Hilbert space. Let $\mathcal{B}(\mathcal{H})$ be the group of bounded linear operators with bounded inverse – boundedness is equivalent to continuity (Section 1.3.1). A *representation* or *module* of G on $\mathcal{H}$ is a homomorphism $\pi : G \to \mathcal{B}(\mathcal{H})$ such that the map $G \to \mathcal{H}$, defined by $g \mapsto \pi(g)v$, is continuous for every $v \in \mathcal{H}$. We call two modules π, π' *equivalent* if there is a bounded operator $A : \mathcal{H} \to \mathcal{H}'$, with bounded inverse, such that $A^{-1}\pi'(g)A = \pi(g)$ for all $g \in G$. The module π is *unitary* if each operator $\pi(g)$ is unitary, that is surjective and

$$\langle \pi(g)v, \pi(g)v' \rangle = \langle v, v' \rangle, \qquad \forall v, v' \in \mathcal{H}.$$

The module π is *irreducible* if there is no closed nontrivial subspace V, such that $\pi(g)V \subseteq V$ for all $g \in G$. Most important are the irreducible unitary modules, and these together form a topological space called the *unitary dual* $\widehat{G}$ of G.

For example, all one-dimensional modules of the additive group $G = \mathbb{R}$ are of the form $x \mapsto e^{\mathrm{i}\alpha x}$ for any $\alpha \in \mathbb{C}$; it will be unitary iff $\alpha \in \mathbb{R}$. The map $x \mapsto \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$ is a representation of $\mathbb{R}$ that is not irreducible (consider $V = \mathbb{C} \times \{0\} \subset \mathbb{C}^2 = \mathcal{H}$). The one-dimensional modules of the group $G = S^1$ are $e^{\mathrm{i}\theta} \mapsto e^{\mathrm{i}n\theta}$ for $n \in \mathbb{Z}$, and all are unitary. The unitary duals of $\mathbb{R}$ and S^1 are $\mathbb{R}$ and $\mathbb{Z}$, respectively.

Continuity is an important requirement. For instance, let $\{b_\beta\}_{\beta \in B}$ be a basis for $\mathbb{R}$ treated as a vector space over $\mathbb{Q}$ (so B is uncountable). Then for any choice of complex numbers α_β, the assignment $\sum_\beta r_\beta b_\beta \mapsto \prod_\beta e^{\mathrm{i}r_\alpha \alpha_\beta}$ defines a (rather chaotic) group (for $r_\beta \in \mathbb{Q}$) homomorphism $\mathbb{R} \to \mathbb{C}^\times$. Continuity of π is needed in order to obtain from π a module of the Lie algebra $\mathfrak{g}$ of G.

Call a vector $v \in \mathcal{H}$ *smooth* if $g \mapsto \pi(g)v$ is a smooth function from G to $\mathcal{H}$. The space $\mathcal{H}_\infty$ of smooth vectors forms a dense G-invariant subspace of $\mathcal{H}$; if $\mathcal{H}$ is finite-dimensional, $\mathcal{H}_\infty$ equals $\mathcal{H}$. Recall that the Lie algebra $\mathfrak{g}$ is the tangent space $T_e G$, and the exponential map exp sends $\mathfrak{g}$ to G. For any $v \in \mathcal{H}_\infty$ and $x \in \mathfrak{g}$, define

$$\delta\pi(x)v = \frac{\mathrm{d}}{\mathrm{d}t}\left(\pi(e^{tx})v\right)_{t=0}. \tag{1.5.14}$$

This defines a $\mathfrak{g}$-module on $\mathcal{H}_\infty$ called the *derived module*. Of course, a (complex) module of the real Lie algebra $\mathfrak{g}$ lifts to a complex module of its complexification $\mathfrak{g}_\mathbb{C} := \mathbb{C} \otimes_\mathbb{R} \mathfrak{g}$.

The theory simplifies enormously if G is compact (for simplicity we also assume connectivity). Then G is a subgroup of the unitary group $U_n(\mathbb{C})$. Moreover:

Theorem 1.5.5 (Peter–Weyl) *Let G be a connected compact finite-dimensional Lie group. Any module π of G is equivalent to a unitary one, is completely reducible, and*

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

$\delta\pi$ is a module of the reductive Lie algebra $\mathfrak{g}_{\mathbb{C}} = \mathbb{C} \otimes \mathfrak{g}$. Any irreducible G-module is finite-dimensional, and the derived module for $\mathfrak{g}_{\mathbb{C}}$ is also irreducible as a Lie algebra module.

The unitary dual $\widehat{G}$ is thus a countable discrete space. The key to proving Theorem 1.5.5 is that it is possible to average (integrate) over the group. This G-invariant *Haar measure* plays the role here of the ubiquitous $\sum_{g\in G}$ in the finite group theory. For example, a G-invariant Hermitian form on $\mathcal{H}$ is obtained by averaging any given Hermitian form over its translates – compactness of G is needed to show that integral converges. See, for example, chapter II.9 of [**92**] for an elementary proof of Theorem 1.5.5.

If G is simply-connected as well as compact and connected, then any irreducible module of $\mathfrak{g}_{\mathbb{C}}$ lifts to one of G. Otherwise, $G = \widetilde{G}/Z$, where $\widetilde{G}$ is the universal cover and Z is some discrete subgroup of $\widetilde{G}$ (Theorem 1.4.3), and a $\mathfrak{g}_{\mathbb{C}}$-module will lift to one on G iff, once it is lifted to $\widetilde{G}$, it is trivial on Z. If it isn't trivial on Z, it would be a *projective representation* for G (Section 3.1.1).

An elementary example of this is provided by the modules of $\mathbb{R} \cong \widetilde{S^1}$ and $S^1 \cong \mathbb{R}/\mathbb{Z}$, given earlier. More interesting is to compare the universal cover $\mathrm{SU}_2(\mathbb{C})$ of the group $\mathrm{SO}_3(\mathbb{R}) \cong \mathrm{SU}_2(\mathbb{C})/\left\langle \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\rangle$. Their complexified Lie algebra $\mathfrak{g}_{\mathbb{C}}$ is $\mathfrak{sl}_2(\mathbb{C})$, whose irreducible modules correspond to highest weights $\lambda \in P_+ = \{0, \omega_1, 2\omega_1, \ldots\}$. Each of these exponentiates to an irreducible module of $\mathrm{SU}_2(\mathbb{C})$. In particular, the $\mathrm{SU}_2(\mathbb{C})$-module corresponding to highest weight $\lambda = n\omega_1$ can be realised as the space of homogeneous polynomials $p(z_1, z_2)$ of degree n, with $\mathrm{SU}_2(\mathbb{C})$ action given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot p(z_1, z_2) = p(az_1 + cz_2, bz_1 + dz_2). \tag{1.5.15}$$

This will be a module of $\mathrm{SO}_3(\mathbb{R})$ iff $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ acts trivially, i.e. iff $p(z_1, z_2) = p(-z_1, -z_2)$ for all p, i.e. iff n is even. Physicists call $n/2$ the 'spin', and the modules with n odd are called 'spinors'. See, for example, chapter 20 of [**214**] for more on this. More generally, the dominant weight $\lambda = \sum_{i=1}^{n-1} \lambda_i \omega_i \in P_+$ gives a module of $\mathrm{PSL}_n(\mathbb{C}) \cong \mathrm{SL}_n(\mathbb{C})/\mathbb{Z}_n$ iff n divides $\sum i\lambda_i$.

Let G be any compact simply-connected connected Lie group, and $\mathfrak{g}$ its (real) Lie algebra. The simply-connected connected complex Lie group associated with the complex Lie algebra $\mathfrak{g}_{\mathbb{C}}$ is called the complexification $G_{\mathbb{C}}$ of G. For example, the complexification of $\mathrm{SU}_n(\mathbb{C})$ is $\mathrm{SL}_n(\mathbb{C})$. *Weyl's unitary trick* says that the irreducible modules of G, $\mathfrak{g}$, $\mathfrak{g}_{\mathbb{C}}$ and $G_{\mathbb{C}}$ are all in natural bijection, using the derived module, complexification of the algebra module, 'exponentiation' of an algebra module to a simply-connected Lie group and restriction. Depending on the context, it is sometimes more convenient to look at the modules of G, $\mathfrak{g}_{\mathbb{C}}$ or $G_{\mathbb{C}}$.

All of the irreducible modules of a compact connected Lie group G are constructed explicitly by the *Borel–Weil Theorem*. It suffices of course to consider simply-connected G. Take $G = \mathrm{SU}_n(\mathbb{C})$ for concreteness. Let B be the upper-triangular matrices in $G_{\mathbb{C}} = \mathrm{SL}_n(\mathbb{C})$. It is called the *Borel subgroup* and is a maximal solvable subgroup in $G_{\mathbb{C}}$. Given

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

a dominant integral weight $\lambda = \sum \lambda_i \omega_i$, put $t = \lambda_1 + 2\lambda_2 + \cdots + (n-1)\lambda_{n-1}$ and

$$\mu = \left(\sum_{i=1}^{n-1} \lambda_i - \frac{1}{n}t, \sum_{i=2}^{n-1} \lambda_i - \frac{1}{n}t, \ldots, \lambda_{n-1} - \frac{1}{n}t, -\frac{1}{n}t \right) \in \mathbb{R}^n.$$

Let $\Gamma(\lambda)$ be the space of holomorphic functions $f(g)$ on $G_{\mathbb{C}}$ (regarded as a complex manifold) such that

$$f(gb) = b_1^{\mu_1} \cdots b_n^{\mu_n} f(g), \qquad \forall g \in G_{\mathbb{C}},\ b = \begin{pmatrix} b_1 & * & * \\ 0 & \ddots & * \\ 0 & 0 & b_n \end{pmatrix}. \tag{1.5.16}$$

Then this is a $G_{\mathbb{C}}$-module (namely, one induced from a one-dimensional B-module), and it is easy to identify its weights since the maximal torus T (the exponentiation of the Cartan subalgebra $\mathfrak{h}$ of $\mathfrak{g}_{\mathbb{C}}$, i.e. the diagonal determinant-1 matrices) is contained in B: we find that $\Gamma(\lambda)$ is the contragredient of the highest-weight representation $V(\lambda)$. From this picture, the Weyl character formula arises through fixed-point formulae for the $G_{\mathbb{C}}$-action on $G_{\mathbb{C}}/B$ [**83**].

The geometry of this construction is quite pretty (see e.g. section 23.3 of [**219**] or [**83**]). Geometrically, the space $G_{\mathbb{C}}/B \cong G/T$ is a flag variety whose points are the various choices $0 \subset V_1 \subset \cdots \subset V_{n-1} \subset \mathbb{C}^n$ of subspaces, where $\dim V_i = i$. Then $\Gamma(\lambda)$ is the space of holomorphic sections of a line bundle $G_{\mathbb{C}} \times_B \mathbb{C}$ on $G_{\mathbb{C}}/B$ naturally associated with λ. Similar comments apply to any other G. Something similar happens for the Virasoro algebra, where the flag manifold is replaced by the moduli space of curves (Section 3.1.2).

As discussed in Section 1.1.3, the natural analogue of the group algebra for a Lie group G is the space $L^2(G)$ of functions $f : G \to \mathbb{C}$, with convolution product. The main importance of these spaces of functions is that they are natural G-modules, using right translation: $(h.f)(g) := f(gh)$. For example, consider $G = S^1$, so $f \in L^2(S^1)$ can be regarded as a function $f(x)$ with period 2π. We find that $L^2(S^1)$ decomposes into the infinite direct sum

$$L^2(S^1) = \oplus_{n \in \mathbb{Z}} V(n)$$

of irreducible one-dimensional modules $V(n)$. More precisely, $L^2(S^1)$ will be a completion of this algebraic direct sum. This means that any 'vector' $f \in L^2(S^1)$ can be written as $\sum_{n \in \mathbb{Z}} f_n$ where each summand $f_n \in V(n)$. Now, $V(n)$ consists of those functions f_n on which $e^{\mathrm{i}y} \in S^1$ acts as $(e^{\mathrm{i}y}.f_n)(x) := f_n(x+y) = e^{\mathrm{i}yn} f_n(x)$ – in other words $f_n(x) = c_n\, e^{\mathrm{i}nx}$ for some complex number c_n. Using the orthogonality of the $e^{\mathrm{i}nx}$, we can explicitly construct the projection operator $L^2(S^1) \to V(n)$, and we find

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, e^{-\mathrm{i}nx}\, \mathrm{d}x,$$

which we recognise as the Fourier transform $\widehat{f}(n)$ of f.

More generally, for arbitrary compact G, the Peter–Weyl Theorem tells us that the matrix entries $\pi(g)_{ij}$ of the irreducible representations of G are dense in the space

of functions on G. More precisely, the *Fourier transform* associates with a function $f \in L^2(G)$, a matrix-valued function $\widehat{f}(\pi)$ on the unitary dual $\widehat{G}$, defined by

$$\widehat{f}(\pi) = \int_G f(g)\,\pi(g)\,\mathrm{d}g,$$

where as usual we're using the Haar measure on G, normalised so that the volume of G is 1. Then for any $f \in L^2(G)$,

$$f(g) = \sum_{\pi \in \widehat{G}} \dim \pi \ \mathrm{tr}\left(\widehat{f}(\pi)\,\pi(g)^\dagger\right).$$

As is familiar from the abelian case, the convolution product is sent to the ordinary (matrix) product: $\widehat{f_1 * f_2}(\pi) = \widehat{f_1}(\pi)\,\widehat{f_2}(\pi)$. We also get a unitary isomorphism between $L^2(G)$ and what we can call $L^2(\widehat{G})$ (the space of these matrix-valued $\widehat{f}$), called the Plancherel formula:

$$\int_G |f(g)|^2\,\mathrm{d}g = \sum_{\pi \in \widehat{G}} (\dim \pi)\,\mathrm{tr}\left(\widehat{f}(\pi)\widehat{f}(\pi)^\dagger\right).$$

The representation theory of noncompact Lie groups is completely different. This can already be seen for the additive group $G = \mathbb{R}$, which has a continuum of irreducible unitary modules (namely $e^{\mathrm{i}\alpha x}$ for all $\alpha \in \mathbb{R}$). The unitary dual $\widehat{G}$ can involve both continuous and discrete parts, and can have a wild topology. Once again, a unitary module is completely reducible into irreducible unitary ones, but for a general noncompact G a direct integral (Section 1.3.1), rather than a direct sum, will be needed, and for wild groups the uniqueness of this decomposition will be lost.

Any connected Lie group is (up to central extensions) the semi-direct product of a solvable Lie group with a semi-simple Lie group – this is the Levi decomposition (see e.g. appendix B in [**348**]). The representation theory of solvable groups is quite well understood, using the *orbit method*. It relates the unitary dual to certain orbits of G on the dual $\mathfrak{g}^*$ of the Lie algebra $\mathfrak{g}$ of G (see [**346**] for an excellent introduction, although section 2 of [**563**] may be more accessible to physicists). Physically, this is just geometric quantisation: G is a symmetry of a physical system; the classical phase space is a symplectic manifold on which G acts (these are essentially the coadjoint orbits); quantum mechanically we would like this to correspond to a Hilbert space carrying a unitary representation of G. Geometric quantisation tries to do for quantum theories what the symplectic geometry of Hamiltonian mechanics does for classical ones: provide an elegant and natural mathematical formulation.

The effect of the semi-direct product on the unitary dual is also under control. However, the representation theory of the (noncompact real) semi-simple groups is poorly understood. See [**349**] for a modern review.

For example, the Heisenberg group H consisting of all matrices

$$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}, \qquad \forall a, b, c \in \mathbb{R},$$

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

is simply-connected and solvable. Its irreducible unitary modules are given in Theorem 2.4.2 below, and we can naturally identify its unitary dual with the xy-plane in $\mathbb{R}^3$ together with the z-axis. On the other hand, $\mathrm{SL}_2(\mathbb{R})$ is a semi-simple noncompact group, topologically equivalent to the interior of a solid torus; its unitary irreducible modules are described in Section 2.4.1, and its unitary dual consists of three one-dimensional families (the principal, spherical principal, and complementary series) and a countable family (the discrete series).

Question 1.5.1. Interpret the trigonometric identities given at the beginning of this section, in terms of the character theory of A_1.

Question 1.5.2. Classify all two-dimensional representations of the abelian Lie algebra $\mathfrak{g} = \mathbb{C}^2$. Which of these are completely reducible?

Question 1.5.3. Let $\mathfrak{g} = \mathfrak{sl}_n(\mathbb{C})$. From first principles, compute the Killing form $\kappa(A_a|E_{cd}), \kappa(A_a|A_b), \kappa(E_{ab}|E_{cd})$.

Question 1.5.4. In effect, Question 1.4.7 defines a representation $\mathfrak{g}$ of $\mathfrak{sl}_3(\mathbb{C})$.
(a) Find the weight-space decomposition of this representation of $\mathfrak{sl}_3(\mathbb{C})$, as well as the corresponding character.
(b) Find the root-space decomposition of $\mathfrak{sl}_3(\mathbb{C})$, i.e. the weight-space decomposition of the adjoint representation of $\mathfrak{sl}_3(\mathbb{C})$. Also compute the character.

Question 1.5.5. Recall the Verma modules $M(\lambda)$ for A_1 constructed in Section 1.5.1.
(a) Prove that each $M(\lambda)$ is indecomposable (i.e. cannot be written as the direct sum of two submodules).
(b) When $\lambda \notin \mathbb{N}$, prove that $M(\lambda)$ is irreducible. Thus $L(\lambda) = M(\lambda)$ for these λ.
(c) When $\lambda = n \in \mathbb{N}$, find all submodules. Verify that the maximal one has highest weight vector x_{n+1}.

Question 1.5.6. Let $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{C})$.
(a) Set $C := ef + fe + \frac{1}{2}h^2 \in U(\mathfrak{g})$. Show that C is in the centre of $U(\mathfrak{g})$. (C is called the *quadratic Casimir* of $\mathfrak{g}$; there is an analogue for any semi-simple $\mathfrak{g}$.)
(b) Given any irreducible module π of $\mathfrak{g}$, prove that $Z := 2\pi(f)\pi(e) + \pi(h) + \frac{1}{2}\pi(h)^2$ is a scalar multiple of the identity.

Question 1.5.7. Let $G = \mathrm{SU}(2)$. Then $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{C})$ (which is the complexification of the Lie algebra of G) acts naturally on the space $\mathbb{C}^\infty(G)$ of all smooth complex-valued functions on G. In particular, $\mathfrak{g}$ can be identified as the space of all left-invariant first-order differential operators. Prove that $U(\mathfrak{g})$ can be identified with the space of all left-invariant finite-order differential operators on $C^\infty(G)$.

Question 1.5.8. (a) Verify the claim in Section 1.5.4 that $\mathfrak{g} = D_4$ with $L(1, 0, 0, 0)$ has twisted character restricting to B_3-character $\mathrm{ch}_{(1,0,0)} - \mathrm{ch}_{(0,0,0)}$ and C_3-character $\mathrm{ch}_{(1,0,0)}$.
(b) Repeat this calculation for $\mathfrak{g} = A_4$ and $\lambda = (1, 0, 0, 1)$.

Question 1.5.9. (a) In Section 1.5.5 we gave a module of $\mathrm{SU}_2(\mathbb{C})$ using degree n polynomials. Find the derived module for the Lie algebra $\mathfrak{sl}_2(\mathbb{C})$, find its weight-spaces, and prove the equivalence with $L(n\omega)$.
(b) Work out the Borel–Weil representation $\Gamma(\lambda)$ for $\mathrm{SU}_2(\mathbb{C})$, for any $\lambda = n\omega_1$, $n \in \mathbb{N}$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

1.6 Category theory

The only difficulty in understanding categories is in realising that they have no real content. They're just a language, highly abstract like the more familiar set theory, but one that can be both natural and suggestive. It tries to deflect some of our instinctive infatuation with objects (nouns), to the mathematically more fruitful one with structure-preserving maps between objects (verbs).

Category theory is intended as a universal language of mathematics, so all concepts should be translated into it. Much as beavers, who as a species hate the sound of running water, plaster a creek with mud and sticks until alas that cursed tinkle stops, so do category theorists devise elaborate and obscure definitions in an attempt to capture a concept that to most of us seemed perfectly clear before they got to it. But at least sometimes this works admirably – for instance no one can be immune to the charm of treating knot invariants with braided monoidal categories.

1.6.1 General philosophy

A *category* **C** consists of two kinds of things. One are the *objects*, and the other are the *arrows* (or *morphisms*). An arrow, written $f : A \to B$, has an initial and a final object (A and B, respectively). We let $\mathrm{Hom}(A, B)$ denote all arrows $A \to B$ in the category. Arrows f, g can be composed to yield a new arrow $f \circ g$, if the final object of g equals the initial object of f. Maps between categories are called *functors* if they take each object (respectively, arrow) of one to the objects (respectively, arrows) of the other, and preserve composition. A gentle introduction to the mathematics of categories is [**370**]; the standard reference is [**397**].

The standard category is called **Set**, where the 'objects' are sets, and the arrows from A to B are functions $f : A \to B$. Many algebraic categories are of that form, with objects being sets with certain structure, and the arrows being structure-preserving maps. A typical example is **Vect**, where the objects are vector spaces over some fixed field and the arrows are linear maps. A rather trivial example of a functor $\mathcal{F}$:**Vect** $\to$ **Set** sends a vector space to its underlying set – $\mathcal{F}$ simply 'forgets' the vector space structure on V and ignores the fact that the arrows f in **Vect** are linear.

Geometric categories often employ the idea of cobordism. For instance, fix a manifold M; let the objects be points $p \in M$, and the arrows $p \to q$ be homotopy equivalence classes of paths σ in M from p to q. Composition of arrows is given by (1.2.5). This category is called the *fundamental groupoid* of M – note that $\mathrm{Hom}(p, p) = \pi_1(M, p)$. A higher-dimensional example is called **Riem**: its objects are disjoint unions of (parametrised) circles S^1, and the arrows are (conformal equivalence classes of) *cobordisms*, that is (Riemann) surfaces whose boundaries are those circles. Composition of arrows in **Riem** amounts to sewing the surfaces along the appropriate boundary circles. A final example of a geometric category is **Braid**: its objects are any finite number (possibly 0) of 'hooks', $\mathrm{Hom}(m, n)$ is empty unless $m = n$, in which case the arrows are the n-braids $\beta \in \mathcal{B}_n$. Such categories, where arrows consist of equivalence classes, are called *quotient categories* [**397**].

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Fig. 1.23 The definition of product and sum.

For a baby example of the translation of the familiar into category theory, consider the usual definition of a one-to-one function: $f(x) = f(y)$ only when $x = y$. Category theory replaces this with the right cancellation law: call an arrow $f : A \to B$ 'one-to-one' if for any object C and any arrows $g, h \in \mathrm{Hom}(C, A)$, $f \circ g = f \circ h$ implies $g = h$. The reader can easily verify that in **Set** this agrees with the usual definition. What does this redefinition gain us? It certainly doesn't seem any simpler. But it does change the focus from the *argument* of f to the *global* functional behaviour of f, and a change of perspective can never be bad. It allows us to transport the idea of one-to-one-ness to arbitrary categories. For instance, in the category **Riem**, all arrows are 'one-to-one'.

Or consider the notion of *product*. In category theory, we say that the triple (P, a, b) is a product of objects A, B if $a : P \to A$ and $b : P \to B$ are arrows, and if for any $f : C \to A, g : C \to B$, there is a unique arrow $h : C \to P$ such that $f = a \circ h$ and $g = b \circ h$. See the left diagram in Figure 1.23. This notion unifies several constructions (each of which is the 'product' in an appropriately chosen category): Cartesian product of sets; intersection of sets; multiplication of numbers; the logical operator 'and'; direct product; infimum in a partially ordered set; etc. *Sum* can be defined similarly, by reversing the orientation of all the arrows in the diagram for product (see the right diagram in Figure 1.23). This unifies the constructions of disjoint union, 'or', addition, tensor product, direct sum, supremum, etc. Of course the specific construction of sum and product depends sensitively on the category. For example, in the category **Ab-Group**, where objects are abelian groups and arrows are homomorphisms, the sum of the cyclic groups $\mathbb{Z}_2$ and $\mathbb{Z}_3$ is their direct product $\mathbb{Z}_2 \times \mathbb{Z}_3 \cong \mathbb{Z}_6$, while in the category **Group**, where objects are groups and arrows homomorphisms, the direct sum of $\mathbb{Z}_2$ and $\mathbb{Z}_3$ is $\mathrm{PSL}_2(\mathbb{Z})$! See Question 1.6.3.

This generality of course comes with a price: it can wash away all of the endearing special features of a favourite theory or structure. There certainly are contexts where, for example, all human beings should be considered equal, but there are other contexts where the given human is none other than your mother and must be treated as such.

1.6.2 Braided monoidal categories

This book tries to identify the natural context for Moonshine. Categories more than sets provide the most appealing language for this context. The starting point for this formulation is braided monoidal categories. Standard references include chapter 1 of [**534**], chapter 1 of [**32**] and chapter XIII of [**338**].

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

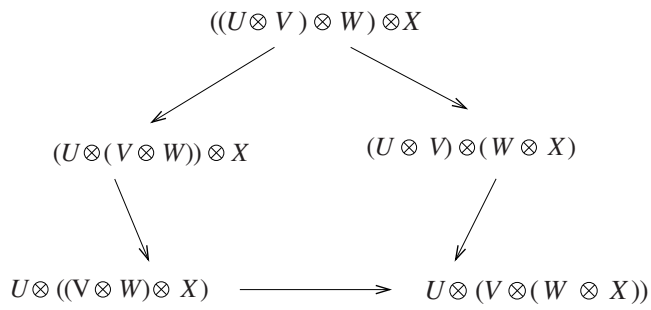

Fig. 1.24 The associativity pentagon.

Let us try to translate the vector space tensor product into category theoretic language. The result, called a *monoidal* or *tensor category*, was obtained by MacLane (1963).

Let $U_i, V_i, i = 1, 2, 3$, be vector spaces, and choose any linear maps $f_j : U_j \to U_{j+1}$, $g_j : V_j \to V_{j+1}$, $j = 1, 2$. Then the composition of the tensor product maps $f_j \otimes g_j : U_j \otimes V_j \to U_{j+1} \otimes V_{j+1}$ is given by $(f_2 \otimes g_2) \circ (f_1 \otimes g_1) = (f_2 \circ f_1) \otimes (g_2 \circ g_1)$. This is exactly the same as saying that '$\otimes$' is a functor between the categories **Vect** $\times$ **Vect** and **Vect**, where the Cartesian product of categories has the obvious meaning.

The tensor product should be associative up to isomorphism: for any objects U, V, W, there should be an isomorphism $a_{UVW} : (U \otimes V) \otimes W \to U \otimes (V \otimes W)$ (called the *associativity constraint*). It should obey a consistency condition coming from the isomorphism $((U \otimes V) \otimes W) \otimes X \cong U \otimes (V \otimes (W \otimes X))$; that is, there are two ways of computing that isomorphism in terms of associativity, and the resulting isomorphisms should agree:

$$(id_V \otimes a_{VWX}) \circ a_{U,V\otimes W,X} \circ (a_{UVW} \otimes id_X) = a_{U,V,W\otimes X} \circ a_{U\otimes V,W,X}. \qquad (1.6.1)$$

This is called the *pentagon axiom*, thanks to its depiction in Figure 1.24.

Moreover, tensoring any object V with the one-dimensional vector space (call it '1') must give back V, so there are isomorphisms $l_V : 1 \otimes V \to V, r_V : V \otimes 1 \to V$. These are required to be consistent with the associativity constraint, by requiring the *triangle axiom*

$$r_V \otimes id_W = (id_V \otimes l_W) \circ a_{V1W}. \qquad (1.6.2)$$

A monoidal category [**397**] is any category **C** possessing such a functor $\otimes$, with unit 1 and invertible arrows l_V, r_V, a_{UVW} satisfying (1.6.1) and (1.6.2). Of course **Vect** with tensor products is monoidal, as is **Set** with disjoint union. **Braid** is monoidal; the tensor product of an n-braid with an m-braid is the $(n + m)$-braid obtained by placing the two braids side-by-side. There are numerous other examples. The word 'monoidal' comes from 'monoid', meaning a group-like structure without inverses.

MacLane proved two things. The first is *coherence*, which says that (1.6.1) and (1.6.2) are sufficient. Remarkably, any other consistency condition we may care to write down will be redundant. To give a random example, the identity involving a's, l's and r's

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

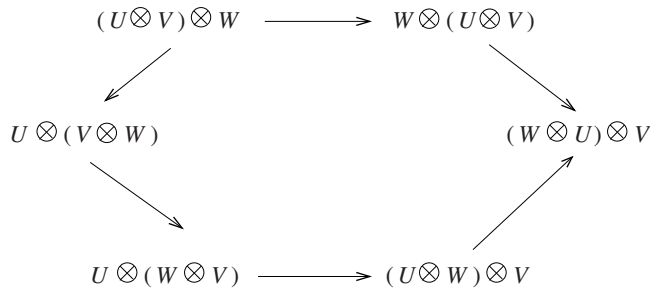

Fig. 1.25 The hexagon equation.

saying that the isomorphisms coming from $U \otimes ((V \otimes W) \otimes (1 \otimes (X \otimes Y))) \cong (U \otimes (V \otimes W)) \otimes (X \otimes Y)$ must agree can be derived from the pentagon and the triangle.

Secondly, MacLane proved that any monoidal category $\mathbf{C}$ is (monoidally) equivalent to a monoidal category $\mathbf{C}^{strict}$ where the associativity constraints are identity maps. Such a monoidal category is called *strict*; in it we can drop all associativity constraints as trivial, and with them all braces '(' and ')' in our tensor products.

Now that we've handled associativity of the tensor product, let's turn next to commutativity. We can't expect anything like MacLane's strictness to apply here – although the vector spaces $U \otimes V$ and $V \otimes U$ are naturally isomorphic, they are not equal. We proceed though in the same way.

For any objects U, V, we have an invertible arrow (called a *commutativity constraint*) $c_{UV} : U \otimes V \to V \otimes U$. Some natural relations are

$$c_{U'V'} \circ (f \otimes g) = (g \otimes f) \circ c_{UV}, \tag{1.6.3a}$$

$$c_{VU} \circ c_{UV} = id_{U \otimes V}, \tag{1.6.3b}$$

$$c_{U,V \otimes W} = c_{UV} \circ c_{UW}. \tag{1.6.3c}$$

The isomorphism $(U \otimes V) \otimes W \cong (W \otimes U) \otimes V$, or more explicitly the equation

$$(c_{UW} \otimes id_V) \circ (id_U \otimes c_{VW}) = c_{U \otimes V, W}, \tag{1.6.3d}$$

is called the *hexagon axiom* (see Figure 1.25).

Any monoidal category with commutativity constraints c_{UV} obeying (1.6.3) is called a *symmetric monoidal category* (MacLane, 1965). **Vect** is an example. Another is the categories Rep $\mathfrak{g}$ or Rep G of finite-dimensional $\mathfrak{g}$- or G-modules, for a Lie algebra $\mathfrak{g}$ (or Lie group G), with tensor product. In fact, *Tannaka–Krein duality* states that a monoidal category with both product and sum, that looks like Rep G (e.g. it has a unit object 1, a contragredient, and all objects decompose intoa sum of simple ones), *is* Rep G for a unique such group G. See, for example, section 9.4 of [**398**] for details and a generalisation.

In 1985, Joyal and Street [**321**] suggested to drop the symmetry condition (1.6.3b). The resulting categories they call *braided monoidal*, for reasons that will be clear shortly. They also pointed out that there is a very convenient graphical calculus in such categories,

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

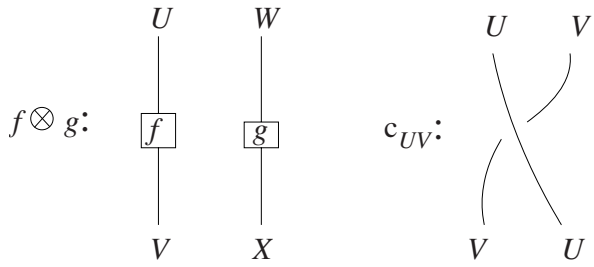

Fig. 1.26 The graphical calculus.

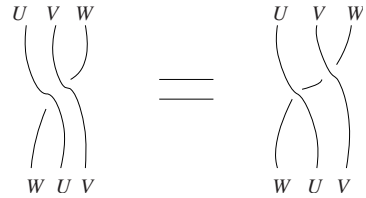

Fig. 1.27 The hexagon axiom revisited.

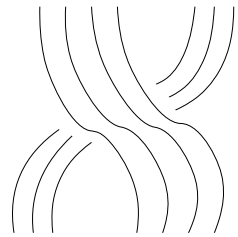

Fig. 1.28 The commutativity constraint c_{43} in **Braid**.

which elegantly keeps track of all relations. Namely, write arrows vertically and tensor products horizontally. Composition is given by vertical concatenation. The left-most diagram in Figure 1.26 represents the arrow $f \otimes g$ where $f \in \mathrm{Hom}(U, V)$ and $g \in \mathrm{Hom}(W, X)$, while the commutativity constraint c_{UV} is depicted as in the right-most. The associativity constraint a_{ABC} is ignored as we identify it with the identity. So we label strands with objects, which can change labels only at a box ('coupon'). The hexagon axiom takes the form of Figure 1.27, which we recognise as two equivalent braids. One immediate consequence is that the category **Braid** described last subsection is braided monoidal, provided we define c_{mn} as in Figure 1.28.

In terms of the graphical calculus, MacLane's symmetry condition (1.6.3b) would permit us to slip one strand through another, reducing the content of a braid (i.e. some combination of commutativity constraints) to that of its underlying permutation.

Joyal–Street also proved coherence for braided monoidal categories, that is equations (1.6.2a), (1.6.2c) and (1.6.3) are sufficient to establish the well-definedness of other

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

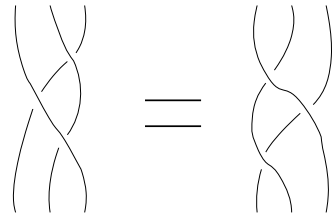

Fig. 1.29 The Yang–Baxter equation.

isomorphisms involving associativity and commutativity. For a famous example, $U \otimes V \otimes W \cong W \otimes V \otimes U$ yields the *Yang–Baxter equation*

$$(c_{VW} \otimes id_U) \circ (id_V \otimes c_{UW}) \circ (c_{UV} \otimes id_W) = (id_W \otimes c_{UV}) \circ (c_{UW} \otimes id_V) \circ (id_U \otimes c_{VW}), \tag{1.6.4}$$

which corresponds graphically to the braid equivalence of Figure 1.29 (compare Figure 1.2). We return to the Yang–Baxter equation in Section 6.2.3.

It's not a coincidence that Figure 1.29 is a braid equivalence – it *must* be, since **Braid** is a braided monoidal category. Conversely, any braid equivalence yields an equation holding in any braided monoidal category. **Braid** is the least-common divisor of all braided monoidal categories, the one with commutativity constraints and nothing else, obeying the minimum possible relations – it is *universal* or *free*. More precisely:

Theorem 1.6.1 [321] *Let* $\mathbf{C}$ *be any (strict) braided monoidal category, and* A *any object in it. Then there exists a unique braided monoidal functor* F : **Braid** $\to \mathbf{C}$ *with* $F(1) = A$ *and* $F(c_{1,1}) = c_{A,A}$.

A 'braided monoidal' functor is one preserving the braided monoidal structure in the obvious way. The object '1' of **Braid** denotes one hook, which generates via tensoring all other objects in **Braid**. This important theorem relates topology and algebra.

The simplest example (in fact too simple) of such universality is the freeness of $\mathbb{Z}$: given any group G with one generator g, there is a unique group homomorphism $\varphi : \mathbb{Z} \to G$ sending $1 \in \mathbb{Z}$ to $g \in G$. Any such G defines an *invariant* for $\mathbb{Z}$: the integer n is assigned the invariant $\varphi(n)$. We call it an invariant, because equal integers must get assigned the same G-value, even if they look different (e.g. 3 and $2 - 1 + 2$ superficially look different, but will be assigned the same G-value $\varphi(3) = \varphi(2 - 1 + 2)$). For example, the invariant φ for $G = \mathbb{Z}_2 = \{[0], [1]\}$ assigns [0] to any even $n \in \mathbb{Z}$ and [1] to any odd n. Because φ is structure-preserving, computing this invariant is relatively easy. Of course integer invariants are not terribly exciting, because it is so easy to determine if two integer expressions (involving arbitrary sums and subtractions) are equal.

Likewise, the universality of **Braid** means that, given any braided monoidal category $\mathbf{C}$ and any braid $\beta \in \mathcal{B}_n$, we get a braid-invariant $F(\beta) \in \mathrm{Hom}_{\mathbf{C}}(A^{\otimes_n}, A^{\otimes_n})$. Here the object $A^{\otimes_n}$ of $\mathbf{C}$ means $A \otimes \cdots \otimes A$ (n times). It is not so difficult to determine directly whether two braids are the same (ambient isotopic) – for example, by 'combing the braid' (see e.g. pages 24–5 of [**59**]) – and thus these braid invariants are also not intrinsically valuable. But they are a stepping stone to something that is.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

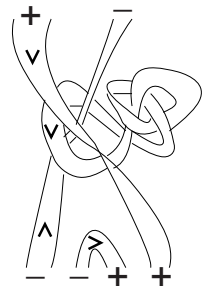

Fig. 1.30 A typical ribbon in Hom $((+,-),(-,-,+,+))$.

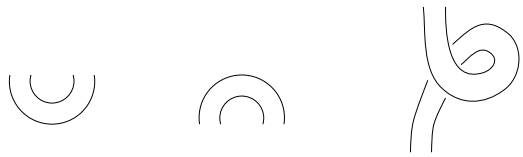

Fig. 1.31 Evaluation, coevaluation and twist.

Theorem 1.6.1 implies that, in any braided monoidal category $\mathbf{C}$, the braid group $\mathcal{B}_n$ acts on both $\mathrm{Hom}_{\mathbf{C}}(U^{\otimes_n}, V)$ and $\mathrm{Hom}_{\mathbf{C}}(U, V^{\otimes_n})$ and the pure braid group $\mathcal{P}_n$ acts on both $\mathrm{Hom}_{\mathbf{C}}(U_1 \otimes \cdots \otimes U_n, V)$ and $\mathrm{Hom}_{\mathbf{C}}(U, V_1 \otimes \cdots \otimes V_n)$ (why?). Thus the groups governing braided monoidal categories are the braid groups $\mathcal{B}_n$ and $\mathcal{P}_n$, while those of symmetric monoidal categories are the symmetric groups $\mathcal{S}_n$ (hence their names).

If we continue with our project of categorising tensor product, we will be rewarded. We can introduce the notion of duals A^* of objects (in the sense of the dual vector space), duals of arrows f^* (the analogue of transpose of matrices), the evaluation map $A^* \otimes A \to 1$ (the evaluation $f(a)$ of a functional $f \in A^*$ on a vector $a \in A$), coevaluation $1 \to A^* \otimes A$ (let b_i be a basis of vector space A and $b_i^* \in A^*$ the dual basis, then the element $\sum_i b_i^* \otimes b_i \in A^* \otimes A$ is independent of the choice of basis). These obey the obvious relations (see, for example, chapter 1 of [**534**]) and the result is called a *ribbon category*; in place of the formal definition it suffices to give the universal ribbon category.

The objects of **Ribbon** are ordered n-tuples $A = (a_1, \ldots, a_n)$ of signs, $a_i = \pm$, for $n \geq 0$ ($n = 0$ is the empty object $\emptyset$). $\mathrm{Hom}(A, B)$ consists of isotopy classes of knotted linked twisted oriented strips, called ribbons. A strip can start at position i on the top (or position j on the bottom) only if $a_i = +1$ (or $b_j = -1$, respectively); similarly, it can end at i or j only if $a_i = -1$ or $b_j = +1$ – see Figure 1.30. Braiding is as before. The dual of $(a_1, \ldots, a_n)$ is $(-a_n, \ldots, -a_1)$, and the dual of a ribbon is given by rotation through 180°. The evaluation and coevaluation are given in Figure 1.31.

We use ribbons (strips) rather than links (strands) because the 360° turn depicted on the right of Figure 1.31 cannot be straightened without introducing a twist in the strip. Up to isotopy, a ribbon can be thought of as braided knotted strands (the spine of each strip) together with an integer assigned to each strand (saying how much that strip is twisted).

As on the left of Figure 1.26, it is very useful to colour ribbons. Let S be any set; by $\mathbf{Ribbon}_S$ we mean the category with objects $((A_1, s_1), \ldots, (A_k, s_k))$ for $A_i \in S, s_i \in \{\pm\}$. The arrows are as before except now they are coloured with $A \in S$; if the ribbon has endpoints, they must be of the form (A, s) and (A, s') where the signs s, s' are as before.

Two isotopic ribbons define an identity holding in any ribbon category:

Theorem 1.6.2 [473] *Let* $\mathbf{C}$ *be a (strict) ribbon category and let* S *be the set of its objects. Then there exists a unique ribbon functor* $F : \mathbf{Ribbon}_S \to \mathbf{C}$ *such that* $F(A, +) = A$ *and* $F(A, -) = A^*$.

By the usual arguments, any ribbon category $\mathbf{C}$ gives us (isotopy) invariants of braided knotted ribbons. The most interesting (because it is the simplest) special case concerns any ribbon $\mathcal{R} \in \mathrm{Hom}_{\mathbf{Ribbon}}(\emptyset, \emptyset)$ without ends: the invariant $F(\mathcal{R})$ will lie in $\mathrm{Hom}_{\mathbb{C}}(F(\emptyset), F(\emptyset))$. This gives an invariant for any link, by drawing its ribbon with zero twist for each strip. Of course some ribbon categories give a complete link invariant (why?).

Unlike for braids, we have no effective way to determine if linked ribbons or links are ambient isotopic (but see [**283**]), so these invariants are topologically interesting. For example, they permit an easy proof that the trefoil and its mirror image are not ambient isotopic, something that took a clever argument from Dehn to do originally. The functoriality property of F makes them relatively easy to compute.

It is far from obvious that there are any nontrivial calculationally practical examples of ribbon categories, independent of **Ribbon**. Fortunately though there are: although **Ribbon** is geometric, there are several ribbon categories coming from algebra (namely representation theory). In fact, there are now so many that the main value of Theorem 1.6.2 is organisational, conceptually gathering together a plethora of link invariants that have been accumulating since the 1980s, starting with the Jones polynomial.

This treatment can and should be pushed much further, starting with the direct sum $U \oplus V$ of objects. See [**534**], [**398**], [**353**] for more details and developments. The refinement called *modular category* is the one of greatest relevance to the mathematics and physics related to Moonshine. We return to categories in Section 4.4.1.

Vogel [**547**] defined a monoidal category $\mathcal{D}'$, which looks like the category of modules of a Lie algebra. He calls it the *Universal Lie algebra*, since given any simple Lie (super)algebra $\mathfrak{g}$, there is a unique functor from $\mathcal{D}'$ to the category of $\mathfrak{g}$-modules satisfying certain natural properties. Roughly, Vogel assigns each such Lie (super)algebra a different point on the projective plane, from which much of its data can easily be computed. For example, the A-series corresponds to the projective coordinates $[n, 2, -2]$, while the exceptional series (the bottom row of Figure 1.20) falls on the line $[-2, a + 4, 2a + 4]$. The 'universal decompositions' and dimension formulae described in Section 1.5.2 arise because they hold for $\mathcal{D}'$.

Question 1.6.1. A variety is a solution set to a system of polynomial equations over some ring R. Interpret this as a functor from a category of rings to a category of sets.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Question 1.6.2. (a) Find what (if anything) product and sum (in the sense of Figure 1.23) are in the category **Set**.
(b) Same question for the category **Riem**.

Question 1.6.3. (a) Show that in the category **Ab-Group** (where objects are abelian groups and arrows are group homomorphisms), sum and product are identical.
(b) Show that in the category **Group**, product is direct product, but sum is not.

Question 1.6.4. Let L be any lattice (Section 1.2.1). Define a category whose objects are elements of L, with $\mathrm{Hom}(v, v) = \mathbb{C}$ and $\mathrm{Hom}(v, w) = \{0\}$ whenever $v \neq w$. Composition of arrows is multiplication. Complete the construction of a ribbon category for this category, where the braiding $c_{v,w}$ is $e^{\mathrm{i}v\cdot w}$.

1.7 Elementary algebraic number theory

The coefficients of the McKay–Thompson series T_g are always integers, as are the fusion multiplicities $\mathcal{N}_{ab}^{c}$ in RCFT. But non-integers often lurk in the shadows, secretly watching their more arrogant brethren the integers strut. One of the consequences of their presence can be the existence of certain Galois symmetries. The Galois theory of cyclotomic fields plays a background role in Moonshine, much as it does for finite groups and modular forms. We sketch the basics in this section.

Galois automorphisms are a generalisation of complex conjugation. If in your problem complex conjugation seems interesting, then there is a good chance other Galois automorphisms will play a role.

1.7.1 Algebraic numbers

Euler and Lagrange were the first to show that 'weird' (complex) numbers could tell us about the integers, but it took Gauss (*c.* 1831) to do this with care and subtlety. For an example of this idea, suppose we are interested in the equation $n = a^2 + b^2$. Consider for concreteness $5 = 2^2 + 1^2$. We can write this as $5 = (2 + \mathrm{i})(2 - \mathrm{i})$, so we are led to consider complex numbers of the form $a + b\mathrm{i}$, for $a, b \in \mathbb{Z}$. These are now called 'Gaussian integers'.

Fact *Let $p \in \mathbb{Z}$ be any prime number. Then p factorises (i.e. is composite) over the Gaussian integers iff $p = 2$ or $p \equiv 1$ (mod 4).*

Now suppose $p \not\equiv 3 \pmod 4$ is prime, and factorise it $p = (a + b\mathrm{i})(c + d\mathrm{i})$. Then $p^2 = (a^2 + b^2)(c^2 + d^2)$, so $a^2 + b^2 = c^2 + d^2 = p$. Conversely, suppose $p = a^2 + b^2$, then $p = (a + b\mathrm{i})(a - b\mathrm{i})$. Thus:

Consequence[15] *Let $p \in \mathbb{Z}$ be any prime number. Then $p = a^2 + b^2$ for $a, b \in \mathbb{Z}$ iff $p = 2$ or $p \equiv 1$ (mod 4).*

[15] This result was first stated by Fermat in one of his infamous margin notes (another is discussed shortly), and was finally proved a century later by Euler. For a one-line proof see Question 1.7.1.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Now we can answer the question: when can n be written as a sum of two squares $n = a^2 + b^2$? Write out the prime decomposition $n = \prod p^{a_p}$. Then $n = a^2 + b^2$ has a solution iff a_p is *even* for every $p \equiv 3 \pmod 4$. For instance, $60 = 2^2 \cdot 3^1 \cdot 5^1$ cannot be written as the sum of two squares, but $90 = 2^1 \cdot 3^2 \cdot 5^1 = \{(1 + \mathrm{i})3(1 + 2\mathrm{i})\}\{(1 - \mathrm{i})3(1 - 2\mathrm{i})\}$ can (e.g. $90 = (-3)^2 + 9^2$). In fact we can find and count all solutions.

More generally, let $\mathbb{K}$ be any subfield of $\mathbb{C}$ (usually we take $\mathbb{K} = \mathbb{Q}$) and $\alpha_1, \ldots, \alpha_n$ be any complex numbers. We discussed 'field' in Section 1.1.1. By $\mathbb{L} = \mathbb{K}(\alpha_1, \ldots, \alpha_n)$ we mean the smallest field containing $\mathbb{K}$ and all α_i. In other words, $\mathbb{L}$ consists of all rational functions poly/poly of the α_i, with coefficients in $\mathbb{K}$. Then $\mathbb{L}$ can be thought of as a vector space over $\mathbb{K}$; write $[\mathbb{L} : \mathbb{K}] \leq \infty$ for the dimension of that vector space. We say $\mathbb{L}$ is an *extension* of the *base-field* $\mathbb{K}$ of *degree* $[\mathbb{L} : \mathbb{K}]$. The most interesting case is when the degree $[\mathbb{L} : \mathbb{K}]$ is *finite*. In this case we can find a single number $\alpha \in \mathbb{L}$ such that $\mathbb{L} = \mathbb{K}[\alpha]$, where as always we write $R[x]$ for all polynomials in x with coefficients in R. Then α will be a zero of a monic polynomial $p(x) \in \mathbb{K}[x]$ and of degree $[\mathbb{L} : \mathbb{K}]$, called the *minimal polynomial* of α. Such α are called *algebraic*, and such extensions $\mathbb{K}[\alpha]$ are called *finite*. The finite extensions most relevant for this book are discussed in Section 1.7.3.

Numbers of course arise throughout science in their role as coordinates; less appreciated is that observing the specific kinds of numbers that arise can provide profound structural information. *This is very much how algebraic number theory impinges on the areas considered in this book.* For an elementary example, recall that Euclid's books are filled with geometric constructions, particularly those involving straight-edge (i.e. drawing the line passing through two points) and compass (i.e. drawing the circle with given centre and radius). The reader can discover for herself how to trisect line segments and double the area of a square, using only straight-edge and compass. But some problems weren't solved back then: for example, how to trisect an angle or double the volume of a cube. To solve these, consider coordinates. Suppose we start with N points (x_i, y_i). We can construct the line joining any two of those points, and the circle centred at some (x_i, y_i) with some radius $|(x_j, y_j) - (x_k, y_k)|$; we can construct new points only as intersections of these lines and circles. Now, if we let $\mathbb{K}$ denote the field generated from $\mathbb{Q}$ by all $2N$ coordinates x_i, y_i, then the equations of our lines and circles will have coefficients belonging to $\mathbb{K}$. The coordinates of the intersection of any two such lines will lie in $\mathbb{K}$, while that of the intersection of a line with a circle, or of two circles, will lie in an extension $\mathbb{L}_1$ of $\mathbb{K}$ of degree $[\mathbb{L}_1 : \mathbb{K}] = 2$. Continuing in this way, we see that any construction, no matter how involved, can only construct points whose coordinates lie in some extension $\mathbb{L}$ of $\mathbb{K}$ of degree a power of 2. Now, given an angle θ, defined by points (0,0), (1,0) and $(\cos(\theta), \sin(\theta))$, trisecting θ means constructing the point $(\cos(\theta/3), \sin(\theta/3))$. But $\alpha = \cos(\theta/3)$ obeys $\cos(\theta) = 4\alpha^3 - 3\alpha$, i.e. $\cos(\theta/3)$ lies (generically) in a degree-3 extension of $\mathbb{K} = \mathbb{Q}[\cos(\theta), \sin(\theta)]$. Thus we cannot trisect that angle, using only a compass and straight-edge, for most θ (e.g. $\theta = 60^\circ$).

The degree $[\mathbb{L} : \mathbb{K}]$ is a (rather crude) invariant of the field extension $\mathbb{L} \supset \mathbb{K}$. We have just seen the power of this simple invariant; in the next subsection we refine it considerably, giving it a group structure.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Consider 'Fermat's Last Theorem', which asserts that there are no positive integer solutions to the equation $x^n + y^n = z^n$, for $n > 2$. It is tempting, as Fermat himself probably did, to factorise this into

$$\prod_{j=0}^{n-1} \left(x + \xi_{2n}^{2j+1} y\right) = z^n,$$

where $\xi_m = \exp[2\pi \mathrm{i}/m]$, and to try to show from this that each $a + \xi_{2n}^{2j+1} b$ has an 'integral' nth root, if $x = a$, $y = b$, $z = c$ is an integral solution. We return to Fermat's Last Theorem in Section 2.2.1.

These examples should give the reader some appreciation for the value of using non-integers to study integers, and also provide some impetus for extending the tools and notions of high school number theory (primes, divisibility, etc.) to complex numbers. The result is *algebraic number theory*. A classic introduction is [**282**]; the book [**515**] is filled with concrete examples.

Euler worked with numbers of the form $\ell + m\sqrt{n}$, for $\ell, m, n \in \mathbb{Z}$, and regarded them as generalised integers, carrying over (without proof) their divisibility laws, etc. from the usual integers. However, it was soon learned that care must be taken. For a simple example, the factorisation $2 = (n - \sqrt{n^2 - 2})(n + \sqrt{n^2 - 2})$ holds for all $n \in \mathbb{Z}$, so what should the 'unique prime factorisation' of 2 be?

The basic theory was developed in the nineteenth century, by Kummer, Dedekind, Frobenius and others. Take the base field $\mathbb{K}$ to be $\mathbb{Q}$ for convenience, and fix a finite extension $\mathbb{L} = \mathbb{Q}[\alpha]$. Any $z \in \mathbb{L}$ is algebraic, i.e. satisfies $a_m z^m + a_{m-1} z^{m-1} + \cdots + a_0 = 0$ for some $a_i \in \mathbb{Z}$ (not all zero). The $\mathbb{L}$-*integers* are those numbers $z \in \mathbb{L}$ that satisfy $z^m + a_{m-1} z^{m-1} + \cdots + a_0 = 0$ for some $a_i \in \mathbb{Z}$ (i.e. $a_m = 1$). The sum and products of $\mathbb{L}$-integers are $\mathbb{L}$-integers, and so we call the set $R_{\mathbb{L}}$ of all these $\mathbb{L}$-integers the *ring of integers*. For example, when $\mathbb{L} = \mathbb{Q}$, $\mathbb{Q}[\mathrm{i}]$, $\mathbb{Q}[\sqrt{2}]$ and $\mathbb{Q}[\sqrt{5}]$, respectively, the ring of integers are $\mathbb{Z}$, $\mathbb{Z} + \mathrm{i}\mathbb{Z}$, $\mathbb{Z} + \sqrt{2}\mathbb{Z}$, and

$$\{(m + n\sqrt{5})/2 \,|\, m, n \in \mathbb{Z},\ m - n \in 2\mathbb{Z}\},$$

respectively. All elements of $\mathbb{L}$ are quotients of $\mathbb{L}$-integers, just as all $r \in \mathbb{Q}$ equal a/b for $a, b \in \mathbb{Z}$.

What should prime mean here? The obvious guess would be any number $\gamma \in R_{\mathbb{L}}$ whose only divisors β are trivial, i.e. the only $\beta \in R_{\mathbb{L}}$ with $\gamma/\beta \in R_{\mathbb{L}}$ are units or γ times units. Units are the analogue here of ± 1: an $\mathbb{L}$-integer u is a unit iff u^{-1} is also an $\mathbb{L}$-integer. The only problem with this definition of prime is that unique factorisation is usually lost. For example, in $\mathbb{L} = \mathbb{Q}[\sqrt{-26}]$, the $\mathbb{L}$-integers are $\mathbb{Z} + \sqrt{-26}\mathbb{Z}$; we have the equation

$$3^3 = 27 = (1 + \sqrt{-26})(1 - \sqrt{-26})$$

and yet, as the reader can easily verify, both 3 and $1 \pm \sqrt{-26}$ are primes by our definition. Incidentally, most finite extensions $\mathbb{L}$ have infinitely many $\mathbb{L}$-units (e.g. $(1 + \sqrt{2})^n$ is a unit of $\mathbb{Q}[\sqrt{2}]$ for any $n \in \mathbb{Z}$).

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

The correct definition of prime (Dedekind, 1871) is a gem. Replace the single $\mathbb{L}$-integer $\gamma \in R_{\mathbb{L}}$ with the set of all multiples $R_{\mathbb{L}}\gamma =: (\gamma)$ of that number. This washes away the irritating ambiguity due to units. Any subset $I \subseteq R_{\mathbb{L}}$ closed under $R_{\mathbb{L}}$-linear combinations (i.e. for which $\sum a_i z_i \in I$ for all $z_i \in I$ and $a_i \in R_{\mathbb{L}}$) is called an *ideal* of $R_{\mathbb{L}}$. For example, (γ) is always an ideal, though for typical rings $R_{\mathbb{L}}$, most ideals won't have a single generator. Consider any ideals I, J of $R_{\mathbb{L}}$. By the product of ideals we mean

$$IJ = \left\{ \sum a_i b_i \mid a_i \in I,\ b_i \in J \right\}.$$

A prime ideal is defined to be any nonzero ideal $P \neq R_{\mathbb{L}}$ such that $IJ = P$ for ideals I, J only if $I = R_{\mathbb{L}}$ or $J = R_{\mathbb{L}}$. In $R_{\mathbb{L}}$, any prime ideal P is maximal (and conversely): the only ideals I satisfying $P \subset I \subset R_{\mathbb{L}}$ are $I = P, R_{\mathbb{L}}$. Although unique factorisation usually won't hold for $\mathbb{L}$-integers, it always holds for ideals: any nonzero ideal I of the ring $R_{\mathbb{L}}$ of integers can be written uniquely as a product of prime ideals.

For example, the prime ideals of $\mathbb{Z}$ are (p) for p prime, and this reduces to the usual unique factorisation of integers. The unique factorisation of the ideal (27) in the field $\mathbb{Q}[\sqrt{-26}]$ is $(27) = P_+^3 P_-^3$, where $P_\pm := (3, 1 \pm \sqrt{-26}) = (3) \cap (1 \pm \sqrt{-26})$. Thus neither $(3) = P_+ P_-$ nor $(1 \pm \sqrt{-26}) = P_\pm^3$ are prime.

We are thus led to picture $\mathbb{L}$-integers as ideals of the ring $R_{\mathbb{L}}$. In fact the name 'ideal', now standard in algebra, was chosen because it corresponds to an *ideal* – as opposed to *true* – number.

This reinterpretation of integers as ideals has a striking geometric parallel. We are taught to study a geometric space X through the functions $f \in \mathbb{C}[X]$ that live on it. In this language, what should play the role of a point $x \in X$? Given any point $a \in X$, we can evaluate these functions $f(x)$ at $x = a$. Algebraically, this corresponds to a homomorphism $\mathbb{C}[X] \to \mathbb{C}$. Those homomorphisms, via their kernels, are essentially in one-to-one correspondence with ideals of the ring $\mathbb{C}[X]$, and thus we should identify points $x \in X$ with certain ideals in $\mathbb{C}[X]$. Looking at concrete examples such as $X = \mathbb{C}^n$, we find that ideals correspond more generally to submanifolds (subvarieties) in X, and that maximal ideals correspond to points. This unexpected and deep connection between number theory and geometry is a great illustration of the effectiveness of abstract algebra.

1.7.2 Galois

Evariste Galois was a brilliantly original French mathematician. Born shortly before Napoleon's ill-fated invasion of Russia, he died shortly before the ill-fated 1832 uprising in Paris. His last words: 'Don't cry, I need all my courage to die at 20'.

Galois grew up in a time and place confused and excited by revolution. He was known to say 'if only I were sure that a body would be enough to incite the people to revolt, I would offer mine'. On 2 May 1832, after frustration over failure in love and failure to convince the Paris mathematical establishment of the depth of his ideas, he made his decision. A duel was arranged with a friend, but only his friend's gun would be loaded. Galois died the day after that bullet perforated his intestine. At his funeral it was

discovered that a famous general had also just died, and the revolutionaries decided to use the general's death rather than Galois' as a pretext for an armed uprising. A few days later the streets of Paris were blocked by barricades, but not because of Galois' sacrifice: his death had been pointless [**529**].[16]

Galois theory in its most general form is the study of relations between objects defined implicitly by some conditions.[17] For example, the objects could be the solutions to a given differential equation. Or the objects could be the different points $\pi^{-1}(p) \subset Y$ sitting above a given point $p \in X$ in a cover $\pi : Y \to X$. In the most familiar incarnation of Galois theory, the objects are the zeros of certain polynomials.

Look at complex conjugation: $\overline{wz} = \overline{w}\,\overline{z}$ and $\overline{w+z} = \overline{w} + \overline{z}$. Also, $\overline{x} = x$ for any $x \in \mathbb{R}$. So we can say that $z \mapsto \overline{z}$ is a structure-preserving map $\mathbb{C} \to \mathbb{C}$ (called an *automorphism* of $\mathbb{C}$) fixing the reals. We say that complex conjugation belongs to the Galois group $\mathrm{Gal}(\mathbb{C}/\mathbb{R})$ of $\mathbb{C}$ over $\mathbb{R}$; apart from complex conjugation, it contains only the identity automorphism.

A way of thinking about the automorphism $\overline{z}$ is that it says that, as far as the real numbers are concerned, i and $-$i are identical twins. Algebra alone can't tell that i is in the upper half-plane, or that going from 1 to i is going counterclockwise about 0, while 1 to $-$i is clockwise.

Let $\mathbb{L}$ be any field containing $\mathbb{Q}$. The *Galois group* $\mathrm{Gal}(\mathbb{L}/\mathbb{Q})$ is the set of all automorphisms=symmetries of $\mathbb{L}$ that fix all rationals.

For example, $\mathbb{L} = \mathbb{Q}[\sqrt{5}]$ is the field of all numbers of the form $a + b\sqrt{5}$, where $a, b \in \mathbb{Q}$. Let's try to find its Galois group. Let $\sigma \in \mathrm{Gal}(\mathbb{F}/\mathbb{Q})$. Then $\sigma(a + b\sqrt{5}) = \sigma(a) + \sigma(b)\sigma(\sqrt{5}) = a + b\,\sigma(\sqrt{5})$, so once we know what σ does to $\sqrt{5}$, we know everything about σ. But $5 = \sigma(5) = \sigma(\sqrt{5}^2) = (\sigma(\sqrt{5}))^2$, so $\sigma(\sqrt{5}) = \pm\sqrt{5}$ and again there are precisely two possible Galois automorphisms here (one is the identity). As far as the arithmetic of $\mathbb{Q}$ is concerned, $\pm\sqrt{5}$ are interchangeable.

Consider more generally any extension $\mathbb{L}$ of the base field $\mathbb{K}$ of degree $n = [\mathbb{L} : \mathbb{K}] < \infty$. As mentioned in the last subsection, these are always of the form $\mathbb{L} = \mathbb{K}[\alpha]$, where α is the root of a monic polynomial $p(x)$ of degree n with coefficients in $\mathbb{K}$. This means any $z \in \mathbb{L}$ is expressible as a polynomial in α with coefficients in $\mathbb{K}$, of degree $< n$. Hence, any automorphism $\sigma \in \mathrm{Gal}(\mathbb{L}/\mathbb{K})$ is uniquely specified by the value $\sigma(\alpha) \in \mathbb{L}$. Since $\sigma(p(x)) = p(\sigma x)$, σ must send α to one of the n roots of $p(x)$. Thus $\|\mathrm{Gal}(\mathbb{L}/\mathbb{K})\| \leq [\mathbb{L} : \mathbb{K}]$. Extensions $\mathbb{L}$ for which $\mathrm{Gal}(\mathbb{L}/\mathbb{K})$ is maximally large (i.e. of order n) are the most interesting and are called *Galois*: they are the extensions for which all roots of $p(x)$ are in $\mathbb{L}$.

[16] Apparently this treatment of Galois' life has been disputed. But surely the main purposes of history are for supplying a context and motivation, for its sheer entertainment value, and for drawing Lofty Morals. And at least when they are successful, it is probably wisest if neither motivation nor entertainment nor Morality be investigated too closely...

[17] This is the dynamic point of view, but the reader should be warned that there is an alternate interpretation. Abstracting out the more structural side of Galois theory, many authors regard Galois theory as ultimately a contravariant functorial correspondence associating to some objects $A, B, \ldots$ (e.g. groups) other objects $K, L, \ldots$ (e.g. fields invariant under the group action) in such a way that $A \subset B$ corresponds to $K \supset L$.

Let $\mathbb{L} \supset \mathbb{K}$ be a finite Galois extension, and write $G := \mathrm{Gal}(\mathbb{L}/\mathbb{K})$. The classical Galois Theorem sets up a natural bijection between fields $\mathbb{J}, \mathbb{L} \supset \mathbb{J} \supset \mathbb{K}$, and subgroups H of G. In particular, to the field $\mathbb{J}$ associate the subgroup $H = \mathrm{Gal}(\mathbb{L}/\mathbb{J})$, and to the subgroup H associate the space (in fact field) $\mathbb{J} = \mathbb{L}^H$ of all elements $z \in \mathbb{L}$ fixed by all $\sigma \in H$. Then $[\mathbb{J} : \mathbb{K}] = \|G/H\|$, and the extension $\mathbb{J} \supset \mathbb{K}$ is Galois iff H is normal in G, in which case $\mathrm{Gal}(\mathbb{J}/\mathbb{K}) \cong G/H$.

We saw earlier the power of the numerical invariant $[\mathbb{L} : \mathbb{K}]$. We should think of $\mathrm{Gal}(\mathbb{L}/\mathbb{K})$ as a group-valued refinement of degree. For an application, suppose for contradiction that we have a general formula for the zeros of any polynomial $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ of degree n. For $n = 2$ we have the quadratic formula (which involves square-roots), and we've all seen the formula for $n = 3$ (which involves square-roots and cube-roots). Does there exist a formula for any n, involving taking arbitrary nested roots of rational expressions in the coefficients a_i? Let $\mathbb{K} = \mathbb{Q}[a_0, \ldots, a_n, \xi_1, \xi_2, \ldots]$ – we include in $\mathbb{K}$ all roots of unity so that all extensions below will be Galois. Then the first kth root we come to in our formula will move us into a Galois extension $\mathbb{K}_1$ of $\mathbb{K}$, with Galois group $\mathrm{Gal}(\mathbb{K}_1/\mathbb{K}) \cong \mathbb{Z}_k$. If the hypothetical formula involves a second radical, requiring us to take say an ℓth root of a rational expression in $\mathbb{K}_1$, then this takes us into a Galois extension $\mathbb{K}_2$ of $\mathbb{K}_1$, with Galois group $\mathrm{Gal}(\mathbb{K}_2/\mathbb{K}_1) \cong \mathbb{Z}_\ell$ – that is, $\mathrm{Gal}(\mathbb{K}_2/\mathbb{K})$ is an extension of the cyclic group $\mathbb{Z}_\ell$ by $\mathbb{Z}_k$. Continuing in this way until all roots in our hypothetical formula are exhausted, we would find that the zeros of the general degree-n polynomial would lie in a Galois extension $\mathbb{L}$ of $\mathbb{K}$ whose Galois group is obtained by repeatedly extending by cyclic groups. Such a group is called *solvable* (Section 1.1.3) for this reason. It is easy to see that $\mathrm{Gal}(\mathbb{L}/\mathbb{K})$ here is in fact the symmetric group $\mathcal{S}_n$, and that $\mathcal{S}_n$ is solvable iff $n \leq 4$ (recall that $\mathcal{A}_5$ is simple!). Thus a general formula for the roots of a general polynomial of degree n, involving nested radicals, can exist only for $n \leq 4$.

Every area of mathematics has a Galois-type theory. In geometry, for instance, covers $f : M \to N$ of a fixed manifold N are in one-to-one correspondence with subgroups $H \cong \pi_1(M)$ of the fundamental group $G := \pi_1(N)$; $\gamma \in \pi_1(N)$ belongs to H iff γ lifts to a closed loop in M. When the subgroup H is normal, G/H is naturally isomorphic to the group of all homeomorphisms $\alpha : M \to M$ satisfying $f \circ \alpha = f$ (these α are called *covering transformations*). See the beautiful book [**363**]. The question 'What is the Galois theory for von Neumann algebras?' led Jones to subfactor theory $M \supset N$ – for instance, his index $[M : N] \in \mathbb{R} \cup \{\infty\}$ plays the role of the degree $[\mathbb{L} : \mathbb{K}] \in \mathbb{Z} \cup \{\infty\}$. Just as the degree $[\mathbb{L} : \mathbb{K}]$ can be refined into the Galois group $\mathrm{Gal}(\mathbb{L}/\mathbb{K})$, the Jones index can be refined into a topological field theory (see Section 6.2.6).

Galois theory is reminiscent, at least qualitatively, of Gödel's Incompleteness Theorem. In mathematics we generally start with a model (e.g. Euclidean geometry or the natural numbers) that we try to capture implicitly by an axiomatic system. Gödel's Theorem tells us that there are infinitely many different models compatible with the given axiomatic system, regardless of how many axioms we include. Each of these is obtained by realising in incompatible ways the undefined terms of the axiomatic system.

Of course it is the *model* and not the *axiomatic system* in which most mathematics occurs. For example, we don't criticise Wiles' work on Fermat's Last Theorem on the grounds that his proof assumes $\mathbb{N}$ is embedded in $\mathbb{C}$, even though this transcendental interpretation of $\mathbb{N}$ surely is not a consequence of Peano's axioms (the axiomatic system describing the natural numbers). Likewise, [**459**] gives a simple statement about $\mathbb{N}$; it is easy to prove using standard arguments involving $\mathbb{R}$, but neither it nor its negation can be proved using only Peano's axioms.

1.7.3 Cyclotomic fields

We are primarily interested in a simple class of numbers: those in the *cyclotomic extensions* of $\mathbb{Q}$. These are the fields $\mathbb{Q}[\xi_n]$, consisting of all polynomials $a_m\xi_n^m + a_{m-1}\xi_n^{m-1} + \cdots + a_0$ in the root of unity $\xi_n := \exp[2\pi\mathrm{i}/n]$, for all $a_i \in \mathbb{Q}$. For instance, $\cos(\pi r)$, $\sin(\pi r)$ and $\sqrt{r}$ are cyclotomic numbers for any $r \in \mathbb{Q}$. In particular,

$$\cos\left(2\pi\frac{m}{n}\right) = \frac{\xi_n^m + \xi_n^{-m}}{2}, \tag{1.7.1a}$$

$$\sin\left(2\pi\frac{m}{n}\right) = \frac{\xi_n^m - \xi_n^{-m}}{2\mathrm{i}}, \tag{1.7.1b}$$

$$\sqrt{p} = c_p \sum_{n=0}^{p-1} \xi_p^{n^2}, \tag{1.7.1c}$$

for any nonzero $m, n \in \mathbb{Z}$, and any odd prime p, where $c_p = 1$ or $-\mathrm{i}$ for $p \equiv \pm 1 \pmod 4$, respectively ((1.7.1c) is called a *Gauss sum*). Only countably many complex numbers are cyclotomic, i.e. lie in $\cup_{n=1}^{\infty}\mathbb{Q}[\xi_n]$, so almost every complex number is not cyclotomic.

Cyclotomic numbers are the numbers in the character tables of finite groups, the values of Lie group characters at elements of finite order, the values of quantum-dimensions in RCFT, and the matrix entries in the $\mathrm{SL}_2(\mathbb{Z})$-representation coming from rational VOAs. The theory is deeply entwined with that of modular forms and functions, as we see in Section 2.3.3. The key property of cyclotomic numbers, which accounts for their ubiquity, has to do with their Galois groups.

As usual, an automorphism $\sigma \in \mathrm{Gal}(\mathbb{Q}[\xi_n]/\mathbb{Q})$ is uniquely determined by what it does to the generator ξ_n. Since $\xi_n^n = 1$, we see that σ must send ξ_n to another nth root of 1, ξ_n^ℓ say; in fact $\sigma(\xi_n)$ must be another 'primitive' nth root of 1, that is ℓ must be coprime to n. So $\mathrm{Gal}(\mathbb{Q}[\xi_n]/\mathbb{Q})$ is isomorphic to the multiplicative group $\mathbb{Z}_n^\times$ of numbers between 1 and n coprime to n. To see what σ does to some $z \in \mathbb{Q}[\xi_n]$, we find the $\ell \in \mathbb{Z}_n^\times$ corresponding to σ and write z as a $\mathbb{Q}$-polynomial $p(\xi_n)$: then $\sigma z = p(\xi_n^\ell)$. For example,

$$\sigma\big(\cos(2\pi a/n)\big) = \sigma\left(\frac{\xi_n^a + \xi_n^{-a}}{2}\right) = \frac{\xi_n^{a\ell} + \xi_n^{-a\ell}}{2} = \cos(2\pi a\ell/n).$$

The defining property of cyclotomic numbers is a central result of classical number theory:

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Theorem 1.7.1 (Kronecker–Weber) *Let $\mathbb{L}$ be a finite Galois extension of $\mathbb{Q}$ with abelian Galois group* $\mathrm{Gal}(\mathbb{L}/\mathbb{Q})$. *Then $\mathbb{L}$ is contained in some cyclotomic extension $\mathbb{Q}[\xi_n]$.*

The proof is quite complicated. Conversely, any cyclotomic extension $\mathbb{Q}[\xi_n]$ of $\mathbb{Q}$ is finite Galois and has abelian Galois group. In fact, the degree of $\mathbb{Q}[\xi_n]$ is given by Euler's ϕ-function:

$$[\mathbb{Q}[\xi_n] : \mathbb{Q}] = \phi(n) := n \prod_{p|n} \frac{p-1}{p}.$$

The minimal polynomial of ξ_n is called the nth *cyclotomic polynomial*; a manifestly integral construction for it is given in [**64**]. Its zeros are ξ_n^i for each i coprime to n.

The ring of *cyclotomic integers* $R_{\mathbb{Q}[\xi_n]}$ is simply $\mathbb{Z}[\xi_n]$. For all $n \neq 1, 2, 4$, $\mathbb{Q}[\xi_n]$ has infinitely many units: for example, $(\xi_n^i - 1)/(\xi_n - 1)$ is a unit of infinite order, for any $1 < i < n - 1$ coprime to n. Unique factorisation at the level of numbers (as opposed to ideals, which always holds) fails in all but 30 cyclotomic fields ($\mathbb{Q}[\xi_{23}]$ is the first field for which it fails).

Kronecker's *Jungentraum* ('dream of youth') [**546**] proposes that just as all abelian extensions of $\mathbb{Q}$ are obtained by adjoining to $\mathbb{Q}$ the values of a transcendental function (namely $\exp[2\pi i z]$) at certain algebraic numbers (namely $z \in \mathbb{Q}$), something similar should happen for abelian extensions of other finite extensions $\mathbb{K}$. This is still far from understood in general, but we know that any abelian extension of $\mathbb{K} = \mathbb{Q}[\sqrt{-d}]$ is contained in an extension of $\mathbb{K}$ by a root of unity, square-roots of integers, and the j-function (0.1.8) evaluated at $(a + \sqrt{b})/2$ for some $a, b \in \mathbb{Z}$.

Question 1.7.1. [**572**] (a) Show that a prime $p \equiv 3 \pmod 4$ cannot be written in the form $a^2 + b^2$ for integers a, b.
(b) Let $p \equiv 1 \pmod 4$ be prime. Define

$$S_p = \{(x, y, z) \in \mathbb{Z}^3 \mid x > 0, y > 0, z > 0,\ x^2 + 4yz = p\}.$$

Verify that for any $(x, y, z) \in S_p$, both $x \neq y - z$ and $x \neq 2y$. Define a map L on S_p by

$$L(x, y, z) = \begin{cases} (x + 2z, z, y - x - z) & \text{if } x < y - z \\ (2y - x, y, x - y + z) & \text{if } y - z < x < 2p \\ (x - 2y, x - y + z, y) & \text{if } x > 2y \end{cases}$$

Verify that L is an involution (i.e. $L(L(x, y, z)) = (x, y, z)$), and that L has exactly one fixed point. Show that this implies that the cardinality $\|S_p\|$ must be odd, and thus that the involution $(x, y, z) \mapsto (x, z, y)$ must also have a fixed point. Conclude that any prime $p \equiv 1 \pmod 4$ has a solution $p = a^2 + b^2$.

Question 1.7.2. Suppose we are given two points P, Q in the plane, distance 1 apart. Determine whether it is possible, using only a straight-edge and compass, to construct a point R collinear with P and Q such that the distance between P and R is $2^{-1/3}$. What if the distance between P and R is instead required to be $2^{-1/4}$?

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press

Question 1.7.3. Let $\mathbb{K} = \mathbb{Q}[2^{1/3}]$. Show that $\mathrm{Gal}(\mathbb{K}/\mathbb{Q})$ is trivial.

Question 1.7.4. Let $\mathbb{L} = \mathbb{Q}[\sqrt{2}, \sqrt{3}]$.
(a) Find an α such that $\mathbb{L} = \mathbb{Q}[\alpha]$.
(b) Find $\mathrm{Gal}(\mathbb{L}/\mathbb{Q})$. Is $\mathbb{L}$ Galois?
(c) For each subgroup H of $\mathrm{Gal}(\mathbb{Q}[\sqrt{2}, \sqrt{3}]/\mathbb{Q})$, find the corresponding extension $\mathbb{J}$.

Question 1.7.5. (a) Show that the values $\mathrm{ch}(g)$ of characters are always cyclotomic integers. After reading this section, can you add anything to your answer to Question 1.1.5?
(b) Let G be any finite group. Prove: G is simple iff for all irreducible characters ch of G, $\mathrm{ch}(a) = \mathrm{ch}(e)$ only when $a = e$.

Question 1.7.6. Find all rational numbers r such that $\cos(2\pi r) \in \mathbb{Q}$.

https://doi.org/10.1017/9781009401548.002 Published online by Cambridge University Press