4
Conformal field theory: the physics of Moonshine

This chapter presents the physical context for Moonshine. Rather than diving into a conventional discourse of conformal field theory (CFT), it might be more helpful to take several steps back and begin with Galileo. Physics even more than mathematics is interwoven with history. Our treatment of CFT is sketchy but should supply the reader with all that is necessary to appreciate the absolutely profound role physics has played in Moonshine and other aspects of 'pure' mathematics in recent years. It is hoped that this chapter will make it easier for the interested reader to pursue more standard treatments of CFT and string theory. It is written primarily with the mathematician in mind.

The third section explores the physics of CFT, and the fourth describes some mathematical formulations. CFT is to a generic quantum field theory what finite-dimensional semi-simple Lie algebras are to generic Lie algebras. Background for both sections is provided by the review of classical and quantum physics sketched in the first two sections.

For a mathematician studying physics, important to keep in mind is that physics has been driven historically more by its predictive power than by conceptual concerns (with a few remarkable exceptions, such as Einstein's general relativity). Given enough time, however, the theory becomes polished to a state of pristine mathematical elegance, as classical mechanics amply demonstrates. In particular, one has the sense that quantum theory is *ad hoc* and rather unsound – and it is both – but these features are due to the historical accident that we were born too close to its inception. Much more important is what it can teach mathematics, which is considerable. *The essence of quantum field theory is completely accessible to mathematicians and,* as mathematics of the late twentieth century shows, *should at least in its broad strokes be part of their standard repertoire*.

A special feature of classical physics is that the behaviour of a system – for example, its trajectory in phase space – becomes much simpler when looked at infinitesimally. The simple universal regularities are captured by differential equations; the complicated incidental features of a specific situation are relegated to the initial conditions. Among mathematicians, this central role of partial differential equations in classical physics was responsible for what had been a near-identification of their study with the subject they call mathematical physics. It was largely with the arrival of string theory that a much richer range of mathematics became relevant to physics, and it is this happy development that made this book possible.

Almost every facet of Moonshine fits comfortably into CFT, where it often was discovered first. Some have questioned though the necessity of involving such a complicated

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

beast, or the closely related 'vertex operator algebras' of the next chapter, in our mathematical explanation of Moonshine. Although CFT has been an invaluable guide so far, they would argue, perhaps we are a little too steeped in its lore. Undoubtedly there is truth in this, but CFT still has new insights to share. It is an integral part of Moonshine's future as much as its past. Sections 4.3 and 4.4 are central to the whole book.

4.1 Classical physics

4.1.1 Nonrelativistic classical mechanics

Temporarily forget what you know of physics. One of the most blatant empirical facts must be that anything in motion on Earth eventually slows to a stop. On the other hand, stars and planets clearly behave otherwise, therefore earthly laws can't apply directly to the Heavens. Those observations are fundamental to Aristotelian physics. The starting point, however, for classical physics is *Newton's First Law*: the remarkable thought (due to Galileo, 1632) that anything anywhere will continue to move in a straight line and at constant speed, unless something (by definition a *force*) acts on it. Although in isolation it has no real content, it presents a powerful strategy for analysing Nature. For example, to first approximation the Moon travels in a circle about the Earth; rather than trying to conceive of some strange mechanism responsible for pushing or dragging the Moon in its nonlinear orbit, the First Law instead leads us to imagine some 'force' that always pulls the Moon towards the Earth. This second possibility is much more promising of course, and led Newton to his theory of gravitation.

Classical mechanics describes systems with finitely many degrees of freedom. The *configuration* (snapshot, instantaneous state) of a classical system at an instant t of time can be identified with the precise values of all degrees of freedom (e.g. position coordinates) at that time. The basic challenge is to predict the configuration at later times. This amounts to setting up and solving a system of differential equations, called the equations of motion of the system.

Consider a system of N particles, with positions $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$. The $3N$ degrees of freedom are the position coordinates x_{ij}. The equations of motion, which determine the trajectories of the N particles by giving their response to the stimulus, are

$$m_i \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x}_i = \mathbf{F}_i, \tag{4.1.1}$$

where $\mathbf{F}_i$ is the net force experienced by the ith particle and the proportionality constant m_i is called its mass. Dots are used to denote time derivatives: for example, velocity is $\dot{\mathbf{x}}$ and acceleration is $\ddot{\mathbf{x}}$. Note that (4.1.1) is compatible with Newton's First Law.

In general the force $\mathbf{F}_i$ can be a function of all positions $\mathbf{x}_j$, velocities $\mathbf{v}_j$ and time t – for example, air resistance is approximately proportional to $\mathbf{v}_i^2$. We will restrict attention to the typical ones (from which can be derived all others), which are of the form

$$(\mathbf{F}_i)_j = -\frac{\partial}{\partial x_{ij}} V(\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

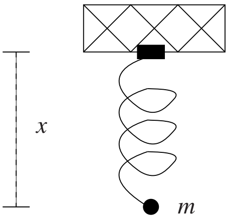

Fig. 4.1 The harmonic oscillator.

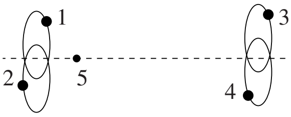

Fig. 4.2 Singular motion of five gravitationally interacting particles.

for some real-valued function V called the *potential*. These are called *conservative* forces because they conserve (keep constant) energy. The potential has units of energy, and the sign is introduced so that V contributes positively to total energy. In quantum mechanics the potential V is more fundamental than the force $\mathbf{F}$.

For example, Newton's gravitational potential is $V = -\sum_{i<j} G \frac{m_i m_j}{|\mathbf{x}_i - \mathbf{x}_j|}$, where G is a positive constant. Einstein found it profoundly significant that the gravitational 'charge' m_i here is numerically (though certainly not conceptually) identical to the 'inertial' mass m_i in (4.1.1) (see Section 4.1.2).

For a one-dimensional example, consider a *harmonic oscillator* – for example, the spring in Figure 4.1. Hooke's Law says that the force $F = -k\,(x - x_0)$, where k is a positive constant and x_0 is the resting length of the spring. Hence $-k\,(x - x_0) = m\ddot{x}$, so

$$x = x_0 + a \cos\left(\sqrt{\frac{k}{m}}\,t\right) + b \sin\left(\sqrt{\frac{k}{m}}\,t\right) = x_0 + A \cos\left(\sqrt{\frac{k}{m}}\,t + B\right). \tag{4.1.2}$$

This force is conservative, with potential $V = \frac{1}{2}k\,(x - x_0)^2$. This elementary system is fundamental to theoretical physics, as it describes small oscillations about stable equilibrium states (i.e. points at which all forces $\mathbf{F}_i$ vanish). Indeed, if $\mathrm{d}V/\mathrm{d}x$ vanishes at $x = x_0$, for some potential V, then the Taylor expansion of $V(x)$ would begin like $a_0 + a_2(x - x_0)^2$, and so it would behave like a harmonic oscillator. We encounter the harmonic oscillator repeatedly in the following pages; in classical field theory these humble oscillations describe, for example, sound waves, and in quantum field theory they are the particles.

The mathematical difficulties faced by quantum field theory are notorious, but remarkably singular behaviour occurs in classical mechanics as well. For one example, consider five point particles interacting gravitationally, positioned as in Figure 4.2. Particle 5 moves horizontally between the orbiting pairs 1 and 2, and 3 and 4. It is possible [**485**] to arrange for particle 5 to zip back-and-forth between those pairs, picking up speed, until in a finite time it reaches infinite speed without ever colliding with the other particles.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

Many other examples of singular behaviour in classical mechanics are possible [**485**]; it is not known yet how typical they are among all possible motions.

Later in this section and the next, we touch on other mathematical difficulties plaguing our physical theories. Generally speaking, these difficulties of classical and quantum physics have to do with probing space to arbitrarily high precision. Whenever we push scientific theories far beyond their established realm of reliability, our arrogance inevitably gets us punished.[1] The infinitesimal structure of space and time is surely such an unjustified speculative extrapolation. Unfortunately, all our physics is built on it. It is tempting to guess that when we understand how the illusion of a macroscopic four-dimensional space-time continuum arises from more fundamental concepts, these mathematical difficulties should become more tractable.

We know from our childhood that global properties can arise from second-order differential equations ('The shortest distance between two points is a straight line'). *Hamilton's principle* says that the solution to the equation of motion $m\ddot{x} = -\frac{\mathrm{d}}{\mathrm{d}x}V$, subject to the boundary conditions $x(t_1) = x_1, x(t_2) = x_2$, is the path $t \mapsto x(t)$ obeying the given boundary conditions, for which the *action*

$$S := \int_{t_1}^{t_2} \left(\frac{1}{2}m\,\dot{x}(t)^2 - V(x(t))\right) \mathrm{d}t \tag{4.1.3}$$

is stationary (minimal if $|x_1 - x_2|$ and $|t_1 - t_2|$ are both small). The integrand is called the *Lagrangian* $L = T - V$, where $T = \frac{1}{2}m\dot{x}^2$ is the *kinetic energy*. The combination $T + V$ for the stationary path $x(t)$ will be independent of the time t, and is called the *energy*. Historically, a hard lesson to learn (even for men like Gauss and Hertz) was that energy is an abstract mathematical notion and not a measure of some physical quantity (see the excellent discussion in chapter 4, vol. I of [**188**]).

This observation leads to a formulation of classical physics called Lagrangian mechanics, which will be central to our discussion of quantum field theory in Section 4.2 (in quantum theory concepts like force, velocity and acceleration cease to play fundamental roles). The possible configurations of our physical system can be regarded as forming a manifold, called the *configuration space* $\mathcal{M}$. For example, for a rigid body such as a potato, the configuration space is $\mathbb{R}^3 \times \mathrm{SO}_3(\mathbb{R}) \cong \mathbb{R}^3 \times \mathbb{P}^3(\mathbb{R})$:$\mathbb{R}^3$ gives its centre-of-mass, and $\mathbb{P}^3(\mathbb{R})$ its orientation. The behaviour of a system is regarded geometrically as a parametrised path $t \mapsto q(t)$ on $\mathcal{M}$, called the trajectory. Let q_i be a complete set of local coordinates on $\mathcal{M}$, obtained by restricting to some open set $U_\alpha \subset \mathcal{M}$ (recall Definition 1.2.3). The q_i represent the degrees of freedom of the system. The Lagrangian $L = T - V$ is a function of q_i and $\dot{q}_j$ – that is, a function on the tangent bundle $T\mathcal{M}$. In particular, in order to capture the kinetic energy T, which usually will be quadratic in the $\dot{q}_i$, we typically want $\mathcal{M}$ to be Riemannian, with T proportional to the norm-squared $\dot{q} \cdot \dot{q}$. The potential V will be a differentiable function on $\mathcal{M}$. The equations of motion

[1] Examples abound. There is, for instance, the famous remark of Lord Kelvin in 1899 that all of physics has been finished. Socrates' theory near the end of *Phaedo* as to the nature of the Earth makes a merry read. In mathematics recall the humbling experiences of Russell's Paradox and Gödel's Incompleteness Theorem.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

in the coordinate patch U_α are the *Euler–Lagrange equations*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i}, \tag{4.1.4}$$

which say that the action (4.1.3) is stationary for the physical solutions $q_i(t)$. Equation (4.1.4) is obtained from the calculus of variations by varying q_i.

To solve a physical system in Lagrangian mechanics, the first task would be to choose good local coordinates q_i on the configuration space $\mathcal{M}$, then to express the kinetic and potential energies in terms of q_i and $\dot{q}_i$, and finally to write down and solve the corresponding partial differential equations (4.1.4). Lagrangian mechanics (and Hamiltonian mechanics, to be discussed shortly) are essentially equivalent to Newtonian mechanics (4.1.1). Their appeal though should be clear to any mathematician: by freeing the formulation from adherence to a specific choice of coordinates, the formal structure of classical mechanics becomes more evident. This is especially valuable when extensions of the theory are needed – for example, when handling enormous numbers of particles in statistical mechanics, or when we were struggling to obtain the laws of quantum mechanics.

Returning to the harmonic oscillator, take $q = x - x_0$. Then $L = T - V = \frac{1}{2}m\,\dot{q}^2 - \frac{1}{2}k\,q^2$ and the Euler–Lagrange equation (4.1.4) yields the differential equation $m\ddot{q} = -k\,q$. The configuration space is $\mathbb{R}$, and trajectories consist of segments $[-A, A]$ traversed periodically. Energy $T + V = \frac{1}{2}kA^2$ is constant on each trajectory.

The pervasive habit of writing physical quantities with ‘units’ (metres, seconds, . . .) leads us into thinking of those mysterious entities as real and indispensable. In fact, many would regard as profound, or at least meaningful, the following question: What is the number of fundamental units in physics? However, Lagrangian mechanics should have led us to a somewhat more sophisticated understanding of units. Units themselves have no fundamental significance; choosing units is a special case of selecting a coordinate patch on the configuration space (together with a choice of time parameter). The common and useful practise of rejecting or anticipating formulae based on unit considerations (‘dimensional analysis’) merely captures some homogeneity information stored in the Lagrangian, and is the analogue here of the conservation laws of the following paragraphs. In particular, suppose we’ve selected a coordinate patch $\varphi : U \to \mathbb{R}^n$, $q \mapsto (q_i)$, and we want to change the scales (i.e. units) on each coordinate axis (which as expressions of nationalistic pride is fairly common). That is, we choose nonzero constants λ_i and consider the rescaling $q_i \mapsto q'_i = \lambda_i q_i$ of local coordinates, as well as $t \mapsto t' = \lambda_0 t$. This has two consequences. Firstly, we can write locally $L(q'_i, \dot{q}_j{}', t') = L'(q_i, \dot{q}_j, t)$, that is, we can continuously deform the Lagrangian. Inevitably, some choices of units will simplify L and hence ease the resulting arithmetic. Secondly and more importantly, it typically will be possible to absorb the rescalings λ_i into the various ‘physical constants’, that is, the parameters in L, which will tell us invariance properties of L and hence of the equations of motion (4.1.4). This is how to obtain the convenient and well-known meta-theorem that says the units of each term of any physical expression should agree.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

For example, note that the harmonic oscillator Lagrangian is invariant under the rescalings $q \mapsto \lambda_1 q$, $t \mapsto \lambda_0 t$, $k \mapsto \lambda_1^{-2} k$ and $m \mapsto \lambda_0^2 \lambda_1^{-2} m$; we see that each term of the solution (4.1.2) has a well-defined and consistent scaling behaviour (as they must). Also, for a preferred choice of λ_i, the Lagrangian simplifies to $\dot{q}^2 - q^2$. For another example, note that the gravitational Lagrangian $L = \frac{1}{2} m \, (\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2) + G \frac{mM}{r}$ is invariant under the rescaling $x_i \mapsto \lambda x_i$, $t \mapsto \lambda_0 t$, provided m rescales like $\lambda^{-2}\lambda_0^2 m$ and GM rescales like $\lambda^3 \lambda_0^{-2}$. In all cases, this scaling behaviour can be taken as defining the 'units' of the corresponding quantity – our definition here that the units of L be trivial differs from the usual one (where L has units of 'energy'), but this is merely a matter of convention.

This discussion should lead us to suspect that other invariance properties of L may yield other 'meta-theorems', generalising in a way the dimensional analysis. Indeed that is beautifully the case. By a *symmetry* of our system, we mean a diffeomorphism α of the configuration space $\mathcal{M}$ respected by the physics:

$$L(\alpha(q), \hat{\alpha}(\dot{q})) = L(q, \dot{q}),$$

where $\hat{\alpha}(\dot{q})$ is the induced map (derivative) on the tangent space with ith component $\sum_j \frac{\partial \alpha(q)_i}{\partial q_j} \dot{q}_j$. Note that, unlike the rescalings considered in the previous paragraph, here we're requiring that L and hence all the physical constants be unchanged by α. Then $q(t)$ is a possible trajectory (i.e. a solution of (4.1.4)) iff $\alpha(q(t))$ is.

Now, suppose we have a *continuous* family α_s of symmetries, that is a one-parameter subgroup $s \mapsto \alpha_s$ in the Lie group of symmetries. This symmetry can be used to vary the coordinates $q_i, \dot{q}_j$ – and hence the action S (4.1.3) – infinitesimally. What does Hamilton's principle ($\delta S = 0$) tell us here? The answer (*Noether's Theorem*[2]) is remarkable: continuous symmetries yield conservation laws! Define the quantity ('charge')

$$Q := \frac{\partial L}{\partial \dot{q}} \left(\frac{\partial \alpha_s(q)}{\partial s} \right) \in \mathbb{R}.$$

This expression is meaningful because the 'generalised momentum' $p := \frac{\partial L}{\partial \dot{q}}$ is a section of the cotangent bundle $T^*\mathcal{M}$, while the derivative $\frac{\partial \alpha_s(q)}{\partial s}$ of the path $\alpha_s(q)$ (q fixed) defines a section of the tangent bundle $T\mathcal{M}$. Less formally, suppose α_s sends q_i to $q_i + s\, f_i(q, \dot{q}, t)$, keeping only first order in the parameter s; then $\hat{\alpha}_s$ sends $\dot{q}_i$ to $\dot{q}_i + s\, \frac{\mathrm{d} f_i}{\mathrm{d}t}$, to first order, and $Q = \sum_i p_i f_i$. In either case, an easy calculation from (4.1.4) shows that Q is constant along each trajectory, that is Q is 'conserved'. (A deeper reason for this is that the Poisson bracket (4.1.6a) gives the space of solutions to (4.1.4) a symplectic structure.)

For example, the gravitational potential $V = -G \frac{m_1 m_2}{|\mathbf{x}_1 - \mathbf{x}_2|}$ is invariant with respect to translations $\alpha_s(\mathbf{x}) = \mathbf{x} + s\mathbf{a}$ for any fixed vector $\mathbf{a} \in \mathbb{R}^3$. The charge Q here is $\mathbf{a} \cdot \mathbf{p}$ where $\mathbf{p}$ is the 'total momentum' $m_1 \frac{\mathrm{d}\mathbf{x}_1}{\mathrm{d}t} + m_2 \frac{\mathrm{d}\mathbf{x}_2}{\mathrm{d}t}$. Varying $\mathbf{a}$, we find that momentum is conserved. We could say that the independence of the physics on absolute position

[2] As is typical, this designation is a little unfair: Noether published this in 1918, but Jacobi already knew in 1842 the connection between translation symmetry and momentum conservation, and rotational symmetry and angular momentum.

implies conservation of momentum. Likewise, independence of the physics on absolute time implies conservation of energy. In classical mechanics, Poincaré showed that all conservation laws are due to an underlying symmetry: if Q is conserved, then the Poisson bracket $\{Q, q\}_P$ of (4.1.6a) generates the corresponding symmetry. What is fundamental here isn't the Lie group action on $T\mathcal{M}$, but rather the infinitesimal generators (Lie algebra action), which need not be derived from a Lie group symmetry.

Another formulation of classical physics, useful for extensions to statistical and quantum mechanics, is Hamiltonian mechanics. Recall the generalised momenta $p_i = \frac{\partial L}{\partial \dot{q}_i}$. Together, the variables q_i, p_j parametrise a $2n$-dimensional manifold, the cotangent bundle $T^*\mathcal{M}$, called *phase space*. The *Hamiltonian* $H(q_i, p_j)$ is the quantity $\sum_i p_i \dot{q}_i - L$, expressed in variables q_i, p_i. Typically, it equals the total energy. The equations of motion here, obtained by varying both q_i and p_j, are *Hamilton's equations*:

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \tag{4.1.5}$$

that is $2n$ first-order differential equations, rather than the n second-order differential equations of Lagrangian mechanics (4.1.4). Although Hamiltonian mechanics is not always equivalent to Lagrangian mechanics, it is for typical systems. Because Hamilton's equations (4.1.5) are first-order, the configuration of the physical system at any time t is uniquely determined by the point in phase space it occupies at a given instant t_0. Thus phase space serves as a moduli space for physics. A more careful treatment of Hamiltonian mechanics requires the language of symplectic geometry – see, for example, [**15**] for details.

In classical mechanics the *observables*, that is the physically measurable quantities such as position, momentum or energy, are by definition real-valued smooth functions $A(q, p)$ on phase space. It is through the observables that a physical theory is compared to experiment. The observables $C^\infty(T^*\mathcal{M})$ form an infinite-dimensional Lie algebra, with bracket (in local coordinates) given by the *Poisson bracket*

$$\{A, B\}_P := \sum_i \left(\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i} \right) \tag{4.1.6a}$$

(see Question 4.1.2). Then Hamilton's equations (4.1.5) imply

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \{A, H\}_P, \tag{4.1.6b}$$

where on the left A is evaluated on a trajectory $(q(t), p(t))$. The term 'first integral' refers to any observable that is constant along each trajectory; the first integrals form a Lie subalgebra of dimension $< 2\mathrm{dim}\mathcal{M}$ in the observables $C^\infty(T^*\mathcal{M})$. Equation (4.1.6a) may seem obscure, but it is essentially equivalent to the natural bracket $[X, Y]$ of vector fields on a manifold – see corollary 5, page 217 of [**15**] for details. As we see in the next section, algebra arises in quantum field theory through the analogue there of Poisson bracket.

For example, recall the harmonic oscillator. The generalised momentum $p = m\dot{q}$ is the usual momentum. The Hamiltonian $H = \frac{1}{2m}p^2 + \frac{1}{2}kq^2$ is the energy. Hamilton's

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

equations tell us $\dot{q} = p/m$ and $\dot{p} = kq$. Phase space is the plane $\mathbb{R}^2$, with ellipses as trajectories. The basic Poisson bracket $\{q, p\}_P = 1$ says the observables $q, p, 1$ span $\mathfrak{Heis}$ (recall (1.4.3)).

4.1.2 Special relativity

The fundamental theoretical advance of the nineteenth century was Maxwell's electromagnetism (Section 4.1.3), which unified light, electricity and magnetism. Although both Newtonian mechanics and Maxwell's theory were enormously successful, they were in some conflict. For instance, in Maxwell's theory is obtained the formula

$$c := \text{speed of light} = \frac{1}{\sqrt{\epsilon_0 \mu_0}},$$

where ϵ_0, μ_0 are numerical constants associated with the vacuum. This seems to suggest that the speed of light is itself a constant, independent of the observer. However Newton – and common sense – would have us believe that the speed at which light, or anything else, travels is variable. If light is emitted from a headlight with speed c, and a bug approaches the oncoming car with speed v, then surely to it that light travels with speed $v + c$.

The standard resolution in the nineteenth century was to regard Maxwell's equations as valid only with respect to a substance called the aether. The aether would be the stuff in which light-waves wave (propagate) – it would be to light what air is to sound. This aether concept was getting increasingly awkward as the century turned. Einstein's act of genius here was to flip the logic and trust Maxwell's message. Thus, the speed of light is the same for all observers: the light from that approaching car strikes the bug with the same speed c it left the headlights. Special relativity consists of the modifications this message implies for Newtonian physics. Indeed what we call magnetism can be thought of as a relativistic correction to the electrostatic force; Maxwell's electromagnetism was the first relativistic theory, created years before Einstein's birth.

The word 'special' in 'special relativity' arises because the equations are simplest and fundamental only for a certain class of privileged observers called 'inertial' – uniformly moving observers for which Newton's First Law holds. A car rounding a corner is certainly not inertial, but a coasting isolated spaceship could be treated as one to good approximation. Special relativity also applies to accelerating observers, provided one works infinitesimally. Physically speaking, *general* relativity (Section 4.1.3), which removes this preferential treatment of inertial observers, is a mathematically elegant global integration of the equivalence principle and locally applied special relativity.

An inertial observer is simply a choice of fixed basis in $\mathbb{R}^4$; the coordinates $(\mathbf{x}, t)$ with respect to this basis, of a point ('event') x in $\mathbb{R}^4$ ('space-time'), have the physical interpretation to that observer as space and time coordinates. Not every choice of basis is permitted: we require them to be orthonormal in the sense that the straight-line trajectory ('world-line') $(\mathbf{x}(t), t)$ traced in space-time $\mathbb{R}^4$ by a beam of light is required to satisfy $(\mathbf{x}(t) - \mathbf{x}(0)) \cdot (\mathbf{x}(t) - \mathbf{x}(0)) = c^2 t^2$ – this is what we mean by the speed of light

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

being constant. Thus we are led to endow space-time $\mathbb{R}^4$ with the indefinite Minkowski metric $\eta = (\eta_{\mu\nu}) = \mathrm{diag}(1, 1, 1, -c^2)$. We write x^2 for $x \cdot x = \sum_{\mu,\nu=1}^{4} x_\mu x_\nu \eta_{\mu\nu}$ and $\mathbf{x}^2 = \sum_{\mu,\nu=1}^{3} \mathbf{x}_\mu \mathbf{x}_\nu$. Basis transformations between inertial observers belong to the Lie group $\mathrm{O}_{3,1}(\mathbb{R})$. As mentioned in Section 1.4.2, it has four connected components; the component containing the identity is the *Lorentz group* $\mathrm{SO}^+_{3,1}(\mathbb{R})$. Its universal cover $\mathrm{SL}_2(\mathbb{C})$ and their semi-direct products with translations $\mathbb{R}^4$ (the *Poincaré group* and its double-cover) also arise in physics. Thus in special relativity space and time are coupled, just as in Euclidean geometry the x, y, z coordinates are coupled (i.e. their independent objective significance is denied). The disturbing dissimilarity between our qualitative experiences of time and space is ignored by Einstein's theory. Discovering what relation this dissimilarity has to the different signs in the metric, or to the apparent magnitude of c, clearly should be a fundamental task. By contrast, in Newtonian mechanics space-time $\mathbb{R}^4$ factorises globally as $\mathbb{R}^3 \times \mathbb{R}$, and the basis transformations are taken from $\mathrm{O}_3(\mathbb{R}) \times \{\pm 1\}$.

That Maxwell's equations are invariant under the Lorentz group was known before Einstein. Einstein's contribution was to interpret the Lorentz group as giving the transformation of physical space and time. For example, the space-time transformation Λ between two observers with parallel spatial coordinate axes but travelling with uniform relative velocity $\mathbf{v} = (v, 0, 0)$, according to Einstein and Newton, is

$$\Lambda = \begin{pmatrix} \frac{1}{\sqrt{1-v^2/c^2}} & 0 & 0 & \frac{v}{\sqrt{1-v^2/c^2}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{v/c^2}{\sqrt{1-v^2/c^2}} & 0 & 0 & \frac{1}{\sqrt{1-v^2/c^2}} \end{pmatrix}, \tag{4.1.7a}$$

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & v \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{4.1.7b}$$

respectively. Note that in the limit $c \to \infty$, (4.1.7a) tends to (4.1.7b). Physically, matrix (4.1.7a) says that the lengths of moving objects shrink, and their clocks run more slowly. This is *not* some illusion, optical or otherwise. For example, the muon is an unstable elementary particle with an average lifespan of 2×10^{-6} seconds when at rest. When travelling at speed v, it will last on average $2 \times 10^{-6}/\sqrt{1 - v^2/c^2}$ seconds. It will travel further than it would have if (4.1.7b) had been the correct transformation, and because of that will be able to participate in interactions that would have been too distant for a muon behaving nonrelativistically. Other physical quantities transform similarly – for example, the parameter m playing the role of relativistic mass equals $m_0/\sqrt{1 - v^2/c^2}$, for some constant m_0 called *rest-mass*. Now, expand this out using the binomial series:

$$m = m_0 + \frac{1}{2} m_0 \frac{v^2}{c^2} + \frac{3}{8} m_0 \frac{v^4}{c^4} + \cdots$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

Multiplying by c^2, we recognise the second term as kinetic energy and we are led to suspect that mc^2 is the relativistic analogue of kinetic energy – that is, $E = mc^2$ for a free particle.[3]

In order to compare observations, we need to understand how the physical quantities change when we switch inertial observers, that is, how they transform with respect to the Lorentz group $\mathrm{SO}^+_{3,1}(\mathbb{R})$. Typically, they transform like matrix entries of $\mathrm{SO}^+_{3,1}$-representations. For example, the 4-vector $(\mathbf{x}, t)$ transforms with respect to the defining representation of the Lorentz group, as does the energy-momentum 4-vector $(\mathbf{p}, E/c^2)$, and thus its Minkowski norm-squared $\mathbf{p}^2 - E^2c^{-2}$ is an observer-independent quantity (a *Lorentz scalar*) and equals $-m_0^2c^2$. It is conventional to denote with superscripts the components of any such 4-vector: for example, $(\mathbf{x}, t) = (x^1, x^2, x^3, x^4)$.

Writing equations of motion presents us with a challenge: in Newtonian physics we always want to differentiate or integrate with respect to time; however, relativity teaches that we shouldn't treat time distinctly from the spatial coordinates. Moreover, '$\mathrm{d}t = \mathrm{d}x^4$' transforms like a component of a 4-vector, which isn't necessarily what we want. The solution is that the infinitesimal norm-squared $\mathrm{d}\mathbf{x}^2 - c^2\mathrm{d}t^2 =: -c^2\mathrm{d}\tau^2$ is $\mathrm{O}_{3,1}$-invariant, defining the 'proper time' τ, and so we should differentiate/integrate with respect to τ. Physically, τ is the time coordinate in the (usually only infinitesimally inertial) reference frame in which the particle is at rest. The Lagrangian L is a Lorentz scalar, and the action (4.1.3) becomes $\int L\,\mathrm{d}\tau$. For example, the Lagrangian for a free particle ('free' means no forces act on it, so the potential V is 0) can be taken to be

$$L = \frac{1}{2}m_0\left(\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tau}\right)^2 - c^2\left(\frac{\mathrm{d}x^4}{\mathrm{d}\tau}\right)^2\right).$$

The Hamiltonian, being energy, transforms like time.

But what if there are several particles: which proper times τ_i do we use? The τ for the centre-of-mass, perhaps? In fact, this is a serious problem. The 'No-Interaction Theorem' (beginning with [**124**]) says that there can be no direct Lorentz-invariant interaction between particles, except through forces localised at a point causing an instantaneous change of velocity that don't change the number of particles. As there do seem to be unstable elementary particles (e.g. the muon) and gravity for instance isn't localised to a point, we have a problem. The obvious solution is to copy the first relativistic interaction theory, namely Maxwell's, and use *fields* (Section 4.1.3).

Special relativity says that the speed of *light* is fundamental to space-time. Modern physics helps us to accept this seeming glorification of light, by saying that there is a special speed c, and any particle with zero rest-mass m_0 (such as the photon, which mediates light) will always travel at that speed. But perhaps more can be said. Surely space-time is not a fundamental physical quantity; eventually it will be recognised as a fairly macroscopic epiphenomenon, and it will be understood how it arises operationally.

[3] The equivalence of matter and energy was proposed 50 years before Einstein, by Mendeleev, the father of the periodic table. Although his reasons were correct, his proposal was ignored and forgotten.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

For instance, we can measure distance using rigid bodies called metersticks and time using quartz watches, but both this rigidity and periodicity are electromagnetic phenomena. Perhaps the constancy of the speed of, for example, light will be understood ultimately as a reflection of this circularity.

Einstein found the special treatment of inertial observers quite artificial. But it seems that accelerating observers can experience interesting phenomena. For instance, consider an observer S standing at the North Pole and an inertial observer T hovering above her, so T watches S uniformly spinning at the rate of one cycle every 24 hours. Let's assume for simplicity that the Earth's equator is a perfect circle; to T, the ratio of its circumference to the diameter of the Earth at the equator should be π. However, if S was to measure precisely the circumference and the diameter, she would find their ratio for this 'circle' to be (very slightly) greater than π. The reason for this is because S's observations must be consistent with T's: (4.1.7a) tells us that lengths parallel to the motion (such as S's metersticks along the equator as seen by T) will dilate by some factor $\sqrt{1-v^2/c^2}$, while lengths perpendicular to the motion (e.g. the diameter) will remain unchanged. Likewise, S will find that her wristwatch will tick more quickly than a clock placed on the equator, even though both are at rest relative to her. Thus both geometry and physics change for non-inertial observers! (For a fairly convincing argument that gravity requires *curved* space-time, see section 7.3 of [**422**].)

In fact relaxing the inertial observer restriction provided Einstein with the key to his remarkable explanation of gravity. As mentioned earlier, the gravitational 'charge' numerically equals the mass m seen in formulae such as $\mathbf{F} = m\mathbf{a}$ or $T = \frac{1}{2}mv^2$ – this is precisely what Galileo's Pisa experiment was designed to verify. There are other 'forces' with this same property, for example the pull we feel when riding a merry-go-round. This got Einstein thinking: perhaps gravity is as fictitious as a centrifugal force? When we are in free-fall – whether in an orbiting spaceship or in an elevator suddenly decoupled from its cable – it is as if we are free of gravity, much as we are suddenly free of the centrifugal force when we step off the merry-go-round. This is the *equivalence principle*, which constitutes the only new physical content of general relativity. We are led to the thought that the gravitational 'force' experienced while sitting in a chair isn't due to the matter in the Earth pulling us towards it, but rather merely a consequence of the chair interfering with our natural inertial motion, just as does a car rounding a corner. All observers are physically valid, but awkward choices (such as me in a chair or in a turning car) introduce fictitious forces such as gravity. Everything tries to move in as straight a line, and with as constant a speed, as possible (at least if it's not under the influence of a true force like magnetism); that astronomical effect we call 'gravity' is merely a consequence of the fact that 'straight' has only a *local* significance. Space-time is not the vector space $\mathbb{R}^4$, but rather a nontrivial (curved) four-dimensional pseudo-Riemannian manifold. Gravity is the convergence or twisting of nearby geodesics; what we perceive as the elliptical revolution of the Earth about the Sun is merely the gentle entwining of the Earth's geodesic with the Sun's (Figure 4.3). General relativity, which we discuss briefly at the end of the next subsection, makes these thoughts mathematically precise.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

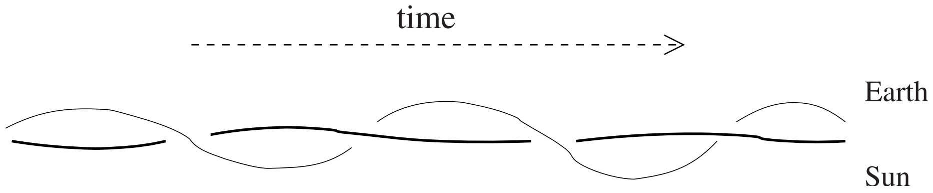

Fig. 4.3 The revolution of the Earth about the Sun.

4.1.3 Classical field theory

In physics, a ‘field’ is as in ‘vector field’ rather than ‘number field’. It means a function of (usually) space-time, or more precisely a section of some vector bundle whose base is space-time. The most familiar example is Newton’s gravitational field, namely the gravitational potential $V(\mathbf{x}, t)$. Another example is Maxwell’s electromagnetic field $F(\mathbf{x}, t)$, which is matrix-valued.

Until now, we’ve been interested in particle dynamics, and the fields were auxiliary. To analyse how object A gravitationally influences object B, we first calculate how A influences the gravitational field, and then how the gravitational field influences B. In classical field theory, the field is a mechanical system in its own right – for example, it carries energy much like a fluid. It allows us to avoid the No-Interaction Theorem of relativistic dynamics. In quantum field theory discussed in the next section, the field is primary and the particle becomes an auxiliary phenomenon called a quantum, apparent only asymptotically.

A cherished physical principle, going back at least to Faraday, is called *locality*. The idea is that the only way we can *directly* affect something, is by nudging it. In order to influence something not touching us, we must propagate a disturbance from us to it, such as a sound-wave in air or a ripple in water. Special relativity sharpened locality into the requirement that no disturbance or influence can travel faster than light, so that space-time points $(\mathbf{x}, t)$, $(\mathbf{x}', t')$ that are *space-like separated* (i.e. obey $(\mathbf{x} - \mathbf{x}')^2 > c^2 (t - t')^2$) are causally independent.[4] As Faraday himself noted, locality leads to the concept of *field*. This is the main purpose for both classical and quantum fields – they provide a natural vehicle for realising locality.

Before, configuration space was finite-dimensional, with coordinates $(q_1, \ldots, q_N)$. Now our coordinates have a continuous index, $q_x = q(x)$, and configuration space is a space of functions. The Lagrangian in particle dynamics looks like $\sum_i T_i - \sum_{i,j} V_{ij}$. Now the sums are replaced by integrals and the Lagrangian becomes

[4] Strictly speaking this isn’t a consequence of relativity, and in fact some physicists have entertained the possible existence of particles (‘tachyons’) that travel faster than light. These would behave curiously (e.g. they slow down the more energised they become), but like us they would require infinite energy to reach the speed of light – sadly, once a tachyon, always a tachyon. The difficulties facing the existence of tachyons are causality paradoxes. If P and Q are two space-like separated events, then there are reference frames in which P occurs before Q, and others in which Q occurs before P (why?). Hence if we had a gun that shot tachyonic bullets, then to some observers our victim would die before we pulled the trigger. Though not a logical contradiction, it is distinctly odd. Almost all physicists dismiss tachyons and faster-than-light influences as science fiction.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

$L = \int\int\int \mathcal{L}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ for some function $\mathcal{L}$ called the Lagrangian density. $\mathcal{L}$ is a function of the fields $\phi(x, y, z, t)$ and their partial derivatives $\partial_x \phi$, etc. (together with contributions from particles). In field theory, $\mathcal{L}$ is more elementary and fundamental than L. Locality takes the form here of requiring that $\mathcal{L}$ only involves one space-time point. For each field ϕ^a there is a field equation

$$\frac{\partial}{\partial t}\frac{\partial \mathcal{L}}{\partial(\partial_t \phi^a)} + \sum_i \frac{\partial}{\partial x_i}\frac{\partial \mathcal{L}}{\partial(\partial_i \phi^a)} = \frac{\partial \mathcal{L}}{\partial \phi^a}, \tag{4.1.8}$$

which describes the behaviour of the field. Additional equations (4.1.4) exist for each particle degree-of-freedom q_i present. The easiest example is the one-dimensional continuous Hooke's Law (e.g. vibrations in a rod). Our field here will be the amplitude $\phi(x, t)$ of the vibration at a point x on the rod. The Lagrangian density is

$$\mathcal{L}(x,t) = \frac{1}{2}\left\{\mu\left(\frac{\partial \phi}{\partial t}(x,t)\right)^2 - y\left(\frac{\partial \phi}{\partial x}(x,t)\right)^2\right\},$$

where μ is a constant called the mass density and y is a constant playing the role here of k. The first term is the kinetic energy density and the second (up to a sign) is the strain, or potential energy, in the rod. The field equation (4.1.8) gives us $\mu\,\frac{\partial^2\phi}{\partial t^2} - y\,\frac{\partial^2\phi}{\partial x^2} = 0$. This is easy to solve; physically it corresponds to a wave propagating with speed $v = \sqrt{y/\mu}$.

Define the momentum $\pi(\mathbf{x}, t) = \frac{\partial \mathcal{L}}{\partial(\partial_t \varphi)}$ conjugate to each field $\varphi(\mathbf{x}, t)$. Then the field equations (4.1.8) can be written as Poisson brackets involving Dirac deltas:

$$\{\varphi(\mathbf{x}, t), \pi(\mathbf{x}', t)\}_P = \delta(\mathbf{x} - \mathbf{x}'), \tag{4.1.9a}$$

$$\{\varphi(\mathbf{x}, t), \varphi(\mathbf{x}', t)\}_P = \{\pi(\mathbf{x}, t), \pi(\mathbf{x}', t)\}_P = 0. \tag{4.1.9b}$$

In special relativity, the Lagrangian density $\mathcal{L}$ transforms trivially (i.e. is a 'scalar') under the Lorentz group, and the fields ϕ^a span various representations R of the Lorentz group: that is, $\phi'^a(x') = \sum_b R(\Lambda)_{ab}\phi^b(x)$ where primes denote quantities in the reference frame (or $\mathbb{R}^4$-basis) obtained from the unprimed one using Lorentz transformation Λ.

An example important to physics (but not to us) is electromagnetism. The electromagnetic field has components $F_{\mu\nu} := \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}$, where A_4 is the electric potential and $\mathbf{A} = (A_1, A_2, A_3)$ is the magnetic potential. This field F transforms in a six-dimensional representation of the Lorentz group. The Lagrangian density is

$$\mathcal{L} = \frac{-1}{4}\sum_{\mu,\nu,\alpha,\beta} F_{\mu\nu}\,F_{\alpha\beta}\,(\eta^{-1})_{\mu\alpha}(\eta^{-1})_{\nu\beta} - \frac{1}{c}\sum_{\mu} j_\mu\,(\eta^{-1})_{\mu\nu}\,A_\nu =: \frac{-1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{c}j_\mu A^\mu,$$

where j is the electric current 4-vector describing the distribution and motion of charged particles. The matrix η^{-1} arises here in its Riemannian role defining an inner-product. The second expression is much more transparent, and uses $\eta^{\pm 1}$ to lower/raise indices, and summing over repeated indices. Of course to the Lagrangian must be added the (relativistic) kinetic energy of the particles or fields. The resulting field equations, called

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

Maxwell's equations, tell us for instance how charged particles create an electromagnetic field.

We see in Section 4.1.1 that even the simplest classical systems can have singular solutions, so the situation for classical field theory can only be worse. Most famous is the self-energy of charged particles in electromagnetism, discussed beautifully in chapter 28, vol. II of [**188**]: a charged particle localised to a point has infinite mass coming from the electromagnetic field. To see this, imagine that we hold half an electron in our left hand and the other half in our right; to make the electron whole we would have to connect these two repulsive halves, and an easy calculation (namely the integral $-\int_1^0 r^{-1}\mathrm{d}r = \infty$) says this requires infinite energy. This problem persists in its quantisation.

A remarkable classical field theory is Einstein's general relativity, in which space-time is a pseudo-Riemannian manifold with metric tensor $g(x)$, locally (but not globally) equivalent to the Minkowski metric η. Ignoring for convenience other forces, the Lagrangian density for a single particle is

$$\mathcal{L}(x) = \frac{1}{2}m_0 \sum_{\mu,\nu} g_{\mu\nu}(x)\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\,\delta^4(x - x(\tau)) + \frac{c^3}{16\pi G}\sqrt{-\det g}\,R, \tag{4.1.10}$$

where G is Newton's gravitational constant and R is a geometric quantity (a measure of the radius of curvature of space-time at x). δ^4 is the highly singular Dirac delta. The numerical constant $c^3/16\pi G$, establishing the coupling strength between space-time and matter, is chosen so that Einstein's theory agrees with Newton's in the appropriate limit. Varying the particle's coordinates x^μ yields the geodesic equation

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \sum_{\nu,\kappa} \Gamma^\mu_{\nu\kappa}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\,\frac{\mathrm{d}x^\kappa}{\mathrm{d}\tau} = 0,$$

describing the straightest possible curves in the manifold ($\Gamma^\mu_{\nu\kappa}$ are the Christoffel symbols). Varying the metric g yields Einstein's field equations

$$R_{\mu\nu} - \frac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}. \tag{4.1.11}$$

$R_{\mu\nu}$ are components of the Ricci tensor and $T_{\mu\nu}$ are those of the stress-energy tensor defined below. The left side is geometrical, depending on first and second partial derivatives of $g_{\mu\nu}$, while the right side is physical, depending on the matter fields. Einstein's field equations (4.1.11), which tell us how matter and energy curve space-time, consist of 10 coupled nonlinear second-order partial differential equations for the components $g_{\mu\nu}$.

The relation between symmetries and conserved quantities in field theory takes the following form (generalised in Question 4.1.1). Suppose the Lagrangian density $\mathcal{L}$ is invariant under a continuous symmetry α_s. Associate with α_s the 4-vector

$$j^\mu(x) = \frac{\partial\mathcal{L}}{\partial(\partial\phi/\partial x^\mu)}\left(\frac{\partial\alpha_s(\phi)}{\partial s}\right), \tag{4.1.12a}$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

for $\mu = 1, 2, 3, 4$, called the 'current'. Then $j(x)$ is *conserved*, that is it obeys

$$\partial_\mu j^\mu := \sum_{\mu=1}^{4} \frac{\partial j^\mu}{\partial x^\mu} = 0. \tag{4.1.12b}$$

This equation tells us to think of $j^4(x)$ as the density of some abstract fluid, and $\mathbf{j}(x) = (j^1(x), j^2(x), j^3(x))$ as its velocity at each space-time point x. Equation (4.1.12b) tells us that this 'fluid' is neither created nor destroyed, so that the total quantity ('charge') $Q(t) = \int j^4(x)\,\mathrm{d}x^1\,\mathrm{d}x^2\,\mathrm{d}x^3$ (if the integral exists) is constant: $\frac{\mathrm{d}Q}{\mathrm{d}t} = 0$.

For example, the invariance of the Lagrangian density $\mathcal{L}$ with respect to time and space translations $x^\nu \mapsto x^\nu + a^\nu$ gives us the 'current' $T^{\mu\nu}(x)$ (one for each ν) called the stress-energy tensor. The 'charges' Q^ν here are the total momentum and energy. Or consider the full Lagrangian density for the coupling of the electromagnetic field F to a complex scalar field ϕ with mass m, charge e and potential V:

$$\mathcal{L} = \frac{-1}{4}\sum F_{\mu\nu}F^{\mu\nu} + \sum\left(\frac{\partial}{\partial x^\mu} - \mathrm{i}eA_\mu\right)\phi\left(\frac{\partial}{\partial x_\mu} + \mathrm{i}eA^\mu\right)\phi - m^2\phi^*\phi - V(\phi^*\phi). \tag{4.1.13}$$

The terms only involving ϕ and ϕ^* form the Lagrangian for the field ϕ alone, while the terms involving both ϕ and A define the interaction. Note that there is a U_1 group symmetry of $\mathcal{L}$, which acts trivially on F and A but acts on ϕ by $\alpha_s(\phi) = e^{\mathrm{i}e\alpha}\phi$. Then Q is indeed proportional to e. We return to this example next section.

Question 4.1.1. (a) Prove the following generalisation of Noether's Theorem. Suppose we have a continuous family α_s of diffeomorphisms of configuration space such that

$$L(\alpha_s(q), \hat{\alpha}_s(\dot{q})) = L(q, \dot{q}) + \frac{\mathrm{d}}{\mathrm{d}t}\Delta(q, \dot{q}),$$

for some function Δ. First, verify that $q(t)$ is a possible trajectory iff $\alpha_s(q(t))$ is. Next, verify that the quantity

$$Q = \frac{\partial L}{\partial \dot{q}}\left(\frac{\partial \alpha_s(q)}{\partial s}\right) - \Delta$$

is constant along any trajectory.
(b) The Lagrangian for a free Newtonian particle is $L = \frac{1}{2}m\dot{\mathbf{x}}^2$. Take $\alpha_s(\mathbf{x}) = \mathbf{x} + s\,\mathbf{a}$ for some constant vector $\mathbf{a} \in \mathbb{R}^3$. Find Δ here, and verify that the 'charge' Q is $m\,\mathbf{x}(0)$.

Question 4.1.2. Verify that the space $C^\infty(T^*\mathcal{M})$ of observables, with bracket given by (4.1.6a), defines a Lie algebra, and that the first integrals form a Lie subalgebra.

4.2 Quantum physics

We tend to have a naive view of progress in science, namely that the old theory gets superseded by a new theory that is better in every meaningful respect: any phenomenon the older theory could explain, and any question the older theory could answer, the new theory would explain and answer at least as accurately; moreover, there would

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

be phenomena and questions that the older theory avoids but the newer, better theory handles adroitly. In reality, progress in science (in contrast to progress in technology) has much in common with progress in popular music or in, say, America's ability to elect great presidents. Copernicus' circular orbits match observation worse than Ptolemy's epicycles. More significantly, Copernicus required the Earth to move at incredible speeds, which mysteriously no experiment could ever detect (e.g. when we jump straight up, we come straight down). Ptolemy himself rejected the heliocentric hypothesis for these and several other good reasons. It was only after Galileo explained the role of inertia, *after* Copernicus' time, that Copernicus' unoriginal idea became scientifically reasonable. Of course to us today all motion is relative and the proceedings of that Great Debate belong in the voluminous Library-of-Dead-Religions. For another example, Aristotelian physics regarded friction as fundamental and the pendulum as complicated derived motion, whereas Newtonian physics regarded the pendulum as simple and friction as compound. In fact, classical physics never successfully explained friction – our present explanation requires quantum mechanics to correctly handle the relevant molecular forces (namely the van der Waals forces, which are residuals of the underlying electromagnetic forces). At least in part, 'progress' in science is a sociological phenomenon, a mantra bubbling on the lips of scientists as they pursue questions they are willing and able to address.

In any case, the conceptually and mathematically elegant classical mechanics has been superseded by the fairly incoherent quantum physics. A century has passed since the birth of the quantum, and although almost all physicists today regard quantum theory as having successfully transcended classical physics, it is dangerous to conclude much from this. But one thing is certain: mathematics has been a great beneficiary of this 'transcendence'.

4.2.1 Nonrelativistic quantum mechanics

For fixed time t, the state of a single particle in quantum mechanics can be captured by a complex-valued *wave-function* $\mathbf{x} \mapsto \psi(\mathbf{x}, t)$. Its interpretation is rather different from 'state' in classical physics: the quantity $|\psi(\mathbf{x}, t)|^2$ is the probability density that the particle is at position $\mathbf{x}$ at time t. Probability arises here *not* because of uncertainty of our knowledge, *nor* because of unavoidable disturbances caused by our heavy-handed measuring processes. Rather, it is a fundamental ingredient of quantum reality. God's analysis too would stop at this probability.

Recall the discussion of Hilbert spaces in Section 1.3.1, in particular the rigged Hilbert space $\mathcal{S}(\mathbb{R}^n) \subset L^2(\mathbb{R}^n) \subset \mathcal{S}(\mathbb{R}^n)^*$, where the Schwartz space $\mathcal{S}(\mathbb{R}^n)$ consists of all smooth functions falling off with their derivatives to 0 quickly as $|x| \to \infty$ and where the Hilbert space $L^2(\mathbb{R}^n)$ consists of the square-integrable functions with inner-product

$$\langle \phi, \psi \rangle := \int_{\mathbb{R}^n} \overline{\phi(\mathbf{x})}\, \psi(\mathbf{x})\, \mathrm{d}^n\mathbf{x}.$$

For each time t, the span of the possible time-slices (states) $\psi(\star, t)$ form the Schwartz space $\mathcal{S} = \mathcal{S}(\mathbb{R}^3)$, while their topological span forms the Hilbert space $\mathcal{H} = L^2(\mathbb{R}^3)$.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

We require the wave-function ψ to be normalised: $\langle\psi,\psi\rangle(t)=1\ \forall t$. Observables here correspond to self-adjoint operators $\widehat{A}:\mathcal{S}\to\mathcal{S}$. For example, the operator associated with measuring the ith coordinate of position takes $\psi\mapsto x_i\psi$, while energy is associated with the operator $\mathrm{i}\hbar\frac{\partial}{\partial t}$ (we can use (4.2.1) below to express it using spatial derivatives) and the ith component of momentum with the operator $-\mathrm{i}\hbar\frac{\partial}{\partial x_i}$.

The role of phase space is (loosely) played here by the projectification $\mathcal{S}/\mathbb{C}$, since the physical states corresponding to nonzero multiples $c\psi$ are the same. This is significant because it tells us that groups can act on $\mathcal{S}$ via *projective* representations, and still be well-defined. This persists in all quantum theories and has many consequences. Not all $\psi\in\mathcal{S}$ though are actually physical states – for example, it appears that every physical state must have a definite electric charge, that is be an eigenvector of some charge operator, and of course most $\psi\in\mathcal{S}$ aren't.

There are two independent ways the wave-function evolves in time. The first way is through Schrödinger's equation, which is the linear partial differential equation

$$\mathrm{i}\hbar\frac{\partial\psi}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\psi+V(\mathbf{x})\,\psi, \tag{4.2.1}$$

where V is the potential energy (which acts multiplicatively on ψ), $\hbar$ is Planck's constant and ∇^2 is the Laplacian $\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial x_3^2}$. Schrödinger's equation governs the deterministic, unitary evolution of ψ occurring between measurements. It is standard to choose units so that Planck's constant $\hbar$ equals 1 (recall the discussion in Section 4.1.1); however, in units natural to our familiar macroscopic world (e.g. metres, kilograms and seconds) its magnitude (about 10^{-34}) emphasises just how invisible quantum effects are to us.

Schrödinger's equation can be formally integrated, and we obtain

$$\psi(\mathbf{x},t)=U(t)\,\psi(\mathbf{x},0), \tag{4.2.2}$$

where $U(t)=\exp[-\mathrm{i}\widehat{H}t/\hbar]$ is a unitary operator on $\mathcal{S}$ (hence $\mathcal{H}$) for the Hamiltonian operator $\widehat{H}$ given by the right side of (4.2.1). Conversely, we could have anticipated (4.2.1) by the following reasoning. The time evolution (4.2.2) should be given by a linear operator $U(t)$ independent of ψ (so $U(s)\,U(t)=U(s+t)$), which preserves the normalisation: $\|U(t)\,\psi(\mathbf{x},0)\|=\|\psi(\mathbf{x},t)\|=1$. This *implies* that $U(t)=\exp[\mathrm{i}H't]$, that is $\partial\psi/\partial t=\mathrm{i}H'\psi$, for some self-adjoint operator H'. For physical reasons we would expect H' to have something to do with energy, that is the classical Hamiltonian H, since energy is the conjugate observable to time just as momentum is to position. Indeed, Schrödinger's equation (4.2.1) comes from the nonrelativistic formula for energy ($E=\frac{1}{2m}\mathbf{p}^2+V$), together with the quantum mechanical substitutions $E\mapsto\mathrm{i}\hbar\frac{\partial}{\partial t}$ and $\mathbf{p}\mapsto-\mathrm{i}\hbar\nabla$.

The second type of wave-function evolution is indeterministic and discontinuous, and occurs at the instant t_0 when a measurement is made. Let $\widehat{A}$ be the self-adjoint operator corresponding to the observable being measured. Assume for simplicity that its spectrum (i.e. its set of eigenvalues) is discrete and nondegenerate. Then there is an orthonormal set $\{\psi_a(\mathbf{x})\}\subset\mathcal{S}$ of eigenvectors spanning $\mathcal{H}$ (topologically). So $\widehat{A}\psi_a=a\psi_a$ and $\langle\psi_a,\psi_b\rangle=\delta_{ab}$. If ψ is the wave-function of the particle being observed, write $\psi(\mathbf{x},t_0)=\sum_a c_a\psi_a(\mathbf{x})$. The result of the observation will be one of the eigenvalues a,

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

a_0 say, but which one cannot be predicted in advance. All that can be said is that $|c_a|^2$ is the probability that a will be the one observed. Nothing was responsible for the given eigenvalue a_0 arising – two completely identical quantum systems can (and usually will) yield different observed values. At time t_0 the wave-function ψ suffers a spontaneous and discontinuous change $\psi \mapsto \psi_{a_0}$ (or more generally the orthogonal projection of ψ into the a_0-eigenspace). For times immediately after t_0, the wave-function then proceeds to evolve by (4.2.1). This second type of evolution is necessary for the experimental consistency of the theory: experimental results can be reproduced! It is a truly physical evolution, and not merely book-keeping reflecting a change in our knowledge of the system.

For example, the simultaneous eigenvalues $\mathbf{p} = (p_1, p_2, p_3) \in \mathbb{R}^3$ of the three momentum operators correspond to eigenfunction $\psi_{\mathbf{p}}(\mathbf{x}, t) = e^{\mathrm{i}\mathbf{p}\cdot\mathbf{x}/\hbar}$, while the simultaneous eigenvalues $\mathbf{a} = (a_1, a_2, a_3) \in \mathbb{R}^3$ of the position operators have eigenfunctions given by the three-dimensional Dirac delta $\delta^3(\mathbf{x} - \mathbf{a})$. These spectra aren't discrete and (generalised) eigenfunctions aren't square-integrable (rather they are tempered distributions – Section 1.3.1), because exact position and momentum observations in quantum theory are nonphysical idealisations (e.g. probing infinitesimal distances requires infinite energy). Moreover, since the position and momentum operators don't share any eigenvectors, it is meaningless to speak simultaneously of the (numerical) position and momentum of a particle: in quantum mechanics a particle cannot have a well-defined trajectory.

This framework generalises in the obvious ways. For n particles, the wave-function ψ looks like $\psi(\mathbf{x}_1, \ldots, \mathbf{x}_n, t)$ and on the right side of (4.2.1) the Laplacian ∇^2 get replaced by the sum of n Laplacians ∇_i^2, one for each $\mathbf{x}_i$.

This treatment of many particles indicates a weak point of quantum mechanics. Experiment tells us that the number of elementary particles can change, for example, a muon can decay into an electron and two neutrinos. It is rather difficult to believe that the fundamental equation of motion in physics changes *discontinuously* with time, but that is how quantum mechanics would model the decay of, for example, the muon: at some time t_0 the wave-function would acquire six more variables and Schrödinger's equation six more terms. The way out (Section 4.2.2) simultaneously handles all numbers of particles.

The fascinating *measurement problem* of quantum physics, present in any quantum theory, is the struggle to understand this dichotomy of wave-function evolutions. What is so special about measurement, that it should obey special laws? After all, surely a measurement is merely a certain kind of physical process. Many remarkable elaborations have been proposed by respected physicists, for example, that the universe splits into different 'parallel universes' after each measurement, or that a measurement involves the imposition of mind on matter. Precisely what constitutes a measurement? Any quantum measurement involves the amplification of a microscopic quantum property or effect to a macroscopic one. What does quantum physics tell us about the macroscopic (classical) world? The linearity of Schrödinger's equation implies that linear combinations ('superpositions') of solutions will again be solutions. Now, *microscopic* superpositions are well-observed and fundamental to the theory; during a quantum measurement (if not at other times) macroscopic superpositions should be unavoidable. However, what would

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

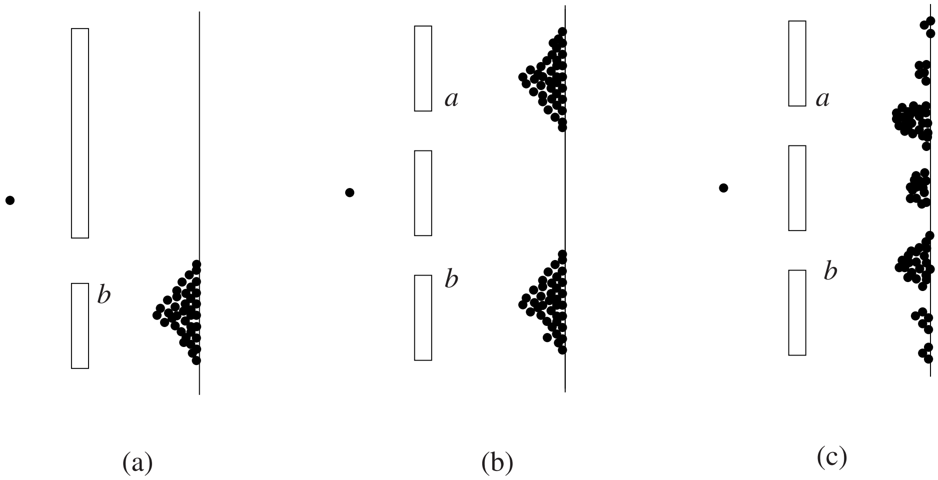

Fig. 4.4 The golf ball experiment.

a macroscopic superposition look like? Why have we never observed the superposition of, for example, a live and dead cat? We are led to the suspicion that quantum physics is incompatible with our most elementary qualitative observations of (macroscopic) physical reality.

To make this more precise, consider the situation depicted in Figure 4.4, where a machine randomly putts golf balls towards two barriers, one behind the other. When a hole is cut into the first barrier, as in Figure 4.4(a), the balls that reach the second barrier (i.e. pass through the hole) will impact it at roughly the same spot – the trajectories of golf balls over short distances are approximately linear. And if we cut two holes into the first barrier, we will get the result depicted in Figure 4.4(b). (We ignore all balls that get stopped by the first barrier.) Now suppose that whenever we avert our eyes for a few minutes, the golf balls make instead the impact pattern of Figure 4.4(c). That unbelievable phenomenon would suggest that changing the nature of our observation can dramatically affect golf ball trajectories. Classically, there is no evidence of this.

Of course that is precisely what occurs in the remarkable two-slit experiment, where electrons are fired at a screen. The electron wave-function ψ is the normalised *superposition* $\frac{1}{\sqrt{2}}(\psi_a + \psi_b)$ of wave-functions corresponding to travel through the a-slit or the b-slit. Individually, the wave-functions $\psi_a(\mathbf{x}, t)$ and $\psi_b(\mathbf{x}, t)$ both give rise to the probability density (for the arrival spot on the screen behind the two slits) we would expect from the golf balls of Figure 4.4(a). However, their superposition ψ gives rise to probabilities $\frac{1}{2}|\psi_a + \psi_b|^2 \neq \frac{1}{2}|\psi_a|^2 + \frac{1}{2}|\psi_b|^2$ – the two possible paths of the electron *interfere* with each other, much as they would if an electron were, for example, a water ripple. If we were to try to detect which slit the electron goes through, say by setting up a detector at each slit (as in Figure 4.4(b)), this additional measurement would first ‘collapse’ ψ into either ψ_a or ψ_b (with equal probabilities). The resulting probability

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

density for the arrival spot would be the particle-like $|\psi_a|^2$ or $|\psi_b|^2$, respectively (or $\frac{1}{2}|\psi_a|^2 + \frac{1}{2}|\psi_b|^2$ if we don't keep track of which slit the electron passed through).[5]

So why can't *macroscopic* states interfere? The special feature ('decoherence') of a macroscopic system seems to be that it is under unavoidable continuous interaction with the environment, through gravity if nothing else. Macroscopically distinct states (e.g. different pointer positions on an instrument, or golf balls rolling through different holes) couple differently to the environment, and so the macroscopic system becomes thoroughly and irreversibly entangled with the environment. This entanglement is essentially irreversible because any interaction that succeeded in untangling the coupling of the state with the environment would require enormous numbers (10^{27} or so) of degrees of freedom to conspire appropriately. This has the effect of making the macroscopic states essentially 'decohere' from each other, that is, the interference terms $\frac{1}{2}\psi_A\overline{\psi_B} + \frac{1}{2}\psi_B\overline{\psi_A}$, when expanded into the disordered microscopic degrees of freedom, get averaged away to zero. To get the flavour of decoherence, consider the wave-functions $\psi_{A,B}$ describing classical objects A, B. They are actually functions of 10^{27} or so space variables x_{ij}, but because they are macroscopic we would expect them effectively to be functions of our familiar three-dimensional space. Moreover, they would be essentially localised in this space, so $|\psi_A(\mathbf{x}, t) + \psi_B(\mathbf{x}, t)|^2 = |\psi_A(\mathbf{x}, t)|^2 + |\psi_B(\mathbf{x}, t)|^2$, provided A and B are situated a macroscopic distance apart (i.e. provided the supports of the effective functions ψ_A and ψ_B are disjoint). This is decoherence.

Of course alone this doesn't resolve the measurement problem. At best decoherence can only explain why macroscopically distinct states in superpositions don't 'see' each other. A (perhaps overly zealous) application of quantum mechanics insists that macroscopic superpositions must occur; from this, the 'Many-Worlds' interpretation is inevitable. The explanation for the mysterious wave-function collapse then would be that measurement entangles the quantum system $\psi^q = \sum c_i \psi_i^q$ with a macroscopic system ψ^c – that is, via Schrödinger's equation, the decoupled wave-function $\psi^q \psi^c$ relevant just prior to measurement would be replaced with the coupled wave-function $\sum c_i \psi_i^q \psi_i^c$ just after. Each coupled state ('world') $\psi_i^q \psi_i^c$ in this superposition would decohere from the others, and so the various quantum states ψ_i^q could no longer 'see' each other. It would be as if at the moment of measurement, the universe split into parallel universes, one for each possible experimental outcome. The 'Many-Worlds' interpretation is quantum mechanics in its purest form; in this framework measurement is a physical process subject only to Schrödinger's equation, and neither wave-function collapse nor the splitting of universes actually occurs. The price of this demystification of measurement is a reality in which almost everything is hidden from us, including infinitely many near-copies of ourselves.[6] A derivation of sorts of the probability rule is also possible within this framework.

[5] We shouldn't over-emphasise this 'wave–particle duality'. 'Waves' and 'particles' are classical metaphors; an electron is neither. Even the name 'wave-function' for ψ is an anachronism going back to de Broglie's hypothesis that an electron behaves like a wave with wavelength h/p.

[6] In defence of this uncomfortable aspect of Many-Worlds, Nature – unlike us – clearly loves enormous numbers of nearly identical copies. Consider blades of grass in a field, or water molecules in a lake (or

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

We've only sketched one possible interpretation. There are many others. For instance, the presence of probability in quantum mechanics strongly suggests that we are ignoring certain degrees of freedom – after all, this is what probability signifies in classical mechanics. It is possible to formulate quantum mechanics as a deterministic classical theory, by introducing 'hidden variables'. In the case of one particle, these hidden degrees of freedom would be the position coordinates $\mathbf{x}(t)$ of the particle. The coordinates $\mathbf{x}(t)$ obey a differential equation involving the wave-function ψ, which in turn obeys Schrödinger's equation. A similar formulation can be made for any number of particles. However, 'Bell's Theorem' says that any multi-particle hidden variables theory must possess the notorious feature called 'nonlocality'. This means that an influence (e.g. measurement) done on one particle can *instantaneously* affect the state of a distant particle. Nonlocality in a theory warns of possible difficulties in making the theory relativistic.

The approaches to the quantum measurement problem illustrate the desperate imagination that squirts from our pores when we're backed into a corner. See the book [**556**] for more details, examples and references to the literature. Like any other metaphysical doctrine, an interpretation is chosen not for its approximation to Truth, but because we find intriguing (and publishable!) the avenues of study it suggests.

For a one-dimensional example of a quantum system, consider once again the harmonic oscillator. The potential is $V = -\frac{k}{2}x^2$, so Schrödinger's equation here reads

$$\mathrm{i}\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + \frac{k}{2}x^2\,\psi. \tag{4.2.3a}$$

Because the potential V is independent of time, this is separable into energy eigenstates: write $\psi(x,t) = e^{-\mathrm{i}Et/\hbar}\psi_E(x)$, where

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi_E}{\mathrm{d}x^2} + \left(\frac{k}{2}x^2 - E\right)\psi_E = 0. \tag{4.2.3b}$$

In order for ψ to be normalisable, we require the boundary conditions $\psi(x,t) \to 0$ as $|x| \to \infty$; this implies (with a little work) that $E = (n+\frac{1}{2})\hbar\sqrt{\frac{k}{m}}$ for $n \in \mathbb{N}$, that is energy is quantised and bounded from below.

A useful idealisation is the step-function potential $V(x) = \begin{cases} 0 & \text{if } x < 0 \\ V_0 & \text{otherwise} \end{cases}$, where V_0 is constant. Solving the corresponding one-dimensional Schrödinger's equation with the requirement that both ψ and its derivative $\frac{\partial\psi}{\partial x}$ be continuous at $x = 0$, we obtain

$$\psi(x,t) = e^{-\mathrm{i}Et/\hbar}\begin{cases} A\exp(\mathrm{i}p_+x/\hbar) & \text{for } x > 0 \\ \exp(\mathrm{i}p_-x/\hbar) + B\exp(-\mathrm{i}p_-x/\hbar) & \text{for } x < 0 \end{cases},$$

where $p_+ = \sqrt{2m\,(E - V_0)}$ and $p_- = \sqrt{2m\,E}$ are the classical momenta (at least for $E > V_0$), and $A = 2\frac{p_-}{p_+ + p_-}$ and $B = \frac{p_- - p_+}{p_+ + p_-}$. Physically, this describes a wave (energy

perhaps research publications?). Or more to the point, consider the uncountably many moments making up each life.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

eigenstate) travelling to the right from $x = -\infty$, with energy $E > 0$; it hits the wall at $x = 0$, part of it continuing to positive x and some of it reflecting back to negative x. If we were to measure whether or not reflection happened, we would find that reflection happened with probability $|B|^2 = 1 - |A|^2$. Note that we get some very nonclassical behaviour: classically, when $E > V_0$ the whole wave would be transmitted to positive x, but here some of the wave is reflected, even when $V_0 < 0$! It is as if we are about to tumble over Niagara Falls in a barrel, only to bounce back the instant we reach the precipice. Related to this is quantum tunnelling (Question 4.2.2).

Quantum mechanics was born around 1926 when Schrödinger obtained (4.2.1) and, simultaneously, when Heisenberg and others developed an equivalent formulation. Unlike Schrödinger's picture, in Heisenberg's the state Ψ of the system is regarded as constant in time, and the time-evolution is carried by the observables $\widehat{A}$. It is completely analogous to the two attitudes towards observables carried in classical mechanics: we can view an observable $A(q, p)$ as a time-independent C^∞-function on phase space, or we can regard it as a function $A(q(t), p(t))$ of time. The equivalence between these two pictures of quantum mechanics is straightforward: the Heisenberg state $\Psi \in \mathcal{S}$ can be taken to be the wave-function $\psi(\star, 0)$ at time $t = 0$, while the Heisenberg operator $\widehat{A}(t)$ corresponds to Schrödinger's operator $\widehat{A}$ via the relation $\widehat{A}(t) = U(t)^{-1}\widehat{A}U(t)$, where $U(t) = \exp[-\mathrm{i}\widehat{H}t/\hbar]$ as before. Differentiating, we find that the equation of motion in Heisenberg's picture is given by commutation with $\widehat{H}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{A}(t) = -\frac{\mathrm{i}}{\hbar}\left[\widehat{A}(t), \widehat{H}\right]. \tag{4.2.4}$$

In relativistic quantum theory, Heisenberg's picture is more convenient because time doesn't play as privileged a role. In particular, just as $U(t)$ describes translations in time, a unitary operator $V(\mathbf{x})$ describes translations in space, and so we can regard the state Ψ as independent also of space. More generally, we have a unitary (projective) representation $(a, \Lambda) \mapsto U_{(a,\Lambda)}$ of the Poincaré group, acting on the infinite-dimensional space of states.

Equation (4.2.4) should look familiar: it is formally identical to the classical evolution (4.1.6b) of observables, provided we replace the Poisson bracket of classical observables there with the commutator of the quantum observables (up to the factor $\mathrm{i}\hbar$). Other examples of this are the calculations $\{x, p\}_P = 1$ and $[\widehat{x}, \widehat{p}] = \mathrm{i}\hbar I$. In other words, the process ('quantisation') of going from classical mechanics to the corresponding quantum mechanics defines a representation of the Lie algebra $C^\infty(T^*\mathcal{M})$ (with Poisson bracket) into the Hilbert space $\mathcal{H}$. However, this quantisation is clouded somewhat by the observation that the classical space $C^\infty(T^*\mathcal{M})$ is also an associative commutative algebra using pointwise product $(fg)(y) = f(y)\,g(y)$ of the functions, and that this product is also important as it is how we can build up general observables from the elementary ones x_i, p_j. Unfortunately, there is no direct analogue of this second product for the space of self-adjoint operators on $\mathcal{H}$ (or $\mathcal{S}$). The closest would be the operation $A * B = \frac{1}{2}(AB + BA)$, which makes the space of quantum operators into a (non-associative) *Jordan* algebra, originally named after the quantum physicist Pascual Jordan but now part of standard algebraic repertoire.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

An alternate, rather intriguing approach to quantisation seeks to formulate quantum mechanics in terms of a one-parameter deformation of the pointwise product algebra $\mathcal{A} = C^\infty(T^*\mathcal{M})$ (see [**141**] for a review). In particular, let $\mathcal{A}[[\lambda]]$ denote the space of all formal power series in λ with coefficients in $\mathcal{A}$. We add these power series term by term in the obvious way, but the product in $\mathcal{A}[[\lambda]]$ is more complicated (though necessarily associative). Expand out the product: $f \star g = \sum_{k=0}^{\infty} C_k(f, g)\,\lambda^k$, where for each $f, g, C_k(f, g) \in \mathcal{A}$. Because it is a deformation we require $C_0(f, g)$ to equal the usual pointwise product fg. In order to relate this to quantum mechanics, we also require that the coefficient $C_1(f, g) - C_1(g, f)$ of the leading term in the commutator $f \star g - g \star f$ be the Poisson bracket $2\{f, g\}_P$. We think of the deformation parameter λ as equalling $\mathrm{i}\hbar/2$. The main appeal of this approach to quantum mechanics is that classical and quantum mechanics are placed on the same page, so rigorous sense can be made of the statement that we recover classical physics from the $\hbar \to 0$ limit. However, it can be criticised for making classical mechanics logically prior to quantum mechanics, when the reverse would seem more natural. Also there are some quantum mechanical systems that don't seem to have a classical analogue. Kontsevich was awarded his Fields medal in 1998 in part for his proof that such a deformation exists not only for any phase space $X = T^*\mathcal{M}$ (this was known before), but more generally for any differentiable manifold X on which can be defined a Poisson bracket (a Lie algebra structure for $C^\infty(X)$).

Consider the harmonic oscillator in Heisenberg's picture. The possible states span a space $\mathcal{S}$, dense in a Hilbert space $\mathcal{H}$. Define the operators

$$\widehat{a} = \frac{(km)^{1/4}}{\sqrt{2\hbar}}\left[\widehat{x} + \frac{1}{\sqrt{km}}\mathrm{i}\widehat{p}\right], \qquad \widehat{a}^\dagger = \frac{(km)^{1/4}}{\sqrt{2\hbar}}\left[\widehat{x} - \frac{1}{\sqrt{km}}\mathrm{i}\widehat{p}\right] \tag{4.2.5}$$

acting on $\mathcal{S}$. These are called annihilation and creation operators, respectively. Note that $[\widehat{a}, \widehat{a}^\dagger] = I$, the identity operator. Hence $I, \widehat{a}, \widehat{a}^\dagger$ define a representation of $\mathfrak{Heis}$ (1.4.3) on the infinite-dimensional space $\mathcal{S}$. Let's find a more explicit realisation of this representation. This requires identifying the *vacuum state* $|0\rangle \in \mathcal{S}$, that is an eigenvector of the Hamiltonian $\widehat{H}$ with minimal eigenvalue (i.e. a state with lowest energy), normalised so that $\||0\rangle\| = 1$. Physically, the vacuum denotes the ground state, containing no particles. The energy operator, that is the Hamiltonian, becomes $\widehat{H} = \frac{\widehat{p}^2}{2m} + \frac{\widehat{x}^2}{2} = (\widehat{a}^\dagger\widehat{a} + \frac{1}{2})\,\hbar\sqrt{\frac{k}{m}}$ (as usual it is time-independent). The vacuum obeys $\widehat{a}|0\rangle = 0$ (why?) and has energy $E_0 = \frac{1}{2}\hbar\sqrt{\frac{k}{m}}$ (i.e. that is its $\widehat{H}$-eigenvalue). Assume that the vacuum is nondegenerate, that is the eigenspace associated with energy E_0 has dimension 1 – a degenerate vacuum would correspond to a number of non-interacting equivalent oscillators working in parallel. This assumption implies that the vacuum vector will be unique up to a phase $e^{\mathrm{i}\alpha}|0\rangle$ (choose one), and that the vacuum state is well-defined. Define vectors $|n\rangle := (n!)^{-\frac{1}{2}}(\widehat{a}^\dagger)^n|0\rangle$. This curious notation is due to Dirac: the functional $\langle\star| \in \mathcal{S}^*$ is called a *bra*, the vector $|\star\rangle \in \mathcal{S}$ a *ket*, and the evaluation $\langle\star|\star\rangle \in \mathbb{C}$ a *bra(c)ket*. This bracket also captures inner-products, using the adjoint $|\star\rangle^\dagger = \langle\star|$. Note that $|n\rangle$ has norm 1, and it is an eigenvector of $\widehat{H}$ with eigenvalue $E_n := (2n+1)E_0$. Construct the operator $\widehat{N} = \widehat{a}^\dagger\widehat{a}$, then $\widehat{N}|n\rangle = n\,|n\rangle$. We are to think of $\widehat{N}$ as a number operator, as

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

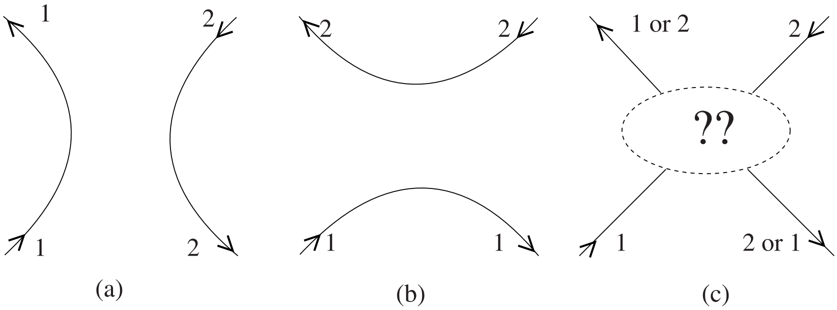

Fig. 4.5 A collision of two identical particles.

it counts the number of *quanta* (or excitations or quantum particles) in the given state. We say that the operator $\widehat{a}^\dagger$ *creates* a quanta, and $\widehat{a}$ *annihilates* a quanta. The vectors $|n\rangle$ $(n = 0, 1, 2, \ldots)$ form an orthonormal set; the state space $\mathcal{S}$ here consists of all $\sum_{n=0}^{\infty} c_n\,|n\rangle$ with $\sum n^m\,|c_n| < \infty$ for all m, while the Hilbert space $\mathcal{H}$ here consists of all $\sum c_n\,|n\rangle$ with $\sum |c_n|^2 < \infty$. In this algebraic way we can recover all of the physics.

When our system consists of a number of subsystems (e.g. different particles), the collective Hilbert space $\mathcal{H}$ will be given by the tensor product $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_n$ of the individual Hilbert spaces (this was implicit in our treatment of measurement, where the two subsystems were the observed and the observer). Given vectors $v_i \in \mathcal{H}_i$, we are to think of the 'diagonal' vector $v_1 \otimes \cdots \otimes v_n =: |v_1, \ldots, v_n\rangle$ as describing the situation where subsystem i is in state v_i. However, as we know, a typical vector u in the tensor product $\mathcal{H}$ won't be of this diagonal form. Only for such states $|v_1, \ldots, v_n\rangle$ do the subsystems themselves possess well-defined states. Even if the system begins in diagonal form (e.g. we start with two distant particles), it will lose this as soon as the subsystems interact. In this way, interacting systems lose their independent existence. This entangling of quantum subsystems doesn't occur in classical mechanics.

Something special, and also nonclassical, happens when the subsystems are *identical* (i.e. the subsystems obey identical laws, and differ only in incidental characteristics such as position). The collective Hilbert space $\mathcal{H}$ now is smaller than the full tensor product: it will be the symmetric product of n copies of the subsystem $\mathcal{H}_1$. More precisely, $\mathcal{H}$ is spanned by 'symmetric' vectors of the form $|v_1, \ldots, v_n\rangle := \frac{1}{\sqrt{n!}} \sum_{\sigma \in \mathcal{S}_n} v_{\sigma 1} \otimes \cdots \otimes v_{\sigma n}$. The physical reason for this is given in Figure 4.5. The first two diagrams represent classically distinct scatterings, but in quantum mechanics trajectories don't exist and we can't tell whether it is particle 1 or rather particle 2 moving northwest after the collision – Figure 4.5(c) applies. The labels '1' and '2' have no physical significance here: the vectors $|v_1, v_2\rangle$ and $|v_2, v_1\rangle$ now correspond to the same state – namely, the one where one of the particles (we cannot ask which) is in state v_1 and the other is in state v_2 – and should be identified. Perhaps we can say that here is the precise pen with which This August Personage signed That Important Document, but we cannot say (pointing) that this electron here was part of the pen at that Propitious Moment. An easy combinatorial consequence of this is that the identical particles here (but not those in the

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

next paragraph!) tend to clump into similar states. This is responsible, for instance, for the existence of the laser.

Recall however that proportional vectors in the space $\mathcal{S}$ correspond to physically equivalent states. Thus it merely suffices to identify, for example, $|v_1, v_2\rangle$ and $|v_2, v_1\rangle$ *up to a scalar factor*. The preceding paragraph describes the *bosons* like photons of light (named after S. N. Bose, who with Einstein first considered their statistical mechanics). The next simplest possibility, describing the *fermions* such as electrons, obeys $|v_1, v_2\rangle = -|v_2, v_1\rangle$. Their Hilbert space is spanned by antisymmetric vectors of the form $|v_1, \ldots, v_n\rangle := \frac{1}{\sqrt{n!}} \sum_{\sigma \in \mathcal{S}_n} (-1)^\sigma v_{\sigma 1} \otimes \cdots \otimes v_{\sigma n}$, where '$(-1)^\sigma$' equals ± 1 for an even/odd permutation σ, respectively. Note that antisymmetry forbids two fermions from sharing the same state. This simple fact is directly responsible for the remarkable diversity of chemical compounds, for if electrons obeyed instead the bosonic possibility $|v_1, v_2\rangle = +|v_2, v_1\rangle$, then there wouldn't be a chemical difference between the elements hydrogen, helium, lithium, . . . It is also responsible for large-scale structure, for example, why we don't fall through the floor.

These bosonic and fermionic 'statistics' correspond to the two one-dimensional representations of the symmetric group $\mathcal{S}_n$, but there are other possibilities (e.g. parastatistics, which involves higher-dimensional representations of $\mathcal{S}_n$, and braid statistics, which can occur when space-time is two-dimensional – both are discussed in, for example, chapter IV of [**269**]). However, only bosons and fermions seem to arise in Nature (except perhaps for some compound systems). Assuming this, a deep result of quantum field theory (Fierz and Pauli's Spin-Statistics Theorem – for a proof see section 4-4 of [**518**]) relates statistics to the Poincaré group. In particular, particles in relativistic quantum mechanics carry a representation of the universal cover of the Poincaré group. When that representation reduces to a representation of the Poincaré group itself, that is when spatial rotations through 2π correspond to the identity (we say the 'spin' is an integer), then the particle is a boson. Otherwise, that is when rotations through 2π correspond to $-I$ (so the spin is a half-integer), the particle will be a fermion. A connection between spin and statistics can be anticipated by the observation that the simple exchange of locations of two objects involves an implicit rotation by 2π of one relative to the other. We discuss this further in Section 4.3.5 below.

An important formulation of quantum physics is due to Feynman, and starts from an observation of Dirac: the infinitesimal quantum mechanical amplitude is governed by the value of the classical action (4.1.3). Suppose we know the wave-function $\mathbf{x} \mapsto \psi(\mathbf{x}, t_i)$ at some fixed initial time t_i. Then ψ at some other time t_f is given by

$$\psi(\mathbf{x}'', t_f) = \int K(\mathbf{x}'', \mathbf{x}'; t_f - t_i)\, \psi(\mathbf{x}', t_i)\, \mathrm{d}^3\mathbf{x}', \tag{4.2.6a}$$

where K, called the 'propagation kernel', is the amplitude for a particle to go from position $\mathbf{x}'$ at time t_i to position $\mathbf{x}''$ at time t_f. The point is that K is given by the 'path integral' $\int \exp(\mathrm{i}\, S(\mathbf{x})/\hbar)\, \mathcal{D}\mathbf{x}$ over all paths $\mathbf{x} : t \mapsto \mathbf{x}(t)$ with endpoints $\mathbf{x}(t_i) = \mathbf{x}', \mathbf{x}(t_f) = \mathbf{x}''$. For each choice of path $\mathbf{x}(t)$, $S(\mathbf{x})$ here is the classical action $\int_{t_i}^{t_f} L(\mathbf{x}, \dot{\mathbf{x}})\, \mathrm{d}t$. Integrals over spaces of paths arise here for much the same reason that the entries of

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

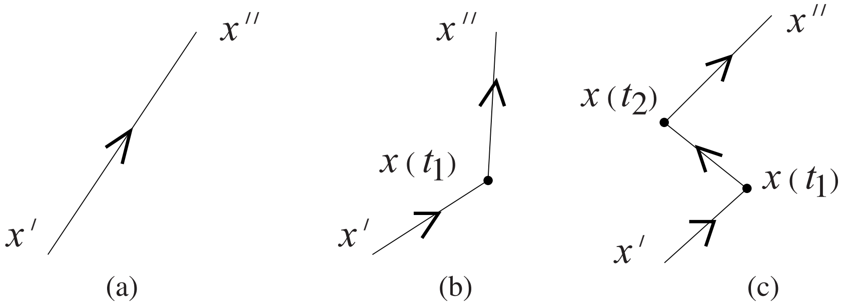

Fig. 4.6 Feynman diagrams in quantum mechanics.

powers A^n of a matrix could be described as sums over length-n walks through the entries of A. The path integral formulation intuits that the particle takes every conceivable trajectory from $(\mathbf{x}', t_i)$ to $(\mathbf{x}'', t_f)$, and each of these (appropriately weighted) contributes to the amplitude K and hence probability $|K|^2$. The precise mathematical meaning of Feynman's path integral is a little elusive, but attempts to define it in terms of, for example, Wiener integrals have been made. It is probably simplest though to regard it heuristically, as is done in Section 4.4.1.

Consider the classical limit $\hbar \to 0$ of (4.2.6a): using the stationary phase approximation, the dominant path $\mathbf{x}'(t)$ in the Feynman integral is one that satisfies the Euler–Lagrange equation (4.1.4). This provides an explanation for the mysteriously teleological Hamilton's principle of classical mechanics, discussed in Section 4.1.1.

The perturbative approach to quantum theories is particularly transparent in the path integral formalism. Write the Lagrangian as the sum $L = L_0 + \lambda L_{int}$ of the free part L_0 and the interaction part $\lambda L_{int} = -\lambda V$; the 'coupling constant' λ is a numerical constant (hopefully small), and we aim to expand the kernel K (and hence the wave-function ψ) in a Taylor expansion in λ. Explicitly, we have

$$\begin{aligned} K(\mathbf{x}'', \mathbf{x}'; t_f - t_i) &= \int \exp\left[\frac{\mathrm{i}}{\hbar}\int_{t_i}^{t_f} (L_0 - \lambda V)\,\mathrm{d}t\right] \mathcal{D}\mathbf{x} \\ &= \int \exp\left[\frac{\mathrm{i}}{\hbar}\int_{t_i}^{t_f} L_0\,\mathrm{d}t\right] \sum_{n=0}^{\infty} \frac{(-\mathrm{i}\lambda/\hbar)^n}{n!} \left(\int_{t_i}^{t_f} V(\mathbf{x}(t))\,\mathrm{d}t\right)^n \mathcal{D}\mathbf{x}. \end{aligned} \tag{4.2.6b}$$

We can represent this pictorially. The $n = 0$ term describes a particle propagating freely from $(\mathbf{x}', t_i)$ to $(\mathbf{x}'', t_f)$; the Feynman diagram for this term is given in Figure 4.6(a). The $n = 1$ term describes a particle propagating freely from $(\mathbf{x}', t_i)$ to some intermediate point $(\mathbf{x}, t_1)$, at which instant the potential V acts multiplicatively, and then the particle resumes free propagation to the final position $(\mathbf{x}'', t_f)$; we then integrate over all intermediate times (and finally over all paths $\mathbf{x}(t)$). The Feynman diagram is given in Figure 4.6(b), where the integration over t_1 is implicit. The kink there is called a 'vertex' – this is the same word as in *vertex* operator algebra. Likewise, the λ^n term corresponds to a Feynman diagram with n vertices, corresponding to the n integrals $\int V\,\mathrm{d}t_j$ in (4.2.6b). The factor $n!$ in (4.2.6b)

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

is removed by taking these intermediate times in the order $t_i < t_1 < \cdots < t_n < t_f$, as the diagrams suggest. In this way, we have replaced the actual physical situation, where of course the interaction V is always present, with a situation where the interaction is only present at discrete moments of time. It is as if the particle only interacts with V at the vertices. These are called *virtual* interactions, as they are mathematical artifacts and don't correspond directly to actual events in Nature.

We'll say more about perturbations and Feynman diagrams later. Typically, the sum (4.2.6b) won't converge, but the first few terms (when interpreted correctly) give good comparison with experiment. Conformal field theory – the physics of Moonshine – arises from the perturbative expansion of the quantum field theory called string theory.

Its treatment of measurement demonstrates that quantum mechanics is heuristic and idealised, and not at all in its finished form. But just as classical physics achieved a profound understanding of the concept of 'rest', and relativity provided a deep reanalysis of space and time, so is quantum mechanics forcing us to reconsider the seemingly harmless notion of observation. After all, we never observe an object, but rather the interaction between objects. Also profound, quantum mechanics teaches us that interacting subsystems become entangled, and physically this means that the whole is indeed much more than the disjoint union of its parts.

4.2.2 Informal quantum field theory

It is surprising that the next three natural tasks – namely, to bring in special relativity, to handle the experimental fact that the number of elementary particles can change, and to quantise classical field theories – are all accommodated by quantum field theories, the quantum theories of systems with infinitely many degrees of freedom. The sketch we provide here won't seem very satisfactory, but this is roughly the treatment to be found in physics textbooks. We avoid as too tangential most calculational issues and many technicalities (e.g. the quirks of fermions). Section 4.2.4 provides a more careful axiomatic treatment of quantum field theory, but knowing the informal physics background, at least in its broader strokes, is essential. A dated though otherwise excellent treatment of quantum field theory, somewhat in our style, is [**479**]; modern and masterful is [**555**].

To the working physicist, quantum field theory is the following conceptual hierarchy.

(i) *Experiment*. The experimenter measures half-lives of particles and scattering cross-sections. How well does experiment compare to theory?

(ii) *Amplitudes*. These observable quantities depend on the magnitude-squared of the appropriate transition amplitude $|\mathrm{in}\rangle \to |\mathrm{out}\rangle$. Unfortunately, transition amplitudes are too hard to calculate from the theory, except in infinite time ($t \to \pm\infty$) limits, which by definition are the entries of the S-matrix. Those limits, though mathematically dubious, are physically intuitive. So the theoretician needs to compute the S-matrix.

(iii) *Correlation functions*. The typical way to compute S-matrix entries is using correlation functions, via the so-called reduction formulae. So the theoretician wants to compute correlation functions.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

(iv) *Feynman diagrams*. Typically, correlation functions are calculated ‘perturbatively’ by Taylor-expanding in some coupling constant. Each term in this (usually divergent) infinite series is computed separately using Feynman diagrams.

Moonshine is interested in the correlation functions of a class of extremely symmetrical and well-behaved quantum field theories called rational conformal field theories – these theories are so special that their correlation functions can be computed exactly and perturbation is not required. But before we turn to them, let’s flesh out some of this hierarchy.

It would seem trivial to make quantum mechanics consistent with special relativity. Consider, for simplicity, a free particle of mass m. Recall that Schrödinger’s equation (4.2.1) corresponds to the nonrelativistic energy $E = \frac{1}{2m}\mathbf{p}^2$. Since relativistic energy satisfies $E^2 - \mathbf{p}^2c^2 = m^2c^4$, the natural guess for the relativistic Schrödinger equation would be

$$\left(\hbar^2\frac{\partial^2}{\partial t^2} - \hbar^2c^2\nabla^2 + m^2c^4\right)\phi(\mathbf{x}, t) = 0. \tag{4.2.7}$$

This is called the Klein–Gordon equation, and was proposed independently by Schrödinger, Klein and Gordon shortly after (4.2.1) was written down.[7] They expected it to describe the relativistic wave-function ϕ of a free ‘scalar’ particle (i.e. $\phi(x)$ is invariant under the action of the Lorentz group $\mathrm{SO}^+_{3,1}(\mathbb{R})$ on x), but such a theory is sick (see Question 4.2.4): for example, it suffers from negative probabilities and the energy eigenvalues have no lower bound (this means that we won’t have a vacuum state $|0\rangle$, which is bad). The way to make (4.2.7) into a sensible physical theory is to interpret it as a quantum field theory.

Quantum field theory is far deeper than quantum mechanics, both physically and mathematically. Witten predicts [**566**] that one of the major themes of twenty-first century mathematics will involve coming to grips with quantum field theory.

Let $\Omega \subset \mathcal{H} \subset \Omega^*$ be a rigged Hilbert space; Ω is the span of the states in the theory, and is constructed below, while $\mathcal{H}$ is their topological span. We obtained nonrelativistic quantum mechanics by replacing classical observables by operators, so we would expect that the fields $\varphi(x)$ in quantum field theories are operator-valued functions of space-time. Unfortunately this is too optimistic, even in the simplest free theories. Rather, the correct statement is that quantum fields φ are operator-valued *distributions* of space-time: for any states $u, v \in \Omega$, the matrix entries $\langle u, \varphi v\rangle$ of φ are tempered distributions of space-time. In other words, the Schwartz space $\mathcal{S} = \mathcal{S}(\mathbb{R}^4)$ is a space of test functions of space-time that ‘smear’ the fields; the values $\varphi(f)$, for each $f \in \mathcal{S}$, are (unbounded) linear operators $\Omega \to \Omega$. Nevertheless, it is traditional to write $\varphi(x)$, as if the fields were functions of space-time, and informally think of $\varphi(f)$ as the integral $\int_{\mathbb{R}^4} f(x)\varphi(x)\,\mathrm{d}^4x$. Unlike the wave-functions of quantum mechanics, a quantum field is not directly a probability

[7] Apparently, Schrödinger first derived the relativistic equation, noticed that it didn’t work but that its nonrelativistic approximation (4.2.1) looked good, and so first published the approximation! See the historical discussion on page 4, vol. I of [**555**].

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

amplitude; rather, it is a linear combination of operators that increase or decrease by one the numbers of particles in any state.

Let $\varphi_1, \ldots, \varphi_n$ be the complete list of quantum fields in the theory. All operators (e.g. observables) occurring in the theory are constructed from these fields. More precisely, locality says that any operator at a given space-time point x is a function of fields and their derivatives, all evaluated at that point.

The mathematical meaning of a theory being (special-)relativistic is that its quantities transform nicely with respect to (i.e. in projective representations of) the Lorentz and Poincaré groups $\mathrm{SO}^+_{3,1}$ and $\mathbb{R}^4 \rtimes \mathrm{SO}^+_{3,1}$. As in Theorem 3.1.1, those projective representations are true representations of the universal covers $\mathrm{SL}_2(\mathbb{C})$ and $\mathbb{R}^4 \rtimes \mathrm{SL}_2(\mathbb{C})$, respectively. Firstly, the state space $\mathcal{H}$ carries a unitary representation $(a, \Lambda) \mapsto U_{(a,\Lambda)}$ of the universal cover of the Poincaré group. These operators $U_{(a,\Lambda)}$ send the state space Ω onto itself; on Ω, we can write $U_{(a,I)} =: \exp[-\mathrm{i} \sum_\mu a_\mu P^\mu/\hbar]$, where the self-adjoint operators P^μ are the observables for momentum and (up to a constant) energy. In particular, $c^2 P^4$ is the Hamiltonian density. The absence of tachyons (footnote 4 in this chapter) says that the simultaneous eigenvalues $(\mathbf{p}, p^4)$ of the energy-momentum operators P^1, P^2, P^3, P^4 all have nonpositive Minkowski norm-squared $\sum_\mu p_\mu p^\mu = \mathbf{p}^2 - c^2(p^4)^2 =: -m^2c^2$. This parameter m is constant in any irreducible representation of $\mathbb{R}^4 \rtimes \mathrm{SL}_2(\mathbb{C})$, and is called the *(rest-)mass*.

The span of the fields φ_i carries a projective representation of all symmetries of the theory. In particular, there is an n-dimensional representation V of $\mathrm{SL}_2(\mathbb{C})$, governing how the n fields transform relativistically: that is,

$$U_{(a,\Lambda)}\, \varphi_i(f)\, U^{-1}_{(a,\Lambda)} = \sum_{i=1}^{n} V(\Lambda^{-1})_{ij}\, \varphi_j((a, \Lambda)^{-1}.f) \tag{4.2.8a}$$

holds in Ω, where the Poincaré transformation $(a, \Lambda) \in \mathbb{R}^4 \rtimes \mathrm{SL}_2(\mathbb{C})$ acts on test functions by $((a, \Lambda).f)(x) = f(\Lambda x + a)$. The inverses on the right side are needed in order for (4.2.8a) to be consistent with $U_{(a',\Lambda')} \circ U_{(a,\Lambda)} = U_{(a',\Lambda')\circ(a,\Lambda)}$. Restricting to translations $\mathbb{R}^4$, the derived representation of (4.2.8a) becomes the important equation of motion

$$\partial_\mu \varphi(x) = \frac{\mathrm{i}}{\hbar}[P_\mu, \varphi(x)]. \tag{4.2.8b}$$

Since the finite-dimensional representations of $\mathrm{SL}_2(\mathbb{C})$ are completely reducible, we can collect the fields together that form irreducible representations, parametrised by Dynkin label $\lambda_1 = \mathbb{N}$. Mysteriously, physicists prefer to use *spin* $s = \lambda_1/2$.

In classical field theory, the particles and fields are phenomenologically independent even though they mutually influence each other. In quantum field theory, particles are secondary, arising from fields, as we see shortly. A great definition, due to Wigner, is:

Definition 4.2.1 *A particle is an irreducible projective representation of the Poincaré group, with real mass m and energy $c^2 p^4 \geq 0$, in the space $\mathcal{H}$ of states of the theory.*

More precisely, the spectra $(\mathbf{p}, p^4)$ of the energy-momentum operators P^μ in an irreducible representation are required to obey $\mathbf{p}^2 \leq c^2(p^4)^2$; the mass $m \geq 0$ is the constant

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

$\sqrt{c^2(p^4)^2 - \mathbf{p}^2}$. Only the vacuum has 0 energy. Unlike the mass, the energy varies within the irreducible representation, and for a particle of mass m is never less than mc^2.

Subatomic experiments suggest that there are elementary (i.e. noncomposite) particles, for instance electrons. Each species of elementary particle in the theory arises from an irreducible $\mathrm{SL}_2(\mathbb{C})$-module in the span of the fields φ_i. In particular, a particle with spin $s \in \frac{1}{2}\mathbb{N}$ requires $2s + 1$ fields $\varphi_{i_1}, \ldots, \varphi_{i_{2s+1}}$, called its components. Other symmetries of the theory combine with $\mathrm{SL}_2(\mathbb{C})$ to form higher-dimensional representations. For example, in *quantum electrodynamics*,[8] 'parity' (i.e. the space-reflection $\mathbf{x} \mapsto -\mathbf{x}$) collects the two-component 'left-' and 'right-handed' electrons into an irreducible four-dimensional representation, while in the Standard Model parity is no longer a symmetry, but the left-handed electron and neutrino transform together as components in a four-dimensional representation of the symmetry group $\mathrm{SU}_3 \times \mathrm{SU}_2 \times \mathrm{U}_1$, while the right-handed electron forms a two-dimensional representation by itself.

A Lagrangian density $\mathcal{L}(x)$ here is a self-adjoint operator, invariant under $\mathrm{SL}_2(\mathbb{C})$, built up polynomially from the various φ_i and $\partial_\mu \varphi_i$, all evaluated at the same space-time point x. Each field φ_i obeys the corresponding Euler–Lagrange equation (4.1.8). As in classical field theory, define the 'canonical momentum field' $\pi_i(x) = \partial \mathcal{L} / \partial(\partial_4 \varphi_i)$ (not to be confused with the momentum operators P^μ). The *equal-time commutation relations*

$$[\varphi_i(\mathbf{x}, t), \pi_j(\mathbf{x}', t)] = \mathrm{i}\hbar\, \delta_{ij} \delta(\mathbf{x} - \mathbf{x}'), \tag{4.2.9a}$$

$$[\varphi_i(\mathbf{x}, t), \varphi_j(\mathbf{x}', t)] = [\pi_i(\mathbf{x}, t), \pi_j(\mathbf{x}', t)] = 0 \tag{4.2.9b}$$

are obtained from the classical Poisson brackets (4.1.9) via standard ('canonical') quantisation. When both φ_i, φ_j are fermionic (i.e. have fractional spin), then (4.2.9) should be replaced with anti-commutation relations. For simplicity, we consider only bosonic fields.

Because disturbances shouldn't travel faster than light, measurements occurring at space-time points x, x' that are space-like separated (i.e. $(x - x')^2 > 0$) should be independent. Quantum theory translates this into the statement that the corresponding observables $\mathcal{O}(x), \mathcal{O}'(x')$ should commute: $[\mathcal{O}(x), \mathcal{O}'(x')] = 0$ when $(x - x')^2 > 0$. Since the observables are built out of the fields φ_i, this is closely related to the commutation relations (4.2.9). Nevertheless, the relations (4.2.9) are controversial, as we'll see.

To see how to use the field equations and (4.2.9), consider for example the density

$$\mathcal{L}(x) = \frac{-1}{2}\left(m^2 c^4 \hbar^{-2} \phi(x)^2 + c^2 \partial_\mu \phi(x)\, \partial^\mu \phi(x)\right), \tag{4.2.10a}$$

where $\phi = \phi^\dagger$ is self-adjoint. (We will see shortly that this $\mathcal{L}$ has to be modified slightly to be physically sensible.) The field equation here is the Klein–Gordon equation (4.2.7). It can be solved by a trick: the Fourier transform of ϕ from 'position-space' into

[8] Quantum electrodynamics ('QED' for short) is the quantum theory of Maxwell'selectromagnetism applied to electrons, positrons (the anti-particle of the electron) and photons (the particle of light). QED is subsumed by the *Standard Model*, the quantum field theory describing all known physics except for gravity.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

‘momentum-space’ converts the Klein–Gordon equation into decoupled classical simple harmonic oscillator equations, so the field ϕ can be formally written

$$\phi(\mathbf{x},t) = \int \sqrt{\frac{\hbar}{(2\pi)^3 2\omega_{\mathbf{p}}}} \left[\widehat{a}(\mathbf{p}) \exp\left[\frac{\mathrm{i}}{\hbar}\,\mathbf{p}\cdot\mathbf{x} - \mathrm{i}\omega_{\mathbf{p}} t\right] + \widehat{a}(\mathbf{p})^{\dagger} \exp\left[-\frac{\mathrm{i}}{\hbar}\,\mathbf{p}\cdot\mathbf{x} + \mathrm{i}\omega_{\mathbf{p}} t\right]\right] \mathrm{d}^3\mathbf{p}, \tag{4.2.10b}$$

where $\omega_{\mathbf{p}} = \hbar^{-1} p^4 = c\hbar^{-1}\sqrt{\mathbf{p}^2 + m^2c^2}$. If ϕ were a real-valued function, (4.2.10b) would give the general solution, for arbitrary coefficients obeying $\overline{\widehat{a}(\mathbf{p})} = \widehat{a}(\mathbf{p})^{\dagger} \in \mathbb{C}$. Here the coefficients are operators, with $\widehat{a}(\mathbf{p})^{\dagger}$ the adjoint of $\widehat{a}(\mathbf{p})$ (hence the notation). The canonical momentum is $\pi = \partial_4\phi$. Solving (4.2.10b) for $\widehat{a}(\mathbf{p})$ and $\widehat{a}(\mathbf{p})^{\dagger}$ in terms of ϕ, equations (4.2.9) become

$$\left[\widehat{a}(\mathbf{p}), \widehat{a}(\mathbf{p}')^{\dagger}\right] = \delta^3(\mathbf{p}-\mathbf{p}'), \qquad \left[\widehat{a}(\mathbf{p}), \widehat{a}(\mathbf{p}')\right] = \left[\widehat{a}(\mathbf{p})^{\dagger}, \widehat{a}(\mathbf{p}')^{\dagger}\right] = 0. \tag{4.2.10c}$$

This trick of switching from position variables to momentum variables is common in field theory, and it isn’t surprising that it should simplify the mathematics: the momentum degrees of freedom are uncoupled because the theory is translation-invariant (Noether’s Theorem!). If instead ϕ is not self-adjoint, then we should expand ϕ into independent coefficients $a(\mathbf{p})$, $b(\mathbf{p})^{\dagger}$.

How do we accommodate particles in quantum field theory? First note that the particle interpretation pertains directly to state vectors $v \in \Omega$, and not the fields – for example, our universe corresponds to some vector $|\mathit{universe}\rangle \in \Omega$. There are, for example, only four electron fields (i.e. one component for each internal degree of freedom); all of the nearly infinitely many electrons in the universe are *created* by those fields in a way we’ll describe shortly. The *number* of electrons is an observable quantity, and hence an eigenvector of the ‘electron-number’ operator $\widehat{N}_e$. Thus a typical vector $v \in \Omega$ will *not* have a well-defined number of (say) electrons.

The most important vector in Ω is the vacuum state $|0\rangle \in \Omega$, which contains zero particles of each type. It is fixed by the representation of the universal cover of the Poincaré group, i.e. $U_{(a,\Lambda)}|0\rangle = |0\rangle$, so in particular the state $|0\rangle$ has total momentum $\mathbf{0}$ and energy 0. As before, it is unique up to scalar multiplication, nondegenerate and has norm 1: $\langle 0|0\rangle := \| |0\rangle \|^2 = 1$. (Actually, in quantum field theories with spontaneous symmetry breaking, such as the Standard Model, the vacuum will be degenerate, but we will ignore this possibility here.)

The particle interpretation is simplest in the free scalar field theory (4.2.10). Equations (4.2.10b) and (4.2.10c) tells us to think of the free field ϕ as infinitely many independent quantum harmonic oscillators (4.2.5), one for each possible momentum. The analogue of the one-particle state $|1\rangle$ there should be the one-particle state $|\mathbf{p}\rangle$ with momentum $\mathbf{p}$ and energy $\omega_{\mathbf{p}}\hbar$, defined by $|\mathbf{p}\rangle := \widehat{a}(\mathbf{p})^{\dagger}\,|0\rangle$. The problem is that its normalisation

$$\| |\mathbf{p}\rangle \|^2 = \langle 0|\,\widehat{a}(\mathbf{p})\widehat{a}(\mathbf{p})^{\dagger}|0\rangle = \delta(0),$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

obtained using (4.2.10c), is infinite. This is why a quantum field ϕ is an operator-valued *distribution*. The one-particle states can't have well-defined momenta, but rather are 'wave-packets', linear combinations ('superpositions') of those momentum states $|\mathbf{p}\rangle$ constructed using test functions f. In particular, let f be in the Schwartz space $\mathcal{S}(\mathbb{R}^{3k})$. The k-particle states in Ω are of the form

$$|f\rangle := \int \cdots \int f(\mathbf{p}_1, \ldots, \mathbf{p}_k)\,\widehat{a}(\mathbf{p}_1)^\dagger \cdots \widehat{a}(\mathbf{p}_k)^\dagger\,|0\rangle\,\mathrm{d}^3\mathbf{p}_k \cdots \mathrm{d}^3\mathbf{p}_1.$$

The state $|f\rangle$ is an eigenvector of the number operator $\widehat{N}_\phi = \int \widehat{a}(\mathbf{p})^\dagger \widehat{a}(\mathbf{p})\,\mathrm{d}^3\mathbf{p}$, with eigenvalue k. The operators $\widehat{a}(\mathbf{p})$ again are annihilation operators and take a k-particle state to a $(k-1)$-particle state. Together, all these k-particle states, for $k = 0, 1, \ldots,$ span the space Ω. The commutation relation $[a^\dagger, a^\dagger] = 0$ means that the particles obey bosonic statistics, that is both $f \in \mathcal{S}(\mathbb{R}^{3k})$ and its symmetrisation $\frac{1}{k!}\sum_{\sigma\in\mathcal{S}_k} f(\mathbf{p}_{\sigma 1}, \ldots, \mathbf{p}_{\sigma k})$ define physically identical states.

Just as a pendulum in classical mechanics undergoes small oscillations about its (vertical) stationary equilibrium position, so does the vacuum in quantum field theory. The oscillations of the quantum vacuum are the electrons, photons, etc. observed in Nature. This particle concept is the kinematics of quantum field theory.

In these free theories, the k particles in $|f\rangle$ move independently and freely. The notion of wave-packets explains the tracks of particles in the cloud chambers of high-energy experiments: such tracks seem to indicate that the particle has, to a good approximation, both a well-defined position and momentum. By contrast, the (nonphysical) momentum eigenstates $|\mathbf{p}\rangle$ are diffused throughout the universe.

Similarly, particles in any *free* quantum field theory arise by interpreting the Fourier coefficients of the fields as creation and annihilation operators (theories with interactions are considered shortly). Now, any operator can be expressed as an integral of sums and products of these creation and annihilation operators (see section 4.2 of [**555**] for a proof). For example, the free scalar theory (4.2.10) has energy–momentum operators

$$P^\mu = \frac{1}{2}\int p^\mu\,\left(\widehat{a}(\mathbf{p})^\dagger \widehat{a}(\mathbf{p}) + \widehat{a}(\mathbf{p})\widehat{a}(\mathbf{p})^\dagger\right)\,\mathrm{d}^3\mathbf{p}.$$

Since $[\widehat{N}_\phi, P^4] = 0$, we see from (4.2.4) that in this *free* theory the number of particles won't change. It can change only when we include interactions.

Note that in the free scalar theory $P^\mu|0\rangle = 0$ for $\mu = 1, 2, 3$, as it should, but $P^4|0\rangle$, which gives the energy of the vacuum, is

$$P^4|0\rangle = \int \hbar\omega_{\mathbf{p}}\left(\widehat{a}(\mathbf{p})^\dagger\,\widehat{a}(\mathbf{p}) + \frac{1}{2}\right)|0\rangle\,\mathrm{d}^3\mathbf{p} = 0 + \frac{\hbar}{2}\int \omega_{\mathbf{p}}\mathrm{d}^3\mathbf{p}|0\rangle,$$

so is divergent. This is a typical infinity in quantum field theory, but is easy to remedy, as it tells us that the Hamiltonian density $\mathcal{H}(\mathbf{p})$ (hence our original Lagrangian density $\mathcal{L}(x)$) is off by an additive (infinite) constant. It isn't surprising in hindsight that the naive guess (4.2.10a) for $\mathcal{L}(x)$ runs into problems: for one thing, classical energy is only defined up to an additive constant; for another, the order in which the numerical coefficients $a, a^\dagger$ appear in classical expressions for energy doesn't matter, while the order of the operators

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

$\widehat{a}, \widehat{a}^\dagger$ in quantum mechanics certainly does. Replacing $\mathcal{L}(x)$ and $\mathcal{H}(\mathbf{p})$ with their 'normal orders' $:\mathcal{L}:$ and $:\mathcal{H}:$, respectively, gives the vacuum zero energy and doesn't otherwise change the physics. The normal order $:\mathcal{O}:$ of an operator $\mathcal{O}$ given by an integral over $\mathbf{p}$'s of a product of $\widehat{a}(\mathbf{p})$'s and $\widehat{a}(\mathbf{p})^\dagger$'s is obtained by moving all annihilation operators $\widehat{a}(\mathbf{p})$ to the right of all creation operators $\widehat{a}(\mathbf{p})^\dagger$. This has the effect of making the evaluation of operators on states as simple as possible. For example, the Hamiltonian density becomes

$$:P^\mu:= \int p^\mu\, \widehat{a}(\mathbf{p})^\dagger \widehat{a}(\mathbf{p})\, \mathrm{d}^3\mathbf{p}.$$

The same procedure works in any quantum field theory to give the vacuum zero energy, with a minor change when there are fermions. We also used normal-ordering in, for example, (3.2.14a) to remove an analogous infinity in Lie theory.

The existence of negative energy states, which we recall was a serious sickness for relativistic quantum mechanics, is handled naturally in quantum field theory. Return for simplicity to the scalar theory, but now with $\phi \neq \phi^\dagger$. The positive energy coefficients $a(\mathbf{p})$ of ϕ annihilate a positive energy particle; the negative energy coefficients $b(\mathbf{p})^\dagger$ create a positive energy particle. The particle annihilated by the field ϕ is not quite the same as the particle created by ϕ: The various parameters describing particles will either be the same (e.g. mass) or opposite (e.g. electric charge), for these two kinds of particles. That is, the pair $\phi, \phi^\dagger$ of fields is associated with *pairs* of particles; one of these we arbitrarily call the *anti-particle*. Physically, an anti-particle can be interpreted as the corresponding particle 'travelling backwards in time with negative energy', and that is how it is depicted in Feynman diagrams. When $\phi = \phi^\dagger$, the particle is its own anti-particle.

This is how particles arise in *free* quantum field theories. The physically interesting quantum field theories have interactions, that is additional terms in $\mathcal{L}(x)$ corresponding to potential energy. Experiments (e.g. the cloud chambers) tell us that a particle interpretation is still appropriate there. A typical experiment begins and ends with several particles separated by macroscopic distances; interactions occur only at intermediate times when some particles are microscopically separated. What we observe are the initial ('incoming') and final ('outgoing') states, and the transition probabilities $|\langle \text{out}|\text{in}\rangle|^2$. Now, macroscopically separated particles should behave independently to good accuracy. Thus these initial and final states are described by the corresponding *free* theory, at least in the limits $t \to \mp\infty$. A particle interpretation applies directly only to these asymptotic states.

In particular, to each field φ_i in a quantum field theory[9] there are fields φ_i^{in} and φ_i^{out}. The field equations (4.1.8) for the φ_i of course include interaction effects, whereas the asymptotic fields $\varphi_i^{\text{in}}, \varphi_i^{\text{out}}$ obey the free field equations, such as the Klein–Gordon equation (4.2.7). Because $P^4|0\rangle = 0$, the vacuum is constant in time ('stable') and is its

[9] Many of the following comments assume the associated particle is stable and can exist in isolation of the other particles, at least asymptotically. This is the case, for example, for an electron, but not the muon or quark, which are also elementary and have their own fields. See the literature for the necessary modifications.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

own incoming and outgoing asymptotic state. All other incoming states are built up from the vacuum $|0\rangle$ and φ^{in} by the process described earlier. The collection of all incoming states spans the space Ω. Similarly, $|0\rangle$ and φ_i^{out} create all outgoing states, and these also span Ω. Thus the 'in-fields' φ_j^{in} describe the (hypothetical) physics that would occur if the initial particles never interacted; the field φ_j interpolates between these free initial and final asymptotic situations (up to a multiplicative constant, as we'll see), and embodies the true physics by carrying the dynamical information of the system.

As mentioned earlier, experiments obtain information on the transition amplitudes $\langle \text{out}|\text{in}\rangle$ between (prepared) initial states and the (observed) final states, and the complicated machinery of quantum field theory is designed to compute these. These inner products can be thought of as matrix entries of an operator S, the *S(cattering)-matrix*, which defines the equivalence $\varphi^{\text{out}} = S^{-1}\varphi^{\text{in}}S$ between the algebras of in-fields and of out-fields, and the equivalence $|\text{in}\rangle = S|\text{out}\rangle$ between the corresponding incoming and outgoing states. Without going into the technical details, the so-called 'Lehmann–Symanzik–Zimmermann reduction formulae' (see e.g. section 7.2 of [**479**], or section 5-1-3 of [**310**]) express the transition amplitudes in terms of an n-fold integral $\int \mathrm{d}^4x_1 \cdots \mathrm{d}^4x_n$ over spacetime, of 'n-point (correlation) functions', or 'Green's functions', or 'vacuum-to-vacuum expectation values' of 'time-ordered products' of the physical fields:

$$\langle \varphi_{j_1}(x_1)\cdots\varphi_{j_n}(x_n)\rangle := \langle 0|T(\varphi_{j_1}(x_1)\cdots\varphi_{j_n}(x_n))|0\rangle. \tag{4.2.11}$$

We will usually use the statistical term 'correlation function', standard in conformal field theory. The symbol 'T' here reorders the fields $\varphi_{j_i}(x_i)$ in increasing order of the time x_i^4, and is needed to guarantee convergence. The number n here is the total number of particles in $|\text{in}\rangle$ and $|\text{out}\rangle$ together.

In classical physics, Noether's Theorem associates with a continuous symmetry a conserved current $j^\mu(x)$ and a conserved charge Q. Now, a symmetry of a classical system may become broken in quantisation – this is called an *anomaly* (see e.g. section 11-5 of [**310**]). Usually an anomaly is bad news, but a harmless anomaly important to us is the soft breaking of the conformal symmetry in CFT. It is measured by a parameter called the *central charge* or *conformal anomaly c* (Section 4.3.1).

When a symmetry survives quantisation, the analogue of Noether's Theorem here is the *Ward identities* (see e.g. section 10.4 of [**555**]), which are differential equations satisfied by the correlation functions. They take the form

$$\frac{\partial}{\partial x^\mu}\langle j^\mu(x)\,\varphi_{j_1}(x_1)\cdots\varphi_{j_n}(x_n)\rangle = -\mathrm{i}\sum_i \delta(x-x_i)\,\langle\varphi_{j_1}(x_1)\cdots G_i\varphi_{j_i}(x_i)\cdots\varphi_{j_n}(x_n)\rangle, \tag{4.2.12}$$

where G_i is the associated representation of the symmetry on the field φ_{j_i}.

The typical, and only general, way to compute correlation functions is perturbation theory. The correlation functions (4.2.11) play the role here of the propagation kernel K in (4.2.6a); their path integral expression looks like

$$\langle\varphi_{j_1}(x_1)\cdots\varphi_{j_n}(x_n)\rangle = \frac{1}{\mathcal{Z}}\int \phi_{j_1}(x_1)\cdots\phi_{j_n}(x_n)\,\exp[\mathrm{i}S(\phi)/\hbar]\,\mathcal{D}\phi, \tag{4.2.13a}$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

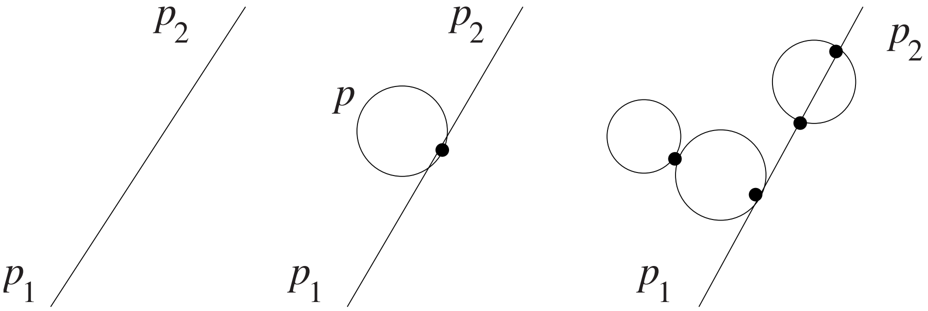

Fig. 4.7 Some two-point Feynman diagrams in the ϕ^4 model.

where S is the classical action (4.1.3) and the integral $\int \mathcal{D}\phi$ is over the space of complex-valued functions $\mathbb{R}^3 \to \mathbb{C}$ (one such 'wave-function' for each field φ_i in the theory). The normalisation factor $1/\mathcal{Z}$ in (4.2.13a) is

$$\mathcal{Z} = \int \exp[\mathrm{i}S(\phi)/\hbar]\,\mathcal{D}\phi, \tag{4.2.13b}$$

called a *partition function* for statistical reasons. We're glossing over technicalities, but the technicalities are (too) easily found in the literature. Once again the mathematical meaning (such as it is) of (4.1.13a) is best ignored; more important are the heuristics it suggests for perturbation.

For that purpose consider a toy model: a single self-adjoint scalar field $\phi = \phi^\dagger$, with ϕ^4 interaction term: $\mathcal{L} = -\frac{1}{2}\sum_\mu \partial_\mu\phi\partial^\mu\phi - \frac{1}{2}m^2 - \frac{\lambda}{4!}\phi^4$ (for typographical clarity we adopt here the usual conventions $c = \hbar = 1$). As always, the equations are simpler if we Fourier-transform to momentum space. The two-point function yields

$$\langle\phi(p_1)\phi(p_2)\rangle = (2\pi)^4\delta^4(p_1+p_2)\left\{\frac{\mathrm{i}}{p_1^2-m^2}\right.$$
$$\left. -\lim_{\epsilon\to 0}\frac{\lambda}{\left(p_1^2-m^2\right)^2}\int\frac{1}{(2\pi)^4}\frac{\mathrm{d}^4p}{p^2-m^2+\mathrm{i}\epsilon} + O(\lambda^2)\right\}. \tag{4.2.13c}$$

The Dirac delta factor expresses momentum conservation. The integral in (4.2.13c) doesn't converge – this infinity is analogous to the infinite self-energy of the electron in classical electromagnetism (Section 4.1.3), and provides the first example of renormalisation, as we will see shortly. The first two terms within the braces of (4.2.13c) correspond to the first two diagrams in Figure 4.7. The second diagram can be interpreted as a particle emitting a pair of virtual particles, which then annihilate themselves. The four-point function $\langle\phi(p_1)\,\phi(p_2)\,\phi(p_3)\,\phi(p_4)\rangle$, computed to λ^1 accuracy, includes the diagrams of Figure 4.8.

The *Feynman rules* describe how to go from the finitely many Feynman diagrams at each perturbation order λ^k, to the corresponding integral expressions. Any book on quantum field theory (e.g. [**310**] or [**555**]) describes them in detail, as they are how the theory makes practical contact with experiment. We will make only general remarks.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

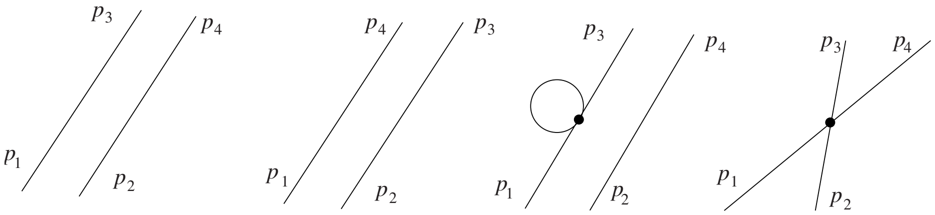

Fig. 4.8 Some four-point Feynman diagrams in the ϕ^4 model.

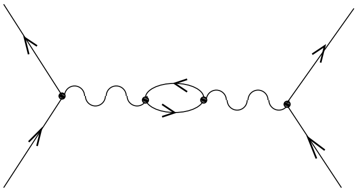

Fig. 4.9 A typical fourth-order term in the scattering of two electrons.

We can write (4.2.13a) symbolically as

$$\int \phi_{j_1}(x_1)\cdots\phi_{j_n}(x_n)\ \exp[\mathrm{i}S(\phi)/\hbar]\,\mathcal{D}\phi = \sum_{\mathcal{G}} c(\mathcal{G}) \int \prod_e \mathrm{d}p_e \prod_v \vartheta_v\delta, \tag{4.2.13d}$$

where the sum is over all Feynman diagrams $\mathcal{G}$ with the external lines (i.e. edges with a free endpoint) corresponding to the fields ϕ_{j_i} in the n-point function. The numerical quantity $c(\mathcal{G})$ is combinatorial. For each internal edge e there is a 'propagator', a momentum p_e and an integral over p_e. At each vertex v there is an operator ϑ_v, which is proportional to the coupling constant, as well as a Dirac delta δ, which expresses momentum conservation at that vertex. Thus each vertex contributes a factor of the coupling constant (which is assumed to be small). The vertices in Figures 4.7 and 4.8 are all of valence 4, because the only interaction term in the Lagrangian density $\mathcal{L}$ here is ϕ^4. More interesting (and physically relevant) quantum field theories involve several types of particles, with several different interaction terms in the Lagrangian, and so the corresponding Feynman diagrams have several types of edges (one for each kind of particle) and several kinds of vertices (one for each term in the interaction Lagrangian). For example, in QED (footnote 8 in this chapter) the interaction term is $-e\overline{\psi}\,\not{A}\psi$, where e is the coupling constant (proportional to the charge of the electron) and where ψ is the (multi-component) field of the electron, $\overline{\psi}$ (essentially the adjoint of ψ) can be thought of as the positron field and $\not{A}$ can be identified with the photon field. A vertex here must consist of three particles: a single incoming or outgoing photon, with an incoming and outgoing electron or positron. A typical Feynman diagram involved in the calculation of the four-point function $\langle\psi(p_1)\,\psi(p_2)\,\overline{\psi}(p_1')\,\overline{\psi}(p_2')\rangle$ is shown in Figure 4.9. It describes the virtual event where the incoming electrons (the bottom two solid lines) exchange a virtual photon (the horizontal wavy line), which in transit spontaneously breaks into an electron–positron pair, which then annihilate, returning the photon. All vertices in

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

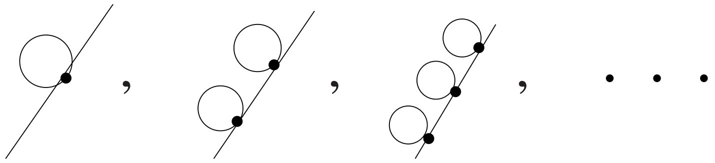

Fig. 4.10 Feynman diagrams contributing to the mass shift.

Figure 4.9 are consistent with the interaction term; as there are four of them there, that diagram contributes to the e^4 term.

In order for an expansion in λ^n (or e^n) to make sense, the individual terms should tend to 0 with n. Embarrassingly, in a typical quantum field theory most individual terms are infinite! A simple example is the two-point function (4.2.13c) at one loop – the problem there is that the integrand doesn't go to 0 fast enough for large p. A different infinity provides a clue how to make sense of these perturbative expansions.

We know from free field theory that the term $-\frac{1}{2}m^2\phi^2$ in the ϕ^4 Lagrangian is a kinetic energy term, and so it is tempting to identify m there with the mass of the ϕ particle. However, that parameter m is not directly observable. The (squares of the) true masses of the particles are defined to be the corresponding eigenvalues of the operator $\sum_\mu P^\mu P_\mu$ (again ignoring $\hbar$'s and c's). The easiest way to compute these eigenvalues is through the two-point function $\langle \phi(p_1)\,\phi(p_2)\rangle$ (called the *propagator* of ϕ): by nonperturbative arguments (see e.g. section 10.2 of [**555**]), the propagator of ϕ should equal the Dirac delta $(2\pi)^4\delta^4(p'+p'')$ times a meromorphic function with a simple pole at $p'^2 = m_\phi^2$ (the physical mass-squared of the particle)[10] with residue i. In the ϕ^4 theory, the propagator to zeroth order (corresponding to the free theory) is $\mathrm{i}/(p'^2-m^2)$, ignoring the Dirac delta factor. However, the perturbative expansion contains geometric series that change the pole. In particular the sequence of diagrams in Figure 4.10 contributes to shifting the denominator, and hence the pole, of the propagator. We call the nonphysical parameter m appearing in the Lagrangian the 'bare mass', in contrast to the true observed mass $m_\phi = m - \delta m$ that is 'dressed' with the cloud of virtual particles arising by virtue of the interaction terms.

The actual values of m and δm can be ignored, since in any physically relevant expression they appear only in the combination $m-\delta m$, which can be replaced by the measured

[10] There is some evidence (by studying the 'running coupling constant') that the propagator of the photon in QED has, in addition to the pole at mass zero (corresponding to the massless photon), a pole at *imaginary mass*. This would correspond to a tachyon (footnote 4 in this chapter) called the Landau ghost, which presumably shouldn't exist. This calculation could indicate a fundamental inconsistency with QED at high energies, but more conservatively may merely indicate a collapse of the perturbative approximation at high energies. Even if each term in the perturbative expansion of QED can be made finite and well-defined (which at present requires *ad hoc* constructions like 'infrared cut-offs'), the full sum over all perturbative orders probably won't converge in any sense. Indeed, the perturbative expansion is a power series in the coupling constant e; if it converged for some small (positive) value of e, then it should also converge for some negative values of e, which for physical reasons is impossible. More generally, many suspect that a consistent quantum field theory must be 'asymptotically free' (i.e. the particles act as if they are free of interactions when the momenta are large). QED is not asymptotically free, but the Standard Model is. However, the Standard Model has other problems (due to the Higgs scalar field) and many suspect that it too is inconsistent.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

value m_ϕ of the physical mass. This is an example of *renormalisation*, and in itself is a standard and uncontroversial ingredient in any physical theory.

However, the mass shift δm can be calculated perturbatively, and in a typical quantum field theory is infinite. Thus in order to account for the observed masses of the particles, the mass parameters in the Lagrangian would also be infinite, which is silly. Nevertheless, the renormalisation scheme given in the previous paragraph works to give sensible and accurate answers.

Likewise, the fields ϕ and coupling constants λ – in short, everything! – appearing in the Lagrangian are also unobservable. The coupling constants λ are renormalised analogously to mass, using the observed strengths of the corresponding interaction, and as usual the rescaling is by an infinite factor. The physical 'renormalised' fields, properly interpolating between the incoming and outgoing free fields, are scalar multiples $Z_\phi^{1/2}\phi$ of the Lagrangian 'bare' fields. This follows, for example, by the residue (call it $Z_\phi \mathrm{i}$) of the propagator: it must equal i, but in a theory with interactions we'll have $Z_\phi \neq 1$ (in fact typically Z_ϕ is infinite). In short, the equal-time commutation relation (4.2.9a) (obeyed by the bare fields) and the residue i of the propagator (necessarily satisfied by the physical fields) are incompatible, and so the bare fields aren't physical. Once again it is not surprising that we must renormalise; what is disturbing is that the renormalisation is infinite.

Quantum field theory makes sense of (i.e. systematically removes) the infinities arising in perturbation theory by a combination of two procedures. The first, called regularisation (Section 4.2.3), introduces some new parameter, call it Λ, and replaces the divergent quantity by a limit as Λ goes to ∞, say, of finite quantities. This nonphysical parameter Λ may be a large momentum cutoff (which corresponds to a small distance cutoff), although more sophisticated cutoffs are common. As long as Λ is finite, the calculation will also be finite, but it will depend on Λ (as well as the various parameters $m, \lambda, \ldots$ in the Lagrangian). However, if we choose ('renormalise') those parameters $m, \lambda, \ldots$ so as to depend on Λ in such a way that the physically relevant quantities are independent of Λ (or at least have a finite limit), we can then take the limit $\Lambda \to \infty$ and get a sensible answer (even though the bare parameters $m, \lambda, \ldots$ will diverge in that limit). We then take those 'sensible answers' to be the predictions of the theory.

In order to remove all infinities, it may be necessary to introduce new bare parameters by adding new terms to $\mathcal{L}$. A quantum field theory is called *renormalisable* if this procedure terminates, that is if all Feynman diagrams will be finite after introducing only finitely many regularisors Λ_i and renormalising the finitely many Lagrangian parameters appropriately. The ϕ^4 model, QED and the Standard Model are all renormalisable. On the other hand, a quantum field theory for gravity in four dimensions, in the spirit of general relativity, is doomed to be nonrenormalisable. Renormalisability is a strong constraint on a theory – for example, it forbids fields with high spin and interaction terms involving many derivatives or products of many fields. For example, the only interaction terms allowed in the Lagrangian of a renormalisable four-dimensional quantum field theory of a single self-adjoint scalar ϕ are ϕ^i for $1 \leq i \leq 4$ and $\sum \partial_\mu \phi\, \partial^\mu \phi$.

A nonrenormalisable theory can always be renormalised (i.e. its divergences all removed) by adding infinitely many new terms to the Lagrangian (along with infinitely

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

many regularisors Λ_i). The problem is that to fix the renormalised values of all those new coupling constants, we would need to perform infinitely many experiments. It would thus appear (and is often argued) that renormalisability would be a necessary condition for a physically relevant, predictive quantum field theory. Such a nonrenormalisable theory would display behaviour that is sensitive to the detailed structure at a much more microscopic level. This behaviour would appear random at the scale on which we are trying to focus. For a macroscopic example, consider the propagation of cracks in glass.

On the other hand, it is possible that all but finitely many of those new parameters will arise in perturbation terms that will be insignificant until the energies of the particles are sufficiently large (e.g. they could involve new particles with very large masses). That is, the contributions from all but finitely many of those parameters could be exponentially suppressed and thus be ignored. Such a theory would be essentially predictive as long as we kept the energies of the collisions far less than the masses of these new and irrelevant particles. Such a nonrenormalisable theory would describe the low energy limit of a more fundamental theory – its nonrenormalisability arises because there is pertinent physics that is not yet accounted for, which occurs at a smaller, deeper scale. For example, quantum gravity could be the low-energy limit of string theory.

In other words, nonrenormalisability could be the norm, as presumably all of our theories are merely limits of deeper ones. A renormalisable theory is merely one in which the deeper physics involves a much higher energy scale (equivalently, a smaller distance scale) than the ones attained in our present experiments. It is a happy accident that the Standard Model is renormalisable. For example, QED applied to a hydrogen atom (an electron moving about a proton) is renormalisable, but is nonrenormalisable when applied instead to a deuteron (an electron moving about a proton–neutron nucleus). The difference is that the physics describing the single proton concerns much smaller distances (approximately 10^{-13} cm) and higher energies than that describing the electron's motion in hydrogen (which involves distances on the order of 10^{-8} cm), while the physics describing the deuteron nucleus also occurs at roughly the same 10^{-8} cm scale.

On a conceptual level, this renormalisation scheme is clearly unsatisfactory. The infinities appearing throughout renormalisation tell us that the fields and parameters appearing in $\mathcal{L}$ are not only nonphysical, but are also nonmathematical. The former is not surprising; the latter gives powerful evidence that the Lagrangian approach to quantum field theory should be avoided. Nevertheless, it works: not only does it permit unambiguous numerical predictions from the Standard Model, but those predictions match up admirably with experiment.

It is easy to get the impression that, whatever its value may be to the pragmatic working physicist, renormalisation should best be avoided by the much more delicately disposed mathematician. Indeed much effort, though with comparatively little success, has been directed at nonperturbative quantum field theory. However, there are many situations where the mathematics arising in perturbation is fascinating. For example, the modular forms arising in string theory, and the Riemann surfaces of conformal field theory, arise directly in the perturbation expansion of string theory. Kreimer, Broadhurst and Connes (see [**105**], [**361**] and references therein) are studying the knot theoretic, Hopf algebraic and number theoretic structure arising in perturbative quantum field theory. Perturbative Chern–Simons theories give both Vassiliev link invariants [**38**] and Gromov–Witten invariants (see e.g. the review [**403**]), depending on how it is perturbatively expanded. We know that what we call perturbative quantum field theory has direct relevance to both mathematics and physics; what hasn't been worked out yet in a conceptually satisfying manner is its precise relationship with 'true' quantum field theory (whatever that is).

This relationship is still mysterious after half a century of work. But recall that Newton's calculus took well over a century to make *mathematical* sense, even though it gave good physics from the beginning. Dirac's use of his delta functions was a much humbler example, but still took several years before Schwartz mathematically legitimised them as distributions. Attempts to make direct sense of quantum field theory are discussed in Section 4.2.4. We are not merely discussing here the rigorous proof of physical conjectures that are almost certainly true – the importance of that activity is easy to overestimate. Rather, we are speaking of making coherent, of finding the meaning of, quantum field theory. There have also been several proposals for a new mathematics underlying quantum field theory. For example, we have the Barrett and Crane interpretation of Feynman diagrams as morphisms in a tensor category (dynamics here comes from representations of the Poincaré group thought of as a 2-category), or Connes' noncommutative geometry (where the geometry of space-time is replaced with an algebra of functions). Some of these approaches are discussed in [**28**].

Of course quantum field theory cannot be identified with perturbative quantum field theory. There are important nonperturbative effects, which cannot be seen in the perturbative expansion. Typical examples are quantum effects due to topologically nontrivial extended solutions to the classical field theory, such as magnetic monopoles (particles carrying magnetic charge) and instantons (solutions concentrated near a point in spacetime rather than along a world-line as happens for particles).

There are other challenges to the coherence of quantum field theory as it is practised today. A famous example is *Haag's Theorem* (1955), which is rigorously proved in the context of the Wightman axioms (see e.g. [**518**]). It says that, given the assumptions built into the picture of quantum field theory sketched above, the S-matrix is very ill-defined unless the theory is free (which isn't physically interesting). We know (Theorem 2.4.2) that there is a unique irreducible unitary representation of the finite-dimensional Heisenberg algebras, but this breaks down for infinite-dimensional ones (Question 2.4.2). Thanks to the equal-time commutation relations (4.2.9), the space-smeared fields $\varphi_j(f)$ of a quantum field theory define at each time t a unitary representation of an infinite-dimensional Heisenberg algebra (just use countably many test functions f with disjoint support). For a fixed quantum field theory, the representations at different times t are unitarily equivalent via the time-evolution operator $U(t) := e^{-\mathrm{i}Ht/\hbar}$, so each theory defines a unique fixed representation. Haag's Theorem tells us that the representations for different values of the coupling constant will be equivalent only if the theories are equivalent. So if our theory is nontrivial, its Heisenberg representation will be different from that of the free theory, that is from that of our so-called asymptotic $t \to \pm\infty$ theories. Thus the limits $U(\pm\infty)$ can't be well defined, and *the justification*

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

for quantum field theory as interpolating between incoming and outgoing states must be dropped (or at least seriously weakened).

One escape is to throw away the equal-time commutation relations (Section 4.2.4). After all, we know that the renormalised (physical) fields won't satisfy them. Also, it seems highly dubious to claim that (4.2.9) are physically relevant, if (4.2.9) permits us to smear fields only in the space direction. We should also smear in the time direction, which means we can no longer speak of *equal-time* relations and the simplicity of (4.2.9) will be lost. On the other hand, (4.2.9) are important, for example, for the usual interpretation of the number operator, and hence are central to the particle interpretation.

The attitude taken by most practitioners of quantum field theory towards these various mathematical difficulties is much like that taken by the author of this book towards most of Life's Little Crises: avoidance. 'Tomorrow they may just go away.' After all, this strategy worked fine with those monsters haunting the night-time shadows of our childhood.

There are formal similarities between quantum field theory and (classical) statistical mechanics. More precisely, path integral expressions in quantum field theory in d-dimensional space-time are the same as, or at least analogous to, thermal averages in statistical mechanics in $d + 1$ space-time dimensions, when the time t is replaced by $-\mathrm{i}k/T$ where k is Boltzmann's constant and T is the temperature. The weak coupling limit in quantum field theory corresponds to the high-temperature limit. Quantum fluctuations about a classical solution correspond to statistical fluctuations about a thermodynamic equilibrium. We won't have much more to say about this connection, though it has been extremely fruitful. For example, spontaneous symmetry breaking in the Standard Model, needed to give masses to particles like the electron, is a phase transition. The Klein–Gordon equation, governing as we know scalar fields, also describes excitations of a dense plasma, or of vortex motions in liquid helium. Conformal field theories, as we shall see next section, can arise both from quantum field theories (string theories) and from statistical mechanics. Incidentally, the transition to imaginary time has an important place in quantum field theory, where it is called 'Wick rotation', and is related to the holomorphicity of the Wightman functions discussed in Section 4.2.4.

The operators in both classical and quantum mechanics form an algebra. This cannot be directly true in quantum field theory, because the product of distributions is not usually a distribution. *It does not make mathematical sense to multiply fields $\varphi_1(x)$, $\varphi_2(y)$ at the same space-time point $x = y$*. Nevertheless, the Lagrangian density, as well as the equal-time commutation relations and many other familiar expressions in quantum field theory, do precisely that. Kenneth Wilson proposed the *operator product expansion* (OPE) as a way to make sense of this. As it is a standard tool of conformal field theory, we defer its treatment to Section 4.3.2. Wilson intended this OPE to be an alternative to the problematic (4.2.9), but as too often happens, his attempt at reformation was absorbed into The System and has become one of its standard tools. The other way to make the operators into an algebra is to smear them, and that is the approach taken by Wightman.

Modern quantum field theory is based on the notion of a *gauge symmetry*. To help understand this important concept, consider the following toy model: a two-dimensional

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

classical particle $(x(t), y(t))$, with equations of motion

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}x(t) + u(t)\,x(t) = 0 = \frac{\mathrm{d}^2}{\mathrm{d}t^2}y(t) + v(t)\,y(t), \tag{4.2.14a}$$

for some fixed functions u, v. Writing $z = x + \mathrm{i}y$ and $w = u + \mathrm{i}v$, this becomes the simpler

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}z(t) + w(t)\,z(t) = 0. \tag{4.2.14b}$$

Of course, this system has a $\mathrm{U}_1(\mathbb{C})$ symmetry, corresponding to a rotation of the z-plane: for any fixed $e^{\mathrm{i}\theta} \in \mathrm{U}_1(\mathbb{C})$, $z(t)$ is a solution of (4.2.14b) iff $e^{\mathrm{i}\theta}z(t)$ is a solution. We call this a *global* (as opposed to *local*) symmetry, because $e^{\mathrm{i}\theta}$ must be constant if it is to define a symmetry of (4.2.14b). However, we can rewrite our system so that $\mathrm{U}_1(\mathbb{C})$ becomes a *local* (time-dependent) symmetry. Introduce a function $A(t)$ (which will serve as a book-keeping or compensating device) and replace each derivative $\mathrm{d}/\mathrm{d}t$ in (4.2.14b) with the differential operator $\mathrm{d}/\mathrm{d}t - \mathrm{i}A(t)$, so (4.2.14b) becomes

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} - \mathrm{i}A(t)\right)\left(\frac{\mathrm{d}}{\mathrm{d}t} - \mathrm{i}A(t)\right) z(t) + w(t)\,z(t) = 0. \tag{4.2.14c}$$

This system (4.2.14c) has a *local* $\mathrm{U}_1(\mathbb{C})$ symmetry: for any smooth function $\theta : \mathbb{R} \to \mathrm{U}_1(\mathbb{C})$, $(z(t), A(t))$ is a solution to (4.2.14c) iff $(e^{\mathrm{i}\theta(t)}z(t), A(t) + \frac{\mathrm{d}}{\mathrm{d}t}\theta(t))$ is a solution to (4.2.14c). Physically, this local symmetry corresponds to the freedom of rotating the system (or the observer) differently at each moment of time. We know from elementary physics that doing this requires introducing the centrifugal forces intimate to all amusement park aficionados. Indeed, we can think of (4.2.14c) as being the equation of motion of a particle z under the influence of a new external force described by A, in addition to the original force described by w. *This is the origin of the 'new external force' A.*

For historical reasons, local symmetries such as the $\mathrm{U}_1(\mathbb{C})$ of (4.2.14c) are called 'gauge symmetries' (gauge here means calibration or scaling). What is significant here is that 'gauging' a global symmetry associates with it a new force; changing the gauge (e.g. rotating the z-plane) is indistinguishable from the action of an apparent force (e.g. a centrifugal one). In the trivial example given above, the force is globally 'fictitious' and the gauging process (4.2.14b) → (4.2.14c) involves no new physics, since we can always solve $A(t) + \dot{\theta}(t) = 0$ for θ and thus 'gauge away' the force A.

Remarkably, all fundamental forces in Nature (namely, gravity, electromagnetism, and the strong and weak nuclear forces) can be obtained by gauging a global symmetry. Consider first special relativity (Section 4.1.2) and for simplicity a single free particle $x(t)$. There, the Poincaré group acts as a global symmetry. It says that the laws of physics shouldn't depend on the choice of origin and inertial observer (coordinate axes). It is a global symmetry, in the sense that once those two choices are made, all observers (regardless of the space-time point x they animate) must agree to use that same origin and coordinate axes in comparing their observations, in order to have a symmetry. This rigidity, this global collaboration, seems physically artificial. What happens if we gauge this symmetry? That is, permit each observer (i.e. each space-time

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

point) to independently choose an origin and coordinate axes. What does that anarchy mean for our description of the relativistic particle? Simply that its coordinates will have changed: $x(t) \mapsto x'(t) = \alpha(x(t))$ where $\alpha : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1}$ encapsulates our new gauge. We require this global change of variables to be invertible, that is to be a diffeomorphism of Minkowski space. So our choice of gauge reduces to a choice of diffeomorphism α. Making the equation of motion independent of that choice α requires introducing book-keeping functions, $A^{\lambda}_{\mu\nu}$, so that the original equation of motion $\mathrm{d}^2x^{\lambda}/\mathrm{d}t^2 = 0$ becomes

$$\frac{\mathrm{d}^2x'^{\lambda}}{\mathrm{d}t^2} - \sum_{\mu,\nu} A^{\lambda}_{\mu\nu} \frac{\mathrm{d}x'^{\mu}}{\mathrm{d}t}\frac{\mathrm{d}x'^{\nu}}{\mathrm{d}t} = 0.$$

Requiring this equation to be equivalent to the original one, we recognise that the components $A^{\lambda}_{\mu\nu}$ are (up to a sign) none other than the Christoffel symbols $\Gamma^{\lambda}_{\mu\nu}$, and that the equation of motion is simply the geodesic equation. The new force corresponding to these A's is identified by Einstein's equivalence principle with gravity. The question of whether gravity can be 'gauged away', that is whether it is globally fictitious and our calculations have been merely a formal mathematical game, reduces to the question of whether space-time is globally flat. It is here – allowing for the suddenly natural possibility that space-time is not flat – that new physics enters. *The real purpose of gauging the symmetry of Minkowski space-time* (Einstein's requirement of 'general covariance') *was to lead us to the idea of curved space-time and the associated force* (which by independent reasoning we identify with gravity). More generally, gauging is a guide for introducing a new force into a theory with a global symmetry: the so-called *principle of minimal interactions*.

Gauging works similarly in quantum field theory. QED results from gauging the $\mathrm{U}_1(\mathbb{C})$ symmetry of free theories. The global U_1 symmetry, $\psi(x) \mapsto e^{\mathrm{i}\theta}\psi(x)$, corresponds to the ambiguity of defining the phase of, for example, the electron field ψ. Once we make the choice at one space-time point, then we must be consistent at all other points. Incidentally, that global symmetry leads to the conservation of global electric charge, by Noether's Theorem. Gauging it means the phase can be changed arbitrarily at each point, that is θ can depend on x. The associated book-keeping field $A_{\mu}(x)$ corresponds to the force we call electromagnetism, and the gauge symmetry implies *local* conservation of charge. For example, in the case of a charged scalar particle, the Klein–Gordon equation (4.2.7) gauges to

$$\sum_{\mu,\nu} \eta^{\mu\nu}(\partial_{\mu} - \mathrm{i}A_{\mu})(\partial_{\nu} - \mathrm{i}A_{\nu})\phi - m^2\phi = 0.$$

It is straightforward to construct a Lagrangian from the original (free) one, which yields the new equations of motion: for example, the free Lagrangian $\sum(\partial_{\mu}\phi^{\dagger})(\partial^{\mu}\phi) + m^2\phi^{\dagger}\phi$ for a scalar field with charge e yields

$$\sum_{\mu}(\partial_{\mu} + \mathrm{i}eA_{\mu})\phi^{\dagger}\,(\partial^{\mu} + \mathrm{i}eA^{\mu})\phi + m^2\phi^{\dagger}\phi.$$

But how should we think of A_{μ}? As another elementary field in the theory. But that means we should add a new term to the gauged Lagrangian, containing partial derivatives

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

of A (otherwise the Euler–Lagrange equations (4.1.8) would be trivial). The simplest gauge-invariant, Lorentz-invariant way to do this is (4.1.13) (with $V = 0$),where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is called the field strength. This is the correct Lagrangian describing the QED of a charged scalar particle. Changing the gauge is indistinguishable from the matter field moving through an electromagnetic field. The associated perturbation theory involves, in Feynman's language, the exchange of virtual particles associated with this new A_μ field – those new particles are called photons.

General relativity tells us to expect a geometric picture here, and indeed that is the case. We think of the matter fields as being sections of a fibre bundle with base $\mathbb{R}^{3,1}$ and fibre $\mathrm{U}_1(\mathbb{C})$; the electromagnetic field A_μ defines a connection for this bundle and $F_{\mu\nu}$ is the curvature tensor.

Similarly, the Standard Model is a gauge theory associated with the gauge group $\mathrm{SU}_3(\mathbb{C}) \times \mathrm{SU}_2(\mathbb{C}) \times \mathrm{U}_1(\mathbb{C})$. SU_3 here corresponds to the strong nuclear force, responsible, for example, for the binding of quarks together to form protons and neutrons, and the binding of protons and neutrons together to form nuclei. $\mathrm{SU}_2 \times \mathrm{U}_1$ describes a unification of electromagnetism with the weak nuclear force (which describes, for example, the decay of the neutron). What this symmetry group $\mathrm{SU}_3 \times \mathrm{SU}_2 \times \mathrm{U}_1$ means physically is less clear than it was for general relativity (or QED), and so the Standard Model lacks the conceptual clarity of Einstein's masterpiece. For example, many believe a deeper quantum field theory will involve a larger gauge group, such as E_6.

Describing other important ingredients of the Standard Model – the fundamental fields and how they transform under $\mathrm{SU}_3 \times \mathrm{SU}_2 \times \mathrm{U}_1$ – would drag us even further from the main thread of this book. For detailed treatments of the Standard Model see, for example, [**310**], [**555**]. Although its comparison with experiment has been fabulous, it is surely not the 'final theory'. For one thing, it suffers from all the conceptual and mathematical flaws mentioned in this subsection. Also, it has 18 free parameters – for example, the electron mass – which must be experimentally determined and (depending on how one counts) there are 61 'elementary' particles in the theory. The Standard Model is an effective theory, valid only for a relatively narrow range of physics. The question is, how different from it will the theory superseding it look?

Quantum field theory challenges our concept of matter. In Newtonian physics reality obtained its solid objective structure from an inert unanalysable 'stuff', from which all substance came; though it could change form (e.g. ice to water), it was the clay on which the Laws of Physics acted. As we moved into the twentieth century we learned that this clay could be transformed into energy ('$E = mc^2$'), and that it is composed of atoms that are mostly empty space. Quantum field theory goes a step beyond: the particles composing atoms are to empty space like sound waves are to air. Bertrand Russell was more accurate than he thought when, in 1956, he compared matter to Lewis Carroll's Cheshire Cat which gradually faded until nothing was left but the grin – matter's grin, Russel speculated, was caused by amusement at those who still think it's there.

Likewise, our notion of force has changed from Newton's definition $\mathbf{F} = m\mathbf{a}$, to something that more generally changes the state of a particle, and that is due not to an active agent but to an indirect effect like a well-hidden symmetry – a further movement of physics away from the prerelativistic infatuation with intuitive space and time.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

4.2.3 The meaning of regularisation

The mathematics of classical physics (symplectic geometry) is well understood, while that of quantum field theory isn't. But it's already clear that, mathematically speaking, quantum field theory is by far the more profound. Much as mechanics helped develop calculus, our standard tool for studying finite-dimensional systems, we can expect quantum field theory to supply us one day with sophisticated new tools for studying infinite dimensions. We are already seeing hints of this.

To a theoretical physicist, quantum field theory is a recipe book, an infinite sequence of finite calculations. To a mathematician, these recipes seem *ad hoc*, and surprisingly classical and finite-dimensional for something that is emphatically neither. A hundred years from now we'll look back at that recipe book much as a modern doctor reflects on medieval medicine: this herb is antiseptic, that incantation is mostly harmless, but leeches and blood-letting were simply bad ideas.

Of all these recipes, those connected with renormalisation and regularisation generate the most ire. For example, even mathematical stoics cannot be unmoved by the substitution (2.3.1). Yet it is in these places where most of the magic lives, as for example the derivation of the Atiyah–Singer Index Theorem from anomaly cancellation indicates.

It isn't difficult for a mathematician to appreciate the inevitability of some form of renormalisation. Consider, for example, the two-body Lagrangian

$$L = \frac{1}{2}m_1\dot{\mathbf{x}}_1{}^2 + \frac{1}{2}m_2\dot{\mathbf{x}}_2{}^2 + G\frac{m_1m_2}{|\mathbf{x}_1 - \mathbf{x}_2|}. \tag{4.2.15a}$$

We can integrate out one of the particles, since the centre-of-mass $m_1\mathbf{x}_1 + m_2\mathbf{x}_2$ is constant (without loss of generality, say it equals $\mathbf{0}$). The resulting one-particle system is

$$L = \frac{1}{2}m\dot{\mathbf{x}}^2 + \frac{k}{|\mathbf{x}|}, \tag{4.2.15b}$$

where $m = m_1(m_1 + m_2)/m_2$ and $k = Gm_1m_2^2/|m_2 + m_1|$. We say that the mass and coupling constants – the 'bare' parameters in (4.2.15a) – have been 'renormalised'.

Something similar happens whenever we integrate away degrees-of-freedom, or account for some effect (e.g. the unavoidable geometric series in Figure 4.10): the new parameters will be readjusted or *renormalised* compared to the old ones. This is completely noncontroversial. What is disturbing about renormalisation in quantum field theory is that you are asked to add/subtract/multiply/divide *infinite* quantities. Regularisation is the procedure of obtaining precise numbers from such an ill-defined operation.

In some sense, regularisation also arises in mathematics. We see it in our Dedekind eta calculation in (2.2.9), or the Virasoro action on affine algebra modules in (3.2.13). Sometimes analytic concerns become significant (e.g. the natural integrals or series one would naively write down turn out to diverge). If those concerns are ignored, we obtain incorrect answers (such as $\eta(-1/\tau) = \eta(\tau)$, or an action of the Witt algebra on affine algebra modules). Of course what we must do is go back and do the analysis properly. Regularisation is merely a symptom of sloppy analysis. It isn't supposed to be the place

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

where the magic appears. The magic was there all along. But the penalty of pretending that (semi-)classical calculations can capture quantum field theory is the introduction of regularisation schemes. The classical calculations fail to pick up that magic, which is then forced to arise in that final step. It's like trying to straighten a Möbius band: as you move your hand around the strip, trying to keep the paper vertical, the twist is relegated to a smaller and smaller portion of paper until eventually the paper tears. That tear is called regularisation. The problem isn't inherent to quantum field theory, the problem is with the fantasy that we can treat quantum field theory semi-classically.

Feynman once asked why the same tricks work over and over in physics. Regularisation is Nature's way of telling us that they don't quite. Unfortunately, we don't yet know how to go back and do the quantum field theory calculations properly. But regularisation must supply some deep hints. For instance, the presence of infinite renormalisation seems to suggest that quantum field theory should be formulated without Lagrangians. Perhaps another hint is that the point ∞ is the difference between the (Riemann) sphere and the (complex) plane, suggesting that regularisation can be interpreted as a (global) topological effect. In [**105**], [**106**], a projective limit of certain Lie groups, corresponding to the Hopf algebra of Feynman graphs, acts on the coupling constants of renormalisable quantum field theories, and contains the renormalisation group as a one-parameter subgroup; dimensional regularisation can in some theories be interpreted as the index theorem in noncommutative geometry.

4.2.4 Mathematical formulations of quantum field theory

Making rigorous sense of quantum field theory is very difficult, as several comments made earlier should indicate. Even the free theories are very subtle; theories with interactions are filled with unresolved problems (Section 4.2.2). One thing is clear: quantum field theory as it is typically practised today (i.e. the informal theory) is mathematically incoherent.

However, quantum field theory *is* a part of mathematics in the sense that important aspects of it have been encoded axiomatically and several examples (mathematically if not physically interesting) have been rigorously constructed. Mathematicians underappreciate just how accessible quantum field theory is. The purpose of this subsection is to briefly describe two of the most influential of these mathematical treatments. These lead to two different formulations of conformal field theories, which we study in later chapters. The fundamental difficulty in the subject lies in rigorously constructing nontrivial examples of quantum field theories within these formulations. Only the very simplest theories (e.g. the free ones) have been rigorously constructed.

The simplest and best-known mathematical treatment of quantum field theory, the Wightman axioms [**518**], was first formulated in the 1950s by Gårding and Wightman. Lagrangians and the equal-time commutation relations (4.2.9) are avoided, and instead attention is focused on the interpolating renormalised 'physical' fields. This makes rigour much easier to attain, but contact with the particle interpretation is more difficult. One unexpected gain is the holomorphicity of the vacuum-to-vacuum expectation values.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

According to Wightman, a quantum field theory consists of the data collected in the following seven axioms W.I–W.VII. For convenience, put $c = \hbar = 1$. Naturally, there is much overlap with the preceding material – the main clarification provided here is what from Section 4.2.2 can (and should?) be avoided.

W.I. (*relativistic state space*) Let $\mathcal{H}$ be a separable Hilbert space, carrying a continuous unitary representation $U_{(a,\Lambda)}$ of the universal cover $\mathbb{R}^4 \rtimes \mathrm{SL}_2(\mathbb{C})$ of the Poincaré group. Define the self-adjoint operators P^μ by $U_{(a,I)} = \exp[\mathrm{i} \sum_\mu P^\mu a_\mu]$; they mutually commute so we can speak of simultaneous eigenstates. All the (simultaneous) eigenvalues p^μ of P^μ are required to satisfy the conditions $p^4 \geq 0$ and $\sum_\mu p_\mu p^\mu \leq 0$.

W.II. (*vacuum*) There is a state $|0\rangle \in \mathcal{H}$, unique up to scalar multiple, invariant under all $U_{(a,\Lambda)}$.

W.III. (*fields*) There is a space $\mathcal{D} \subset \mathcal{H}$, dense in $\mathcal{H}$ and containing $|0\rangle$. There are a finite number $\varphi_1, \ldots, \varphi_M$ of operator-valued tempered distributions over space-time $\mathbb{R}^4$, such that for any 'test function' $f \in \mathcal{S}(\mathbb{R}^4)$, each $\varphi_i(f)$ is an operator from $\mathcal{D}$ to $\mathcal{D}$. The set of fields φ_i is closed under adjoint (i.e. $\varphi_i^\dagger$ equals some φ_j).

W.IV. (*covariance of fields*) For all $(a, \Lambda) \in \mathbb{R}^4 \rtimes \mathrm{SL}_2$, $U_{(a,\Lambda)}(\mathcal{D}) = \mathcal{D}$. Equation (4.2.8a) holds in $\mathcal{D}$, and so the matrices $V(\Lambda)$ define an M-dimensional $\mathrm{SL}_2(\mathbb{C})$-representation.

Physically, the vectors in $\mathcal{H}$ (or rather the rays) are interpreted as the possible states of the theory, and the φ_i are the (renormalised interpolating) quantum fields. We discuss tempered distributions and the Schwartz space $\mathcal{S}$ in Section 1.3.1, and the Poincaré and Lorentz groups and their doubles in Section 4.1.2. If there are any other symmetries of the theory, then $\mathcal{H}$ will also carry a unitary projective representation of those groups. The energy–momentum operators P^μ, generating space-time translations, exist because of the assumed unitarity of the U's. They mutually commute because their exponentiations $U_{(a,I)}$ do. Up to a factor of c^2, the eigenvalue p^4 is the energy of the state and $\sqrt{-\sum p_\mu p^\mu}$ its mass m. We call the vector $|0\rangle \in \mathcal{D}$ in W.II the *vacuum*, and normalise it so that $\langle 0|0\rangle = 1$.

Postulating a common domain $\mathcal{D}$ is necessary because (Section 1.3.1) unbounded operators on a Hilbert space aren't defined everywhere (think of differentiation on the space of square-integrable functions $L^2(\mathbb{R})$). We see from W.III that $\mathcal{D}$ certainly contains the vectors obtained from the vacuum $|0\rangle$ by applying all polynomials in the smeared fields $\varphi_i(f)$, and we learn in W.VI below that those vectors $p(\varphi(f))|0\rangle$ are indeed dense in $\mathcal{H}$. To some approximation, $\mathcal{D}$ can be identified with that subspace (see page 98 of [**518**]).

W.V. (*local commutativity*) For any pair of test functions $f, g \in \mathcal{S}(\mathbb{R}^4)$ satisfying $f(x)\, g(x) = 0$ whenever $(x - y)^2 \geq 0$ (in other words, the supports of f and g are space-like separated), then for any fields φ_i, φ_j, a sign $\pm$ (depending on i, j) can be chosen so that on $\mathcal{D}$

$$[\varphi_i(f), \varphi_j(g)]_\pm := \varphi_i(f)\, \varphi_j(g) \pm \varphi_j(g)\, \varphi_i(f) = 0.$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

W.VI. (*completeness*) The vacuum is cyclic for the smeared fields. That is, polynomials in the smeared fields $\varphi_i(f)$, applied to the vacuum $|0\rangle$, form a subspace dense in $\mathcal{H}$.

Completeness W.VI implies irreducibility of the smeared field operators, in the following sense (inspired by Schur's Lemma): if $B : \mathcal{D} \to \mathcal{D}$ is a bounded operator satisfying

$$\langle u, B\varphi_i(f)v\rangle = \langle \varphi_i(f)^* u, Bv\rangle, \qquad \forall u, v \in \mathcal{D},\ \forall f \in \mathcal{S}(\mathbb{R}^4),\ \forall i = 1, \ldots, M$$

(so in this weak sense B commutes with all φ_i), then B is a constant multiple of the identity. Completeness corresponds here to the remark in Section 4.2.2 that any operator in the theory can be expressed as a function of the smeared fields.

Physically, local commutativity W.V concerns the quantum mechanical fact that measurements localised at space-time points x and y should commute (i.e. be simultaneously measurable without mutual interference) when x and y are space-like separated. It is a consequence of the axioms which sign to take, as is discussed below.

A final axiom is needed to make content with particles (that is to say, with experiment). As it is more technical, it is often avoided in treatments of Wightman's axioms, and we too will be sketchy. The basic idea is that any single particle state $|\lambda\rangle \in \mathcal{H}$ (as usual, $\lambda = \lambda(p)$ describes the decomposition of the state into momentum eigenstates $|p\rangle$) will be an eigenvector for the operator $\sum_\mu P^\mu P_\mu$, with eigenvalue $-m^2c^2$ independent of λ (m is the mass of the particle). On the other hand, eigenstates $|\lambda_1, \ldots, \lambda_n\rangle$ of $\sum P^\mu P_\mu$ corresponding to $n > 1$ particles will have eigenvalue varying continuously with the λ_i. In other words, considering the spectral decomposition of the self-adjoint operator $\sum P^\mu P_\mu$ in $\mathcal{H}$, the single particle states $|\lambda\rangle$ correspond to the discrete part of the spectrum. Call $\mathcal{H}^{(1)}$ the Hilbert space they span – it is a proper subspace of $\mathcal{H}$. There need be no direct relation between the number of elementary fields φ_i and the types of single particles. For example, in the Standard Model quarks correspond to elementary fields but not particles, and protons are particles without a corresponding elementary field. We can now construct incoming $|\lambda_1, \ldots, \lambda_n\rangle^{\text{in}}$ and outgoing $|\lambda_1, \ldots, \lambda_n\rangle^{\text{out}}$ n particle states, corresponding in the $t \to \mp\infty$ limits to tensor products $|\lambda_1\rangle \otimes \cdots \otimes |\lambda_n\rangle$ – see section II.V of [**269**] for the detailed construction. Then the final axiom is:

W.VII. (*asymptotic completeness*) The incoming particle states $|\lambda_1, \ldots, \lambda_n\rangle^{\text{in}}$ topologically span $\mathcal{H}$, as do the outgoing particle states $|\lambda_1, \ldots, \lambda_n\rangle^{\text{out}}$.

Unfortunately, this treatment requires all particles in the theory to have nonzero mass, and so isn't realistic. For example, in quantum electrodynamics the photon is massless and the electron is always surrounded by a cloud of photons, so the single electron states don't belong to a discrete eigenspace of the operator $\sum P^\mu P_\mu$, but rather the eigenvalue varies continuously with upper bound $-m^2c^2$ corresponding to the mass of the electron. For a more sophisticated treatment of the particle concept within quantum field theory, see chapter VI in [**269**].

The role of the n-point functions (4.2.11) are played here by the *Wightman functions*, which are also vacuum-to-vacuum expectation values but aren't time-ordered. Let

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

$\varphi_{i_1}, \ldots, \varphi_{i_n}$ be n fields, not necessarily distinct. Define W_n to be the inner-product

$$W_n(x_1, \ldots, x_n) := W_{\varphi_{i_1},\ldots,\varphi_{i_n}}(x_1, \ldots, x_n) := \langle 0|\varphi_{i_1}(x_1)\cdots\varphi_{i_n}(x_n)|0\rangle.$$

Of course, to make sense of this expression we must smear the points x_i, that is, replace them with test functions f_i. Thus W_n is a complex-valued function of $\mathcal{S}(\mathbb{R}^4) \times \cdots \times \mathcal{S}(\mathbb{R}^4)$. Thanks to Schwartz's Nuclear Theorem, W_n has a unique extension to a tempered distribution on $\mathcal{S}(\mathbb{R}^{4n})$, and it is this extension that is studied. Nevertheless, the inaccurate and occasionally misleading notation $W_n(x_1, \ldots, x_n)$ is too standard to change.

It is possible to convert the data and properties in W.I–W.VII into constraints on the Wightman functions. For example, the relativistic invariance of the vacuum leads to the expression, valid for any (Λ, a),

$$\begin{aligned}\sum_{j_1,\ldots,j_n=1}^{M} V_{i_1 j_1}(\Lambda)\cdots V_{i_n j_n}(\Lambda)\, W_{\varphi_{j_1},\ldots,\varphi_{j_n}}(x_1, \ldots, x_n)&\\ = W_{\varphi_{i_1},\ldots,\varphi_{i_n}}(\Lambda x_1 + a, \ldots, \Lambda x_n + a).&\qquad (4.2.16)\end{aligned}$$

As always of course, everything should be smeared, that is evaluated at $f_i \in \mathcal{S}(\mathbb{R}^4)$ (or $f \in \mathcal{S}(\mathbb{R}^{4n})$). In its unsmeared form, (4.2.16) suggests that W_n is actually a 'generalised function' $w_n(\xi_1, \ldots, \xi_{n-1})$ of the differences $\xi_i = x_i - x_{i+1}$; the precise statement and proof for smeared W_n is given in pages 39–40 of [**518**].

A central result (due to Wightman) is the Reconstruction Theorem: these vacuum-to-vacuum functions W_n uniquely determine the quantum field theory. More precisely, if a collection of tempered distributions W_n satisfies all of the 'obvious' properties (such as the covariance (4.2.16)) that the set of all Wightman functions *should* obey, then the Hilbert space $\mathcal{H}$ and the various fields φ_i obeying axioms W.I–W.VI can be constructed, and moreover any quantum field theory realising the given Wightman functions will be equivalent to the one constructed. The general proof is notationally laborious though fairly straightforward (it is closely related to the Gel'fand–Naimark–Segal construction of a Hilbert space $\mathcal{H}_\rho$ and a representation π_ρ of a C^*-algebra $\mathcal{A}$, associated with a functional $\rho : \mathcal{A} \to \mathbb{C}$). See section 3-4 of [**518**] for the explicit statement and proof for the theory of a single free boson. The Reconstruction Theorem does not tell us when W.VII (i.e. the particle interpretation) holds.

Wightman also proved another remarkable property of his functions: each 'generalised function' $w_n(\xi_1, \ldots, \xi_{n-1})$ is the limit as $z_i \to \xi_i$ of a *holomorphic function* $w_n(z_1, \ldots, z_{n-1})$ of complex variables $z_i \in \mathbb{C}^4$. The domain of holomorphicity contains the following points: $\mathrm{Re}(z_i)$ can be arbitrary but $y_i := \mathrm{Im}(z_i)$ lies in the forward light-cone (i.e. $y_i^4 > 0$ and $y_i \cdot y_i < 0$). So the *distributions* $W_n(x_1, \ldots, x_n)$ are boundary values of the holomorphic *functions* $w_n(z_1, \ldots, z_{n-1})$. The proof of this is not difficult, and involves writing $w_n(z_1, \ldots, z_{n-1})$ as the Laplace transform of the Fourier transform of $w_n(\xi_1, \ldots, \xi_{n-1})$. Physically, this amounts to holomorphically extending from real time (i.e. the Minkowski space-time of physics) to imaginary time (i.e. Euclidean space-time, with better analytic properties).

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

As mentioned earlier, the choice of sign in w.v is fixed. In particular, if φ_1 and φ_2 have spins s_1 and s_2, then we take the sign $-(-1)^{(2s_1)(2s_2)}$. In small space-time dimensions alternatives to bosons and fermions are possible – see section 4.3.5 below – but these exotic possibilities are precluded here by the local commutativity axiom.

Apart from free theories, very few quantum field theories obeying the Wightman axioms have been constructed. In 1953, Thirring rigorously constructed the first interacting theories, but these live in two-dimensional space-time. In the 1960s and 1970s several nontrivial theories with interactions (e.g. a single scalar with ϕ^4 interaction term) were constructed in three and especially two space-time dimensions. One of the \$1 million Clay Institute problems (see http://www.claymath.org/) is to rigorously construct four-dimensional gauge quantum field theories. Quite probably there are easier ways of becoming a millionaire.

In the 1960s Haag and Kastler proposed a different axiomatic approach to quantum field theory, which although more abstract and complicated, appears to be more flexible. We will only sketch it here – see the excellent book [**269**] for a complete treatment, as well as several insights into general quantum field theory. This approach avoids fields, focusing instead on the algebra of observables – as the existence of very different-looking but equivalent field theories emphasises, it is the observables and not the fields that have a direct physical meaning. Remarkably, the entire physical content of the theory can be recovered from these algebras of observables.

Their starting point is to associate with each bounded open set $\mathcal{O}$ in space-time $\mathbb{R}^{3,1}$, a von Neumann algebra $\mathcal{A}(\mathcal{O})$ of bounded operators on a fixed Hilbert space $\mathcal{H}$. This is the same state space $\mathcal{H}$ as in the Wightman axioms, but its role here is much more minor. The self-adjoint elements in $\mathcal{A}(\mathcal{O})$ correspond to the measurements performable within the region $\mathcal{O}$, and so $\mathcal{O}_1 \subset \mathcal{O}_2$ implies $\mathcal{A}(\mathcal{O}_1) \subset \mathcal{A}(\mathcal{O}_2)$. If fields φ were present, $\mathcal{A}(\mathcal{O})$ would be obtained from polynomials in the smeared fields $\varphi(f)$, for test functions with support in $\mathcal{O}$. Conversely, one may hope to define fields $\varphi(x)$ by sending $\mathcal{O} \to \{x\}$. Thus this approach is related to that of Wightman, and it shares with the latter the near-absence of nontrivial examples.

Question 4.2.1. The nonrelativistic analogue of the Poincaré group is the Galilei group, generated by all translations $(\Delta\mathbf{x}, \Delta t)$, all rotations $R \in \mathrm{SO}_3$ and all 'boosts' in velocity $\Delta\mathbf{v} \in \mathbb{R}^3$, as in (4.1.7b). Galilean invariance for nonrelativistic quantum mechanics says that, for any element $\alpha = (R, \Delta\mathbf{v}, \Delta\mathbf{x}, \Delta t)$ of the Galilei group, a wave-function $\psi(x)$ satisfies Schrödinger's equation (4.2.1) iff the corresponding transformed wave-function $\psi'(x')$ (whatever that is) satisfies

$$\mathrm{i}\hbar\frac{\partial\psi'(x')}{\partial t'} = -\frac{\hbar^2}{2m}\nabla'^2\psi'(x') + V(\mathbf{x}')\,\psi'(x'),$$

where $x' = \alpha.x = (t\Delta v + R\mathbf{x} + \Delta\mathbf{x}, t + \Delta t)$ as usual. Show that the obvious transformation formula $\psi'(x') = \psi(x)$ (corresponding to a nonrelativistic scalar) fails here. Rather than transforming in a representation of the Galilei group, ψ must transform in a *projective* representation. Show that the transformation law $\psi'(x') = \exp[\mathrm{i}\Delta_\alpha(x)/\hbar]\psi(x)$ works, where $\Delta_\alpha(x) = m\,(\Delta\mathbf{v})\cdot\mathbf{x} + \frac{m}{2}(\Delta\mathbf{v})^2\,t$.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

Question 4.2.2. Let V_0 be a constant. Solve the one-dimensional Schrödinger equation (4.2.1) for the potential $V(x) = \begin{cases} V_0 & \text{for } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$, with the condition that both ψ and $\partial\psi$ be continuous at $x = \pm 1$.

Question 4.2.3. (a) The vacuum $|0\rangle$ for the harmonic oscillator is the state with minimum possible energy. Find its normalised wave-function $\phi(x, t)$. (See equations (4.2.3).)
(b) Use your answer in (a) to find the average value (expectation value) $\int \psi^* \hat{x}^4 \psi$ of the observable $\hat{x}^4$ in the vacuum.
(c) Now do the same calculation using the Heisenberg picture (4.2.5): calculate the expectation value $\langle 0|\hat{x}^4|0\rangle$ using creation/annihilation operators.

Question 4.2.4. (a) In nonrelativistic quantum physics, the current density is $\mathbf{j}(x) = \frac{\mathrm{i}}{2m}(\overline{\psi}\nabla\psi - (\nabla\overline{\psi})\psi)$ and the probability density is $\rho(x) = |\psi(x)|^2$. Verify that they obey the equation of continuity $\partial\rho/\partial t + \nabla \cdot \mathbf{j} = 0$. (The equation of continuity says that the spatial integrals $\int \rho(\mathbf{x}, t)\, \mathrm{d}^3\mathbf{x}$ are independent of t.)
(b) Suppose ϕ was a wave-function obeying the Klein–Gordon equation (4.2.7). The relativistic version of $(\mathbf{j}, \rho)$ is $j^\mu(x) = \frac{\mathrm{i}}{2m}(\overline{\phi}\, \partial^\mu \phi - (\partial^\mu \overline{\phi})\phi)$. Verify that this obeys the relativistic equation of continuity $\sum_\mu \partial_\mu j^\mu = 0$, but that the corresponding probability density j^4 is not positive. (This is the first sickness of relativistic quantum physics based on the Klein–Gordon equation. The reason for these negative probabilities is that j^4 involves a time derivative, due to the Klein–Gordon equation being second order in time.)
(c) Verify that $\phi_k(x) = \exp[-\mathrm{i} \sum k_\mu x^\mu]$ satisfies the Klein–Gordon equation and is also an eigenfunction of energy and momentum, provided k and m are related in a certain way. Verify that negative energy solutions to the Klein–Gordon equation do exist. (This is the second, related sickness.)

Question 4.2.5. Mathematically speaking, bounded operators are much nicer than unbounded ones. Explain why, physically speaking, we don't lose any generality restricting to bounded self-adjoint observables.

4.3 From strings to conformal field theory

In this section we introduce rational conformal field theory (RCFT), as it is known in physics. Standard references for this material are the book [**131**] and the review articles [**239**], [**209**], [**224**]. We also touch on one of its motivations: string theory. A more mathematical treatment of RCFT is provided in the following section.

We essentially identify conformal field theory (CFT) and perturbative string theory, but this is an oversimplification. For instance, a string theory exists simultaneously on several Riemann surfaces, and the corresponding amplitudes are added together. These surfaces correspond to the various terms in a perturbative expansion (a Taylor series in the string tension parameter T) of the true physical amplitudes. In string theory, the quantities for each surface are of no direct significance by themselves, any more than the term '196 884q' by itself means anything special to $\mathrm{SL}_2(\mathbb{Z})$. In CFT, on the other hand,

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

the Riemann surface is fixed – for example, the theory on the torus could be realised by a statistical mechanical model on the plane where the fields obey doubly-periodic boundary conditions. In fact, it is the deep connection to string theory that gave conformal field theorists the compulsion to explore their theories in arbitrary genus.

Conformal field theory and string theory have impacted remarkably on mathematics. For instance, five of the twelve Fields medals awarded in the 1990s were to men (Drinfel'd, Jones, Witten; Borcherds, Kontsevich) whose work directly concerned aspects of CFT. Probably no other structure has affected so many areas of mathematics in so short a time. Moonshine (and this book) have been deeply influenced by CFT.

The impact so far on physics has been less profound. String theory is still our best hope for a unified theory of everything, and in particular a consistent theory of quantum gravity. It goes through periods of boom and periods of bust, not unlike the breathing of a snoring drunk, and it is still too early to draw any definite conclusions.

However, recall Dirac's quote in Section 1.2.2 about the deep relation between mathematics and physics. For example, the inverse-square law ('force is proportional to $|x-y|^{-2}$') is so mathematically elegant that it must play a role in physics, at least in certain limiting situations. We see it in Newton's gravitation, and the Coulomb force between electric charges, and we now understand it to be the effective macroscopic theory associated with a massless boson in an abelian gauge theory. The same, it can be argued, should be true with string theory.[11]

4.3.1 String theory

The Standard Model describes the quantum theory of the electromagnetic, weak and strong forces. It ignores the force that to us plodding behemoths is the most blatant: *gravity*. The direct approach to quantising gravity fails: the resulting quantum field theory is easy to write down but it is nonrenormalisable and computationally useless. This strongly suggests that new physics should be entering in at high energies (= small distances). Indeed, naive calculations involving general relativity (which relates energy densities to the space-time metric) suggest that as we zoom in on space-time at distances of around 10^{-33} cm (the so-called *Planck length*), the virtual quantum oscillations will change the *topology* of space-time. Far from being a continuum (manifold), space-time at small scales would seem to be some sort of quantum foam.

Because this issue is so fundamental, there are several approaches to resolving it. One of these is string theory, which was created by accident in 1968, where it was applied to the wrong problem, and gave, it was soon realised, the wrong answers. The explosion of interest in it as a theory of quantum gravity, and everything else, began in 1984.

The electron is a *particle*, that is, it can be localised to a point. The Standard Model, say, contains several other equally fundamental particles, each distinguished by different abstract assignments (e.g. representations) attached to that point. In string theory, the

[11] I owe this thought to Peter Goddard.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

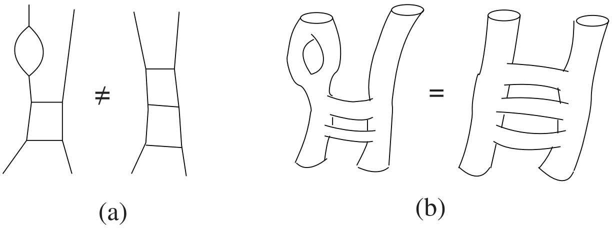

Fig. 4.11 Some two-loop Feynman diagrams of (a) particles and (b) strings.

fundamental object is a string (i.e. a finite curve of length approximately 10^{-33} cm). Depending on the particular theory, this string can be open or closed, oriented or unoriented.

There are several advantages to having extended objects. One is that the particle zoo is simplified, as those abstract assignments can be modelled geometrically using the changing shape of the string. For example, the difference between a string realising an electron, and a string realising a photon, is in how it oscillates. In place of the several dozen 'elementary particles' of the Standard Model, we have only one string, whose precise physical properties at a given time depend not only on its momentum but also its vibrational mode. Likewise, the possible interactions are simplified. Recall that to each term in a particle Lagrangian $\mathcal{L}$, we have a possible vertex for the Feynman diagrams of perturbation theory. On the other hand the interactions of strings are purely topological: for example, a single string can split into two, or two join into one. Most importantly, a theory of quantum gravity seems to arise naturally and seems far better behaved than other quantum theories of gravity.

The weary reader may wonder whether future physicists could initiate new 'revolutions' by replacing strings with membranes or other higher-dimensional manifolds. Such a reader may find some solace in the No-Go theorem described in chapter 2.1.1 of [**261**]. Nevertheless, modern string theory interprets *D-branes* (membranes where the endpoints of open strings reside) as dynamical objects in their own right, corresponding to higher-energy semi-classical solutions. Just as for low-energy approximations we study perturbations about a vacuum, for higher-energy approximations we need to study perturbations about D-branes. It is hoped (though with little justification) that together those perturbative patches cover all of parameter space.

The Lagrangian of a free particle says that the classical particle travels in such a way that its arc-length is minimised. The natural analogue for a string says that the classical string tries to minimise the area of the surface ('world-sheet') it traces out. This *Nambu–Goto action* describes what we now call the bosonic string. An equivalent formulation, called the *Polyakov action*, expresses it as an integral over moduli space.

We are interested in perturbative string theory. Recall (4.2.13d). Figure 4.11 gives some two-loop Feynman diagrams arising in the scattering of two particles/strings. As usual, we take the incoming and outgoing states to be asymptotic (this simplifies things considerably). For simplicity, make the particle theory ϕ^3 (so the diagrams are trivalent)

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

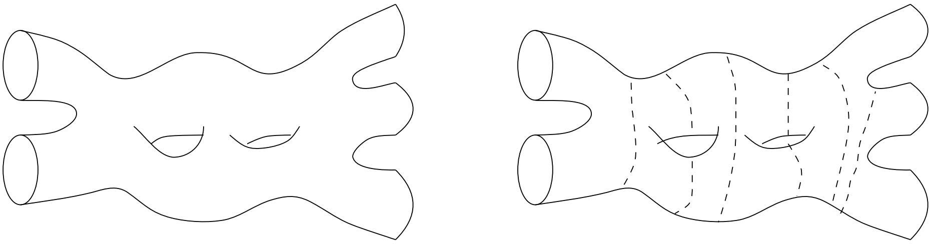

Fig. 4.12 Dissecting a surface into pairs-of-pants.

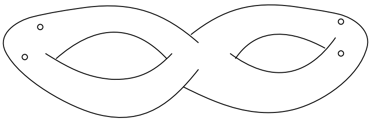

Fig. 4.13 The punctured surface corresponding to Figure 4.11(b).

and the string closed. For the particle, both diagrams in (a) would contribute a term. For the string, the equality in (b) reflects the fact that in Polyakov's formulation, conformally equivalent world-sheets correspond to the same term in the perturbative expansion, and should only be counted once. This is why the Feynman sum reduces to an integral over moduli space (in this case $\mathfrak{M}_{2,4}$).

In any quantum field theory, each vertex v contributes some operator ϑ_v to that perturbation summand. To what does this correspond in (b)? We obtain our 'vertices' by dissecting our world-sheet into spheres with three legs ('*pairs-of-pants*'), as in Figure 4.12. The operator in string theory is called a *vertex (intertwining) operator*. It is a local operator describing the absorption or emission of a string state by another. Surprisingly, these vertex operators are central to the rest of our story.

Because we're really interested in asymptotic $t \to \pm\infty$ initial/final states, the external tubes of the world-sheets are semi-infinite. We can conformally shrink those tubes into punctures (one for each incoming/outgoing string), so Figure 4.11(b) becomes Figure 4.13. The easiest example of this map is also the most important: send a cylindrical world-sheet, with local coordinates $-\infty < t < 0$ and $0 \leq \theta < 2\pi$, to the complex plane using $(t, \theta) \mapsto z = e^{t-\mathrm{i}\theta}$; then the cylinder goes to the unit disc and $t = -\infty$ corresponds to the puncture at $z = 0$. It thus suffices to consider world-sheets that are compact surfaces, with marked points indicating the external lines. The data of those external string states are stored in the appropriate vertex operator attached to that point. This is one of the remarkable features of string theory: that space-time string amplitudes (in, for example, 26 dimensions) can be expressed as correlation functions (4.2.11) in a point-particle quantum field theory in two dimensions, where the fields are vertex operators.

String theory is important to Moonshine because modular functions arise there. That amplitudes in string theory could be modular functions was known almost from the very beginning, and by 1971 we even knew the modern geometric explanation: one-loop vacuum-to-vacuum amplitudes in string theory are path integrals $\int \mathcal{Z}(\text{torus})\, \mathrm{d}[\text{torus}]$ over

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

conformal equivalence classes of tori; because the moduli space of tori is $\mathbb{H}/\mathrm{SL}_2(\mathbb{Z})$ (Section 2.1.4), this makes the modularity of $\mathcal{Z}(\tau) := \mathcal{Z}(\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z}))$ manifest. The meromorphicity of the amplitudes at the cusps follows from the good behaviour ('factorisation') of the amplitudes when the surface is deformed into one with nodes (Section 2.1.4). In short, modular forms and functions appear very naturally in perturbative string theory. Elsewhere, especially Section 7.2.4, we study why this is in more depth.

The modularity (Theorem 3.2.3) of the affine algebra characters χ_λ arises from strings living on the corresponding compact simply-connected Lie group G (this is the so-called Wess–Zumino–Witten model). Likewise, quadratic moonshine (i.e. the modularity of theta functions) arises from the theory of strings living on the torus $\mathbb{R}^n/L$. There is also a string theory responsible for the modularity of the j-function (0.1.8). Much of the remainder of this book tries to explain this.

It is often argued that string theory makes no experimental predictions, other than the dimension of space, which it over-estimates by a factor of 3. This is perhaps a little unfair. String theory predicts a world qualitatively much like that we observe: a world with quantum gravity governed by Einstein's equations at the low-energy, long-distance limit, and gauge groups large enough to include the Standard Model with its zoo of particles. String theory also seems more finite than usual quantum field theories. Unlike the 18 adjustable parameters of the Standard Model, and the fairly arbitrary choices of gauge groups and particles possible in quantum field theories, there is a unique (M-)theory!

But that too is a little dishonest. There are enormous numbers of classical solutions, and each of these serves as a possible vacuum to perturb about. Each choice of vacuum corresponds to a different effective dimension of space-time, gauge group, etc. – different physics. So the problem for the perturbative approach is which vacuum to choose. This isn't so strange: the dynamic role of the vacuum is also important in the Standard Model, where the vacuum is less symmetric than the Lagrangian, and this gives rise to the masses of particles, etc. Also, we know that perturbation theory is only an approximation (probably ill-defined) to the full quantum theory, where for instance we have quantum tunnelling between different vacua. To really understand the effective physics and thus make precise experimental predictions would require a truly nonperturbative treatment of string theory, and this is difficult (D-branes are our most reliable probe for this). In fact, when we have large numbers of strongly interacting strings, the string picture probably ceases as a good way of capturing the physics. But these issues, though important for physics, don't concern Moonshine.

Whether a believer, sceptic or agnostic, one must concede that string theory is truly remarkable. To Witten, physics without strings is like mathematics without complex numbers: just as the particle traces out a real curve (its world-line), the string traces out a complex curve (its world-sheet). Standard string theory books are [**261**], [**463**].

4.3.2 Informal conformal field theory

A conformal field theory is a quantum field theory, usually on a two-dimensional space-time, whose symmetries include the conformal transformations. The first

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

two-dimensional CFT (the $c = 1/2$ free fermion) was constructed by Thirring in 1953. CFT really took off in the 1980s, starting with [**50**]. It arises in string theory, as well as the statistical mechanics describing certain phase transitions. Higher-dimensional CFT appears in the so-called AdS/CFT correspondence (see e.g. [**5**]).

The relation between CFT and string theory is that CFT lives on the world-sheet Σ traced by the strings as they evolve (colliding and separating) through time. Of course, a quantum string only collides and separates in the virtual sense of a Feynman diagram, and so CFT arises in perturbative string theory. More precisely, each term in the Feynman perturbation expansion of S-matrix entries in closed string theory will be a correlation function in a CFT living on the world-sheet. The world-sheets of these scattering strings have a boundary component for every incoming and outgoing string, as in Figure 4.11(b). Any such surface is conformally equivalent to a compact Riemann surface Σ with marked points $p_1, \ldots, p_n$ (one for every incoming and outgoing string), as in Figure 4.13. For reasons we will explain shortly, we also require a choice of local coordinate z_i for each p_i – that is, an explicit identification of a neighbourhood of $p_i \in \Sigma$ with one of $0 \in \mathbb{C}$, so that $z_i = 0$ is the coordinate for p_i. We discuss the moduli space $\widehat{\mathfrak{M}}_{g,n}$ of these 'enhanced surfaces' in Section 2.1.4.

This space-time Σ can be any conformal surface, and we identify conformally equivalent Σ. We restrict to compact orientable Σ, although we don't fix an orientation on it. Because of the string theory interpretation, it is tempting but incorrect to give each such Σ a Lorentzian metric (i.e. locally $\mathrm{d}t^2 - \mathrm{d}x^2$), but for compact Σ such a metric exists only for the torus. Instead, we give each Σ the usual Euclidean signature (i.e. locally $\mathrm{d}x^2 + \mathrm{d}y^2 = \mathrm{d}z\,\mathrm{d}\overline{z}$) of Riemann surfaces. We think of the same CFT as living simultaneously on all such Σ. This leads inevitably to a moduli space formulation.

The simplest indication why two dimensions are so special for CFT is that the space of local conformal transformations, which forms a Lie algebra isomorphic to $\mathfrak{so}_{n+1,1}(\mathbb{R})$ in $\mathbb{R}^n$ for $n > 2$, becomes infinite-dimensional in two dimensions. More precisely, if $f(z)$ is any holomorphic map with nonzero derivative $f'(z_0)$ at some point $z_0 \in \mathbb{C}$, then f is conformal in a neighbourhood of z_0 (the converse is also true – see, for example, theorem 14.2 in [**481**]). Similarly, anti-holomorphic maps preserve the absolute value of angles but reverse the sign. This is essentially the statement that the Lie algebra of conformal Killing vector fields in $\mathbb{R}^n$ is infinite-dimensional iff $n = 2$ (see chapter 1 of [**495**] for a definition and proof); when $n = 2$ it contains two commuting copies of the Witt algebra $\mathfrak{Witt}$ (1.4.9) (one copy for the holomorphic maps and one for the anti-holomorphic ones), arising as dense polynomial subalgebras in this conformal algebra. In our approach, this is how the Virasoro algebra arises. As mentioned in Section 3.1.2, n copies of $\mathfrak{Witt}$ act on the enhanced moduli space $\widehat{\mathfrak{M}}_{g,n}$, either by changing the local coordinate z_i, moving the insertion point p_i or changing the complex structure of Σ.

The CFT literature is very sloppy when discussing the conformal *group* in two dimensions. In spite of numerous published claims to the contrary, it is not the conformal group of $\mathbb{R}^2$ versus that of $\mathbb{R}^n$ ($n > 2$) that singles out two dimensions. The conformal group is isomorphic to the finite-dimensional $\mathrm{SO}_{n+1,1}(\mathbb{R})$ *in any* $\mathbb{R}^n$. Although we can identify $\mathbb{R}^2$ with $\mathbb{C}$, and although holomorphic functions f are locally conformal (provided we

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

avoid the zeros of f'), these f don't form a group. Although the conformal group of $\mathbb{R}^2 \cong \mathbb{C}$ (or its compactification S^2, if we permit poles) is finite-dimensional, the conformal group of 'Minkowski space' $\mathbb{R}^{1,1}$ (or better, its compactification $S^1 \times S^1$ – one S^1 for each null-direction $x^1 \pm x^2$) is infinite-dimensional, and for $S^1 \times S^1$ consists of two copies of $\mathrm{Diff}^+(S^1) \times \mathrm{Diff}^+(S^1)$, where $\mathrm{Diff}^+(S^1)$ isthe oriented diffeomorphism-group of the circle (Section 3.1.2). Thus its Lie algebra is $\mathfrak{Witt} \oplus \mathfrak{Witt}$. If one wants an infinite-dimensional conformal group in CFT, one must put a Minkowski metric on the cylinder or plane.

The subtle and poorly understood role of two dimensions for the conformal group is carefully discussed in [**495**]. Also interesting is how it arises in Segal's picture (Section 4.4.1). For the interplay and representation theories of $\mathfrak{Witt}$, its central extension $\mathfrak{Vir}$ and the real Lie group $\mathrm{Diff}^+(S^1)$, see Section 3.1.2.

On the cylindrical world-sheet in string theory, given a Minkowski metric, the standard light-cone coordinates would be $t \pm x$, where t is time and x is a periodic angle parameter. The solutions to the classical equations of motion on the cylinder would be functions of $t \pm x$ (i.e. left- and right-moving disturbances travelling at the speed of light). As always, the Hamiltonian is proportional to the generator $\partial/\partial t$ of time translations. The Euclidean version (which is what we use) is $w, \overline{w} = t \mp \mathrm{i}x$, and so the left- and right-movers become holomorphic/anti-holomorphic functions of the cylindrical coordinate $w = t - \mathrm{i}x$. As is traditional but slightly disturbing, w and $\overline{w}$ are usually to be treated as independent complex variables; we will return to this subtle point shortly. By a formal application of the chain rule, the Hamiltonian in the Euclidean picture will be

$$\frac{\partial}{\partial t} = \frac{\partial w}{\partial t}\frac{\partial}{\partial w} + \frac{\partial \overline{w}}{\partial t}\frac{\partial}{\partial \overline{w}} = \frac{\partial}{\partial w} + \frac{\partial}{\partial \overline{w}}.$$

In CFT, we prefer to use compact surfaces with marked points, so we should conformally map the semi-infinite tubes of the world-sheets to punctures on a compact surface. Locally, such a map looks like $z = \exp(w)$. This conformally maps our Euclidean cylinder to the punctured plane $\mathbb{C} \setminus 0$. Likewise, $\overline{z} = \exp(\overline{w})$ becomes to the right-moving coordinate. We can now write the Hamiltonian $\mathfrak{Witt}$ generators $\ell_n = \overline{z}^{n+1}\partial_z$:

$$\frac{\partial}{\partial t} = z\frac{\partial}{\partial z} + \overline{z}\frac{\partial}{\partial \overline{z}} = -\ell_0 - \overline{\ell}_0.$$

Basic data in the CFT are the quantum fields $\varphi(z, \overline{z})$ – the vertex operators of last subsection – centred at $z = \overline{z} = 0$ on the Riemann sphere $\Sigma = \mathbb{P}^1(\mathbb{C})$. The notation $\varphi(z, \overline{z})$ emphasises that these fields may depend neither holomorphically nor anti-holomorphically on z. These φ are 'operator-valued distributions' on Σ, acting on the space $\mathcal{H}$ of states for the punctured plane (i.e. corresponding to a propagating string); as usual in quantum field theory, they create the various states by acting on the vacuum $|0\rangle \in \mathcal{H}$. As usual, $\mathcal{H}$ comes with a Hermitian product, which allows us to compare $|\mathrm{in}\rangle$ with $|\mathrm{out}\rangle$; in a physical theory it should be positive-definite (a theory without this positive-definiteness is called *non-unitary*). When we say $\varphi(z, \overline{z})$ is 'centred at 0', we mean that the matrix entry $\langle u, \varphi(z, \overline{z})v\rangle$ will be a Laurent polynomial in the local

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

coordinates z and $\overline{z}$, for any $u, v \in \mathcal{H}$, with a singularity only at 0 (unless the outgoing state u isn't the vacuum, in which case infinity can also be singular).

In a CFT, anything that looks like a quantum field is called a quantum field. In the quantum field theories of Section 4.2, only the finitely many generating fields (e.g. the ones appearing in the Lagrangian) are usually called quantum fields.

Any quantum field theory has a *state-field correspondence*: to a field φ is associated its incoming state, that is the $t \to -\infty$ limit of $\varphi|0\rangle$. Typically, different fields can correspond to the same state. In CFT though, this correspondence becomes a bijection: to a given field $\varphi(z, \overline{z})$ on $\mathbb{P}^1(\mathbb{C})$, we associate the state $\varphi(0,0)\,|0\rangle = v \in \mathcal{H}$ (recall that $z = e^{t-\mathrm{i}x}$). Let φ_v denote the unique field corresponding to state v.

As for any quantum field theory, solving a CFT requires calculating all n-point correlation functions (4.2.11):

$$\langle \varphi_{v_1}(z_1, \overline{z}_1)\, \varphi_{v_2}(z_2, \overline{z}_2) \cdots \varphi_{v_n}(z_n, \overline{z}_n)\rangle_{\Sigma; p_1, \ldots, p_n}, \tag{4.3.1a}$$

for any choice of enhanced surface (Σ, p_i, z_i) and states $v_i \in \mathcal{H}$. We think of $\varphi_{v_i}(z_i, \overline{z}_i)$ as being centred at p_i; the local coordinates $z_i, \overline{z_i}$ describe it as an 'operator-valued distribution' on Σ about p_i. Simplest is the sphere $\Sigma = \mathbb{P}^1(\mathbb{C})$, because then we can fix a global variable w, and choose $z_i = w - p_i$. In this case the time-ordering of (4.2.11), necessary for convergence, becomes the radial-ordering

$$|p_1| < |p_2| < \cdots < |p_n|, \tag{4.3.1b}$$

because of our map $e^{t-\mathrm{i}x}$. The interpretation of n-point functions for other surfaces is more subtle and will be discussed shortly.

The partition functions $\mathcal{Z}_\Sigma$ (4.2.13b) correspond to vacuum-to-vacuum string amplitudes, and are functions on the moduli space of Σ. For example, a sphere is the world-sheet traced by a closed string spontaneously created from and then reabsorbed into the vacuum. As usual in quantum field theory, we can organise these amplitudes by how many internal 'loops' are involved (i.e. the *genus* of the surface): topologically, 0-loop (i.e. 'tree-level') world-sheets are spheres, 1-loop world-sheets are tori, etc. The 0-loop contribution isn't very interesting (all spheres are conformally equivalent), but we'll see shortly that the 1-loop partition function contains considerable information.

Next we describe two general tools introduced by Kenneth Wilson in the 1960s (see e.g. [**558**]). The first is the operator product expansion (OPE). The idea is to replace the ill-defined product $\varphi_1(x)\varphi_2(x)$ of quantum fields by

$$\varphi_1(x)\,\varphi_2(x') = \sum_{n=0}^{\infty} C_n(x - x')\,O_n(x), \tag{4.3.2}$$

so the singularity structure as $x' \to x$ becomes manifest. The singular terms of (4.3.2) are physically the relevant ones. Here, the O_n are fields in the theory, and are expressible as polynomials in the fields φ_i and their various derivatives. The coefficients C_n are complex-valued functions with singularities of the form $|x|^{-p}$ (for $p > 0$) or $\log|x|$, with the more singular coefficients C_n corresponding to simpler fields O_n. Equation (4.3.2) is meant to hold for x' close to x, in the weak sense of matrix entries, that is

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

correlation functions (4.3.1a). The significance of (4.3.2) to (4.3.1a) should be clear. A derivation and clarification of this fundamental concept (4.3.2) is made in (5.1.6), in the context of vertex operator algebras. The scalar quantum field theory in four dimensions, with ϕ^4 interaction term, is worked out in detail in section 13-5-1 of [**310**], where we find for example that the only singular coefficient in the OPE of $\phi(x)\phi(y)$ is proportional to $\log(x^2)$. The reader may find helpful the discussion of OPE given in lecture 3 of [**567**].

The OPE can be made more explicit here because CFT (unlike most theories) is scale-invariant, and this is Wilson's second tool. We apply it separately to z and $\overline{z}$. Scale-invariance means we have a unitary representation $s \mapsto U(s)$ of the multiplicative group $\mathbb{R}^\times_>$ of positive real numbers, which is a symmetry of the Lagrangian; an eigenfield φ transforms by $U(s)^{-1}\varphi(z,\overline{z})U(s) = s^h\varphi(sz,\overline{z})$ for some real number h (the 'scaling dimension' or *conformal weight* of φ). Similarly, scaling $\overline{z}$ yields an independent conformal weight $\overline{h}$. Scale-invariance requires that the coefficient C_n in (4.3.2) scales like

$$C_n(sz,\overline{sz}) = s^{-h_1-h_2+h(n)}\overline{s}^{-\overline{h}_1-\overline{h}_2+\overline{h}(n)}C_n(z,\overline{z}),$$

where $h(n)$ is the conformal weight of O_n. Since

$$U(s)^{-1}\partial_z\varphi\, U(s) = \frac{\partial}{\partial z}s^h\varphi(sz,\overline{z}) = s^h s\frac{\partial}{\partial(sz)}\varphi(sz,\overline{z}) = s^{h+1}(\partial_z\varphi)(sz,\overline{z}),$$

the field $\partial_z\varphi$ has conformal weight $h+1$. Thus the possible conformal weights of the fields O_n lie in $\mathbb{N}h_1 + \mathbb{N}h_2$. This means that (4.3.2) involves only finitely many singular coefficients C_n. We see this more explicitly in (5.1.6).

Recall that, classically, a continuous symmetry implies by Noether's Theorem the existence of a conserved current and conserved charges. In the case of the conformal symmetry of CFT, the conserved current is the *stress–energy tensor*, which has nonzero components $T(z) := T_{zz}(z)$ and $\overline{T}(\overline{z}) := T_{\overline{zz}}(\overline{z})$. The conserved charges $L_n := \frac{1}{2\pi \mathrm{i}}\oint T(z)z^{n-1}\mathrm{d}z$ satisfy

$$T(z) = \sum_{n\in\mathbb{Z}} L_n\, z^{-n-2} \tag{4.3.3}$$

(and similarly for $\overline{L_n}$). In a quantum field theory, these arise in the Ward identities (4.2.12). Here these say, roughly, that taking a derivative of a correlation function $\langle\cdots\rangle_\Sigma$ with respect to a component of the metric on Σ is equivalent to inserting some component of $T(z)$ into that correlation function. The OPE of the field $T(z)$ with itself can be computed:

$$T(z)\,T(z') = \frac{c}{2}\,(z-z')^{-4}\,id + 2\,(z-z')^{-2}\,T(z) + \cdots\,, \tag{4.3.4}$$

where we display only the singular terms. The number c is called the (holomorphic) *central charge* of the CFT. From this we obtain (see (5.1.6c)) the commutation relations for the modes L_n, and we recover (3.1.5a). In other words, the modes L_n define a representation of the Virasoro algebra on $\mathcal{H}$. Likewise, the modes $\overline{L_m}$ also define a representation of the Virasoro algebra (say with central charge $\overline{c}$). These two copies of

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

$\mathfrak{Vir}$ commute: $[L_n, \overline{L_m}] = 0$. From the Hermitian product we get that $c, \overline{c}$ and all the conformal weights h are nonnegative real numbers.

Thus, just as the usual quantum field theories (e.g. the Standard Model) carry projective representations of the Poincaré algebra, a CFT carries a projective representation of its conformal algebra, that is, of two commuting copies of the Witt algebra. Hence we get the true representation of $\mathfrak{Vir} \oplus \mathfrak{Vir}$ on $\mathcal{H}$ defined above. A nonzero central charge c (which is typical) amounts physically to a soft breaking of the conformal symmetry – an anomaly – caused by considering CFT on a surface with curvature. More precisely, the correlation functions (4.3.1a) of a CFT will always be invariant under complex diffeomorphisms of the surface Σ, but in genus > 1 when $c \neq 0$ the correlation functions change under local rescalings of the metric. The central charge can be interpreted physically [**3**] as a Casimir (vacuum) energy, something which depends on space-time topology.

As we have seen, everything in CFT comes in a combination of strictly holomorphic (left-moving) and strictly anti-holomorphic (right-moving) quantities. Here, 'holomorphic' is in terms of the two-dimensional space-time Σ (which locally looks like $\mathbb{C}$), or the local parameters on the appropriate moduli space (which usually locally looks like $\mathbb{C}^\infty$). These holomorphic and anti-holomorphic building blocks are called *chiral*. A CFT is studied by first analysing its chiral parts, and then determining explicitly how they piece together to form the physical quantities. For the applications of CFT to Moonshine, the chiral parts and not the full CFT are what's important. More generally, almost all attention in CFT by mathematicians has focused on the chiral data.

Let $\mathcal{V}$ consist of all the holomorphic fields $\varphi(z)$, and $\overline{\mathcal{V}}$ the anti-holomorphic ones. For example, $\mathcal{V}$ contains $T(z)$. Both $\mathcal{V}$ and $\overline{\mathcal{V}}$ are closed under the OPE (4.3.2), and so form algebras called the *chiral algebras* of the theory. In the next chapter these algebras are axiomatised. $\mathcal{V}$ and $\overline{\mathcal{V}}$ mutually commute and the symmetry algebra of the CFT is often identified with $\mathcal{V} \oplus \overline{\mathcal{V}}$. However, the vacuum is not invariant under most of $\mathcal{V} \oplus \overline{\mathcal{V}}$; we say this symmetry is 'spontaneously broken'. Under the state-field correspondence, $\mathcal{V}$ and $\overline{\mathcal{V}}$ correspond to subspaces V and $\overline{V}$ of the state space $\mathcal{H}$. We call the quantum fields $\varphi(z) \in \mathcal{V}$ *(chiral) vertex operators*.

Since L_0 acts like $-z\partial_z$, the scaling operator $U(s)$ defined earlier is s^{-L_0}. The Virasoro operators L_0, $L_{\pm 1}$ are special in that they generate the three-dimensional conformal group $\mathrm{SL}_2(\mathbb{C})$ of the (Riemann) sphere. We have

$$s^{L_0}\varphi_v(z)\,s^{-L_0} = s^h\varphi_v(sz), \tag{4.3.5a}$$

$$e^{xL_{-1}}\varphi_v(z)\,e^{-xL_{-1}} = \varphi_v(z+x), \tag{4.3.5b}$$

$$e^{xL_1}\varphi_v(z)\,e^{-xL_1} = (1-xz)^{-2h}\varphi_v\left(\frac{z}{1-xv}\right), \tag{4.3.5c}$$

for any $v \in V$, provided $L_0 v = hv$ (we say v has conformal weight h) and $L_1 v = 0$. Such states v are called *conformal quasi-primaries*. If in addition v satisfies $L_n v = 0$ for all $n > 0$, then v is called a *conformal primary* state. They are precisely the lowest-weight states (Section 3.1.2) for the irreducible $\mathfrak{Vir}$-submodules of state-space $\mathcal{H}$; $\mathcal{H}$

will be the direct integral (Section 1.3.1) over all conformal primaries of the associated lowest-weight $\mathfrak{Vir}$-modules. Equations (4.3.5) are generalised in (5.3.15).

More generally, the state-space $\mathcal{H}$ carries a representation of the symmetry algebra $\mathcal{V} \oplus \overline{\mathcal{V}}$, and decomposes into a direct integral of irreducible $\mathcal{V} \oplus \overline{\mathcal{V}}$-modules (proposition 3.1 of [**187**]). A *rational conformal field theory* (RCFT) is one whose state-space $\mathcal{H}$ decomposes into a *finite* sum

$$\mathcal{H} = \oplus M \otimes \overline{N}, \tag{4.3.6a}$$

where M and $\overline{N}$ are irreducible modules of the chiral algebras $\mathcal{V}$ and $\overline{\mathcal{V}}$, respectively. One of the summands in (4.3.6a) is $V \otimes \overline{V}$. The rational ones are the CFTs we are interested in; the name 'rational' was chosen because for them the central charge c and all conformal weights h are rational numbers. The chiral algebras of an RCFT will have only finitely many irreducible modules M; for later convenience let $\Phi = \Phi(\mathcal{V})$ denote the set of these. The $M \in \Phi$ are called *chiral primaries* even though they don't necessarily correspond to a unique vector in $\mathcal{H}$. It is more convenient to write (4.3.6a) in the equivalent form

$$\mathcal{H} = \oplus_{M\in\Phi,\overline{N}\in\overline{\Phi}} \mathcal{Z}_{M,\overline{N}}\, M \otimes \overline{N}, \tag{4.3.6b}$$

where $\mathcal{Z}_{M,N}$ are multiplicities (many of which may be 0). It turns out (because $\mathcal{V}$ is maximal) that $\mathcal{Z}$ will be a permutation matrix. This decomposition (4.3.6b) is reminiscent of the decomposition of a group algebra into irreducible modules. A beautiful interpretation in terms of Frobenius algebras in category theory is given in [**211**].

An important class of RCFT are the *Wess–Zumino–Witten* (WZW) models. These correspond to strings living on a compact Lie group G. Their mathematics is especially pretty, and any natural question seems to have an elegant Lie-theoretic answer. The chiral algebra $\mathcal{V}$ is closely related to the affine Kac–Moody algebra $\overline{\mathfrak{g}}^{(1)}$ associated with G (Section 5.2.2); its modules $M \in \Phi$ can be identified with the integrable highest-weight modules $L(\lambda)$ at a level k determined by c and (3.2.9c).

As with everything else in CFT, the correlation functions (4.3.1a) can be expressed in terms of purely chiral quantities called *conformal* or *chiral blocks*

$$\mathcal{F} = \langle \mathcal{I}_1(v_1, z_1)\mathcal{I}_2(v_2, z_2)\cdots\mathcal{I}_n(v_n, z_n)\rangle_{(\Sigma;p_1,\ldots,p_n;M^1,\ldots,M^n)}\,. \tag{4.3.7}$$

Once again, Σ is a compact Riemann surface with marked points p_i; to each point p_i we assign a local coordinate z_i as before, and also a choice of irreducible module $M^i \in \Phi$. The state v_i is taken from M^i, and the fields $\mathcal{I}_i(v_i, z_i)$, centred at p_i, are called *intertwining operators* and generalise the vertex operators $\varphi_v \in \mathcal{V}$. See Definition 6.1.9 (roughly, each $\mathcal{I}_i(v_i, z_i)$ is an operator-valued distribution sending vectors in some module to another). In the case of higher genus Σ, (4.3.7) cannot be taken too literally, and the study of higher-genus chiral blocks is more difficult [**573**], [**296**]; roughly, the points p_i are first taken in the same coordinate patch of Σ; the function is then extended holomorphically. It will need branch-cuts in Σ to be well-defined.

To solve a given RCFT, it suffices to:

(a) construct all possible chiral blocks (4.3.7); and
(b) reconstruct the correlation functions (4.3.1a) from those chiral blocks.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

In its broad strokes, part (a) was explained in work of Moore–Seiberg [**436**] (and more carefully in [**32**]) – see Section 6.1.4. In deep work, Huang is pursuing the explicit solution to (a) for all sufficiently nice chiral algebras $\mathcal{V}$ (see e.g. [**295**] for the genus-0 story and [**296**] for genus-1). Likewise, in a series of papers written by Fuchs, Schweigert and collaborators, topological field theories (Section 4.4.3) are used to find a solution to (b) (see the reviews [**211**], [**496**]).

In CFT the Ward identities (4.2.12) are especially useful, since the symmetries are so considerable. For example, they imply that it suffices to evaluate the chiral blocks (4.3.7) when all v_i are conformal primaries. Recall that $\mathfrak{Witt}$ acts on moduli spaces (Section 3.1.2); this lifts to one of $\mathfrak{Vir}$ on chiral blocks, and the resulting partial differential equations are the KZ equations of Section 3.2.4. Their monodromy is what makes the chiral blocks so interesting, especially to Moonshine.

The most important example of chiral block is for the torus $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$ with one marked point (it doesn't matter where), assigned $\mathcal{V}$-module $M^1 = V$ and state $v_1 = |0\rangle$. Taking any operator $\mathcal{I}_1$ intertwining some $M \in \Phi(\mathcal{V})$ with itself, the corresponding chiral block (up to a constant multiple) will be the *graded dimension*

$$\chi_M(\tau) := \mathrm{tr}_M e^{2\pi \mathrm{i} \tau\,(L_0 - c/24)}, \tag{4.3.8a}$$

where c is the central charge and L_0 is the Virasoro generator corresponding to energy. We explain in Section 5.3.4 how this arises. Using (4.3.6), the 0-point correlation function for the torus – the 1-loop partition function $\mathcal{Z}$ – becomes

$$\mathcal{Z}(\tau,\overline{\tau}) := \mathrm{tr}_{\mathcal{H}} e^{2\pi \mathrm{i}\,[\tau\,(L_0-c/24)-\overline{\tau}\,(\overline{L_0}-\overline{c}/24)]} = \sum_{M\in\Phi,\overline{N}\in\overline{\Phi}} \mathcal{Z}_{M,\overline{N}}\,\chi_M(\tau)\,\chi_{\overline{N}}(\overline{\tau}). \tag{4.3.8b}$$

This is a very typical decomposition of a physical correlation function into chiral blocks.

The reviews [**496**], [**216**] provide careful explanations of why sometimes we treat z and $\overline{z}$ as independent, and other times we must treat one as the complex conjugate of the other. In short, from the point of view of chiral data, the single space-time Σ of the full CFT is really two disjoint copies with opposite orientation (the Schottky double). For example, the torus with modular parameter $\tau \in \mathbb{H}$ is paired with the one with parameter $-\overline{\tau} \in \mathbb{H}$. As in (4.3.8b), the correlation functions of the full CFT involve both modular parameters, but at the chiral level the two tori don't see each other.

In particular, for a given choice $(\Sigma; \{p_i\}; \{M^i\})$, an RCFT assigns a finite-dimensional space $\mathfrak{B}^{(g,n)}_{\{p_i\},\{M^i\}}$ of chiral blocks. Each chiral block depends multi-linearly on the $v_i \in M^i$, and meromorphically on the z_i, though branch-cuts in Σ between p_i will be needed. The dimension of this space $\mathfrak{B}^{(g,n)}_{\{p_i\},\{M^i\}}$ is called the *Verlinde dimension*, and is given by Verlinde's formula (6.1.2) below.

For example, consider a WZW model associated with an affine algebra $\mathfrak{g} = \overline{\mathfrak{g}}^{(1)}$ and level $k \in \mathbb{N}$. Fix an extended surface (Σ, p_i, z_i). We have a copy of $\mathfrak{g}$ at each p_i, built in the usual way (Section 3.2.2) from the loop algebra $\overline{\mathfrak{g}} \otimes \mathbb{C}[z_i^{\pm 1}]$. The chiral primaries $M \in \Phi$ are the integrable highest weights $\lambda \in P^k_+(\mathfrak{g})$; to each point p_i choose some $\lambda^{(i)} \in P^k_+(\mathfrak{g})$. The associated space $\mathfrak{B}$ of chiral blocks is constructed in [**530**], and these

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

have an important geometric interpretation as spaces of generalised theta functions (see chapter 10 of [**495**]).

The affine algebra characters χ_λ of (3.2.9a), as well as the j-function (0.1.8) are examples of chiral blocks. As we see next subsection, the spaces $\mathfrak{B}^{(g,n)}_{\{p_i\},\{M^i\}}$ naturally carry a representation of the mapping class group $\widehat{\Gamma}_{g,n}$, and this is the source of the relation of the braid group to subfactors, as well as the modularity of Moonshine. In particular, the RCFT characters (4.3.8a) transform nicely under $\mathrm{SL}_2(\mathbb{Z})$: for example,

$$\chi_M(-1/\tau) = \sum_{N\in\Phi} S_{M,N}\, \chi_N(\tau), \tag{4.3.9a}$$

$$\chi_M(\tau+1) = \sum_{N\in\Phi} T_{M,N}\, \chi_N(\tau), \tag{4.3.9b}$$

where S, T are finite complex matrices. This T matrix is given by

$$T_{M,N} = e^{2\pi \mathrm{i}(h_M - c/24)}\delta_{M,N}, \tag{4.3.10}$$

where h_M is a real number (called the *conformal weight*) associated with the chiral primary $M \in \Phi$. The matrix S is, however, more complicated (Section 6.1.2). For example, the matrix T for the WZW models involves the quadratic Casimir of $\overline{\mathfrak{g}}$, while the matrix S involves characters of G evaluated at elements of finite order.

The simplest class of RCFT are the *minimal models*, which have the smallest possible chiral algebra (generated only by the identity field and the stress–energy field $T(z)$) and nevertheless still have a finite decomposition (4.3.6a). They are well understood (see e.g. [**131**]).They are the RCFT with central charge $0 < c < 1$, and correspond to the discrete series (3.1.6) of $\mathfrak{Vir}$.

The smallest nontrivial minimal model is the Ising model. It has central charge $c = 0.5$. The associated chiral algebra has three irreducible modules, which we label $\Phi = \{0, \epsilon, \sigma\}$ as in [**131**]. Their graded dimensions (4.3.8a) are

$$\chi_0(\tau) = q^{-1/48}\,(1 + q^2 + q^3 + 2q^4 + 2q^5 + 3q^6 + 3q^7 + \cdots), \tag{4.3.11a}$$

$$\chi_\epsilon(\tau) = q^{23/48}\,(1 + q + q^2 + q^3 + 2q^4 + 2q^5 + 3q^6 + 3q^7 + \cdots), \tag{4.3.11b}$$

$$\chi_\sigma(\tau) = q^{1/24}\,(1 + q + q^2 + 2q^3 + 2q^4 + 3q^5 + 4q^6 + 5q^7 + \cdots), \tag{4.3.11c}$$

where as always $q = e^{2\pi \mathrm{i}\tau}$. From this we can read off the conformal weights $h_0 = 0, h_\epsilon = 1/2, h_\sigma = 1/16$, and hence the T matrix of (4.3.9b):

$$T = \begin{pmatrix} e^{-\pi \mathrm{i}/24} & 0 & 0 \\ 0 & e^{23\pi \mathrm{i}/24} & 0 \\ 0 & 0 & e^{\pi \mathrm{i}/12} \end{pmatrix}. \tag{4.3.11d}$$

The matrix S is more difficult to find, but it equals

$$S = \frac{1}{2}\begin{pmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 0 \end{pmatrix}. \tag{4.3.11e}$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

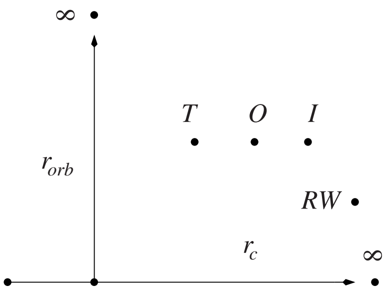

Fig. 4.14 The moduli space of conformal field theories with central charge $c = 1$.

The 1-loop partition function $\mathcal{Z}(\tau)$ of (4.3.8b) is

$$\mathcal{Z}(\tau) = |\chi_0(\tau)|^2 + |\chi_\epsilon(\tau)|^2 + |\chi_\sigma(\tau)|^2.$$

The CFT corresponding to open string perturbation – *boundary CFT* – is also interesting (see e.g. the review [**461**]). In this direction, see the proposals in [**215**], [**458**] (building on the α-induction of subfactors [**65**]). For instance, the 1-loop partition function corresponds to a Frobenius algebra (Section 4.4.3) in the modular category of modules of the associated chiral algebra, and the boundary CFT data arise as a 'category module'. However, boundary CFT isn't so relevant for Moonshine and will mostly be ignored in this book.

The space of CFTs can be probed using 'marginal operators' – fields φ_v with conformal weight $(h, \overline{h}) = (1, 1)$ obeying certain other properties (see e.g. [**137**] and [**246**] section 8.6). A given CFT can be deformed (changing its spectrum but not central charge c), provided it contains such a field. If the given CFT has n marginal operators, then the space of CFTs in its neighbourhood is expected (typically) to look like an n-dimensional real manifold. When the given CFT has more marginal operators than the neighbouring ones, the space of CFTs at that point may look like two manifolds intersecting transversely, or it can mean an orbifold singularity where you get different realisations for the same CFTs. The RCFTs are special points in this space. The space of known $c = 1$ CFTs is drawn in Figure 4.14. Points on the horizontal and vertical lines are parametrised by a radius $\sqrt{2}^{-1} \leq r_{orb}, r_c \leq \infty$; these two half-lines intersect at $r_{orb} = 1/\sqrt{2}, r_c = \sqrt{2}$. The known *rational* $c = 1$ CFT consists of the three isolated theories *T(etrahedral)*, *O(ctahedral)* and *I(cosahedral)*, together with those theories with $r_{orb}^2 \in \mathbb{Q}$ or $r_c^2 \in \mathbb{Q}$. The fourth isolated point, *RW*, is irrational and described in [**483**]. Theories with radii r_c and $r_c' = 1/(2r_c)$ are equivalent, as are those with radii r_{orb} and $r_{orb}' = 1/(2r_{orb})$ (this is an example of 'T-duality', and arises from the extra marginal operator possessed by the $r_c = \sqrt{2}^{-1}$ and $r_{orb} = \sqrt{2}^{-1}$ theories). The intersection point also has two, while the isolated points have no marginal operators, and the remainder have one (which permits r to be continuously varied). The moduli space for CFT with central charge $c < 1$ consists of countably many isolated points [**91**]. Very little is known about the moduli space for $c > 1$.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

4.3.3 Monodromy in CFT

One way to make conformal symmetry manifest is to make the relevant physical quantities be holomorphic functions of (or more precisely, sections of bundles over) the appropriate moduli spaces. Let $\mathcal{V}$ be the chiral algebra of an RCFT and let Φ label its (finitely many) irreducible modules, that is the chiral primaries. Let 0 denote the one corresponding to the subspace V of $\mathcal{H}$. Let's investigate more closely what chiral blocks (4.3.7) are.

In any RCFT, there are differential equations that the chiral blocks must satisfy. The most well known of these are the *Knizhnik–Zamolodchikov* (or KZ) equations. We studied these for WZW models at genus 0 in Section 3.2.4. Good expositions of this material are given in [**355**], [**207**], [**186**]. Differential equations can also be found using null vectors [**50**], and using the Ward identities.

Return to the Ising model, introduced last subsection. We know its chiral blocks in genus-0 with two or three marked points (Question 4.3.5). Consider now four marked points on the Riemann sphere, at positions $w_i \in \mathbb{C} \cup \{\infty\}$. The chiral block will be the product of the quantity

$$\prod_{1\leq i<j\leq 4} (w_i - w_j)^{-h_i-h_j+\frac{1}{3}\sum_k h_k} \tag{4.3.12a}$$

with some function of the cross-ratio

$$w := \frac{(w_1-w_2)(w_3-w_4)}{(w_1-w_3)(w_2-w_4)}. \tag{4.3.12b}$$

We can simplify this using the Möbius symmetry of the Riemann sphere to move w_i to $0, w, 1, \infty$, respectively. If we label all four marked points with the primary field $\sigma \in \Phi$, then the space of chiral blocks is two-dimensional, spanned by

$$\mathcal{F}_1(w) = \frac{\sqrt{1+\sqrt{1-w}}}{\sqrt{2}\,(w(1-w))^{1/8}}, \tag{4.3.13a}$$

$$\mathcal{F}_2(w) = \frac{\sqrt{1-\sqrt{1-w}}}{\sqrt{2}\,(w(1-w))^{1/8}}. \tag{4.3.13b}$$

The fractional powers tell us these chiral blocks have branch-point singularities – that is, to get a holomorphic function on the w-plane, we need to make semi-infinite cuts. Nevertheless, we can analytically continue these functions along any curve. Take a point w_0 so that $0 < |w_0| < 1$, and consider the circle $w(t) = w_0\, e^{2\pi i t}$ for $0 \leq t \leq 1$. Nothing special happens to the numerator of the $\mathcal{F}_i(w)$: its values at $t = 0$ and $t = 1$ are equal. The denominator however picks up a factor $e^{2\pi \mathrm{i}/8}$, and thus both blocks $\mathcal{F}_i(w)$ pick up a net factor of $e^{-2\pi \mathrm{i}/8}$. We call this the *monodromy* about $w = 0$ (Section 3.2.4).

Consider next their monodromy about $w = 1$. Here our circle will be $w(t) = 1 + w_0 e^{2\pi i t}$, again for w_0 small. Note that the numerators of $\mathcal{F}_1$ and $\mathcal{F}_2$ switch, and the denominators again pick up a factor of $e^{2\pi \mathrm{i}/8}$. Thus this monodromy can be written

$$\begin{pmatrix}\mathcal{F}_1(w)\\ \mathcal{F}_2(w)\end{pmatrix} \mapsto \begin{pmatrix} 0 & e^{-2\pi\mathrm{i}/8} \\ e^{-2\pi\mathrm{i}/8} & 0\end{pmatrix}\begin{pmatrix}\mathcal{F}_1(w)\\ \mathcal{F}_2(w)\end{pmatrix}.$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

In Section 3.2.4 we explain how to think of this. Reintroducing the four coordinates w_i, the chiral blocks $\mathcal{F}_i$ will be holomorphic on the universal cover $\widetilde{\mathfrak{C}_4}$ of the configuration space $\mathfrak{C}_4$ of (1.2.6). Analytically continuing along any closed path γ in $\mathfrak{C}_4$ (across any of those branch cuts) defines an action of the fundamental group $\pi_1(\mathfrak{C}_4)$ on the space $\mathfrak{B}^{(0,4)}$ of chiral blocks. This group $\pi_1(\mathfrak{C}_4)$ is the pure braid group of the sphere with four strands. An element β of the full braid group of the sphere maps the space $\mathfrak{B}^{(0,4)}_{m^1,m^2,m^3,m^4}$ to $\mathfrak{B}^{(0,4)}_{m^{\beta 1},m^{\beta 2},m^{\beta 3},m^{\beta 4}}$, where βi is the associated permutation, so in our example (4.3.13) the full braid group acts. We can recover the usual planar braid groups $\mathcal{P}_3$ and $\mathcal{B}_3$ here by fixing one of the four points at say ∞, and letting the others wander around.

Equivalently, as a 'function' on the configuration space, the chiral blocks form (multi-valued) holomorphic sections of a projective flat vector bundle. What this means is that each chiral block satisfies a system of partial differential equations (the KZ equations) describing how to parallel-transport it around the configuration space, and flatness says it will *locally* depend only on the moduli space parameters (and not on the path chosen). *Globally*, however, there will be monodromy [**437**], [**32**], [**355**].

More generally, a chiral block $\mathcal{F}$ on an enhanced surface Σ is a multi-valued function on the corresponding moduli space. To make it well defined, $\mathcal{F}$ can be lifted to the corresponding Teichmüller space. There will be an action of the corresponding mapping class group $\widehat{\Gamma}_{g,n}$ coming from monodromy (a *projective* action, if as usual the central charge c is nonzero). How to centrally extend these $\widehat{\Gamma}_{g,n}$ so that the projective representation becomes a true one is discussed, for example, in [**404**]. This picture, which is explained quite clearly in [**32**] and is developed further in, for example, Section 7.2.4, encompasses not only the braid group monodromy of the KZ equation (Section 3.2.4) but also the modular group action (4.3.9) on the graded dimensions (4.3.8a). It is the source of the modularity in Moonshine.

Although the chiral blocks themselves are multi-valued functions on the moduli spaces $\widehat{\mathfrak{M}}_{g,n}$, conformal invariance requires that the n-point correlation functions (4.3.1) themselves be well-defined functions on $\widehat{\mathfrak{M}}_{g,n}$. For example, even though the graded dimensions χ_M transform as in (4.3.9), the 1-loop partition function in (4.3.8b) is $\mathrm{SL}_2(\mathbb{Z})$-invariant. See also Question 4.3.7.

As we know from Section 2.2.1, there is more to being a modular form or function than transforming nicely with respect to $\mathrm{SL}_2(\mathbb{Z})$. The behaviour at the cusps of $\mathbb{H}$ is also crucial, as it says our function lives on a compact space. Something similar also holds in RCFT. The analogue of cusps for the other moduli spaces – that is, the surfaces corresponding to the extra points needed for compactification – are surfaces with nodes (Section 2.1.4). What we need is nice behaviour of chiral blocks as we move in moduli space towards surfaces with nodes, that is, as we shrink a closed curve about a handle on our surface down to zero radius. This is given by (4.4.3) and is called *factorisation* [**203**], [**539**]. It connects the moduli spaces of different topologies, and tells us CFT is defined on a 'universal tower' of moduli spaces (Sections 3.1.2 and 6.3.3).

Incidentally, it is tempting to try to extend this formalism to the 'surfaces of infinite genus' given by projective limits $\lim_{\leftarrow}\Gamma\backslash\mathbb{H}$ (see Section 2.4.1). The discrete groups Γ appearing in each such limit must all be commensurable (i.e. intersections of any two of them should have finite index in both), in order for the limit to be defined. In Section 2.4.1 we describe the most famous piece of such a limit: the modular tower $\lim_{\leftarrow}\Gamma(N)\backslash\mathbb{H}$, so important to number theory. The assignment of, for example, chiral blocks to such 'surfaces' may be built up from those of each $\Gamma\backslash\mathbb{H}$, in a relatively straightforward way; because of this, perhaps we could interpret the string-theoretic data for $\lim_{\leftarrow}\Gamma\backslash\mathbb{H}$ as the (nonperturbative?) contribution ('sum') associated collectively with all world-sheets $\Gamma\backslash\mathbb{H}$ appearing in that limit. In any case, we are led to speculate from (2.4.3) that both CFT and the theory of vertex operator algebras (and indeed Moonshine itself) may extend quite nicely to the p-adics $\widehat{\mathbb{Q}}_p$. Some moves in this direction are [**562**], [**520**]. To a number theorist, the usual perturbation about a vacuum would correspond to the infinite prime, but would mysteriously ignore the contributions from all the finite primes. It would be interesting to see if nonperturbative phenomena like D-branes can be sensed by these projective limits.

As discussed at the end of Section 2.2.1, the analogue of q-expansions, for chiral blocks and partition functions in higher genus, are expansions about surfaces with nodes. A natural projectively flat connection on these spaces $\mathfrak{B}^{(g,n)}$ of chiral blocks is given by the stress–energy tensor $T(z)$ [**203**], [**530**]; this connection is responsible for the KZ equations, and is the analogue here of the $\mathfrak{Witt}$ action on moduli spaces, and the meaning of $T(z)$ insertions into correlation functions discussed in Section 4.3.2.

4.3.4 Twisted #4: the orbifold construction

To particles, a space-time singularity is a problem; to strings, it is merely a region where stringy effects are large. The most tractable way to introduce such singularities is by quotienting ('gauging') by a finite group. This construction plays a fundamental role for CFTs and vertex operator algebras; it is the physics underlying what Norton calls *generalised Moonshine* (Section 7.3.2). This is where finite group theory touches CFT.

Let M be a manifold and G a finite group of symmetries of M. The set M/G of G-orbits inherits a topology from M, and forms a manifold-like space called an *orbifold*. Fixed points become conical singularities. For example, $\{\pm 1\}$ acts on $M = \mathbb{R}$ by multiplication. The orbifold $\mathbb{R}/\{\pm 1\}$ can be identified with the interval $x \geq 0$. The fixed point at $x = 0$ becomes a singular point on the orbifold, that is, a point where locally the orbifold does not look like some open n-ball (open interval in this one-dimensional case). For other examples, see Question 4.3.8.

Orbifolds were introduced into geometry in the 1950s as spaces with mild singularities; recalling Definition 1.2.3, they are V_α/G_α patched together, where $V_\alpha \subset \mathbb{R}^n$ is open and G_α is a finite group. They were introduced into string theory in [**143**], which greatly increased the class of background space-times in which the string could live and still be amenable to calculation. This subsection briefly sketches the corresponding construction for CFT; our purpose is to motivate Section 5.3.6.

For concreteness think of a closed string whose world-sheet $\Sigma \subset M$ is a torus, since the 1-loop partition function (4.3.8b) is the easiest way to obtain the spectrum (4.3.6)

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

of the theory. Think of Σ being parametrised by $z \in \mathbb{C}/(\tau\mathbb{Z} + 2\pi\mathbb{Z})$, with τ being the time-period of the 1-loop and 2π being the space-period of the closed string. Here, G is a finite group of symmetries of the theory – it acts not only on space-time M, but also on the internal states of the string (i.e. the state-space $\mathcal{H}$ carries a representation of G). Assume for now that G is abelian and that $\mathcal{H} = V \otimes \overline{V}$. For example, this is satisfied by the WZW theory for ${E_8}^{(1)}$ at level 1, or strings living on the torus $\mathbb{R}^n/L$ for an even self-dual n-dimensional lattice L. Consider first the chiral data. The orbifold chiral algebra $\mathcal{V}^{orb}$ is the subalgebra $\mathcal{V}^G$ of $\mathcal{V}$ consisting of all G-invariant fields. More difficult to answer is what the orbifold state-space $\mathcal{H}^{orb}$ looks like.

In the case of a point particle, a 1-loop world-line $\mathbf{x}(t) \in M/G$ would be a circle, the motion $\mathbf{x}(t)$ would be periodic (say with period T); lifting $\mathbf{x}(t)$ to M, we would require that $\mathbf{x}(T) = g.\mathbf{x}(0)$ for some $g \in G$. The closed string also requires this twisted periodicity in the time direction, but being closed it will similarly have a twisted periodicity in the space direction. Thus we are led to consider string processes satisfying the boundary conditions

$$\mathbf{x}(z+\tau) = g.\mathbf{x}(z), \qquad \mathbf{x}(z+2\pi) = h.\mathbf{x}(z). \tag{4.3.14a}$$

The strings satisfying $\mathbf{x}(2\pi) = h.\mathbf{x}(0)$ form the h*-twisted sector* $\mathcal{V}^h$ – these twisted sectors are the special feature of strings living on orbifolds. They don't live in the original chiral space V, and are hard to construct; in particular, there isn't a systematic twisted analogue of the vertex operator construction (i.e. exponentials of free fields) of untwisted sectors.

The contribution of the processes (4.3.14a) to the 1-loop path integral will be

$$\mathcal{Z}_{(g,h)}(\tau) := \mathrm{tr}_{\mathcal{V}^h}\, g\, e^{2\pi \mathrm{i}\tau\,(L_0 - c/24)}, \tag{4.3.14b}$$

for reasons that will become clearer next section (the trace comes from obtaining the torus by sewing together the inner and outer boundaries of an annulus). Each (finite-dimensional) L_0-eigenspace in $\mathcal{V}^h$carries a representation of the group $\langle g\rangle$, so that is the matrix to substitute into the trace (4.3.14b). The modular group $\mathrm{SL}_2(\mathbb{Z})$ acts on the cycles (homology H_1) of the torus in the usual way, which gives the behaviour of $\mathcal{Z}_{(g,h)}$ under modular transformations:

$$\mathcal{Z}_{(g,h)}\left(\frac{a\tau+b}{c\tau+d}\right) = \mathcal{Z}_{(g^a h^c,\, g^b h^d)}(\tau). \tag{4.3.14c}$$

Actually, we will find shortly that in general this transformation has to be modified slightly.

The twisted sector $\mathcal{V}^h$ is an irreducible (twisted) module for the original chiral algebra $\mathcal{V}$ (Section 5.3.6). In terms of the orbifold chiral algebra $\mathcal{V}^G$, $\mathcal{V}^h$ will be a true module, though not an irreducible one. Its decomposition ('branching rules') into irreducible $\mathcal{V}^G$-modules is

$$\mathcal{V}^h = \oplus_\rho \mathcal{V}^h_\rho \otimes \rho, \tag{4.3.15a}$$

where the sum is over all irreducible G-representations ρ (when G is non-abelian, this

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

will be modified slightly). Plugging this into (4.3.14b) gives the equivalent expressions

$$\mathcal{Z}_{(g,h)}(\tau) = \sum_{\rho} \mathrm{ch}_{\rho}(g)\, \chi_{(h,\rho)}(\tau), \tag{4.3.15b}$$

$$\chi_{(h,\rho)}(\tau) := \mathrm{tr}_{\mathcal{V}^h_\rho} e^{2\pi \mathrm{i}\tau\,(L_0-c/24)} = \frac{1}{\|C_G(h)\|} \sum_{g\in G} \overline{\mathrm{ch}_{\rho}(g)}\, \mathcal{Z}_{(g,h)}(\tau). \tag{4.3.15c}$$

The graded dimension $\chi_{(h,\rho)}$, unlike $\mathcal{Z}_{(g,h)}$, has a q-expansion with coefficients in $\mathbb{N}$, but $\mathcal{Z}_{(g,h)}$ has the simpler modular behaviour, in perfect analogy to Θ_{t+L} versus $\Theta_{L;r,s}$ (compare (2.2.11) and (2.3.10)).

An important example of this orbifold construction is the Moonshine module $V^\natural$ (Sections 5.3.6 and 7.2.1). Its starting point is the chiral algebra $\mathcal{V}(\Lambda)$ for the torus $\mathbb{R}^{24}/\Lambda$, where Λ is the Leech lattice. The symmetry group G corresponds to the centre $\{\pm 1\}$ of $\mathrm{Aut}(\Lambda)$. The graded dimension of the untwisted sector $\mathcal{V}(\Lambda)$ is $\mathcal{Z}_{(1,1)}(\tau) = J(\tau) + 24$, and has -1-twisted graded dimension

$$\mathcal{Z}_{(-1,1)}(\tau) = q^{-1} \prod_{n=0}^{\infty} (1-q^{2n+1})^{24} = q^{-1} - 24 + 276q - 2048q^2 + \cdots$$

The -1-twisted sector $\mathcal{V}(\Lambda)^{-1}$ has untwisted/twisted graded dimension

$$\begin{aligned}\mathcal{Z}_{(\pm 1,-1)} &= 2^{12} q^{1/2} \prod_{n=0}^{\infty} \left(1 \mp q^{(2n+1)/2}\right)^{-24} \\ &= q^{1/2} \pm 98304q + 1228800q^{3/2} \pm 10747904q^2 + \cdots\end{aligned}$$

The Moonshine module $V^\natural$ consists of the sectors $\mathcal{V}(\Lambda)^1_+ \oplus \mathcal{V}(\Lambda)^{-1}_+$ and so has graded dimension

$$\begin{aligned}\chi_{V^\natural}(\tau) &= \chi_{(1,+)}(\tau) + \chi_{(-1,+)}(\tau) \\ &= \frac{1}{2}\left(\mathcal{Z}_{(1,1)}(\tau) + \mathcal{Z}_{(-1,1)}(\tau) + \mathcal{Z}_{(1,-1)}(\tau) - \mathcal{Z}_{(-1,-1)}(\tau)\right) = J(\tau).\end{aligned} \tag{4.3.16}$$

So far we have discussed only the *chiral* orbifold CFT – our main interest. The state-space (4.3.6) of the full orbifold CFT can look like

$$\mathcal{H}^{orb} = \oplus\, \mathcal{V}^g_\rho \otimes \overline{\mathcal{V}}^g_\rho. \tag{4.3.17}$$

There are other possibilities for $\mathcal{H}^{orb}$; a systematic but far from exhaustive source is provided by discrete torsion [**136**]. The lattice construction $L\{T\}$ of Section 2.3.3 (applied to indefinite lattices L) is this orbifold construction of $\mathcal{H}^{orb}$, coming largely from discrete torsion. The construction of $V^\natural$ is a heterotic version (i.e. with trivial 'anti-holomorphic' chiral algebra $\overline{\mathcal{V}}$). In any case, the full orbifold theory will typically involve most sectors $\mathcal{V}^g_\rho$. Modular invariance (4.3.14c) is one way to see the necessity of this; another is string dynamics (see figure 8.1 in [**463**], vol. I).

There are three significant generalisations of this orbifold construction as outlined above. *Non-abelian* orbifold groups G are at least as interesting to us (e.g. Maxi-Moonshine concerns $V^\natural/\mathbb{M}$), and introduce new subtleties. For example, using (4.3.14a) to evaluate $\mathbf{x}((z+\tau)+2\pi) = \mathbf{x}((z+2\pi)+\tau)$ requires $hg.\mathbf{x}(z) = gh.\mathbf{x}(z)$. That is, we

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

should limit ourselves to boundary conditions (4.3.14a) whose pairs (g, h) commute. Moreover, consider the h-twisted sector $\mathbf{x}(2\pi) = h.\mathbf{x}(0)$; hitting both sides with $g \in G$ yields $(g\mathbf{x})(2\pi) = (ghg^{-1}).(g\mathbf{x})(0)$, that is, the twisted sectors $\mathcal{V}^h$ and $\mathcal{V}^{ghg^{-1}}$ are naturally isomorphic. In fact, $\mathcal{Z}_{(g,h)} = \mathcal{Z}_{(kgk^{-1},khk^{-1})}$ for any $k \in G$, so we should identify each boundary condition (g, h) with all simultaneous conjugations (kgk^{-1}, khk^{-1}). This will be clearer in Sections 5.3.6 and 6.2.4. The sums in (4.2.15) are over all $g \in C_G(h)$ and all irreducible $C_G(h)$-representations ρ, where $C_G(h)$ is the centraliser of h in G.

For the second generalisation, note that $g \in C_G(h)$ takes the sector $\mathcal{V}^h$ to $\mathcal{V}^{ghg^{-1}} = \mathcal{V}^h$ so (as in Section 1.5.4) we get a linear map $\phi_g^{(h)} : \mathcal{V}^h \to \mathcal{V}^h$. So far we have implicitly assumed that these assignments $g \mapsto \phi_g^{(h)}$ define a representation of $C_G(h)$. But $\mathcal{V}^h$ are chiral data and so group actions, etc. may be projective. That is, we only know that $g \mapsto \phi_g^{(h)}$ defines a *projective* representation of $C_G(h)$. In this case, (4.3.14c) must be replaced by

$$\mathcal{Z}_{(g,h)}\left(\frac{a\tau + b}{c\tau + d}\right) = \gamma\, \mathcal{Z}_{(g^a h^c, g^b h^d)}(\tau), \tag{4.3.18}$$

for some root of unity γ. See [**138**], and Section 5.3.6 below, for details. For example, the Maxi-Moonshine orbifold $V^\natural/\mathbb{M}$ will necessarily be of that projective type [**408**].

For the final generalisation, we have discussed orbifolding the CFTs with one chiral primary (i.e. with $\|\Phi\| = 1$) only because they are simpler. The behaviour of more typical multi-primary orbifolds is analogous (Section 5.3.6). For example, the horizontal line of $c = 1$ CFTs in Figure 4.14 corresponds to bosons compactified on a circle of radius r, while the vertical line there corresponds to bosons on the orbifold $S^1/\mathbb{Z}_2$ (see the treatment in [**246**]); most of these theories have infinitely many chiral primaries (i.e. aren't rational). The WZW theory for $A_1{}^{(1)}$ at level 1 is a $c = 1$ theory with two chiral primaries corresponding to a string living on S^3; we can orbifold this rational theory by any of the finite subgroups of $\mathrm{SU}_2(\mathbb{C})$. These subgroups fall into an *A–D–E* pattern (Section 2.5.2). Orbifolding by the (cyclic) A-series of subgroups gives the $c = 1$ theories $r_c = n/\sqrt{2}$, and by the (dihedral) D-series gives the $c = 1$ theories $r_{orb} = n/\sqrt{2}$. The (tetrahedral) E_6-, (octahedral) E_7- and (icosahedral) E_8-subgroups give us the isolated theories T, O, I of Figure 4.14.

Choose any CFT $\mathcal{H}$ and tensor it with itself n times to get a new CFT $\mathcal{H}^{\otimes n}$. The orbifold $\mathcal{H}^{\otimes n}/\mathcal{S}_n$ is called a *permutation orbifold*. Requiring that $\mathcal{H}^{\otimes n}/\mathcal{S}_n$ possesses the standard CFT properties imposes highly nontrivial conditions on the chiral data of $\mathcal{H}$. See, for example, [**37**] for applications of this powerful theoretical tool.

4.3.5 Braided #4: the braid group in quantum field theory

Much of Moonshine is implicit in two-dimensional CFT. What is the most distinctive physical feature of two-dimensional quantum field theory?

In three or more dimensions, the rotation group $\mathrm{SO}_n(\mathbb{R})$ is non-abelian. We know everything about the finite-dimensional unitary projective representations of this simple Lie group: there are countably many, namely the highest-weight representations of its

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

universal cover $\mathrm{Spin}_n(\mathbb{R})$. Physically, we know these fall into two families ('superselection sectors'), depending on what happens after a rotation by 2π: the true representations of $\mathrm{SO}_2(\mathbb{R})$ (the 'integer-spin' bosons) and those that are merely projective (the 'half-integer spin' fermions).

In two dimensions, this familiar picture collapses, as the rotation group $\mathrm{SO}_2(\mathbb{R})$ is isomorphic to S^1 and has universal cover $\mathbb{R}$. The unitary representations are parametrised by the 'unitary duals' $\widehat{S^1} \cong \mathbb{Z}$ and $\widehat{\mathbb{R}} \cong \mathbb{R}$, respectively. In particular, the element $x \in \mathbb{R}$ is sent to the 1×1 matrix $e^{2\pi i\alpha x}$ for 'spin' $\alpha \in \widehat{\mathbb{R}} = \mathbb{R}$. The behaviour (monodromy) of these representations under rotations by 2π again determines the physics, and instead of the boson/fermion alternative, we get superselection sectors parametrised by $\widehat{\mathbb{R}}/\widehat{S^1} \cong S^1$.

The different physics of bosons and fermions is revealed by the spin–statistics relation. Define as in (1.2.6) the configuration space $\mathfrak{C}_n(\mathbb{R}^d)$ of n distinct points $x^{(i)}$ in $\mathbb{R}^d$, consisting of n copies of $\mathbb{R}^d$ with all diagonals $x^{(i)} = x^{(j)}$ deleted. We are interested in these describing the positions of n identical particles, so for each permutation $\sigma \in \mathcal{S}_n$ identify $(x^{(1)}, \ldots, x^{(n)}) \in \mathfrak{C}_n(\mathbb{R}^d)$ with $(x^{(\sigma 1)}, \ldots, x^{(\sigma n)})$. A closed loop in $\mathfrak{C}_n(\mathbb{R}^d)/\mathcal{S}_n$ corresponds to an explicit rearrangement of the n particles. It is important to note that, for any n, d, the space of trajectories will be disconnected. In Feynman's formalism, this means we have the freedom to introduce relative factors between the corresponding disjoint path integrals. By unitarity these factors should be phases (complex numbers of modulus 1), and consistency requires them to define a representation of the fundamental group $\pi_1(\mathfrak{C}_n(\mathbb{R}^d))$. For $d > 2$ this fundamental group is the symmetric group $\mathcal{S}_n$, and so there are only two possible choices for these relative phases, corresponding to the two one-dimensional representations of $\mathcal{S}_n$: all $+1$'s, or $\det(\sigma)$. The spin–statistics theorem [**518**] tells us that $+1$ corresponds to bosons and $\det(\sigma)$ to fermions.

In two dimensions, the fundamental group is the braid group $\mathcal{B}_n$, and its one-dimensional unitary representations are parametrised by $t \in \mathbb{R}/\mathbb{Z}$ and defined by $\sigma_i \mapsto e^{2\pi i t}$. This t parametrises the different consistent assignments of phases to the disjoint integrals in the Feynman expressions. Again, the spin–statistics theorem relates this phase assignment to spin: this t is the same as the spin α (mod 1). This is called *braid statistics* for obvious reasons. Such particles are called *plektons* (after the Greek word for 'braid') or *anyons* (since they can have *any* spin).

One-dimensional representations of $\mathcal{S}_n$ or $\mathcal{B}_n$ are the simplest. Higher-dimensional representations would indicate an internal structure and are considered in, for example, parastatistics. In Section 4.3.3 we see how higher-dimensional representations arise in a similar way in CFT. See, for example, [**204**], [**191**] for some general treatments of braid statistics in CFT. Possible physical realisations of braid statistics are reviewed in [**557**], [**345**]. In particular, subjecting certain semiconductors to large magnetic fields and cold temperatures yields the so-called fractional quantum Hall effect, and its quasi-particles provide an actual realisation of anyons. Since braid statistics is a topological effect, it is intimately related to the Aharanov–Bohm effect (a notorious topological effect in quantum theories).

So two dimensions are special for quantum field theory. We know four dimensions are special in differential geometry [**195**]. For example, in any $\mathbb{R}^n$ all differential structures

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

are equivalent, except $n = 4$ where there are uncountably many inequivalent ones (Section 1.2.2). Are those two dimensions related to these four dimensions, and are they related to the apparent four-dimensionality of macroscopic space-time? This isn't clear to this author.

The possibility of braid statistics arises in two dimensions because the space-like vectors in two-dimensional space-time are disconnected. The other special features of two dimensions are all related to this. As we discuss in Section 4.3.2, the space of local conformal transformations is finite-dimensional in n dimensions, except for $n = 2$ where it is infinite. The light-cone minus the origin is also disconnected in two dimensions, and this implies the existence of infinitely many conserved currents.

What makes four dimensions special in differential geometry is the behaviour of embedded 2-discs (many proofs in n dimensions are based on understanding that behaviour). A generic map of a disc into an n-manifold has self-intersections that are one-dimensional if $n = 3$, which consist of isolated points if $n = 4$, and are non-existent if $n \geq 5$. Also, the Seiberg–Witten equations (so useful for studying 4-manifolds) exploit the fact that the rotation algebra $\mathfrak{so}_4 \cong \mathfrak{so}_3 \oplus \mathfrak{so}_3$ (corresponding to a group $\mathrm{SO}_4(\mathbb{R})$ homeomorphic to $S^3 \times \mathbb{P}^3(\mathbb{R})$) is nonsimple, while in all other dimensions $n > 2$ $\mathfrak{so}_n$ is simple.

Question 4.3.1. (a) Consider the free scalar theory in d dimensions, given by Lagrangian $\mathcal{L} = -\frac{1}{2}\sum_\mu \partial_\mu\phi\, \partial^\mu\phi$. Assuming scale-invariance of $\mathcal{L}$, deduce the scaling dimension of ϕ.
(b) This theory is massless. What happens when the mass term is introduced?

Question 4.3.2. Prove that when $p + q \neq 2$, the infinitesimal conformal symmetries of $\mathbb{R}^{p,q}$ form a finite-dimensional Lie algebra, but that it is infinite-dimensional when $p + q = 2$. (That is, write $x^\mu \mapsto x^\mu + \epsilon^\mu(x)$; we're interested in those infinitesimal ϵ^μ for which the metric $\mathrm{d}s^2$ goes to a multiple of itself.)

Question 4.3.3. Let Σ be a Riemann surface of genus g with n discs removed. Suppose it is dissected into N 'pairs-of-pants' (i.e. spheres with three discs removed). Prove that this dissection is possible only if $n + 2g > 2$, in which case $N = n + 2g - 2$.

Question 4.3.4. Assuming (4.3.5) and the state-field correspondence, prove $L_1 v = 0$ and $L_0 v = hv$.

Question 4.3.5. Suppose $L_1 v_i = 0$ and $L_0 v_i = h_i v_i$. Compute the chiral blocks

$$\langle \varphi_{v_1}(z_1)\,\varphi_{v_2}(z_1)\rangle = \begin{cases} C_{12}\,|z_1 - z_2|^{-2h_1} & \text{if } h_1 = h_2 \\ \quad 0 & \text{otherwise} \end{cases},$$

$$\langle \varphi_{v_1}(z_1)\,\varphi_{v_2}(z_2)\,\varphi_{v_3}(z_3)\rangle = \frac{C_{123}}{|z_1 - z_2|^{h_1+h_2-h_3}\,|z_2 - z_3|^{h_2+h_3-h_1}\,|z_1 - z_3|^{h_1+h_3-h_2}}$$

for constants C_{12}, C_{123}, using (4.3.5).

Question 4.3.6. Describe the monodromy (if any) about $w = \infty$ of the chiral blocks in (4.3.13).

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

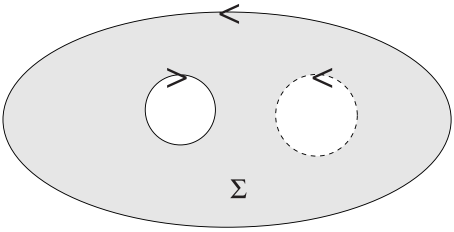

Fig. 4.15 A morphism $\Sigma : C_1 \to C_2$.

Question 4.3.7. Find the sesquilinear combinations $\sum_{i=1,2} c_{ij}\mathcal{F}_i(w)\overline{\mathcal{F}_j(w)}$ of the chiral blocks in (4.3.13), which are invariant under the various monodromies. (The physical correlation functions will be of that form.)

Question 4.3.8. Describe the following orbifolds: (a) $(\mathbb{R}/\mathbb{Z})/\{\pm 1\}$; (b) $(\mathbb{C}/(\mathbb{Z}\tau + \mathbb{Z}))/\{\pm 1\}$; (c) $(\mathbb{C}/(\mathbb{Z} + \mathrm{i}\mathbb{Z}) \setminus \Delta)/\mathbb{Z}_2$, where Δ is the diagonal $x + \mathrm{i}x$, and $\mathbb{Z}_2$ acts by identifying (x, y) and (y, x).

4.4 Mathematical formulations of conformal field theory

In Sections 4.3.2 and 4.3.3 we gave a quick standard sketch of the basics of CFT, introducing the reader to the main notions. In this section, as well as Chapter 5 and Section 6.1, we explore certain aspects of CFT more carefully, clarifying them considerably. Surprisingly, many of these aspects are fundamental to Moonshine.

4.4.1 Categories

A deeply influential formulation of CFT is due to Graeme Segal [**500**], [**502**], [**498**]; see also [**241**]. It is motivated by string theory (Section 4.3.1) and is phrased using category theory (Section 1.6.1). According to Segal, a CFT is a functor $\mathcal{S}$ from a category **C** of Riemann surfaces (the world-sheets) to the category **Hilb** of Hilbert spaces (the state-spaces).

The objects of category **C** are finite disjoint unions C_n of n circles, for all $n \geq 0$. We fix a parametrisation on these circles – that is, a smooth identification t of each circle C with $\mathbb{R}/\mathbb{Z}$; this induces an orientation on C. A morphism $C_m \to C_n$ is a (not necessarily connected) Riemann surface Σ with boundary $\partial\Sigma$ consisting of $m + n$ parametrised circles; exactly n of those boundary circles come with parametrisations consistent with the orientation of Σ induced from its complex structure. We think of these n as 'outgoing' strings and the remaining m as 'incoming' ones. For example, in Figure 4.15 the solid circles are outgoing and the dashed one is incoming. We identify two such morphisms $\Sigma : C_m \to C_n$, $\Sigma' : C_m \to C_n$ if there is a conformal map $f : \Sigma \to \Sigma'$ such that the parametrisations t_i and $t'_i \circ f$ of the boundaries $\partial\Sigma$ and $\partial\Sigma'$ agree.

The space $\mathrm{Hom}(C_m, C_n)$ is topological, with a connected component $\mathbf{C}_\Sigma$ for each homeomorphism class $[\Sigma]$ of (not necessarily connected) surfaces with boundary having

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

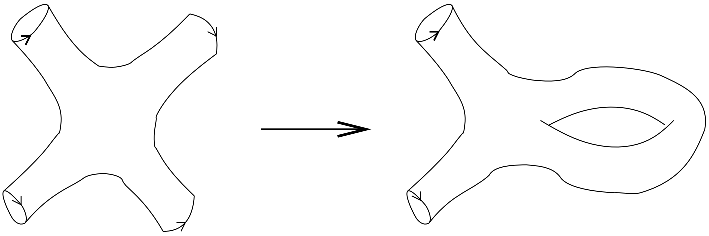

Fig. 4.16 An example of sewing.

$m + n$ components. For example, $\mathrm{Hom}(C_0, C_0)$ has one component for every choice of n_0 spheres, n_1 tori, ..., n_g compact genus-g surfaces, ..., provided $\sum_g n_g < \infty$.

Finally, the composition $\Sigma' \circ \Sigma$ of morphisms $\Sigma : C_m \to C_n$, $\Sigma' : C_n \to C_p$ is obtained by sewing together the surfaces Σ and Σ' along the circles in C_n by using the parametrisation to identify corresponding points on the boundaries. In fact, this sewing construction is the main reason we require these boundary circles to be parametrised.

Recalling Definition 2.1.6, this space $\mathbf{C}_\Sigma$ can be regarded as the quotient of the space of complex structures on Σ, by the group of all diffeomorphisms of Σ that are the identity on the boundary $\partial\Sigma$. Thus, $\mathbf{C}_\Sigma$ is an infinite-dimensional moduli space. Write $\mathbf{C}_{g,k}$ for the component of $\mathrm{Hom}(C_m, C_n)$ corresponding to connected genus-g surfaces (with $k = m + n$ punctures) – this is the most interesting part of $\mathbf{C}_\Sigma$. Recall the enhanced moduli space $\widehat{\mathfrak{M}}_{g,k}$defined in Section 2.1.4; provided only that $k > 0$, $\mathbf{C}_{g,k}$ is a finite-dimensional complex *manifold*, unlike $\widehat{\mathfrak{M}}_{g,k}$, and can be expressed as a bundle over $\widehat{\mathfrak{M}}_{g,k}$ with infinite-dimensional fibre (page 453 of [**502**]). The mapping class group for $\mathbf{C}_{g,k}$ is the $\widehat{\Gamma}_{g,k}$ of Section 2.1.4, that is an extension of $\Gamma_{g,k}$ by k copies of $\mathbb{Z}$.

The most important space is that $\mathbf{C}_{0,2}$ of annuli. We get the easy homeomorphism

$$\mathbf{C}_{0,2} \cong (0, 1) \times (\mathrm{Diff}^+(S^1) \times \mathrm{Diff}^+(S^1))/S^1. \tag{4.4.1}$$

The interval (0,1) arises because any annulus is diffeomorphic to $r \le |z| \le 1$ for some $0 < r < 1$. The two copies of $\mathrm{Diff}^+(S^1)$ correspond to reparametrisations of the two boundary circles – this is where the two copies $L_n, \overline{L}_m$ of $\mathfrak{Vir}$ arise. We factor out by S^1 since rotations are the only holomorphic automorphisms of $r \le |z| \le 1$.

A CFT is (among other things) a projective representation of category $\mathbf{C}$: to each object C_n we assign a vector space $\mathcal{S}(C_n)$, and to each morphism $\Sigma : C_m \to C_n$ a linear map $\mathcal{S} : \mathcal{S}(C_m) \to \mathcal{S}(C_n)$, such that for any objects C_m, C_n, C_p and morphisms $\Sigma' : C_m \to C_n$, $\Sigma : C_n \to C_p$, we obtain the functorial *sewing axiom*

$$\mathcal{S}(\Sigma \circ \Sigma') = c(\Sigma, \Sigma')\,\mathcal{S}(\Sigma) \circ \mathcal{S}(\Sigma') \tag{4.4.2}$$

for some nonzero $c(\Sigma, \Sigma') \in \mathbb{C}$. More precisely, $\mathcal{S}(C_n)$ is the tensor product $\mathcal{H} \otimes \cdots \otimes \mathcal{H} =: \mathcal{H}^{\otimes n}$ of the state-space $\mathcal{H}$ of our CFT, and $\mathcal{H}^{\otimes 0} := \mathbb{C}$. Here, $\mathcal{H}$ is something like the space $L^2(\mathcal{L}M)$ of wave-functions on the loop-space $\mathcal{L}M := \{f : s' \to M\}$, where M is the space-time in which the string lives. Convergence in the Figure 4.16 sewing operation described below requires the operator $\mathcal{S}(\Sigma)$ to be trace class.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

The idea is for $\mathcal{S}(\Sigma)$ to mimic the Feynman path integral (4.2.13a), while avoiding the latter's analytic challenges. In string theory, the incoming state $|\mathrm{in}\rangle$ consists of a choice of string state for each of the m circles, so $|\mathrm{in}\rangle \in \mathcal{H}^{\otimes_m}$; similarly $|\mathrm{out}\rangle \in \mathcal{H}^{\otimes_n}$. Segal's operator $\mathcal{S}(\Sigma)$ is none other than the (finite) scattering matrix, or the time-evolution operator $e^{\mathrm{i}Ht}$ (holomorphically extended to imaginary time): the desired string amplitude is$\langle \mathrm{out}|\mathcal{S}(\Sigma)|\mathrm{in}\rangle$. This is what Segal is trying to capture formally.

If Σ is the disjoint union of surfaces Σ_1 and Σ_2, then $\mathcal{S}(\Sigma) = \mathcal{S}(\Sigma_1) \otimes \mathcal{S}(\Sigma_2)$. That the fundamental identity (4.4.2) should hold can be seen by cutting open a Feynman path integral: an integral over all paths starting from α at time 0 to ω at time 1 can be expressed as the integral over all possible μ of all paths starting from α at $t = 0$ to μ at $t = 0.5$, and all paths from μ at $t = 0.5$ to ω at $t = 1$. This is just matrix multiplication, as (4.4.2) suggests. A physical description of sewing can be found, for instance, in section 9.3 of [**253**]. To construct the projective factor in (4.4.2), Segal uses the 'determinant line bundle' [**192**] (see e.g. [**498**], [**502**] for details). An alternate approach to central charge $c \neq 0$ within the Segal formalism is given in lecture 2 of [**241**].

Another kind of sewing occurs when two oppositely oriented boundary components of Σ are sewn together, increasing the genus by 1, as is illustrated in Figure 4.16. Algebraically, this corresponds to taking a trace or a sum using the Hermitian form. (To see why this is compatible with (4.4.2), interpret matrix multiplication as a trace of the tensor product of the matrices.)

Segal's use of surfaces with boundary differs from that of Section 4.3.2. Usually, quantum field theory restricts to the (easier to calculate) limiting case where the incoming and outgoing states are at $t = \mp\infty$. This is the strategy followed in Section 4.3. Segal is instead trying to capture the string amplitudes for finite times, because it makes the $\mathfrak{Vir}$ action manifest, as we'll see shortly. The relation of Segal's picture with that of enhanced compact surfaces is made in pages 6–7 of [**295**].

The multiplication $\mathbf{C}_{0,2} \times \mathbf{C}_{0,2} \to \mathbf{C}_{0,2}$ makes the annuli space $\mathbf{C}_{0,2}$ into an infinite-dimensional complex Lie semi-group (it has no identity and inverses). Its multiplication is described explicitly in section 9 of [**448**], but to get a taste for it, forget temporarily the parametrisations on the boundary circles: then the sewing of annuli $r < |z| < 1$ and $r' < |z| < 1$ obviously yields the annulus $rr' < |z| < 1$, and so this annulus semi-group is isomorphic to that of the interval (0, 1) under multiplication. Recall from Section 3.1.2 that the complex Lie algebra $\mathfrak{Witt}$ has no Lie group, or equivalently that the real Lie group $\mathrm{Diff}^+(S^1)$ has no complexification. The semi-group $\mathbf{C}_{0,2}$ should be regarded as the complexification of $\mathrm{Diff}^+(S^1)$; it plays the same role for $\mathrm{Diff}^+(S^1)$ that the punctured disc $0 < |z| < 1$ plays for S^1. One hint of this is (4.4.1). Another (proposition 3.1 of [**502**]) is that there is a one-to-one correspondence between positive energy projective representations of $\mathrm{Diff}^+(S^1)$ (recall their definition in Section 3.1.2) and holomorphic projective representations of $\mathbf{C}_{0,2}$. The positive energy representations of $\mathrm{Diff}^+(S^1)$ are the only ones with a hope to extend to $\mathbf{C}_{0,2}$, and all of them are necessarily projective. By a conjecture of Kac, these are all highest-weight modules.

In applications to string theory (namely in the presence of 'ghosts'), the positive-definiteness of the Hermitian product in the Hilbert space $\mathcal{H}$ should be weakened. Also,

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

one may wish to supersymmetrise the state-spaces, that is, give them a $\mathbb{Z}_2$-grading (in order to include fermions). See [**502**] for some comments along these lines.

Note that there is an action of $\mathbf{C}_{0,2}$ on each $\mathbf{C}_\Sigma$ – in fact, one for each boundary circle. This semi-group action amounts to lengthening the arms of each end (equivalently, shrinking the boundary circle); physically, this corresponds to time evolution $t \to \infty$ of outgoing states, or time devolution $t \to -\infty$ of incoming states. We are used to time evolution being a unitary (hence invertible) process, but here time is imaginary, that is, space-time is Euclidean, so time evolution is a contraction. As mentioned in Section 4.2.4, Euclidean space-time is better behaved mathematically than the more physical Minkowski space-time, though in a healthy quantum field theory they should be equivalent.

This semi-group action is the integration of the action of $\mathfrak{Witt}$ on the moduli spaces (Section 3.1.2). By (4.4.2), this action means that each space $\mathcal{S}(\mathbf{C}_\Sigma)$ carries a projective $\mathbf{C}_{0,2}$-representation. In particular, we get an action of $\mathbf{C}_{0,2}$ on the state-space $\mathcal{H}$, projective if $c \neq 0$. This is how we recover the representation of $\mathfrak{Vir} \oplus \mathfrak{Vir}$ on $\mathcal{H}$ that is so important in Section 4.3.2.

The higher-genus behaviour of an RCFT is determined from the lower-genus behaviour, by composition of 'arrows' (i.e. the sewing together of surfaces) in category $\mathbf{C}$, as we see in Figures 4.12 and 4.16. Note that several different sewings can yield the same surface. That they must each give the same answer turns out to be a powerful constraint on CFT, called *duality* (Section 6.1.4).

Thanks to sewing, a CFT is uniquely determined by the chiral algebras $\mathcal{V}, \overline{\mathcal{V}}$; the 1-loop partition function (which gives the spectrum of the theory, i.e. the structure of $\mathcal{H}$ as a $\mathcal{V} \oplus \overline{\mathcal{V}}$-module); and the OPE (4.3.2) (see e.g. section 4 of [**502**]).

The simplest interesting example here is the 'tree-level creation of a string from the vacuum', i.e. $\Sigma : C_0 \to C_1$. In this case the world-sheet looks like a bowl, that is homeomorphic with a disc D, and so is associated with a linear map $\mathcal{S}(D) : \mathbb{C} \to \mathcal{H}$. Equivalently, $\mathcal{S}(D)$ is the assignment of the vector $\mathcal{S}(D)(1)$ in $\mathcal{H}$ to D. In the case of the standard unit disc (i.e. where $D = \{z \in \mathbb{C} \mid |z| \leq 1\}$ and the parametrisation of the boundary S^1 is simply $\theta \mapsto e^{2\pi i\theta}$), this vector is called the *vacuum state* $|0\rangle$. In section 9 of [**502**] it is explained how to recover the stress–energy tensors $T(z), \overline{T}(\overline{z})$, by deforming the complex structure on the disc; this idea is borrowed from CFT.

For another important example, a surface $\Sigma : C_2 \to C_1$, that is a pair-of-pants, corresponds to a bilinear map $\mathcal{H} \otimes \mathcal{H} \to \mathcal{H}$, and makes $\mathcal{H}$ into an algebra. Choosing Σ appropriately, this gives the OPE (4.3.2). A different choice defines the physical vertex operators (this is explicitly given on page 770 of [**241**]).

Finally, suppose the initial and final objects here are both C_0, so the world-sheets Σ are closed Riemann surfaces. Segal's functor $\mathcal{S}(\Sigma)$ is a linear map $\mathbb{C} \to \mathbb{C}$, so is completely determined by its value at $1 \in \mathbb{C}$. This value $\mathcal{S}(\Sigma)(1) =: \mathcal{Z}(\Sigma) \in \mathbb{C}$ is the partition function. Consider now Σ a torus. Up to conformal equivalence, Σ can be written as the quotient $\Sigma_\tau := \mathbb{C}/(\mathbb{Z} + \mathbb{Z}\tau)$, and so the 1-loop partition function $\mathcal{Z}(\Sigma_\tau)$ becomes a function on $\mathbb{H}$. As we know, Σ_τ and $\Sigma_{\alpha.\tau}$ are conformally equivalent when $\alpha \in \mathrm{SL}_2(\mathbb{Z})$, and so $\mathcal{Z}$ must be modular invariant. We can construct a torus by sewing together the two

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

ends of a cylinder, or equivalently an annulus $A_q = \{z \in \mathbb{C} \mid |q| \leq |z| \leq 1\}$ for $q \in \mathbb{C}$ where the boundaries are parametrised by $qe^{2\pi \mathrm{i}\theta}$ and $e^{2\pi \mathrm{i}\theta}$. We know that this recovers Σ_τ up to conformal equivalence, if $q = e^{2\pi \mathrm{i}\tau}$. Then $\mathcal{S}(A_q) = q^{L_0}\overline{q}^{\overline{L_0}}$ and so by the sewing axiom (with $c = 0$ for convenience) the torus partition function becomes

$$\mathcal{Z}(\tau) = \mathrm{tr}_{\mathcal{H}} q^{L_0}\overline{q}^{\overline{L_0}}.$$

It must be invariant under the usual action of $\mathrm{SL}_2(\mathbb{Z})$. Of course, if the central charge is nonzero, then the sewing axiom picks up a multiplicative factor that recovers (4.3.8b). See page 768 of [**241**] for details.

So far Segal is addressing general CFT. He defines an RCFT – our main interest – as a *modular functor* $\mathfrak{B}$. It assigns to each surface Σ its space of chiral blocks (4.3.7). Let Φ be a finite set of labels – this parametrises the irreducible modules of chiral algebra $\mathcal{V}$. One of these labels, call it 0, is distinguished (it corresponds to the vacuum, and was called V in Section 4.3.2). We require that Φ has an involution $i \mapsto i^*$, called *charge conjugation* and related to complex conjugation. By a labelled Riemann surface with boundary (Σ, α) we mean to assign a label $\alpha_i \in \Phi$ to each (parametrised) boundary circle of Σ. These are the objects in a category $\mathbf{Riem}_\Phi$. The morphisms are 'holomorphic collapsing maps' (see section 5 of [**502**]), which sew together pairs of boundary circles in the usual way. The target is the category $\mathbf{Vect}_f$ of finite-dimensional vector spaces, since the spaces of chiral blocks live there; morphisms are linear transformations.

Definition 4.4.1 [**502**] A modular functor *is a functor* $\mathfrak{B}$ *from* $\mathbf{Riem}_\Phi$ *to* $\mathbf{Vect}_f$*, such that:*

(i) $\mathfrak{B}$ *takes the disjoint union* $\Sigma \cup \Sigma'$ *to* $\mathfrak{B}(\Sigma) \otimes \mathfrak{B}(\Sigma')$.

(ii) $\mathfrak{B}(\Sigma) = \mathfrak{B}(-\Sigma)$*, where '*$-\Sigma$*' means that we reverse the orientation of all boundary circles of* Σ *(i.e. interchange incoming with outgoing circles), and also replace each label* α_i *with its conjugate* α_i^*.

(iii) *Suppose surface* Σ *is obtained from surface* Σ' *by cutting along a closed curve. For each label* $i \in \Phi$*, let* Σ_i *be the surface* Σ *labelled the same as* Σ'*, except its two additional circles are both given the label* i*. Then*

$$\oplus_{i \in \Phi} \mathfrak{B}(\Sigma_i) \cong \mathfrak{B}(\Sigma'). \tag{4.4.3}$$

(iv) *If* D *is the standard disc then* $\mathfrak{B}(D)$ *is* $\mathbb{C}$ *if the boundary is labelled 0, and* $\{0\}$ *otherwise.*

(v) *Finally, if* Σ_w *is a family of surfaces varying holomorphically with a parameter* w*, then the spaces* $\mathfrak{B}(\Sigma_w)$ *fit together to form a holomorphic vector bundle.*

We won't spell out precisely what condition (v) means (roughly, it says that the chiral blocks are holomorphic functions on the moduli space), but certainly it implies that the dimension of $\mathfrak{B}(\Sigma)$ only depends on the orientations of the boundary circles and the labels, and not on the complex structure of Σ. We discuss chiral blocks in Section 4.3.3. Their most important property is that they carry a projective representation of the mapping class group of Σ. The definition of modular functor using closed surfaces with marked points, as well as an alternate approach to $c \neq 0$, is given in chapter 5 of [**32**].

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

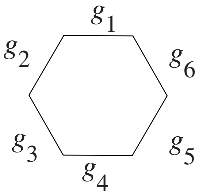

Fig. 4.17 A natural depiction of an identity $g_1g_2g_3g_4g_5g_6 = e$.

There are still no known examples of modular functors, though it is expected that any sufficiently nice vertex operator algebra will yield one. Nevertheless, this picture of RCFT is incomplete, as it only captures some elements of the chiral halves of an RCFT. For instance, the modular functor corresponding to Monstrous Moonshine is trivial. The 1-loop partition function (4.3.8b) is important data for the RCFT, but its presence here is obscure (to this author at least), as more generally is the explicit relation between the full CFT and the two chiral halves.

4.4.2 Groups are decorated surfaces

This short subsection motivates topological field theory and can be skipped on first reading.

Fix a group G. We can think of G as a set of identities $g_1g_2\cdots g_k = e$. Conjugating by g_1, we observe

$$g_1g_2\cdots g_k = e \text{ iff } g_2g_3\cdots g_kg_1 = e. \tag{4.4.4}$$

Thus, an identity '$g_1\cdots g_k = e$' in G really should be written circularly, as in Figure 4.17. In other words, we can think of G as a way to assign to each polygon, whose sides are labelled consecutively by elements g_i of G, a number $\mathcal{P}(g_1, g_2, \ldots, g_k) \in \{0, 1\}$. We assign '1' to a given labelled polygon if, starting anywhere on the circumference and reading counterclockwise, the product of the labels equals e; otherwise assign '0' to it. We get a dihedral symmetry,

$$\mathcal{P}(g_1, g_2, \ldots, g_k) = \mathcal{P}(g_2, \ldots, g_k, g_1), \tag{4.4.5a}$$

$$\mathcal{P}(g_1, g_2, \ldots, g_k) = \mathcal{P}\big(g_k^{-1}, \ldots, g_2^{-1}, g_1^{-1}\big), \tag{4.4.5b}$$

corresponding to the symmetries of the k-gon.

Of course not every assignment of 0's and 1's to labelled polygons will come from groups. Most importantly, we have

$$\mathcal{P}(g_1, \ldots, g_m, h_1, \ldots, h_n) = \sum_{g\in G} \mathcal{P}(g_1, \ldots, g_m, g)\,\mathcal{P}(g^{-1}, h_1, \ldots, h_n). \tag{4.4.5c}$$

This can be depicted pictorially as the dissection rule of Figure 4.18. We also get the normalisation rule

$$\sum_{g\in G} \mathcal{P}(g_1, \ldots, g_k, g) = 1. \tag{4.4.5d}$$

This polygonal definition is completely equivalent to the usual one of a group:

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

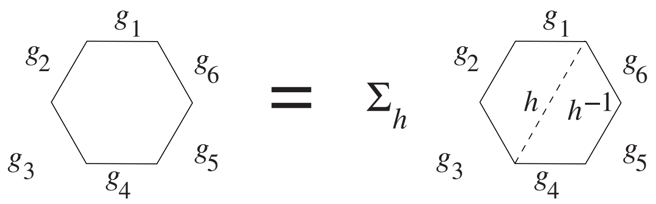

Fig. 4.18 The dissection rule: $g_1g_2g_3g_4g_5g_6 = e$ iff $g_1g_2g_3 = (g_4g_5g_6)^{-1}$.

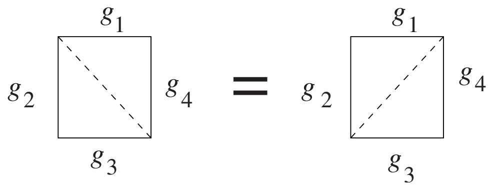

Fig. 4.19 Associativity in a group.

Proposition 4.4.2 *Let S be a set and let $\Pi(S)$ be the set of all polygons labelled with elements of S. Suppose $\mathcal{P} : \Pi(S) \to \{0, 1\}$ obeys all equations (4.4.5), where for $g \in S$,'g^{-1}' denotes the unique element of S satisfying $\mathcal{P}(g, g^{-1}) = 1$. Define $e \in S$ by $\mathcal{P}(e) = 1$ and the multiplication 'gh' by $\mathcal{P}(g, h, (gh)^{-1}) = 1$. Then this defines a group structure on S compatible with the values $\mathcal{P}$ of the polygons in $\Pi(S)$.*

Thus, knowing the values of 1-gons, 2-gons and triangles fixes all other values. Associativity is equivalent to Figure 4.19, and all other generalised associativity relations can be derived from it. The entire group structure is encoded in a few polygons – the rest are redundant – and indeed that is how a group is usually defined. But there is an aesthetic appeal to considering this global (albeit highly redundant) structure provided by all identities in G, and this charm is lost if we focus only on the banal building blocks. It is reminiscent of interpreting the presentation (1.1.9) as a group of braids.

Nevertheless, this rephrasing of the definition of a group is unsatisfactory for several reasons. It seems artificial that the values $\mathcal{P}$ are always either 0's or 1's. Why should we limit the right side of (4.4.4) to being e – for example, any central element will work equally well. Can we consistently sew together two sides of the same polygon, and get more interesting topologies? What does the normalisation condition really mean group-theoretically? These thoughts lead to the following construction.

Fix a group G and irreducible character ch (Section 1.1.3). A polygon whose sides are labelled with elements g_i of G is assigned the complex number $\mathcal{P}(g_1, \ldots, g_k) = \frac{\mathrm{ch}(e)}{\|G\|}\mathrm{ch}(g_1 \cdots g_k)$ (recall that ch(e) is the dimension of ch). Equation (4.4.5a) continues to hold, while (4.4.5b) becomes $\overline{\mathcal{P}(g_1, \ldots, g_k)} = \mathcal{P}(g_k^{-1}, \ldots, g_1^{-1})$. Equation (4.4.5c) follows from the generalised orthogonality relation (theorem 2.13 in [**308**])

$$\frac{1}{\|G\|}\sum_{g\in G}\mathrm{ch}_i(gh)\,\mathrm{ch}_j(g^{-1}) = \delta_{ij}\frac{\mathrm{ch}_i(h)}{\mathrm{ch}_i(e)},$$

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

valid for irreducible $\mathrm{ch}_i, \mathrm{ch}_j$. The ‘normalisation condition’ (4.4.5d) should be replaced by

$$\sum_{g \in G} |\mathcal{P}(g_1, \ldots, g_k, g)|^2 = \frac{\mathrm{ch}(e)^2}{\|G\|},$$

where $\mathrm{ch} = \sum m_i \mathrm{ch}_i$ expresses ch as a sum of irreducible characters. We see that, as before, two consecutive arcs, labelled g, h, can always be replaced by a single arc labelled gh; so a polygon can always be replaced with a disc. Moreover, the label on a disc depends only on the conjugacy class.

More generally, we can use any character of the form $\mathrm{ch} = \sum \mathrm{ch}_i(e)\, \mathrm{ch}_i$, where we sum over any subset of the irreducible characters; then $\mathcal{P} = \mathrm{ch}/\|G\|$ works. For instance, the original assignment (with values in $\{0, 1\}$) corresponds to the character ch of the regular representation of G. The normalisation condition (4.4.5d) is thus seen to be a consequence of orthogonality of characters.

There is no need to stop here. The dissection rule applied to an annulus labelled with conjugacy classes K_g, K_h (h inner, g outer) implies it is assigned $\mathrm{ch}(g)\,\overline{\mathrm{ch}(h)}$; more generally, a disc with n smaller discs removed will have value $\mathrm{ch}(g)\overline{\mathrm{ch}(h_1)} \cdots \overline{\mathrm{ch}(h_n)}$. In these more general settings, the orientation of the boundary circle should be made explicit (here they’re all taken to be counter-clockwise). A torus with a disc removed, and the boundary circle labelled K_g, has value $\frac{\|G\|}{\mathrm{ch}(e)}\mathrm{ch}(g)$.

Likewise, any surface with (oriented) punctures labelled by conjugacy classes can be assigned a well-defined complex number. This is, in fact, a slightly enhanced topological field theory (Question 4.4.4).

4.4.3 Topological field theory

The essence of mathematics involves seeing that two different-looking things are actually (from the appropriate perspective) the same. What are different ways of going from point a to point b? In algebra these are functions, the simplest being linear; in geometry, these are cobordisms; in physics, this is time evolution. A topological field theory is their identification.

This subsection strays a little from the main thread of this book, and so we will only sketch the basic idea. The following definition, that topological field theory is a monoidal functor from the cobordism category to $\mathbf{Vect}_f$, is due to Atiyah and was heavily influenced by Segal’s definition of CFT (Section 4.4.1). Topological field theory is a beautiful language that has elegantly formulated several deep mathematical ideas (e.g. Morse theory, the Jones polynomial, Donaldson invariants) – see the reviews [**25**], [**564**], [**62**], [**534**], [**32**]. The first topological field theories were constructed in physics by Schwarz (1978) and Witten (1982) (see [**62**] for references). Physically, a topological field theory should arise from the large-distance limit of any quantum field theory with mass gap.

Definition 4.4.3 [25] *A* topological field theory *in $d + 1$ dimensions assigns to each compact oriented smooth d-dimensional manifold Σ a finite-dimensional complex vector*

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

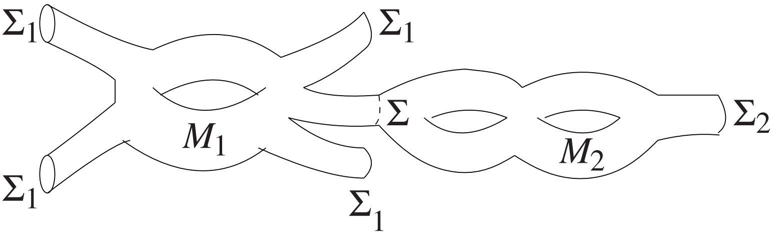

Fig. 4.20 Sewing.

space $\mathcal{T}(\Sigma)$, and to each compact oriented $(d+1)$-dimensional manifold M with boundary Σ, a vector $\mathcal{T}(M) \in \mathcal{T}(\Sigma)$, such that:

(i) *$\mathcal{T}(\Sigma^*) = \mathcal{T}(\Sigma)^*$, where Σ^* denotes Σ with opposite orientation, and $\mathcal{T}(\Sigma)^*$ is the dual space.*

(ii) *$\mathcal{T}$ takes the disjoint union $\Sigma_1 \cup \Sigma_2$ to $\mathcal{T}(\Sigma_1) \otimes \mathcal{T}(\Sigma_2)$.*

(iii) *If $\partial M_i = \Sigma \cup \Sigma_i$ (disjoint union) and M is obtained from M_1 and M_2 by sewing along a common boundary component Σ, as in Figure 4.20, then $\mathcal{T}(M) = \mathcal{T}(M_2) \circ \mathcal{T}(M_1)$.*

(iv) *$\mathcal{T}$ takes the empty d-manifold $\emptyset$ to $\mathbb{C}$.*

(v) *$\mathcal{T}(\Sigma \times I)$ is the identity endomorphism of $\mathcal{T}(\Sigma)$, where I is the unit interval.*

(vi) *If $f : \Sigma \to \Sigma'$ is a homeomorphism, then there is a vector space isomorphism $\mathcal{T}_f : \mathcal{T}(\Sigma) \to \mathcal{T}(\Sigma')$; if $F : M \to M'$ is a homeomorphism, then*

$$\mathcal{T}_{F|_{\partial M}}(\mathcal{T}(M)) = \mathcal{T}(M').$$

Some technicalities are implicit here; see section 4.2 of [**32**] for any needed clarifications. The book [**534**] is also helpful. If the boundary of M is Σ, and we write Σ as the disjoint union $\Sigma_1 \cup \Sigma_2$, then $\mathcal{T}(\Sigma) = \mathcal{T}(\Sigma_1) \otimes \mathcal{T}(\Sigma_2^*)^*$ and thus the 'vector' $\mathcal{T}(M)$ can be regarded as a linear map $\mathcal{T}(\Sigma_2^*) \to \mathcal{T}(\Sigma_1)$. This functional interpretation is implied in (iii) and (v).

M plays the role here of space-time, and Σ that of space (i.e. a space-like time-slice of M). $\mathcal{T}(\Sigma)$ is the space of all states at the given instant, while the map $\mathcal{Z}(M)$ is the time-evolution operator $e^{\mathrm{i}Ht}$. Condition (iv) can be interpreted as saying that the Hamiltonian H is 0, so the only evolution is topological.

Question 4.4.6 asks for a proof of the homotopy invariance of $\mathcal{T}$. This means that the mapping class group of Σ, that is the group of components of the group $\mathrm{Diff}^+(\Sigma)$ of orientation-preserving diffeomorphisms, acts on the space $\mathcal{T}(\Sigma)$. This is obviously important to us.

Condition (iv) is needed to eliminate the trivial theory. If M is a closed manifold (i.e. it has no boundary), then $\mathcal{T}(M) \in \mathcal{T}(\emptyset) = \mathbb{C}$. Thus a topological field theory assigns a numerical invariant to closed $(d+1)$-dimensional manifolds.

Let Σ be any d-manifold and put $M_1 = \Sigma \times I$, $M_2 = \Sigma^* \times I$. Sewing these together along corresponding copies of Σ, we get $M = \Sigma \times S^1$. From (v) we get that $\mathcal{T}(M_i)$ are the identity maps $\mathcal{T}(\Sigma) \to \mathcal{T}(\Sigma)$ and $\mathcal{T}(\Sigma)^* \to \mathcal{T}(\Sigma)^*$, respectively. But we can also

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

think of them as vectors in $\mathcal{T}(\Sigma) \otimes \mathcal{T}(\Sigma)^*$ and $\mathcal{T}(\Sigma)^* \otimes \mathcal{T}(\Sigma)$, so these vectors must be $\sum_i e_i \otimes e_i^*$ and $\sum_j e_j^* \otimes e_j$, respectively, where e_i is any basis of $\mathcal{T}(\Sigma)$ and e_i^* is the dual basis. Thus

$$\mathcal{T}(\Sigma \times S^1) = \left\langle \sum_i e_i \otimes e_i^*, \sum_j e_j^* \otimes e_j \right\rangle = \dim(\mathcal{T}(\Sigma)). \tag{4.4.6a}$$

Now, we know that 'dimension' can be twisted into 'character' whenever a group is present. So let γ lie in the mapping class group and define $\Sigma \times_\gamma S^1$ to be the $(d+1)$-dimensional manifold obtained by sewing $\Sigma \times I$ to $\Sigma^* \times I$ by identifying the boundary $\Sigma^* \times 0$ with $\Sigma \times 0$ and $\gamma(\Sigma) \times 1$ with $\Sigma \times 1$. Repeating the earlier calculation yields

$$\mathcal{T}(\Sigma \times_\gamma S^1) = \left\langle \sum_i \mathcal{T}_\gamma(e_i) \otimes e_i^*, \sum_j e_j^* \otimes e_j \right\rangle = \mathrm{tr}(\mathcal{T}_\gamma). \tag{4.4.6b}$$

Theorem 4.4.4 *A topological field theory in* $1+1$ *dimensions is equivalent to a finite-dimensional commutative associative algebra* A *over* $\mathbb{C}$ *with unit* 1*, together with a linear map* $\mathrm{tr}\colon A \to \mathbb{C}$ *such that the bilinear form* $(a, b) \mapsto \mathrm{tr}(ab)$ *is nondegenerate.*

Nondegenerate here means that the only $a \in A$ with $\mathrm{tr}(ab) = 0$, $\forall b \in A$, is $a = 0$. Such an algebra A is called a *Frobenius algebra* – see, for example, chapter 2 of [**353**]. Frobenius algebras were introduced by Frobenius in 1903. The association of a Frobenius algebra to a (1+1)-dimensional topological field theory is straightforward. The vector space A is given by $\mathcal{T}(S^1)$. The boundary of the disc D can be thought of as $\partial D = S^1$ or $\partial D = \emptyset \cup (S^1)^*$, the former interpretation defines a vector $1 := \mathcal{T}(D) \in A$, while the latter defines the map $\mathrm{tr} := \mathcal{T}(D) : A \to \mathbb{C}$. The product structure on A comes from $\mathcal{T}$ applied to a pair-of-pants, with boundary $S^1 \cup (S^1 \cup S^1)^*$. The various properties obeyed by multiplication, 1 and tr follow inductively from the various pictures – it's a good idea for the reader to work these out. The proof that a Frobenius algebra defines a unique and well-defined topological field theory is based on the fact that any surface can be obtained by sewing together discs, cylinders and pairs-of-pants; the only difficulty is verifying well-definedness: as we know, the same surface can be decomposed this way in many different ways. The details of this proof are given in section 4.3 of [**32**]; see also section 3.3 of [**353**] for a more pedagogical treatment. This proof is practise for Section 6.1.4, where we do the same for RCFT.

Our symbol 'Σ' in Definition 4.4.3 is due to the special importance of $d = 2$. In his analysis of the Jones polynomial, Witten discovered the explicit relation between topological field theory in $2+1$ dimensions and CFT (in the usual two dimensions): the spaces $\mathcal{T}(\Sigma)$ are the spaces $\mathfrak{B}(\Sigma)$ of chiral blocks. The relation between CFT and $(2+1)$-dimensional topological field theory is carefully explained in chapter 5 of [**32**]. In particular, the association of a modular functor with a topological field theory is easy, but (according to [**32**]) the association of a topological field theory with each modular functor is only conjectural at present. $(2+1)$-dimensional topological field theory has been used recently in a series of papers (see [**211**] for a review) for constructing (boundary) RCFT correlation functions.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

4.4.4 From amplitudes to algebra

The final rigorous approach to CFT we sketch reconstructs the chiral theory directly from the vacuum-to-vacuum amplitudes. The physical appeal of this approach is that it starts with ‘observational’ data. For us, it’s excellent motivation for the material of the next chapter. We focus on the chiral halves of RCFT – the parts of CFT of greatest interest to mathematics.

A chiral half of a CFT on a sphere consists of a state-space $\mathcal{H}$ and a collection

$$\langle Y(\psi_1, a_1)\, Y(\psi_2, z_2) \cdots Y(\psi_n, z_n)\rangle \tag{4.4.7}$$

of correlation functions, where z_i lie in the Riemann sphere $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$. To avoid circularity, restrict (4.4.7) to states ψ_i in some (typically finite-dimensional) subspace $\mathcal{H}_{gen}$ that generates $\mathcal{H}$. For now, all we need to know about these correlation functions (4.4.7) is that they are multi-linear in the states ψ_i, symmetric under permutation of ψ_i and analytic in the points z_i, except possibly for poles when $z_i = z_j$. At this point the notation in (4.4.7) is purely formal, so for example ‘$Y(\psi_i, z_i)$’ has no meaning. Our first task is to associate with the states $\psi \in \mathcal{H}_{gen}$, vertex operators $Y(\psi, z)$.

Let O be any open set in $\mathbb{P}^1(\mathbb{C})$, with the property that its complement is path-connected and contains a disc. A counterintuitive result of axiomatic quantum field theory (the Reeh–Schlieder Theorem [**518**], [**269**]) says that the states $\sum \varphi_1(f_1) \cdots \varphi_n(f_n)\,|0\rangle$ generated from the vacuum $|0\rangle$ by fields φ_i smeared by test-functions f_i localised to O, will be dense in $\mathcal{H}$. This observation motivates the following construction.

Define the space $\mathcal{V}_O$ formally spanned by all words $Y(\psi_1, z_1) \cdots Y(\psi_n, z_n)\,|0\rangle$, where $\psi \in \mathcal{H}_{gen}$ and $z_i \in O$, z_i pairwise distinct, and we require any word to be bilinear and symmetric in the ψ_i. We want to complete these infinite-dimensional spaces (i.e. include the limits of certain sequences), topologise them (i.e. decide when vectors are ‘close’) and identify vectors that are physically indistinguishable (i.e. quotient by null-vectors). We can do all three, using the amplitudes (4.4.7) to define a bilinear pairing $\mathcal{V}_O \times \mathcal{V}_{O'} \to \mathbb{C}$, for any open set O' in the complement of O:

$$\begin{aligned}&\left(\sum_i Y\left(\psi_1^{(i)}, z_1^{(i)}\right) \cdots Y\left(\psi_{m^{(i)}}^{(i)}, z_{m^{(i)}}^{(i)}\right)|0\rangle, \sum_j Y\left(\phi_1^{(j)}, w_1^{(j)}\right) \cdots Y\left(\phi_{n^{(j)}}^{(j)}, w_{n^{(j)}}^{(j)}\right)|0\rangle\right)\\ &\quad\mapsto \sum_{i,j} \left\langle Y\left(\psi_1^{(i)}, z_1^{(i)}\right) \cdots Y\left(\phi_{n^{(j)}}^{(j)}, w_{n^{(j)}}^{(j)}\right)\right\rangle\end{aligned} \tag{4.4.8}$$

for all $\psi_k^{(i)}, \phi_\ell^{(j)} \in \mathcal{H}_{gen}$, $z_k^{(i)} \in O$, $w_\ell^{(j)} \in O'$. A topology on say $\mathcal{V}_{O'}$ is obtained by defining this pairing to be continuous. We identify vectors in $\mathcal{V}_{O'}$ by quotienting by those vectors in $\mathcal{V}_{O'}$ that are orthogonal to all of $\mathcal{V}_O$. This pairing (4.4.8) also allows us to complete the space $\mathcal{V}_{O'}$. The resulting space turns out to be independent of O' – call it $\mathcal{V}^O$. See [**227**] for details.

If $O_1 \subset O_2$, then we get a natural continuous embedding of $\mathcal{V}^{O_2}$ into $\mathcal{V}^{O_1}$. The role here of the space $\mathcal{H}$ of states is played by this collection $\mathcal{V}$ of topological vector spaces, just as the role of the algebra $\mathcal{A}$ of observables in quantum field theory is played in *algebraic* quantum field theory by the net $\mathcal{A}(\mathcal{O})$ (Section 4.2.4). However, if $O \subset \mathbb{P}^1(\mathbb{C})$ contains ∞ but not 0, then we can define the modes $\psi_{(n)}$, for $\psi \in \mathcal{H}_{gen}$, in the usual

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

way and from this get a Fock space $\mathcal{H}^O \subset \mathcal{V}^O$ spanned by all $(\psi_1)_{(n_1)} \cdots (\psi_k)_{(n_k)}|0\rangle$. It is easy to see that it is dense in $\mathcal{V}^O$ and independent of the choice of O. This Fock space will be the VOA (Definition 5.1.3) of the CFT.

It is now easy to define the vertex operators $Y(\psi, z)$. Choose any $\psi \in \mathcal{H}_{gen}$ and $z \in O$, and any subset $O' \subset O$ with $z \notin O'$. Then the operator $Y(\psi, z) : \mathcal{V}^O \to \mathcal{V}^{O'}$ is defined by

$$\sum_i Y\left(\psi_1^{(i)}, z_1^{(i)}\right) \cdots Y\left(\psi_{m^{(i)}}^{(i)}, z_{m^{(i)}}^{(i)}\right)|0\rangle$$
$$\mapsto \sum_i Y(\psi, z)\, Y\left(\psi_1^{(i)}, z_1^{(i)}\right) \cdots Y\left(\psi_{m^{(i)}}^{(i)}, z_{m^{(i)}}^{(i)}\right)|0\rangle$$

(there is a little work to see that this operator lifts from $\mathcal{V}_C$ to $\mathcal{V}^O$ – again see [**227**]). Note that we automatically obtain commutativity: the identity

$$Y(\psi, z)\,Y(\phi, w) = Y(\phi, w)\,Y(\psi, z)$$

holds in $\mathcal{V}^O$ provided $z, w \in O, z \neq w, \psi, \phi \in \mathcal{H}_{gen}$ (compare VA4 in Definition 5.1.3).

So far we have assumed only the most basic properties of the amplitudes (4.4.7). The full splendour of CFT begins to reveal itself once we impose Möbius invariance, which says that it shouldn't matter how we identify the sphere $\mathbb{P}^1(\mathbb{C})$ with the complex coordinates $\mathbb{C} \cup \{\infty\}$. This invariance implies the usual Möbius covariance of the amplitudes and vertex operators. It allows us to extend the definition of vertex operators to, for example, $\mathcal{V}^O$, and to establish Jacobi's identity (5.1.7a). Although this is where things start getting interesting, this is where we leave off.

We know the state-space $\mathcal{H}$ of the CFT is a module for the chiral algebra. This is recovered in this formalism through the two-point functions, which are of the form

$$\left\langle Y''(\varphi_2, w_2)\,Y(\psi_1, z_1) \cdots Y(\psi_m, z_m)\,Y'(\varphi_1, w_1)\right\rangle, \tag{4.4.9}$$

where the states ψ_i lie in $\mathcal{H}_{gen}$ as before, and φ_j lie in spaces $\mathcal{W}_j$ (which we can take to be dual to each other, although this isn't necessary). We can construct spaces $\mathcal{W}^O$ much as before, generated by

$$\sum_i Y(\psi_1, z_1) \cdots Y(\psi_m, z_m)\,Y'(\varphi_1, w_1)\,|0\rangle,$$

and interpret the symbol $Y'(\varphi_1, w_1)$ as a vertex operator sending $\mathcal{W}^O \to \mathcal{W}^{O'}$, much as before. This leads quite naturally to the notion of a VOA-module (Definition 5.3.1).

An observation that will be helpful in Section 5.3.2 in motivating Zhu's algebra is that each representation corresponds to a linear functional on the chiral algebra:

Proposition 4.4.5 [429], [227] *The amplitudes (4.4.9) define a representation of the chiral algebra $\mathcal{V}$, provided that for each open O with path-connected complement, and each states $\varphi_j \in \mathcal{W}_j$ and points $w_i \notin O$, there is a state $v = v(\varphi_1, \varphi_2, w_1, w_2) \in \mathcal{V}^O$ satisfying*

$$\langle Y''(\varphi_2, w_2)\,Y(\psi_1, z_1) \cdots Y(\psi_m, z_m)\,Y'(\varphi_1, w_1)\rangle = \langle Y(\psi_1, z_1) \cdots Y(\psi_m, z_m)\,v\rangle$$

for all choices of $z_i \in O$, $\psi_i \in \mathcal{H}_{gen}$.

https://doi.org/10.1017/9781009401548.005 Published online by Cambridge University Press

The proof of the proposition isn't difficult (see theorem 6 in [**227**]). This proposition permits us to characterise the representations of a chiral algebra by states v. It turns out that these v, which can be interpreted as linear functionals on the Fock space $\mathcal{H}^{O'}$ using the pairing (4.4.8), vanish on a certain large subspace $0^{O'}_{w_1,w_2}$ of $\mathcal{H}^{O'}$, and so define linear functionals on the quotient $\mathcal{H}^{O'}/O^{O'}_{w_1,w_2}$. In the case of a *rational* CFT, this quotient space will be finite-dimensional and is called *Zhu's algebra* (Section 5.3.2).

Question 4.4.1. What is the value $\mathfrak{B}(S^2)$ that Segal's functor associates with the sphere?

Question 4.4.2. Suppose labelled surfaces Σ and Σ' are sewed end-to-end (so the corresponding labels match, and the corresponding circle orientations are opposite), to produce a new labelled surface Σ''. Construct a canonical map $\mathfrak{B}(\Sigma) \otimes \mathfrak{B}(\Sigma') \rightarrow \mathfrak{B}(\Sigma'')$. If $\mathfrak{B}(\Sigma)$, $\mathfrak{B}(\Sigma')$, $\mathfrak{B}(\Sigma'')$ are all nonzero, can that map be identically 0?

Question 4.4.3. (a) Let A be any annulus with oppositely oriented boundary circles. Prove $\mathfrak{B}(A) = \{0\}$, unless both circles are given the same label $i \in \Phi$, in which case $\mathfrak{B}(A) = \mathbb{C}$.
(b) If T is any torus, prove that $\mathfrak{B}(T)$ has dimension equal to the cardinality of Φ.

Question 4.4.4. Find a relation between the assignments $\mathcal{P}^{\chi}$ to surfaces with punctures labelled with conjugacy classes of G, and two-dimensional topological field theory.

Question 4.4.5. (a) If M is the disjoint union of M_1 and M_2, what is $\mathcal{T}(M)$ in terms of $\mathcal{T}(M_i)$?
(b) What does $\mathcal{T}$ send the empty $(d+1)$-manifold $\emptyset$ to?

Question 4.4.6. Prove that if $f : \Sigma \rightarrow \Sigma'$ is a homeomorphism homotopic to the identity, then the linear map $\mathcal{T}_f$ of (vi) is the identity.

Question 4.4.7. Classify all topological field theories of dimension $d = 0$.