2

Basic Concepts of Topology and Condensed Matter

> In these days the angel of topology and the devil of abstract algebra fight for the soul of each individual mathematical domain.
>
> Hermann Weyl (1939)

The goal of this chapter is to introduce some of the key concepts and tools that are used repeatedly in the analysis of topological phases of matter. We start with two basic results from nonrelativistic quantum mechanics. The Berry phase is a geometric phase in quantum mechanics that can be introduced simply as resulting from adiabatic evolution in time. An example of its physical meaning is that, when the Hamiltonian of a system is brought adiabatically around a closed path in parameter space, a system initially in a nondegenerate energy eigenstate accumulates a physically meaningful phase relative to its initial wavefunctions, even when there are no transitions to other states. The same mathematical structure appears whenever wavefunctions evolve smoothly as a function of some other parameter, such as momentum or position. Clearly this structure is related to gauge fields and other physical examples of holonomy. As an example of how the Berry phase appears outside its original context of adiabatic evolution, we introduce the coherent-state representation of a quantum spin.

The second basic quantum result introduced is Bloch's theorem, which is the statement that wavefunctions of a single particle in a periodic potential take a particularly simple form: they are products of a plane wave and a part with the same periodicity as the potential. Before considering a periodic potential, we warm up with the specific problem of single-particle energy levels (Landau levels) of a particle in the continuum in a constant magnetic field, as these states wind up being widely used in quantum Hall physics. Bloch's theorem yields two important differences between electrons in a solid and in vacuum. In a solid, the wavefunction has additional material-dependent structure beyond that of a plane wave, and momentum is also modified as only crystal momentum (momentum modulo a reciprocal

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

lattice vector) is defined. In more mathematical language, Bloch's theorem naturally gives to wavefunctions of a particle in a periodic potential (e.g., a crystal) a fiber bundle structure similar to that of a gauge theory: a band structure associates every value of the crystal momentum with a wavefunction with the periodicity of the crystal. We discuss how this works in the tight-binding models beloved by physicists working on topological phases.

We then review the basic picture of "conventional" order in solids, via the Landau theory of symmetry-breaking phases. The concept of an order parameter remains essential for many of the contexts where topology arises in condensed matter physics, particularly for the discussion of topological defects that follows. The next section uses simple examples to introduce two flavors of topology: the cohomology of differential forms and the basics of homotopy. This section is primarily for theoretically inclined readers who would like to understand the mathematical description of topological phases. As there are excellent textbooks specifically devoted to explaining these mathematical methods for physicists (Nakahara, 1998; Stone and Goldbart, 2009), our emphasis is on conveying the minimum necessary to profit from discussions later in the book; we have often preferred the specific example to the general classification and have sought accessibility at some cost in rigor.

Topological defects in ordered phases breaking continuous symmetries were the first nontrivial application of topology to condensed matter, and we discuss vortices in a superfluid or XY model in some detail. Even for readers who skip the introduction to topological methods, the idea of homotopy groups is quite useful for complicated topological defects and also for defect-free topological configurations like skyrmions. The Berezinskii–Kosterlitz–Thouless transition from vortex unbinding in superfluid films is discussed at a nontechnical level (i.e., without renormalization group ideas) in a sidebar, as an example of the emergent physics of many vortices.

2.1 Berry Phases in Quantum Mechanics

We start with a beautiful geometric property of quantum mechanics whose full significance was understood only in the early 1980s: the geometric or Berry phase. In the course of this book we will frequently be interested in how the eigenstates of a Hamiltonian vary as a function of some parameters. The Berry phase is a subtle and physically important consequence of such variation. One way to introduce the geometric phase is by considering adiabatic changes in time of the Hamiltonian's parameters. We will primarily be interested in this book in cases where the same quantity appears even though the change in Hamiltonian parameters is not a function of time. The first examples of this type of geometric phase in physics

were found more than fifty years ago in optics (Pancharatnam, 1956) and chemical dynamics (Longuet-Higgins et al., 1958), and the classic paper of Berry (1984) was the first to identify the concept in its full generality.

An important result from undergraduate quantum mechanics is the adiabatic approximation. Suppose that a system is prepared in a nondegenerate eigenstate of a time-dependent Hamiltonian H. For later reference, we will write H as a function of some parameters λ_i that depend on time: $H(t) = H(\lambda_1(t), \lambda_2(t), \ldots) = H(\boldsymbol{\lambda})$. If the eigenstate remains nondegenerate, then the adiabatic theorem states that if the Hamiltonian changes slowly in time (how slowly depends primarily on the energy gap between adjacent eigenstates), then there are no transitions between eigenstates.

This approximation, when correct, actually only gives part of the story: it describes the probability to remain in the eigenstate that evolved from the initial eigenstate, but there is actually nontrivial information in the *phase* of the final state as well. This result may seem quite surprising because the overall phase in quantum mechanics is in general independent of observable quantities. However, the Berry phase from an adiabatic evolution is observable: more precisely, phase differences are observable, such as the phase difference between one system taken around a closed path in parameter space and another system initially identical to the first whose parameters remain unchanged. Conceptually and mathematically, the Berry phase will turn out to be closely related to the notion of how the electromagnetic vector potential in quantum mechanics gives a relative phase between wavefunctions at different points, with physical consequences such as the Aharonov-Bohm effect.

Berry's result for a closed path is deceptively simple to state. In moving a system adiabatically around a closed path in parameter space, the final wavefunction is in the same eigenstate as the initial one (again, under the assumptions of the adiabatic approximation as stated above), but its phase has changed:

$$|\psi(t_f)\rangle = e^{-(i/\hbar)\int_{t_i}^{t_f} E(t')\,dt'} e^{i\gamma} |\psi(t_i)\rangle. \tag{2.1}$$

Here $E(t')$ means the corresponding eigenvalue of the Hamiltonian at that time, and γ is the Berry phase, expressed as an integral over a path in *parameter* space with no time dependence:

$$\gamma = i \oint \langle \psi(\boldsymbol{\lambda}) | \nabla_{\boldsymbol{\lambda}} | \psi(\boldsymbol{\lambda}) \rangle \cdot d\boldsymbol{\lambda}. \tag{2.2}$$

Note that there are two different arguments of ψ in the above formulas. When ψ has a time argument, it means the wavefunction of the system at that time. When ψ has a parameter argument, it means the reference wavefunction we have chosen for that point in the parameter space of the Hamiltonian. A key assumption of the

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

derivation is that there is some smooth choice of the $\psi(\boldsymbol{\lambda})$ throughout a surface in parameter space with the loop as boundary.

For an open path, we need to describe the phase of the wavefunction relative to this reference set, so the expression becomes more complicated (for the closed path, we could simply compare the initial and final wavefunctions, without needing the reference set at these points). We will show that, assuming $\psi(t_i) = \psi(\boldsymbol{\lambda}(t_i))$ so that the initial wavefunction is equal to the reference state at the corresponding value of parameters,

$$\langle \psi(\boldsymbol{\lambda}(t)) | \psi(t) \rangle = e^{-(i/\hbar)\int_0^t E(t')\,dt'} e^{i\gamma}, \tag{2.3}$$

that is, the Berry phase appears when comparing the actual time-dependent evolved state $\psi(t)$ to the reference state at the point in parameter space $\boldsymbol{\lambda}(t)$. We can take the time derivative of both sides and use the time-dependent Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = H(t)\psi. \tag{2.4}$$

The two sides agree initially if we choose the appropriate boundary condition on the Berry phase. The time derivative of the left side of (2.3) is

$$\langle \psi(\boldsymbol{\lambda}(t)) | \frac{-iE(t)}{\hbar} | \psi(t) \rangle + \frac{d\lambda_j}{dt} \langle \partial_{\lambda_j} \psi(\boldsymbol{\lambda}(t)) | \psi(t) \rangle, \tag{2.5}$$

so writing $e^{i\theta(t)} = \langle \psi(\boldsymbol{\lambda}(t)) | \psi(t) \rangle$, we have computed

$$\frac{d}{dt} e^{i\theta(t)} = \left(\frac{-iE(t)}{\hbar} + \frac{d\lambda_j}{dt} \langle \partial_{\lambda_j} \psi(\boldsymbol{\lambda}(t)) | \psi(\boldsymbol{\lambda}(t)) \rangle \right) e^{i\theta(t)}. \tag{2.6}$$

Note that in the second term on the right side, it is $\psi(\lambda(t))$ that appears in the ket, and we have used $|\psi(\boldsymbol{\lambda}(t))\rangle e^{i\theta(t)} = |\psi(\boldsymbol{\lambda}(t))\rangle \langle \psi(\boldsymbol{\lambda}(t)) | \psi(t) \rangle = |\psi(t)\rangle$, because the projection operator onto $|\psi(\boldsymbol{\lambda}(t))\rangle$ must act trivially on a state already in the same one-dimensional subspace (i.e., differing only by a phase), which we remain in because of the adiabatic approximation. This time evolution of the phase is satisfied if (note that for E we do not need to distinguish between time and λ dependence)

$$\dot{\theta}(t) = -\frac{E(t)}{\hbar} - i \frac{d\lambda_j}{dt} \langle \partial_{\lambda_j} \psi(\boldsymbol{\lambda}(t)) | \psi(\boldsymbol{\lambda}(t)) \rangle, \tag{2.7}$$

which is our desired conclusion after noting that (dropping the explicit t dependence)

$$\frac{d\gamma}{dt} = i \frac{d\lambda_j}{dt} \langle \psi(\boldsymbol{\lambda}) | \partial_{\lambda_j} \psi(\boldsymbol{\lambda}) \rangle = -i \frac{d\lambda_j}{dt} \langle \partial_{\lambda_j} \psi(\boldsymbol{\lambda}) | \psi(\boldsymbol{\lambda}) \rangle. \tag{2.8}$$

The negative sign in the last equality, which shows that the Berry connection or Berry vector potential

$$\mathcal{A}_j = i \langle \psi(\boldsymbol{\lambda}) | \partial_{\lambda_j} \psi(\boldsymbol{\lambda}) \rangle \tag{2.9}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

is real, follows from noting that $\partial_{\lambda_j}\langle\psi|\psi\rangle = 0$ by constancy of normalization. It is crucial for a nonzero Berry phase that the eigenstates of H change somewhere on the path beyond just a phase factor (i.e., they are physically inequivalent vectors in Hilbert space). If this were not the case, as must happen for a Hilbert space of dimension 1, then the Berry phase is just the phase difference between the reference choice at the initial and final points, which is zero for a closed path. The spin example below shows that a Hilbert space of dimension 2 is sufficient for the Berry phase to be significant. So although the rate of change in H dropped out, as long as the system remains adiabatic, and only the path taken by H enters the Berry phase, the eigenstates of H must have a genuine evolution, not just a phase change, for the Berry phase to be nontrivial.

Now one can ask whether the Berry phase is independent of how we chose the reference wavefunctions (in this case, the U(1) degree of freedom in the wavefunction at each λ, where U(1) means the group of one-by-one unitary matrices or phase factors $e^{i\phi}$). While for an open path it clearly is not gauge-independent (here meaning that we change the phase of the wavefunction at different points in parameter space), the Berry phase is gauge-independent for a closed path, for exactly the same reasons as a closed line integral of $\mathcal{A}$ is gauge-independent in electrodynamics: we can integrate the Berry flux or Berry curvature $\mathcal{F}_{ij} = \partial_i\mathcal{A}_j - \partial_j\mathcal{A}_i$ (which the reader can check is gauge-independent, just like the field strength tensor $F_{\mu\nu}$ in electrodynamics, by recalling that gauge transformation adds a gradient to A) on the surface bounded by the path. The Berry curvature is a field strength in parameter space, which can have any number of dimensions, but when working in three dimensions we often convert it to a vector, the curl of $\mathcal{A}$.

Note, however, that this gauge invariance assumed that the loop used to compute the Berry phase encircled a surface on which the Berry curvature $\mathcal{F}$ was defined. If we had a closed loop on a manifold that is topologically nontrivial, in a sense that will be clearer by the end of the chapter, then this argument doesn't work. For an example, consider the circle as an abstract object without an interior, which is the right picture of the Brillouin zone (set of inequivalent momenta) of a one-dimensional crystal. Note that a gauge transformation changes $\mathcal{A}$ by the gradient of a scalar phase, from the definition (2.2). Even if the reference wavefunctions are uniquely defined, however, that phase could change by a multiple of 2π on circling the path, so that only $e^{i\gamma}$ would be well defined; this object is the Berry phase version of a Wilson loop in gauge theories. The possibility of having a Berry vector potential that, when integrated along paths, gives quantities that are almost but not completely gauge-invariant (i.e., gauge-invariant except for discrete jumps) will be explained more mathematically in Section 2.7.

To get some geometric intuition for what the Berry phase means, we explain why the Berry connection $\mathcal{A}$ is sometimes called a connection, and the flux $\mathcal{F}$ is

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

sometimes called a curvature. A connection is a way to compare vector spaces that are attached to different points of a manifold, forming a vector bundle. In our case, there is a one-dimensional complex vector space attached at each point in parameter space, spanned by the local eigenstate. The inner product lets us compare vectors at the same point in parameter space, but the Berry connection appears when we try to compare two vectors from slightly different points.

A useful example of a real vector bundle is the tangent bundle to a Riemannian manifold (say, a sphere), made up of tangent vectors at each point, which have a dot product corresponding to the inner product in quantum mechanics. The connection in this case, which gives rise to parallel transport of tangent vectors, determines the same curvature that appears below in the example of the Gauss–Bonnet theorem. Consider an airplane moving around the surface of the Earth and carrying a gyroscope that is fixed to lie in the tangent plane to the Earth's surface (i.e., free to rotate around the normal axis to the tangent plane). If the airplane follows a great circle, then it will appear to be going straight ahead to a passenger on board, and the gyroscope will not rotate relative to the plane's axis.

However, if the airplane follows a line of latitude other than the equator, or any other path that is not a geodesic (see a differential geometry text for details), it will feel constantly as though it is turning, and the gyroscope will appear to rotate relative to the airplane's direction. After going around a closed path with the airplane, the gyroscope may have rotated compared to a stationary gyroscope (the same physics that underlies Foucault's pendulum). As an exercise, one can work out that the total angle of rotation in circling a line of latitude is $2\pi \sin\phi$, where ϕ is the latitude. At the equator this gives no rotation, while at the North Pole this gives a 2π rotation. This is a geometrical version of the same idea of holonomy (here, failure of a gyroscope to return to its initial direction in going around a closed path) that underlies the Berry phase. Note that a vector potential in a gauge theory and the associated Wilson loop are also examples of the concept of holonomy in a (now complex) vector bundle. Instead of the vector bundle consisting of the tangent vectors through a point, the Berry phase comes from the Hilbert space of wavefunctions at a certain value of some parameters in the Hamiltonian. More precisely, the Berry connection tells us how to compare two states (Hilbert-space vectors) at different places in parameter space, in the same way as parallel transport tells us how to compare two tangent vectors at two points on a curved manifold.

We make two remarks in passing about generalizations. First, the Abelian U(1) Berry phase described above generalizes immediately to a non-Abelian Berry phase factor U(N), that is, an N-by-N unitary matrix, when instead of a single nondegenerate eigenstate, we consider a subspace of dimension N. Here Abelian means that two elements of U(1) commute, while two elements of U(N) for $N > 1$ do not necessarily commute. This has some important applications to topological phases that were discovered only recently. Bloch's theorem, introduced in a

moment, will later be used to discuss how Abelian and non-Abelian connections arise naturally in a solid.

Second, while we have focused on continuous paths in the above, it is possible to make a similarly gauge-invariant object from wavefunctions at as few as three points. Consider the complex number $U = \langle u_1|u_2\rangle\langle u_2|u_3\rangle\langle u_3|u_1\rangle$. This is independent under gauge changes at any of the points; indeed it can be written in terms of the manifestly gauge-invariant projection operators $P_i = |u_i\rangle\langle u_i|$ as $U = Tr(P_1P_2P_3)$. The magnitude of U is generally less than unity, so it is not purely a phase, but as we increase the number of points and make them fill out a closed path, U becomes essentially the Wilson loop $e^{i\gamma}$ made from the Berry phase along that path. We demonstrate a use of this in the explicit example of a Berry phase in Box 2.1.

Box 2.1 The Berry Phase of the Adiabatic Dynamics of a Spin

Let us consider how the state of a quantum spin evolves if the spin always points in the direction of an applied magnetic field and the magnetic field varies smoothly. This will give some intuitive understanding of the Berry phase, and the mathematics involved will be useful later when we discuss the topological underpinnings of the integer and fractional quantum Hall effects. We first need to define a representation of the spin Hilbert space that is more symmetric than the usual S_z basis (the basis of states with definite values of the projection of spin along some fixed axis).

The coherent-state representation of a quantum spin of magnitude s is useful for many spin problems where having a fixed reference axis is undesirable. The basic idea is to represent spins in a basis of states $|\hat{\Omega}\rangle$ that in the classical limit behave like classical vectors: the states are obtained by rotating the North Pole $|s, s\rangle$ with a rotation operator $R(\theta, \phi, \chi)$ that is a function of three Euler angles. The vector $\langle \mathbf{S}\rangle$ of expectation values of the spin operators in the basis state $|\hat{\Omega}\rangle$ points along the unit vector $\hat{\Omega}$.

While χ is essentially an arbitrary phase convention, keeping careful track of the spin wavefunction shows that there is a physically meaningful Berry phase for any smooth choice of the spin reference wavefunctions. Define, for a unit vector $\hat{\Omega} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$,

$$|\hat{\Omega}\rangle = R(\chi, \theta, \phi)|s, s\rangle = e^{iS^z\phi}e^{iS^y\theta}e^{iS^z\chi}|s, s\rangle. \tag{2.10}$$

It is intentional that there are two S^z operators appearing in this formula, as the Euler angles are defined using one rotation around the original z-axis and one around the final one. One can use the explicit representation of the coherent states and some algebra to compute the inner product

$$\langle\hat{\Omega}|\hat{\Omega}'\rangle = \left(\frac{1 + \hat{\Omega}\cdot\hat{\Omega}'}{2}\right)^s e^{-is\psi} \tag{2.11}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

with

$$\psi = 2\arctan\left[\tan\left(\frac{\phi-\phi'}{2}\right)\frac{\cos[\frac{1}{2}(\theta+\theta')]}{\cos[\frac{1}{2}(\theta-\theta')]}\right]+\chi-\chi'. \tag{2.12}$$

For further details of the coherent-state representation, see the book of Auerbach (1994). Since (2.11) becomes zero as $s \to \infty$ unless $\hat{\Omega} = \hat{\Omega}'$, coherent states also justify the claim that as $s \to \infty$, the spins are like classical unit vectors (the classical limit of a spin).

Note that the coherent states are not orthogonal for finite s. The completeness relation is

$$\frac{2s+1}{4\pi}\int d\hat{\Omega}|\hat{\Omega}\rangle\langle\hat{\Omega}| = \mathbf{1}. \tag{2.13}$$

Now specialize to $s = 1/2$ and consider dynamics in the two-by-two Hamiltonian given by

$$H = -\hat{\mathbf{n}}\cdot\boldsymbol{\sigma} \tag{2.14}$$

where $\boldsymbol{\sigma}$ is a vector of Pauli matrices. At time $t = 0$, suppose that the spin is initially prepared in the ground state $|\hat{\Omega}\rangle$, with $\hat{\Omega} = \hat{\mathbf{n}}$. Let the unit vector $\mathbf{n}$ be time-dependent, but assume that it changes sufficiently slowly that there are no transitions out of the lower-energy spin state. In this problem the energies are constant, so we can neglect the energetic part of the phase change around a path, assuming the path is traced out adiabatically.

Note that we follow tradition in using a more compact notation than in the general derivation of Berry's phase: the reference wavefunction $|\psi(\hat{\Omega})\rangle$ is now just written as $|\hat{\Omega}\rangle$. We also give a slightly different perspective on the Berry phase

$$\gamma = i\int\langle\psi(\boldsymbol{\lambda})|\nabla_{\boldsymbol{\lambda}}|\psi(\boldsymbol{\lambda})\rangle\cdot d\boldsymbol{\lambda} = \int_{t_i}^{t_f}\langle\hat{\Omega}(t)|i\frac{d}{dt}|\hat{\Omega}(t)\rangle\,dt. \tag{2.15}$$

Consider the overlap of wavefunctions at two slightly different times t and $t+dt$. The magnitude of this overlap is less than 1, but only by an amount of order dt^2. At order dt, the change in the overlap is purely imaginary. We can use this to build up the Berry phase over a segment of path as a product of a large number of overlaps, using

$$\begin{aligned}\gamma = -\Im\log\Big[&\langle\hat{\Omega}(t_i)|\hat{\Omega}(t_i+dt)\rangle\langle\hat{\Omega}(t_i+dt)|\hat{\Omega}(t_i+2\,dt)\rangle\ldots\\ &\ldots\langle\hat{\Omega}(t_f-2\,dt)|\hat{\Omega}(t_f-dt)\rangle\langle\hat{\Omega}(t_f-dt)|\hat{\Omega}(t_f)\rangle\Big].\end{aligned} \tag{2.16}$$

Recall that the phase factor between original and final states used in our earlier definition of the Berry phase was $\exp(i\gamma)$, which is why the logarithm appears here. Actually, one can start (Vanderbilt, 2018) from this discrete expression as the definition of the Berry phase for any number of intermediate steps, and then obtain the continuum limit we have focused on here by having a large number of steps. Clearly the 2π

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

ambiguity of the imaginary part of the logarithm is consistent with the interpretation of γ as a phase. To see the equivalence, write

$$-\Im \log\langle\hat{\Omega}(t_i)|\hat{\Omega}(t_i+dt)\rangle \approx -\Im \log(1+dt\,\langle\hat{\Omega}|\frac{d}{dt}\hat{\Omega}\rangle)$$
$$\approx -dt\,\Im\langle\hat{\Omega}|\frac{d}{dt}\hat{\Omega}\rangle = i\,dt\,\langle\hat{\Omega}|\frac{d}{dt}\hat{\Omega}\rangle, \qquad (2.17)$$

where we have used that the inner product in the last step is purely imaginary.

Using the explicit coherent state representation, one finds for the overlap (Auerbach, 1994)

$$\langle\hat{\Omega}(t+dt)|\hat{\Omega}(t)\rangle = e^{-is\,dt\,\dot{\phi}\cos(\theta(t))-is\dot{\chi}}, \qquad (2.18)$$

where $\dot{\chi}$ results from whatever phase convention was chosen in the coherent state via

$$\dot{\chi} = \frac{d\chi}{d\hat{\Omega}}\frac{d\hat{\Omega}}{dt}. \qquad (2.19)$$

So the change in the Berry phase is

$$\dot{\gamma} = -s\dot{\chi} - s\,dt\,\dot{\phi}\cos(\theta(t)). \qquad (2.20)$$

Around a closed path $\dot{\chi}$ must integrate to zero, since χ only changes through the change in $\hat{\Omega}$, which returns to its initial value. The other part need not be zero and has a simple geometrical interpretation. We see that for an *open* path the phase change is not directly meaningful since it depends on the arbitrary phase convention in χ.

The phase gained around a closed path $\mathcal{P}$ on the unit sphere is

$$\gamma_{\mathcal{P}} = -s\int_{t_i}^{t_f} dt\,\dot{\phi}\cos\theta = -s\oint_{\phi_i}^{\phi_f=\phi_i} d\phi\,\cos(\theta(\phi)). \qquad (2.21)$$

The point of the second rewriting is that the integral depends only on the closed path traced out by the magnetic field direction, not by how it moves along that path. This integral just measures the area traced out on the unit sphere, since

$$-\oint_{\phi_i}^{\phi_i} d\phi\,\cos(\theta(\phi)) = -\oint_{\phi_i}^{\phi_i} d\phi\,[1-\cos(\theta(\phi))]. \qquad (2.22)$$

Here we assume that the path did not encircle the North Pole so that ϕ remained well defined. Now the integrand is just the integral of $d(\cos\theta)$ as θ runs from the North Pole to the present point.

To summarize, the net effect of taking the spin around a closed path is to induce a phase proportional to s and to the area enclosed. This can be written as the loop integral of a "magnetic monopole" vector potential on the sphere with constant field strength:

$$\gamma = s\int_{t_i}^{t_f} dt\,\mathbf{A}(\hat{\Omega})\dot{\hat{\Omega}}. \qquad (2.23)$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

One gauge choice for this vector potential is

$$\boldsymbol{\mathcal{A}} = -\frac{1-\cos\theta}{\sin\theta}\hat{\phi}, \tag{2.24}$$

which has a singularity at the South Pole ($\theta = \pi$).

A subtlety is that any gauge choice for a nonzero monopole vector potential has to have a singularity somewhere on the sphere, because of the nonzero integral of the curl of $\boldsymbol{\mathcal{A}}$, which is gauge-independent. For example, consider integrating the vector potential in (2.24) over a small circle around the South Pole. The result will be nearly equal to the area of the sphere (because we set up this gauge choice to capture areas starting from the North Pole, as in our explicit calculation), which explains why the vector potential had to diverge in order to get a finite answer over a tiny circle. At the North Pole, there is a Dirac string containing (Berry) magnetic flux entering the sphere, in order for the flux elsewhere to be uniformly directed outward as if a magnetic monopole were located at the center of the sphere. An alternative gauge choice would define areas starting from the South Pole and be singular at the North Pole. Note that the observable Berry phase factor $e^{i\gamma}$ is unchanged under this difference of 4π, though, because even after multiplying by the spin quantum number s, the ambiguity is a multiple of 2π.

We will explore some further interesting consequences of this spin problem later on. There are several problems where the integral of a Berry curvature over a two-dimensional manifold, such as the sphere in the above example, is a topological invariant and physically observable. The Berry phase studied here turns out to be essential in developing the path integral representation of a quantum spin, and the degeneracy of the lowest-energy eigenstates in the monopole field (the lowest Landau level as explained in the following section) on the sphere is $2s + 1$, consistent with the number of states in the spin multiplet.

2.2 One Electron in a Magnetic Field: Landau Levels

The first topological phases to be discovered were found in devices known as quantum wells or two-dimensional electron gases, in which electrons are constrained to move in a plane by a confining potential in the $\hat{\mathbf{z}}$ direction (Box 3.2). Looking ahead to our discussion of energy bands in the following chapter, it is typically a good approximation to assume that the electron density in such materials is small enough that electrons are near a band extremum and have a quadratic dispersion like free electrons. The key discoveries known as the integer and fractional quantum Hall effects emerged when a strong magnetic field was applied perpendicular to the plane of motion of the electrons, and a crucial ingredient we will use in Chapters 3 and 5 is the quantum mechanics of a free particle in a magnetic field.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

To start, the classical equations of motion for a particle of charge q and mass m moving in two dimensions in a magnetic field applied in the perpendicular direction, $\mathbf{B} = B\hat{\mathbf{z}}$, are

$$\begin{aligned} m\frac{d^2x}{dt} &= qB\frac{dy}{dt}, \\ m\frac{d^2y}{dt} &= -qB\frac{dx}{dt}. \end{aligned} \tag{2.25}$$

Defining the cyclotron frequency

$$\omega_c = \frac{qB}{m} \tag{2.26}$$

allows us to write down the general solution of this equation in an elegant form by switching to a complex parameterization of the two-dimensional plane, $z \equiv x + iy$ (with $\dot{z} = dz/dt$):

$$\begin{aligned} \ddot{z}(t) &= i\omega_c \dot{z}(t), \\ \dot{z}(t) &= \dot{z}(0)e^{i\omega_c t}, \\ z(t) &= z(0) + \frac{\dot{z}(0)}{i\omega_c}\left(e^{i\omega_c t} - 1\right). \end{aligned} \tag{2.27}$$

This describes the circular motion of the particle. Note that the frequency of this motion, ω_c, is independent of the radius of the circle, $R_L = \frac{|\dot{z}(0)|}{\omega_c}$: the particle's speed is proportional to R_L. Its kinetic energy is time-independent because the Lorentz force, $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$, only acts in a direction perpendicular to the particle's motion. It is given by

$$E_{\text{kin}} = \frac{q^2B^2}{2m}R_L^2 = \frac{1}{2}m\omega_c^2 R_L^2. \tag{2.28}$$

Such periodic motion with frequency independent of amplitude and an energy proportional to its square is reminiscent of a harmonic oscillator. However, there is a *four*-dimensional phase space, as real space is two-dimensional, but it will turn out that in detail things are not very different from a two-dimensional simple harmonic oscillator: the energy levels are indeed given by $E_n = \hbar\omega_c(n + 1/2)$, with n a nonnegative integer, the Landau level index.

The degeneracy of each level is macroscopic, proportional to the system's area A. This accounts for the remaining pair of degrees of freedom. To obtain this degeneracy, which will be verified in a concrete calculation momentarily, let us posit in an *ad hoc* way that the number of states in the system needs to remain invariant when the field is switched on. In zero field, the number of states is given by $N = \int \frac{d^2p\, d^2x}{h^2}$, where x and p are canonically conjugate position and momentum coordinates. With $p^2/2m = E$, one finds for a system of size $\int d^2x = A$, a

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

density of states $\rho(E) = \frac{1}{A}\frac{dN}{dE} = \frac{2\pi m}{h^2}dE$. A Landau level groups together states over an energy interval $\Delta E = \hbar\omega_c$,

$$\int_E^{E+\Delta E} \rho(E)dE = \frac{qB}{h} = \frac{1}{2\pi\ell^2}, \tag{2.29}$$

where $\ell^2 = \hbar/qB$ is called the magnetic length. It is the fundamental length scale in the problem, which, for example, encodes the size of the smallest wavepacket of an eigenstate that can be constructed in the presence of a field. The areal density of states of each Landau level is $\frac{1}{2\pi\ell^2}$, or one electron in the area that encloses one single-electron quantum $\phi = \frac{h}{e}$ of flux from the applied magnetic field.

To confirm these results formally, and as we will need explicit wavefunctions, let us next derive these properties quantum-mechanically. The Hamiltonian reads, now allowing for an effective mass m^*,

$$\mathcal{H} = \frac{1}{2m^*}(-i\hbar\nabla - q\mathbf{A})^2 \equiv \frac{\hat{p}^2}{2m^*}. \tag{2.30}$$

We will pick gauges for $\mathbf{A}$ and find associated wavefunctions momentarily, but first we point out some general symmetry aspects.

For $\mathbf{A} = 0$, this is the Hamiltonian of a free particle. Translations, generated by $\hat{p}_x$ and $\hat{p}_y$, commute: $[\hat{p}_x, \hat{p}_y] = 0$. The resulting group structure is thus Abelian, and the concomitant irreducible representations are therefore one-dimensional, labeled by the usual momentum $\mathbf{k} = (k_x, k_y)$. In nonzero field, $\nabla \times \mathbf{A} = B\hat{z}$, this is no longer the case:

$$[\hat{p}_x, \hat{p}_y] = -i\hbar^2/\ell^2\ , \tag{2.31}$$

unlike in a circularly symmetric harmonic oscillator, so x- and y- coordinates are no longer independent.

Indeed, a second set of canonical commutation relations is defined for the so-called guiding center coordinates

$$c_x = x - p_y/m^*\omega_c, \quad c_y = y + p_x/m^*\omega_c\ , \tag{2.32}$$

for which $[c_x, c_y] = i\ell^2$. Here, ℓ^2 plays the role of the volume element of phase space normally supplied by h. As we will see below, the Hamiltonian is independent of the value of the guiding center variables, and to quantize them semiclassically, one cuts up the c_x–c_y phase into segments of area $2\pi\ell^2$. There are many ways to do this, the most convenient being ones which respect the symmetry of the Hamiltonian implied by the choice of gauge. In general this symmetry is lower than of the physical problem: in the presence of magnetic field, a symmetry operation in general maps the Hamiltonian onto one which is equivalent *up to a*

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

gauge-transformation. The resulting symmetry group in a field is known as a magnetic translation group (see Eq. 5.10 for more details), an instance of a projective symmetry group.

Two gauge choices are particularly useful: the Landau gauge, which preserves explicit translational symmetry in one direction while apparently discarding isotropy and the perpendicular translations: $\mathbf{A} = (0, x, 0)$; and the circular or rotational gauge, which chooses a preferred origin but preserves isotropy: $\mathbf{A} = (-\frac{y}{2}, \frac{x}{2}, 0)$. The concomitant semiclassical quantizations are, firstly, areas bordered by adjacent parallel lines at $x_n = \frac{2\pi\ell^2}{L_y}$ for the Landau case; or by adjacent concentric circles of radius $\ell\sqrt{2n}$ for the circular gauge. In either case, there is one state for each flux quantum threading the system. An interesting modification occurs when the kinetic term is not quadratic but rather linear and Dirac-like, as for the electrons in graphene (Section 2.5.2).

The eigenfunctions can be constructed by defining the ladder operators

$$\begin{aligned} a^\dagger &= \frac{\ell}{\sqrt{2}\hbar}\left(p_x + ip_y\right) \\ b &= \frac{1}{\sqrt{2}\ell}\left(c_x + ic_y\right) \end{aligned} \tag{2.33}$$

where $[a, a^\dagger] = [b, b^\dagger] = 1$, with $b, b^\dagger$ commuting with $a, a^\dagger$ as well as the Hamiltonian

$$\mathcal{H} = \hbar\omega_c\left(a^\dagger a + 1/2\right) . \tag{2.34}$$

As advertised, the spectrum is that of a harmonic oscillator not depending on the guiding center ladder operators b at all. A complete set of states is then given by

$$|n, m\rangle = \frac{a^{\dagger n} b^{\dagger m}}{\sqrt{n!m!}}|0, 0\rangle . \tag{2.35}$$

2.2.1 Symmetric and Landau Gauge Wavefunctions

For a study of the quantum Hall effect, it is easiest to follow Laughlin's choice of symmetric gauge, which preserves the isotropy obeyed by a radial interaction. For this choice, one obtains $|0, 0\rangle = \frac{1}{\sqrt{2\pi\ell^2}}\exp(-|z|^2/4\ell^2)$ and, for the nth Landau level, an expression involving associated Laguerre polynomials

$$|n, m\rangle = \frac{\left(\frac{z}{\sqrt{2}\ell}\right)^m \exp(-|z|^2/4\ell^2) L_n^m\left(\frac{|z|^2}{4\ell^2}\right)}{\sqrt{2\pi\ell^2(n+m)!/n!}} . \tag{2.36}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

The $n = 0$ case of this formula will be used heavily in Section 3.1 and Section 5.1 for the integer and fractional quantum Hall effects, respectively. In Landau gauge, the unnormalized wavefunctions read

$$|n, k\rangle_L = e^{iky} H_n\left(x + k\ell^2\right) \exp\left[-\frac{1}{2\ell^2}\left(x + k\ell^2\right)^2\right] \tag{2.37}$$

with $k = \frac{2\pi}{L_y} \cdot s$, and $s = 0, 1, \ldots, L_y - 1$ fixing the center of the wavefunctions in the x-direction spaced by $\frac{2\pi}{L_y}\ell^2$ as in the semiclassical quantization. The Hermite polynomial H_n is a constant for the lowest Landau level ($n = 0$). This form is straightforward to see directly from the Hamiltonian in Landau gauge, as the momentum in one direction (here $\hat{\mathbf{y}}$) commutes with the Hamiltonian. Finding simultaneous eigenstates of $\hat{\mathbf{y}}$-momentum and energy then gives plane-wave behavior along y and a harmonic oscillator problem for the x coordinate, consistent with the harmonic oscillator spectrum and the wavefunctions in (2.37).

What if the magnetic field were so strong that we could not get away with using continuum limit of the kinetic term of free electrons? As a starting point to answer that question, let us drop the magnetic field and obtain a general result for one-particle wavefunctions in periodic potentials.

2.3 One Electron in a Crystal: Bloch's Theorem

One of the cornerstones of the theory of solids is Bloch's theorem for electrons in a periodic potential such as provided by the ions of crystal. While stable phases should not depend very much on whether the underlying solid is a perfect crystal, many important developments have started from Bloch's theorem, which naturally gives electronic wavefunctions the structure of a bundle over the set of inequivalent momenta (the Brillouin zone). The classic text of Ashcroft and Mermin is a good source for a detailed treatment. Here we confine ourselves to justifying this theorem in the following form: given a potential invariant under a set of lattice vectors $\mathbf{R}$, $V(\mathbf{r} + \mathbf{R}) = V(\mathbf{r})$, closed under addition, the electronic eigenstates can be labeled by a crystal momentum $\mathbf{k}$ and written in the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}), \tag{2.38}$$

where the function u has the periodicity of the lattice. Note that the crystal momentum $\mathbf{k}$ is only defined up to addition of reciprocal lattice vectors, that is, vectors whose dot product with every one of the original lattice vectors is a multiple of 2π.

We give a quick proof of Bloch's theorem in one spatial dimension, then consider the Berry phase of the resulting wavefunctions. For the former, we essentially use two theorems. First, that Abelian groups (ones where multiplication is commutative) have one-dimensional representations: in this case, lattice translations

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

form an Abelian group as they have the same net result regardless of the order in which they are effected, so that they can be labeled by a crystal momentum k. Second, in quantum mechanics two operators which commute can be simultaneously diagonalized, so that the label k can also be attached to the eigenfunctions of the Hamiltonians. This latter ceases to apply in the presence of a magnetic field, when instead of Bloch states we encounter Landau levels as a result of a non-Abelian (i.e., noncommuting) magnetic translation algebra (Eq. 5.10).

Starting with the second item, we note that in the problem at hand, we have a non-Hermitian operator (lattice translations by the lattice spacing a: $(T_a\psi)(x) = \psi(x+a)$) that commutes with the Hamiltonian. It turns out that only one of the two operators needs to be Hermitian for simultaneous eigenstates to exist if the other operator is normal (commutes with its adjoint), which is true for the translation operator since $T_a^\dagger = T_{-a}$, and two translations commute. Therefore we can find wavefunctions that are energy eigenstates and satisfy

$$(T_a\psi)(x) = \lambda\psi(x). \tag{2.39}$$

Now if the magnitude of λ is not 1, repeated application of this formula will give a wavefunction that either blows up at spatial positive infinity or negative infinity. We would like to find wavefunctions that can extend throughout an infinite solid with bounded probability density, and hence require $|\lambda| = 1$. From that it follows that $\lambda = e^{i\theta}$, and we define $k = \theta/a$, where we need to specify an interval of width 2π to uniquely define θ, say $[-\pi, \pi)$. In other words, k is ambiguous by addition of a multiple of $2\pi/a$, as expected. So we have shown

$$\psi_k(x + a) = e^{ika}\psi_k(x). \tag{2.40}$$

The last step is to define $u_k(x) = \psi_k(x)e^{-ikx}$; then (2.40) shows that u_k is periodic with period a, and $\psi_k(x) = e^{ikx}u_k(x)$.[1]

While the energetics of Bloch wavefunctions underlies many properties of solids, there is also Berry phase physics arising from the dependence of u_k on k that was understood only rather recently. Note that, even though this is one-dimensional, there is a nontrivial closed loop in the parameter k that can be defined because of the periodicity of the Brillouin zone $k \in [-\pi/a, \pi/a)$:

$$\gamma = \oint_{-\pi/a}^{\pi/a} \langle u_k | i\partial_k | u_k \rangle dk. \tag{2.41}$$

How are we to interpret this Berry phase physically, and is it even gauge-invariant? We will derive its physical meaning (connected to electrical polarization) in Chapter 4, but an intuitive clue is provided if we make the replacement $i\partial_k$ by x, as

[1] Readers interested in more information, such as the inclusion of spin and the three-dimensional case, can consult the solid state text of Ashcroft and Mermin (1976).

would be appropriate if we consider the action on a plane wave. This suggests, correctly, that the Berry phase may have something to do with the spatial location of the electrons, but evaluating the position operator in a Bloch state gives an ill-defined answer; for this real-space approach to work, we would need to introduce localized Wannier orbitals in place of the extended Bloch states.

2.4 The Simplest Tight-Binding Model

The tight-binding approximation is a simple version of a method known as linear combination of atomic orbitals (LCAO). The basic concept is provided here because tight-binding models are often used to demonstrate various kinds of topological behavior, but any solid-state textbook will have considerably more details. In this section we treat a single orbital per unit cell, because the new features that emerge when there are multiple orbitals per unit cell are a major subject of Chapter 4.

The goal is to find a way to deal with electronic wavefunctions on an infinite periodic array without having to consider the enormous Hilbert space of all possible wavefunctions. We look for a variational ground state made as a superposition of *one* orbital wavefunction on each atom. Let atom m have one orbital that we keep, $|m\rangle$. For example, for a chain of hydrogen atoms, we would keep the 1s orbital on every atom, and figure that this is probably a good approximation to the low-energy states of the chain, since the gap to the $n = 2$ orbitals is quite large.

In the simplest case of a diatomic molecule made of two identical atoms, and with one orbital on each atom, suppose that the ground state is made from the even combination of those orbitals, while the first excited state is made from the odd combination. The splitting is determined by the magnitude of the hopping t. Writing $|1\rangle$ and $|2\rangle$ for the orbitals kept on atom 1 and 2, the hopping integral is

$$-t = \langle 1|V_2|2\rangle = \langle 1|V_1|2\rangle, \tag{2.42}$$

where V_1 and V_2 are the atomic potentials centered on atom 1 and 2, respectively. For the energy we find

$$E_\pm = \epsilon_0 + V_{\text{cross}} \pm |t| \tag{2.43}$$

where ϵ_0 is the original orbital energy and $V_{\text{cross}} = \langle 1|V_2|1\rangle = \langle 2|V_1|2\rangle$, where we have assumed that the two atoms are identical.

Generalizing this to an infinite chain gives a Hamiltonian with two terms: a diagonal term with the energy of each orbital V_0 (actually computing this is a little subtle for an infinite chain, so we assume some finite value is provided), and a

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

hopping term $-t$ between nearest-neighbor atoms that is essentially similar to the two-atom case. We seek normalized wavelike solutions of the Hamiltonian

$$|\psi\rangle = \sum_{n=1}^{N} \frac{e^{ikan}|n\rangle}{\sqrt{N}}. \tag{2.44}$$

Here the chain of atoms is taken to form a ring, so that the boundary conditions are periodic, and soon we will take $N \to \infty$.

Acting on this wavefunction with the Hamiltonian, which is a tridiagonal matrix in the basis of atomic orbitals, we find that indeed these waves can be eigenstates with energy satisfying

$$E = V_0 - 2t\cos(ka). \tag{2.45}$$

Here and above we are making the approximation that the overlap between different atomic orbitals $\langle i|j\rangle$ can be neglected; including this effect leads to a generalized eigenvalue problem rather than a conventional eigenvalue problem, but the basic physics of the approximation is unchanged.

While the wavefunction in (2.44) is an energy eigenstate, it may not be obvious that it satisfies Bloch's theorem. Let ϕ be the wavefunction of the atomic orbital being used, so that the wavefunction of $|n\rangle$ is $\phi(x - an)$. Then the explicit wavefunction is

$$\psi(x) = \sum_{n=-\infty}^{\infty} e^{ikna}\phi(x - na), \tag{2.46}$$

and

$$\psi(x+a) = \sum_{n=-\infty}^{\infty} e^{ikna}\phi(x-(n-1)a) = \sum_{m=-\infty}^{\infty} e^{ik(m+1)a}\phi(x-ma) = e^{ika}\psi(x), \tag{2.47}$$

where $m = n - 1$.

An interesting question is about what spatial information remains if we only have access to the tight-binding Hamiltonian and not to the specific properties of the orbitals. This will become important when we later consider multiple orbitals within the unit cell, which allows the spatial distribution of the electronic state to change with crystal momentum. As hinted above, the Berry phase of Bloch wavefunctions contains spatial information and allows numerous important properties to be extracted.

The main subtlety arising in the case of multiple orbitals with different locations in the unit cell is that at least two different conventions for the form of the tight-binding basis and Hamiltonian are widely used in the literature ["Convention I" and "Convention II," in the language of the book (Vanderbilt, 2018), which

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

has a lengthy explanation]. While both conventions give the same physical results if properly implemented, the expressions used for Berry phase quantities such as the electric polarization are simplest in Convention I, although the tight-binding Hamiltonian may seem to be unnecessarily complicated in that convention.

The basic difference is whether phase factors appearing in the finite-dimensional tight-binding "Bloch Hamiltonian," defined in a moment, involve distances between orbitals within the unit cell or only lattice vectors. Let us formalize our understanding of tight-binding models a little, and go to three dimensions in writing formulae. Under some simplifying assumptions, a tight-binding model with N orbitals in the unit cell is related to finding eigenvalues of an $N \times N$ matrix. Suppose the orbitals in the unit cell with lattice vector $\mathbf{R}$ are described by states $|\phi_{\mathbf{R}j}\rangle$, $j = 1, \ldots, N$, with wavefunction

$$\phi_{\mathbf{R}j}(\mathbf{r}) = \varphi_j(\mathbf{r} - \mathbf{R} - \mathbf{r}_j). \tag{2.48}$$

Here φ_j contains information about the local orbital character. Assume again that these states are orthonormal; relaxing this assumption just converts the matrix problem to a generalized eigenvalue problem. Assume also that the position operator is diagonal in this basis with elements $\mathbf{R} + \mathbf{r}_j$. Information such as the values of V_0 and t in the above example comes from the real-space Hamiltonian

$$H_{jk}(\mathbf{R}') = \langle \phi_{0j}|H|\phi_{\mathbf{R}'k}\rangle. \tag{2.49}$$

Here $\mathbf{R}'$ is a (relative) lattice vector, possibly zero. In the example above with one orbital per unit cell in one dimension, $H(\pm a) = -t$ and $H(0) = V_0$.

We can construct states of Bloch form from these orbitals,

$$|\psi_j^{\mathbf{k}}\rangle = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot(\mathbf{R}+\mathbf{r}_j)}|\phi_{\mathbf{R}j}\rangle, \tag{2.50}$$

where the sum is over all lattice vectors. Note that the form of the exponent means that these wavefunctions *are not* periodic in $\mathbf{k}$ with the reciprocal lattice, assuming at least one $\mathbf{r}_j$ is nonzero, but pick up phase factors. This is Convention I, also known as periodic gauge; Convention II just consists in dropping the $\mathbf{k}\cdot\mathbf{r}_j$ in the exponent.

These are not yet energy eigenstates, but we can form linear combinations of them that are eigenstates and still satisfy Bloch's theorem. The Hamiltonian in this basis is diagonal in $\mathbf{k}$ and has the form

$$H_{jk}^{\mathbf{k}} = \langle \psi_j^{\mathbf{k}}|H|\psi_k^{\mathbf{k}}\rangle = \sum_{\mathbf{R}'} e^{i\,\mathbf{k}\cdot(\mathbf{R}'+\mathbf{r}_k-\mathbf{r}_j)} H_{jk}(\mathbf{R}'). \tag{2.51}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Box 4.1 discusses a famous one-dimensional example with two orbitals per unit cell using this approach.

Equation (2.45) is the first of many examples in this book of a band structure of electrons. What happened to all the high-energy states of an electron moving in vacuum? We got rid of them by working only with the variational states made up of superpositions of the atomic orbitals. If we excited an electron to high energy in a solid, it would undoubtedly go into one of the higher orbitals that we threw away. But the reason why we can think of electrons under weak fields as moving in the band, and even see Bloch oscillations that directly probe the periodic nature of crystal momentum, is that there is an *energy gap* between the lowest band and higher bands. Under weak, slowly varying fields, the electron cannot gain enough energy to jump across this gap. It is useful in the theory of metals to divide processes into "intraband," at low energies/frequencies, and "interband," once transitions become allowed, as the underlying physics can be quite different.

2.5 Dirac Band Structure of the Honeycomb Lattice

The honeycomb lattice has turned out to be one of the most popular actors in topological physics, repeatedly putting in prominent appearances. In this book, this includes the physics of graphene, Haldane's Chern insulator in Section 3.2 and Kitaev's honeycomb model in Section 7.4. In each of these, the properties of the energies and wavefunctions of noninteracting electrons with nearest, and at most next nearest neighbor, hopping are central ingredients to their topological properties. This section presents a brief account of this band structure for nearest-neighbor hopping, see, for example, the exposition in Goerbig (2011) in the context of graphene for further details.

The first thing to derive are the famous gapless Dirac points with the linear ("relativistic") dispersion in their vicinity, and the symmetric spectrum of valence and conduction band, evoking particles as well as the Dirac sea filled with negative energy antiparticles. This thus presents an instance of an emergent, and stable, relativistic Lorentz invariance.

All these features are present for the nearest-neighbor hopping problem. The honeycomb lattice has two sublattices, labeled A and B, that is, it can be thought of a Bravais lattice with a two-atom basis (Figure 3.8). The corresponding Bravais lattice is a triangular lattice, which has coordination $z = 6$: each site has six nearest-neighbors, displaced by vectors $\pm\mathbf{n}_1$, $\pm\mathbf{n}_2$ and $\pm\mathbf{n}_3 = \pm(\mathbf{n}_2 - \mathbf{n}_1)$, where $\mathbf{n}_{1,2}$ are chosen to enclose an angle $2\pi/6$. These are therefore the vectors linking each honeycomb lattice site with the sites on the same sublattice in adjacent unit cells, that is, its next-nearest neigbors. Importantly, the honeycomb lattice is bipartite, that is, all nearest-neighbor bonds link sites on opposite sublattices. The

Fourier transform of the hopping matrix is hence a purely off-diagonal $n_s \times n_s = 2\times 2$ matrix with entries given by $s_{\mathbf{k}} = |s_{\mathbf{k}}| \exp(i\zeta_{\mathbf{k}}) = 1+\exp(i\mathbf{k}\cdot\mathbf{n}_1)+\exp(i\ \mathbf{k}\cdot\mathbf{n}_2)$

$$\mathcal{H} = t \begin{pmatrix} 0 & s^*(\mathbf{k}) \\ s(\mathbf{k}) & 0 \end{pmatrix} . \tag{2.52}$$

The eigenvalue spectrum of this matrix is

$$\epsilon_{\mathbf{k},\lambda} = \lambda t \sqrt{3 + 2\sum_{i=1}^{3} \cos(\mathbf{k}\cdot\mathbf{n}_i)} , \tag{2.53}$$

where $\lambda = \pm 1$ distinguishes the two bands. The corresponding eigenfunctions reside equally on the two sublattices:

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \lambda \exp(i\zeta_{\mathbf{k}}) \end{pmatrix} . \tag{2.54}$$

2.5.1 Dirac Points and Dirac Equation

The Dirac points correspond to zero energy. This occurs at the six corners of the Brioullin zone, but these form triplets related to each other by addition of reciprocal lattice vectors, so that there are only two inequivalent, but symmetry-related Dirac points. The location is conventionally denoted by $\mathbf{K}$ and $\mathbf{K}'$, where the choice $\mathbf{K}' = -\mathbf{K}$ is possible. This allows the definition of a valley index $\iota = \pm 1$ so that the Dirac points are at $\iota\mathbf{K}$, with $\mathbf{K} = (4\pi/3, 0)$ in Figure 2.1.

One can now carry out a gradient expansion by Taylor expanding $\mathcal{H}$ in Eq. 2.52 in deviations $\mathbf{q}$ from $\iota\mathbf{K}$. This yields the Dirac equation

$$\mathcal{H}^{\iota} = \iota v_F \left(q_x \sigma^x + q_y \sigma^y\right) . \tag{2.55}$$

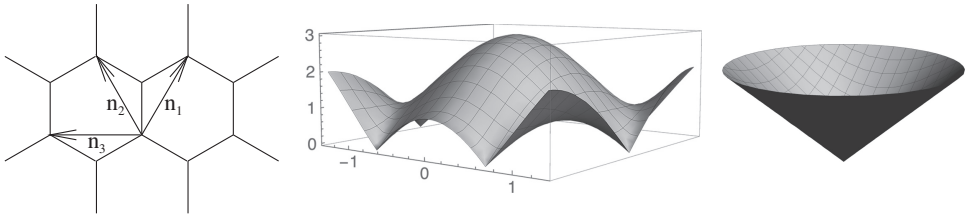

Fig. 2.1 Dirac band structure (middle) of the honeycomb lattice (left); see also Figure 3.8. The arrows denote the basis vectors $\mathbf{n}_{1,2}$ of the underlying triangular Bravais lattice. The Dirac points are visible where the energy vanishes. The Dirac cone is magnified in the right panel; x- and y-axes are the crystal momenta, labeled in units of π. Only the positive energy band is shown.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Here, $\sigma^{x,y}$ are Pauli matrices. In the present convention, the sublattice labels are interchanged between the two valleys, that is, the top spinor components at $\pm\mathbf{K}$ are associated with opposite sublattices; this will be important in the context of strain-induced artificial gauge fields in Chapter 8.

This gradient expansion fixes the effective Fermi velocity in terms of the nearest-neighbor hopping strength t as

$$v_F = \frac{3|t|a}{2\hbar} , \tag{2.56}$$

where in the following, as often in this book, we use a convention which sets $\hbar$ to unity; also, we will drop a in the following. This is the nearest-neighbor distance on the honeycomb lattice, which equals about 0.14 nm for graphene, while v_F turns out to be in the ballpark of 10^6 m/s.

The low-energy spectrum near the Dirac point, valid at energies $\epsilon \ll t$ much less than the total bandwidth, exhibits the desired linear dispersion

$$\epsilon_{\mathbf{q}} = \pm v_F q . \tag{2.57}$$

2.5.2 Relativistic Landau Levels

A magnetic field can now be added via minimal coupling, in the same way as for the nonrelativistic case discussed following Eq. 2.30, that is, by replacing the momentum operator $\mathbf{p}$ by $\mathbf{p} - q\mathbf{A}$. This leads to the reappearance of the ladder operators $a, a^\dagger$ from Eq. 2.33, along with the magnetic length ℓ. Implementing this starting from Eq. 2.55 yields

$$\mathcal{H}^\iota = \iota\sqrt{2}\frac{v_F}{\ell}\begin{pmatrix} 0 & a \\ a^\dagger & 0 \end{pmatrix}. \tag{2.58}$$

The corresponding eigenvalues are the relativistic Landau level energies

$$\epsilon_{n,\lambda} = \lambda\frac{v_F}{\ell}\sqrt{2n} . \tag{2.59}$$

This has the following salient features. First, the Landau level energies are no longer equidistant like in the nonrelativistic case, but now scale as $\sqrt{n}$. The reason they get "closer together" at higher energies is that the density of states grows with energy in the relativistic case. While for the standard case of a parabolic dispersion relation, $\rho(E)$ is constant, the linear dispersion leads to a linear density of states, $\rho(E) \propto E$, so that the integrated density of states up to energy E is $\rho_{\text{tot}}(E) \propto E^2$. The number of Landau levels hosting a fixed number of single particle states thus scales as $n \sim E^2$, that is, $\epsilon_n \sim \sqrt{n}$.[2]

[2] For an anisotropic honeycomb lattice, when the sum of hopping integrals in two directions equals that in the third, the two Dirac cones merge, at which point the dispersion in one direction is linear, and quadratic in the other. It is a simple exercise to work out how the Landau level energies scale with n in this case.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Second, there now are Landau levels with positive and negative energies, $\lambda = \pm 1$. In particular, this means that there is no longer a *lowest* Landau level. Instead, and third, there now exists a special *central* Landau level with $n = 0$. To see what is special about this, consider the Landau level wavefunctions, that is, the eigenvectors of $\mathcal{H}$ in Eq. 2.58. These are given by

$$|n = 0\rangle_{\mathrm{D}} = \begin{pmatrix} 0 \\ |n = 0\rangle \end{pmatrix} \tag{2.60}$$

for $n = 0$, and for $n \neq 0$ by

$$|n \neq 0\rangle_{\mathrm{D}}^{\lambda,\iota} = \frac{1}{\sqrt{2}} \begin{pmatrix} |n-1\rangle \\ \iota\lambda|n\rangle \end{pmatrix} . \tag{2.61}$$

Here, we are using the same wavefunctions as in Eq. 2.35, having suppressed the m-index, which plays the same role in either case. The important feature is that the wavefunction of the central Landau level resides on one sublattice only, A for one of the Dirac points, and B for the other. For all other Landau levels, they reside on both with equal probability.

2.6 Landau Theory of Symmetry-Breaking Phases

Most phases of matter, including solids, magnets, superfluids, and many others, can be understood in terms of "broken symmetry." At high temperature, fluctuations are induced by the requirement to maximize entropy, and these fluctuations tend to destroy order. As temperature is lowered, the energy gain from developing order can overwhelm the entropy cost from lowering disorder. A remarkable fact that only became clear after the solution of the two-dimensional Ising model by Onsager in 1948 is that this energy-entropy competition can lead to a sharp phase transition, described mathematically by a singularity in some derivative of the free energy that emerges in the thermodynamic limit (the limit of an infinite number of degrees of freedom).

For our purposes, we need a way to describe such a breaking of symmetry mathematically. Rather than try to describe every microscopic degree of freedom in a complicated interacting system, we will eventually follow Landau and introduce a classical field theory in terms of some emergent or coarse-grained field that describes the type of order we wish to study. This will lead to a very useful body of ideas collectively known as Landau–Ginzburg theory.

To start, let us first consider an Ising model on a hypercubic lattice (chain, square, cubic lattices in $d = 1, 2$ and 3). Our microscopic description is in terms of a discrete spin variable $s_i = \pm 1$ at each site, with the energy function

$$\mathcal{H} = -J \sum_{\langle ij \rangle} s_i s_j . \tag{2.62}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Here $\langle ij \rangle$ denote nearest-neighbor bonds and J is some interaction strength with units of energy in terms of which we measure the temperature, $T = 1/\beta$. The corresponding partition function is over all microstates, weighted by the Boltzmann factor $\exp[-\beta\mathcal{H}(\{s_i\})]$:

$$Z = \sum_{s_i=\pm 1} \exp\left[-\beta\mathcal{H}(\{s_i\})\right]. \tag{2.63}$$

At temperature $T = \infty$, the system is equally likely to be in any microstate. At $T = 0$, only two microstates occur: one with all spins up and one with all spins down. The surprise is that, if the lattice of spins is in more than one dimension, there is a nonzero temperature T_c, proportional to J, below which the zero-temperature description is qualitatively but not quantitatively correct.

2.6.1 Mean-Field Theory

As an explicit example, one can construct the lattice mean-field description of the Ising model. One assumes that each spin "sees" a mean field due to the average magnetization, m, of its z neighbors, where z is known as the coordination number of the lattice. The total mean-field H_{eff} then consists of the exchange field zJm and the applied field h. This mean field induces a magnetization for the spin in question. Self-consistency is achieved by demanding that the induced magnetization m_i equals that of its neighbors, m: computing the expectation value of the magnetization of a single spin in a field gives

$$m = \tanh(\beta H_{\text{eff}} m) = \tanh(\beta z J m + \beta h m), \tag{2.64}$$

where k_B is Boltzmann's constant. The behavior of this self-consistent equation changes at a special value T_c of temperature with $k_B T_c = zJ$. The full significance of this temperature will require more explanation, but for now, note that for $h = 0$, $T > T_c$ only has the solution $m = 0$, while $T < T_c$ allows for two solutions with lower free energy, $|m| \neq 0$. Thus the mean-field transition temperature is $T_c = zJ/k_B$ or $\beta_c = (zJ)^{-1}$.

Physicists say that the system breaks symmetry below T_c and picks out a particular sign of the average spin m, where the angle brackets denote thermal averaging. Mathematicians have more satisfactory definitions because the average spin strictly speaking is always zero in the canonical ensemble: one can look at either whether there is a nonzero correlation function $\langle s_i s_j \rangle$ as i and j become infinitely far apart, or look for a singularity in some derivative of the free energy (in a first derivative for a first-order transition, in some higher derivative for a second-, or higher-, order transition).

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Even if m is always zero in terms of the Boltzmann sum, physical systems do actually break symmetry, chiefly for dynamical reasons: for example, a bar magnet of iron will in principle explore the whole phase space and flip its north and south poles, but the time it takes to do so may be larger than the age of the universe, let alone the time for which the bar magnet will physically exist. Hence we will mostly be content to discuss broken symmetry as real, for example, $m \neq 0$ in the Ising model, even if that is somewhat sloppy mathematically.

2.6.2 Symmetry Breaking and Universality

The mean-field theory of the previous section appears somewhat ad hoc: Why can one replace all neighbors by a mean-field? How does one guess the nature of candidate symmetry breakings? From here to develop a full description of Landau–Ginzburg theory and the renormalization group is a serious undertaking which we are in large part excused of as this book is concerned with topological physics. However, for two reasons we nonetheless present a conceptual outline here. First, knowledge of Landau–Ginzburg theory is necessary to appreciate how much topological physics is different. Second, it will lead us to encounter Berezinskii–Kosterlitz–Thouless physics, one of the important topics linking local and topological forms of order. The reader interested in a more detailed and comprehensive treatment is referred to Cardy (1996) or other texts on advanced statistical physics.

Our conceptual outline is intended to expose the basic mathematical ingredients and give a couple of examples illustrating in particular how seemingly disparate systems can end up with similar universal properties as suggested in Chapter 1, while seemingly similar systems turn out to behave fundamentally differently. Mathematically, a powerful way to understand the broken symmetry is in terms of two symmetry groups: G, the symmetry group of the high-temperature phase, and H, the residual symmetry group that survives in the low-temperature phase.

For instance, in the case of a magnet on a lattice, such as (2.62), G consists of the full symmetry of the lattice – translations, reflections, rotations, ... – together with those of the internal (e.g., spin) space such as inversion for Ising spins, $\mathcal{I}_I : s \to -s$, or rotations in the plane for XY spins,

$$\mathcal{R}_{XY} : \begin{pmatrix} s_x \\ s_y \end{pmatrix} \to \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} s_x \\ s_y \end{pmatrix} . \tag{2.65}$$

The above case, where it was the Ising inversion symmetry, $\mathcal{I}_I$, which was discarded upon ordering, is a particularly simple case leaving all the lattice symmetries intact. (Generally, there will be symmetry operations combining lattice and internal symmetries, e.g., in the presence of spin-orbit coupling.)

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

One then defines the order parameter manifold as the quotient $M = G/H$, where dividing by H (taking cosets of H in G) means that we identify two elements of G that differ by multiplication by an element of H; note that unlike G and H, M is in general no longer a group. The idea of this coset is that H is still an unbroken symmetry of the low-temperature phase, so we should think of the system in this phase as invariant under H, which is true if the system is described as a set of microscopic states and the set is unchanged under the action of symmetries in H.

The notion of an order parameter is basic in Landau theory: it is what we use to model all the complicated microscopic states in terms of one, or a few, macroscopic variables. The idea of the order parameter manifold is that, for many interesting phenomena, we do not care especially about the magnitude of the order parameter itself. We care instead about the set of distinguishable low-temperature states at an arbitrary temperature in the ordered phase, which is exactly M. Note that M does not have any information about the magnitude of the order parameter: at every temperature in the uniaxial ordered phase of the isotropic (Heisenberg) ferromagnet, for example, the order parameter manifold is the same sphere: the high-temperature symmetry is $G = \mathrm{SO}(3)$ (rotations in 3D), the low-temperature symmetry is $H = \mathrm{SO}(2)$ (rotations around the axis of the order), and the order parameter manifold is $M = \mathrm{SO}(3)/\mathrm{SO}(2) = S^2$, the unit sphere in three dimensions. One reason the set M is important will become clear when we discuss topological defects later in this chapter.

A basic tenet of Landau–Ginzburg (LG) theory states that, if G is a subgroup of H, the transition where the order parameter vanishes can (but does not have to) be continuous. In this case its universal properties are predetermined generically (i.e., with high probability in the mathematical sense: to escape this, one needs to fine-tune parameters so as to end up, e.g., at a multicritical point). If this condition is not met, the transition will generically be first-order. The intuition behind this famous prediction of Landau's, for which there may not be a rigorous proof, is as follows: the two phases have symmetries that are not in a subset/superset relation, which means that we cannot say one has more or less symmetry than the other; hence it would be fine-tuned for the continuous breaking of one symmetry (and appearance of order) to take place at the same point in parameter space as the continuous restoration of the other symmetry (and disappearance of order).

Symmetry considerations enter the construction of the basic quantity underlying the LG analysis, the LG functional $\mathcal{L}_{\mathrm{LG}}$. It starts from the order parameter field, $m(\mathbf{r})$, as a function of spatial location $\mathbf{r}$, and constructs an expression that is fully symmetric under G, the group describing the symmetries of the high-temperature (also known as disordered) phase.

Such a recipe is not very practicable in general, but two approximations are made which make it more tractable. As universality occurs in continuous phase

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

transitions, there is a regime where the order parameter field $m(\mathbf{r})$ is small, so that one can Taylor expand $\mathcal{L}_{\rm LG}$ in powers of m. Also, as one is interested in low-energy/long-wavelength physics, one can use a systematic gradient expansion around uniform solutions, rather than allowing arbitrarily fast spatial variations.

Concretely, in the case of the Ising ferromagnet, the order parameter is a simple scalar, and $\mathcal{I}_I : m \to -m$ forbids terms in $\mathcal{L}_{\rm LG}$ which are odd in m. Then, writing the average value of $m(\mathbf{r})$ as m, and expanding in small fluctuations around it, one obtains

$$\mathcal{L}_{\rm LG} = \int d^d\mathbf{r} \left(a_0 + a_1 m^2 + a_2 m^4 + \ldots + b_0(\nabla m)^2 + \ldots\right) . \tag{2.66}$$

We have assumed rotational symmetry in the form of the gradient term; more generally, we would omit terms with an odd power of spatial derivatives if the system is statistically invariant under spatial inversion $\mathbf{r} \to -\mathbf{r}$. For the XY ferromagnet, little changes: the main difference arises from the fact that $\mathbf{m} = (m_x, m_y) = m(\cos\theta, \sin\theta) \sim me^{i\theta}$ is now a vector, so that m^2 needs to be understood as the expression $|\mathbf{m}|^2$ invariant under $\mathcal{R}_{XY}$. This innocuous change – the mean-field equation (2.64) remains basically unchanged – turns out to change the critical behavior completely, as we will see below.

By contrast, an entirely different Ising problem turns out to be much more like the XY magnet, and the next few lines describe how this comes about while also illustrating the construction of $\mathcal{L}_{\rm LG}$ in a slightly more complex setting.

Consider an Ising antiferromagnet (2.62), $J < 0$, defined on a triangular lattice. There is generally no simple way rigorously to obtain the nature of the ordered phase which is reached as the temperature is lowered, but a guess is provided by the minimum of the interaction energy (2.62) upon Fourier transforming. Indeed, for the case of the ferromagnet on the hypercubic lattice, one finds $\mathcal{H}(q) = -J \sum_{\alpha=1}^{d} \cos(q_\alpha)$, which is manifestly minimized at $q_\alpha \equiv 0$, indicating a preference for a uniform order with every spin having the same average $s_i \sim m$: this is just the ferromagnetic order parameter.

For the triangular Ising antiferromagnet

$$\mathcal{H}(q) = |J| \left[\cos q_x + 2\cos\left(\frac{q_x}{2}\right)\cos\left(\frac{\sqrt{3}q_y}{2}\right)\right], \tag{2.67}$$

which is minimized for $(q_x, q_y) = \pm\mathbf{K} = \left(\pm\frac{4\pi}{3}, 0\right)$. There are thus two minima, $s^{\pm}$, at the corners of the Brillouin zone, which are spatially modulated.

To construct a symmetry-broken spin configuration, one expands $m(\mathbf{r})$ as before, with the requirement of real values for the spin field yielding $m(\mathbf{r}) = m_+(\mathbf{r})s^+ + m_-(\mathbf{r})s^-$ with $m_+ = m_-^* = me^{i\theta}$.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

This feeds into the construction of $\mathcal{L}_{\rm LG}$: inversion again removes odd terms in m. In addition, translations now act nontrivially, with $\mathcal{T}_x : m_\pm \to \exp\left(\pm\frac{4\pi}{3}i\right) m_\pm$ for a unit translation along the $\hat{x}$-axis. The phase factors can only be cancelled for combinations $m_+ m_- = m^2$, or for "cubic" powers $m_\pm^{3\cdot n}$. Here $n = 1$ is forbidden by inversion, so that the lowest spatially uniform terms are

$$\mathcal{L}_{\rm LG} = \int d^2 r \left[a_2 m^2 + a_4 m^4 + a_6 m^6 + b_6 m^6 \cos(6\theta) + \dots \right]. \tag{2.68}$$

Note that we all of a sudden have an XY degree of freedom $me^{i\theta}$ in an Ising system! Indeed, we have ended up with an XY theory with a six fold clock term, $b_6 m^6 \cos(6\theta)$. This is one example of the origin of universality: the long-wavelength properties do not care much about the microscopic nature of the spin – Ising or XY – but rather about the order parameter symmetry, which does inherit items such as inversion symmetry but can nonetheless exhibit entirely different emergent symmetry properties.

2.6.3 Quantum and Statistical Path Integrals

Landau–Ginzburg Theory

With this in hand, we now turn to the construction of field theories to describe conventional symmetry breaking and its universality. There are two field theories we will deal with. The first is Landau–Ginzburg theory, now treated beyond the mean-field approximation. This uses the $\mathcal{L}_{\rm LG}$ derived above to define a partition function

$$Z_{\rm LG} = \int Dm \exp\left[-\mathcal{L}_{\rm LG}\right]. \tag{2.69}$$

This is a statistical integral over the order parameter field $m(r)$. You may be familiar with sums over spin configurations in an Ising model, for example. It is often convenient to work in the continuum and integrate over configurations of a continuous field, and such integrals are very similar mathematically to the path integrals appearing in quantum field theory.

Here the measure of the integral can be defined more precisely in Fourier space, where omitting high-wavevector components is typically necessary for a sensible theory. In condensed matter, this makes physical sense as a short-distance cutoff below which the field m is not meaningful: m cannot vary on a scale shorter than the separation between neighboring sites. This is a vastly simplified version of the original problem in at least two ways: considering the integration over the coarse-grained order parameter field $m(\mathbf{r})$, and we are not doing any microscopic calculation of the coefficients that appear in the expansion.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

A remarkable fact is that the above Landau–Ginzburg theory can be not just qualitatively correct but actually exact for some properties, such as critical exponents near second-order phase transitions, even without a microscopic calculation of the coefficients. Such properties are referred to as universal: they depend on symmetry and dimensionality but little else. For example, the liquid–gas critical point in the phase diagram of water has the same critical exponents as the Ising phase transition in three dimensions. As a mathematical example of where universality comes from, the terms beyond a_2 in the above energy turn out not to impact critical exponents and selected other properties, as long as the lower-order coefficients have appropriate signs. We will study one or two examples of critical points later, concentrating on examples where topological considerations are important.

Nonlinear σ Model

Landau–Ginzburg theory was motivated as a high temperature expansion taking advantage of the smallness of the order parameter m. An alternate field theory, the nonlinear σ-model, can be viewed as an expansion starting from zero temperature. We will concentrate on systems in which the zero-temperature phase breaks a continuous symmetry, so that the set of inequivalent order parameter states M is a continuous manifold; this includes, for example, Heisenberg and XY magnets, in which $M = S^2$ and $M = S^1$, respectively, but not Ising magnets, where M is a set with two elements. Here S^d is the d-dimensional surface of the unit sphere in $d+1$-dimensional real space. For an XY magnet[3], we can label a ground state simply by an angle θ between 0 and 2π (the order parameter has a magnitude Δ as well, but all ground states have the same magnitude of the order parameter by symmetry).

When the temperature is slightly increased, fluctuations of the order parameter will take place. The nonlinear σ-model is a theory that ignores fluctuations of the order parameter *magnitude* but captures fluctuations in its *direction*, which are lower in energy or softer. More precisely, the nonlinear σ-model into a symmetric space $M = G/H$ is defined as a path integral over an M-valued field. For the XY case above, this can be written simply in terms of a spatially varying angle $\theta(\mathbf{r})$:

$$Z_{\mathrm{NL}\sigma\mathrm{M}} = \int (D\theta)\, e^{-\int d^d\mathbf{r}\, g(\nabla\theta)^2}, \tag{2.70}$$

where we have incorporated β into the definition of the coupling g. We will return to this model once we have said a bit more about topological defects in Section 2.8; it will turn out (Box 2.3) that for our XY example in two spatial dimensions, the

[3] We choose the example of $M = \mathrm{U}(1) \cong \mathrm{SO}(2)$ here for a reason. It turns out that, for the nonlinear σ-model to include gapless excitations, the form of the theory can become more complicated. Namely, for Lie groups more complicated than U(1), an additional term of topological origin is required, leading to the Wess–Zumino–Witten model that we discuss in Box 4.2.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

physics depends crucially on vortices, the simplest kind of topological defect, and in fact shows a phase transition that would not be present if, hypothetically, the field θ were not periodic and Eq. 2.70 became just the massless Gaussian model or the free scalar field theory.

We have written both of the above field theories in a classical or Euclidean representation, where there is a positive weight on each field configuration. A natural question is how the partition function integral in such a theory is related to the quantum path integral that may be familiar from an advanced course in quantum mechanics, which in general has a complex integrand. The easiest example of the analytic continuation to imaginary time that connects the two types of path integrals is for the harmonic oscillator. Its partition function at a finite temperature T is

$$Z_{\text{harmonic}} \approx \int dx(\tau)\, \exp\left[-\int_0^\beta d\tau\, \left(\dot{x}^2(\tau)/2m + kx^2/2 \right) \right], \tag{2.71}$$

where there are periodic boundary conditions on $x(\tau)$: $x(\beta) = x(0)$. A worthwhile calculation (hint: simplify the integral by considering Fourier components of $x(\tau)$) leads to the result

$$Z_{\text{harmonic}} = \frac{1}{2\sinh(\beta\hbar\omega/2)} = \sum_{n=0}^{\infty} e^{-\beta\hbar\omega(n+1/2)}, \tag{2.72}$$

where the last expression is what we would calculate from the spectrum. Now analytically continuing this calculation from imaginary time τ gives a trace of the form appearing in a quantum path integral,

$$Z_{\text{harmonic}} = \text{Tr}\, e^{-\beta H} \rightarrow \text{Tr}\, e^{-itH/\hbar} = \int dx_0\, U(x_0, t; x_0, 0), \tag{2.73}$$

where in the last step we have used the position basis to put the result in terms of matrix elements of the unitary time evolution operator U. Now the divergence of Z at real times $t = 2\pi n/\omega$, for integer n, can be simply interpreted: at these times all the energy eigenstates that appear in an arbitrary initial condition appear with exactly the same phases, so the state is (aside from an overall phase factor) exactly the initial state, the time evolution operator is the identity, and the trace diverges.

A variety of field theories of topological phases will appear in Chapter 6, and several books specifically devoted to the many other applications of field theories in condensed matter physics are listed in the appendix. Although we will not discuss the uses of this approach, it turns out that the Berry phase of a spin-half discussed in Section 2.1 is the crucial ingredient in developing a path integral for quantum spins, which can be used to understand, for example, a key difference between integer-spin and half-integer-spin antiferromagnetic chains (Auerbach, 1994).

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

2.7 Two Mathematical Approaches to Topology

This book is primarily about the physics of materials in which some property has a topological character, in the sense of being described by mathematical objects that are robust under continuous deformations. This section introduces two commonly used flavors of algebraic topology, which will be used repeatedly in physics contexts in later sections of the book. The first flavor, cohomology, can be motivated intuitively by the question of which integrals over a path (or surfaces, etc.) are invariant of smooth changes of the path but sensitive to its topology. The second flavor, homotopy, was actually the first use of topology in condensed matter physics: it appears naturally in the classification of topological defects in symmetry-breaking phases, which we review.

Homotopy and cohomology are often sensitive to the same topological character of a manifold, and consequently cohomological integrals are often used by physicists to compute the homotopy class of some material or configuration. A particularly important tool in Chapters 3 and 4 will be integrals constructed from the Berry phase of a material's Bloch states. The particular form of these integrals will make a little more sense once we see how in this chapter certain objects are naturally suited to compute topological properties. As a warmup for our topological discussion, Box 2.2 gives an example of how integrals over a local geometrical object, the Gaussian curvature, become topological. Here the emergence of topology may be a little easier to picture than when the local geometrical object emerges from quantum mechanics via the Berry phase.

Box 2.2 Topology from Geometry: The Gauss–Bonnet Theorem

A topologist has been described as "a mathematician who can't tell the difference between a doughnut and a coffee cup." As a prelude to the connections between geometry and topology, we start by discussing an integral that will help us classify two-dimensional compact manifolds (surfaces without boundaries) embedded smoothly in three dimensions. For our purposes, a manifold is a d-dimensional surface that locally "looks like" $\mathbb{R}^d$. The integral we construct is topologically invariant in that if one such surface can be smoothly deformed into another, then the two will have the same value of the integral. The integral can't tell the difference between the surface of a coffee cup and that of a doughnut, but it can tell that the surface of a doughnut (a torus) is different from a sphere. Similar connections between global geometry and topology appear frequently.

We start with a bit of local geometry. Given our $d = 2$ surface embedded in $d = 3$ Euclidean space, we can choose coordinates at any point on the surface so that the $(x, y, z = 0)$ plane is tangent to the surface, which can locally be specified by a single

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

function $z(x, y)$. We choose $(x = 0, y = 0)$ to be the given point, so $z(0, 0) = 0$. The tangency condition is

$$\left.\frac{\partial z}{\partial x}\right|_{(0,0)} = \left.\frac{\partial z}{\partial y}\right|_{(0,0)} = 0. \tag{2.74}$$

Hence we can approximate z locally from its second derivatives:

$$z(x, y) \approx \frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} \frac{\partial^2 z}{\partial^2 x} & \frac{\partial^2 z}{\partial x \partial y} \\ \frac{\partial^2 z}{\partial y \partial x} & \frac{\partial^2 z}{\partial^2 y} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}. \tag{2.75}$$

The Hessian matrix that appears in the above is real and symmetric. It can be diagonalized and has two real eigenvalues λ_1, λ_2, corresponding to two orthogonal eigendirections in the (x, y) plane. The geometric interpretation of these eigenvalues is simple: their magnitude is an inverse radius of curvature, and their sign tells whether the surface is curving toward or away from the positive z direction in our coordinate system. To see why the first is true, suppose that we carried out the same process for a circle of radius r tangent to the x-axis at the origin. Parameterize the circle by an angle θ that is 0 at the origin and traces the circle counterclockwise, that is,

$$x = r\sin\theta, \quad y = r(1 - \cos(\theta)). \tag{2.76}$$

Near the origin, we have

$$y = r(1 - \cos(\sin^{-1}(x/r)) \approx r - r\left(1 - \frac{x^2}{2r^2}\right) = \frac{x^2}{2r}, \tag{2.77}$$

which corresponds to an eigenvalue $\lambda = 1/r$ of the matrix in Eq. 2.75.

Going back to the Hessian, its determinant (the product of its eigenvalues $\lambda_1\lambda_2$) is called the Gaussian curvature and has a remarkable geometric significance. As in the example above, the Gaussian curvature remains the signed product of inverse radii of curvature along the two perpendicular principal directions. First, consider a sphere of radius r, which at every point has $\lambda_1 = \lambda_2 = 1/r$. Then we can integrate the Gaussian curvature over the sphere's surface,

$$\int_{S^2} \lambda_1\lambda_2\, dA = \frac{4\pi r^2}{r^2} = 4\pi. \tag{2.78}$$

Beyond simply being independent of radius, this integral actually gives the same value for any manifold that can be smoothly deformed to a sphere. (More generally, scale invariance is a necessary but not sufficient for an integral to give a topological invariant.)

However, we can easily find a manifold with a different value for the integral. Consider the torus (Figure 2.2) made by revolving the circle in Eq. 2.76, with $r = 1$, around the axis of symmetry $x = t, y = -1, z = 0$, with $-\infty < t < \infty$. To compute the Gaussian curvature at each point, we sketch the calculation of the eigenvalues of the Hessian as follows. One eigenvalue is around the smaller circle, with radius of

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

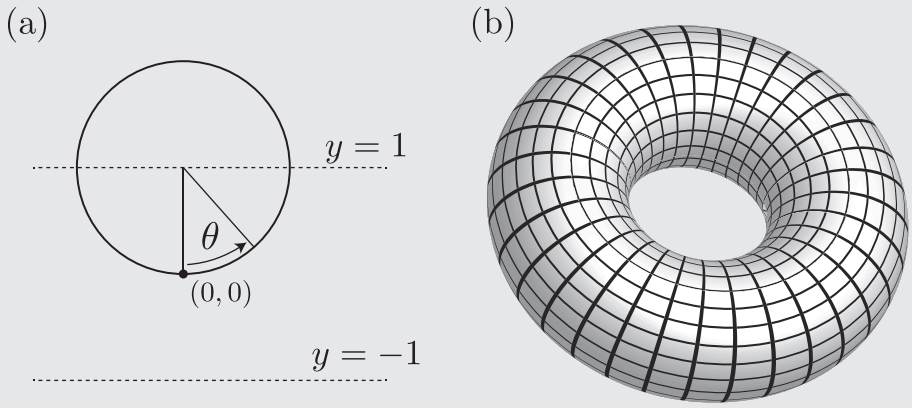

Fig. 2.2 An example of the Gauss–Bonnet theorem. (a) Coordinates of circle of radius 1 as described in text. Revolving this circle around the line $y = 1$ leads to a sphere, while revolution around the line $y = -1$ leads to a torus like that shown in (b). (b) This torus has zero total Gaussian curvature: the negative curvature at points on the inside (i.e., around the hole in the doughnut) compensates the positive curvature on the outside. The circles on the torus's surface indicate the principal directions.

curvature r: $\lambda_1 = 1/r = 1$. Then the second eigenvalue must correspond to the perpendicular direction, which has a radius of curvature that depends on the angle θ around the smaller circle (we keep $\theta = 0$ to indicate the point closest to the axis of symmetry). The distance from the axis of symmetry is $2 - \cos\theta$, so we might have guessed $\lambda_2 = (2 - \cos\theta)^{-1}$, but there is an additional factor of $\cos\theta$ that appears because of the difference in direction between the surface normal and this curvature. So our guess is that

$$\lambda_2 = -\frac{\cos\theta}{2 - \cos\theta}. \tag{2.79}$$

As a check and to understand the sign, note that this predicts a radius of curvature 1 at the origin and other points closest to the symmetry axis, with a negative sign in the eigenvalue indicating that this curvature is in an opposite sense as that described by λ_1. At the top, the radius of curvature is 3 and in the same sense as that described by λ_1, and on the sides, λ_2 vanishes because the direction of curvature is orthogonal to the tangent vector.

Now we compute the curvature integral. With ϕ the angle around the symmetry axis, the curvature integral is

$$\int_{T^2} \lambda_1\lambda_2\, dA = \int_0^{2\pi} d\theta \int_0^{2\pi} (2-\cos\theta)\, d\phi\, \lambda_1\lambda_2 = \int_0^{2\pi} d\theta \int_0^{2\pi} d\phi\, (-\cos\theta) = 0. \tag{2.80}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Again this zero answer is generic to any surface that can be smoothly deformed to the torus. The general result (the Gauss–Bonnet formula) of which the above are examples is

$$\int_S \lambda_1 \lambda_2 \, dA = 2\pi\chi = 2\pi(2-2g), \tag{2.81}$$

where χ is a topological invariant known as the Euler characteristic and g is the genus, essentially the number of holes in the surface.[a] If we go beyond the sphere and torus to a manifold with boundaries, the Euler characteristic becomes $2-2g-b$, where b is the number of boundaries: one can check this by noting that by cutting a torus, one can produce two annuli (by slicing a bagel) or alternately one cylinder with two boundaries (by slicing what one of us calls a bundt cake and the other a Gugelhupf). The Gauss–Bonnet formula and Euler characteristic are examples of a general principle: we will encounter other examples where a topological invariant is expressed as an integral over a local quantity with a geometric significance.

[a] A good question is why we write the Euler characteristic as $2-2g$ rather than $1-g$; one way to motivate this is by considering polygonal approximations to the surface. The discrete Euler characteristic $V-E+F$, where V, E, F count vertices, edges, and faces, is equal to χ. For example, the five Platonic solids all have $V-E+F=2$.

2.7.1 Invariant Integrals along Paths: Exact Forms

As our first example of a topological property, let's ask about making line integrals along paths (not path integrals in the physics sense, where the path itself is integrated over) that are nearly independent of the precise path: they will turn out to depend in some cases on topological properties (homotopy or cohomology). We will assume throughout, unless otherwise specified, that all functions are smooth (i.e., $\mathbb{C}^\infty$, meaning derivatives of all orders exist).

First, suppose that we deal with paths on some open set U in the two-dimensional plane $\mathbb{R}^2$.[4] We consider a smooth path $(u(t), v(t))$, where $0 \le t \le 1$ and the endpoints may be different.[5]

Now let $f(x, y) = (p(x, y), q(x, y))$ be a two-dimensional vector field that lets us compute line integrals of this path:

$$W = \int_0^1 dt \left(p\frac{du}{dt} + q\frac{dv}{dt} \right), \tag{2.82}$$

where p and q are evaluated at $(x(t), y(t))$.

[4] Open set: some region of nonzero size around each point in the set is also in the set.

[5] To make these results more precise, we should provide for adding one path to another by requiring only piecewise smooth paths, and require that u and v be smooth in an open set including $t \in [0, 1]$. For additional rigor, see the first few chapters of Fulton (1995).

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Mathematical note: in more fancy language, f is a differential form, a 1-form to be precise. All that means is that f is something we can use to form integrals over paths that are linear probes of the tangent vector of the path. Another way to state this is that the tangent vector to a path, which we call a vector, transforms naturally in an opposite way to the gradient of a function, which we call a covector (a linear functional on vectors).[6] We will say a bit more about such forms in a moment.

Our first goal is to show that the following three statements are equivalent: (a) W depends only on the endpoints $(u(0), v(0))$ and $(u(1), v(1))$ (Figure 2.3a); (b) $W = 0$ for any closed path; (c) f is the gradient of a function g: $(p, q) = (\partial_x g, \partial_y g)$. The formal language used for (c) is that f is an *exact form*: $f = dg$ is the differential of a 0-form (a smooth function) g.[7]

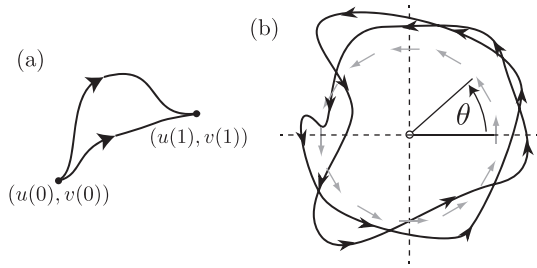

Fig. 2.3 The plane and punctured plane as examples of the cohomology of differential forms. (a) On the plane $\mathbb{R}^2$, the integrals of a gradient along a path depend only on the endpoints, so the two paths shown give the same result. The integral of a gradient along any closed path is zero. (b) On the punctured plane $\mathbb{R}^2 - (0, 0)$, closed paths can give nonzero integrals of the 1-form shown with gray arrows, which is singular at the origin. The path shown circles the origin twice and has winding number 2. As a simple example of continuous versus discrete calculation of topological invariants, the winding number could be computed either using the integral of the form shown, which locally is the gradient of the angle $\theta = \tan^{-1}(y/x)$, or by counting the signed number of crossings of a radius, such as the positive x-axis, by the path.

[6] To convince yourself that this is true, think about how both transform under a rotation $x_i' = R_{ij} x_j$ on the underlying space. A tangent vector transforms by matrix multiplication with R, i.e., $\mathbf{v}' = R\mathbf{v}$, while a gradient vector transforms differently: $\frac{\partial f(\mathbf{x}')}{\partial x_i'} = \frac{\partial f}{\partial x_j}\frac{\partial x_j}{\partial x_i'} = \nabla_x f R^{-1}$. This distinction is very similar to that between kets (state vectors) and bras (linear functionals on state vectors) in quantum mechanics.

[7] As an aside, we note that in an undergraduate thermodynamics course, the reader may have come across some statements about the difference between the left-hand and right-hand sides of the equation describing changes of the internal energy U: $dU = dQ - dW$, where the differentials on the right-hand side are sometimes denoted by slightly different symbols than d. The mathematical statement is that while the function of state dU is an exact form, the heat added and work done, dQ and dW are not. We are not aware of any treatments of topological phenomena in this context.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

Note that (c) obviously implies (a) and (b), since then $W = g(u(1), v(1)) - g(u(0), v(0))$. To show that (b) implies (a), suppose (b) is true and (a) is not. Then there are two paths γ_1, γ_2 that have different integrals but the same endpoints. Form a new path γ so that, as t goes from 0 to $\frac{1}{2}$, γ_1 is traced, and then as t goes from $\frac{1}{2}$ to 1, γ_2 is traced opposite its original direction (now you can see why piecewise smooth paths are needed if one wants to be rigorous). Then this integral is nonzero, which contradicts (b).

It remains to show that (a) implies (c). Define $g(x, y)$ as equal to 0 at $(0, 0)$, or some other reference point in U if U does not include the origin. Everywhere else, set g equal to the W obtained by integrating over an arbitrary path from $(0, 0)$ to the final point, which by (a) is path-independent. (If U is not connected, then carry out this process on each connected component.) We will show that $\partial_x g = p$, and the same logic then implies $\partial_y g = q$. We need to compute

$$\partial_x g = \lim_{\Delta x \to 0} \frac{g(x + \Delta x, y) - g(x, y)}{\Delta x}. \tag{2.83}$$

We can obtain g by any path we like, so let's take an arbitrary path to define $g(x, y)$, then add a short horizontal segment to that path to define the path for $g(x + \Delta x, y)$. The value of the integral along this extra horizontal segment converges to $p(x, y)(\Delta x)$, as needed.

It turns out that the above case is simple because the plane we started with is topologically trivial. Before proceeding to look at a nontrivial example, let us state one requirement on f that is satisfied whenever f is exact ($f = dg$). The fact that partial derivatives commute means that, with $f = dg = (p, q)$, $\partial_y p = \partial_x q$. We can come up with an elegant notation for this property by expanding our knowledge of differential forms.

Before, we obtained a 1-form f as the differential of a scalar g by defining

$$f = dg = \partial_x g \, dx + \partial_y g \, dy. \tag{2.84}$$

Note that we now include the differential elements dx, dy in the definition of f, and that 1-forms form a real vector space (spanned by dx, dy): we can add them and multiply them by scalars. To obtain a 2-form as the differential of a 1-form, we repeat the process: writing $f = f_i dx_i$ (with $x_1 = x, x_2 = y, f_1 = p, f_2 = q$)

$$df = \sum_{i,j} \frac{\partial f_i}{\partial x_j} \, dx_j \wedge dx_i, \tag{2.85}$$

where the $\wedge$ product (the wedge product, or exterior product) between differential forms satisfies the rule $dx_i \wedge dx_j = -dx_j \wedge dx_i$, which implies that if any coordinate

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

appears twice, then we get zero: $dx \wedge dx = 0$. For some intuition about why this anticommutation property is important, note that in our 2D example,

$$df = (\partial_x f_y - \partial_y f_x) dx \wedge dy, \tag{2.86}$$

so that the function appearing in df is just the curl of the 2D vector field represented by f. So our statement about partial derivatives commuting is just the statement that if $f = dg$, then $df = 0$, or that the curl of a gradient is zero. We label any 1-form satisfying $df = 0$ a *closed form*. While every exact form is also closed, we will see that not every closed form is exact, with profound consequences.

2.7.2 Locally Invariant Integrals along Paths: Closed Forms and Cohomology

As an example of nontrivial topology, we would now like to come up with an example where integrals over paths are only path-independent in a limited topological sense: the integral is the same for any two paths that are *homotopic*, one of the fundamental concepts of topology (to be defined in a moment). Basically, two paths are homotopic if one can be smoothly deformed into another. Consider the vector field

$$f = (p, q) = \left(-\frac{y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right) = \frac{-y dx + x dy}{x^2 + y^2}, \tag{2.87}$$

where in the second step we have written it using our 1-form notation. This vector field is well defined everywhere except the origin (Figure 2.3b). This 1-form looks locally like the gradient of $g = \tan^{-1}(y/x)$ (which just measures the angle in polar coordinates), but that function can only be defined smoothly on some open sets. For example, in a disc around the origin, the 2π ambiguity of the inverse tangent prevents defining g globally.

So if we have a path that lies entirely in a region where g can be defined, then the integral of this 1-form over the path will give the change in angle between the starting point and end point $g(u(1), v(1)) - g(u(0), v(0))$. What about other types of paths, for example, paths in $\mathbb{R}^2 - (0, 0)$, the 2D plane with the origin omitted, that circle the origin and return to the starting point? We can still integrate using the 1-form f, even if it is not the gradient of a scalar function g, and will obtain the value $2\pi n$, where n is the winding number: a signed integer that describes how many times the closed path $(u(t), v(t))$ circled the origin as t went from 0 to 1.

Now this winding number does not change as we make a small change in the closed path, as long as the path remains in $\mathbb{R}^2 - (0, 0)$. What mathematical property

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

of f guarantees this? Above we saw that any exact 1-form (the differential of a scalar function) is also closed. While f is not exact, we can see that it is closed:

$$df = \left(\partial_x \frac{x}{x^2+y^2}\right) dx \wedge dy + \left(\partial_y \frac{-y}{x^2+y^2}\right) dy \wedge dx = \frac{2-2}{x^2+y^2} dx \wedge dy = 0. \tag{2.88}$$

In other words, $(-y, x)/(x^2 + y^2)$ is curl-free (irrotational), while $(-y, x)$ has constant nonzero curl. Now suppose that we are given two paths γ_1 and γ_2 that differ by going in different ways around some small patch dA in which the 1-form remains defined. The difference in the integral of f over these two paths is then the integral of df over the enclosed surface by Stokes's theorem, which is zero if f is a closed form.

So we conclude that if f is a closed form ($df = 0$), then the path integral is path-independent if we move the path continuously through a region where f is always defined. For an exact form ($f = dg$), the integral is completely path-independent. In the case of $\mathbb{R}^2 - (0, 0)$, the 1-form in Eq. 2.87 is locally but not completely path-independent. Both closed forms and exact forms are vector spaces (we can add and multiply by scalars), and typically infinite-dimensional, but their quotient as vector spaces is typically finite-dimensional. (The quotient of a vector space A by a vector space B is the vector space that identifies any two elements of A that differ only by an element of B as discussed in the context of local order parameters earlier in this chapter.) A basic object in cohomology is the first de Rham cohomology group (a vector space is by definition a group under addition),

$$H^1(M) = \frac{\text{closed 1-forms on } M}{\text{exact 1-forms on } M} = \frac{Z^1(M)}{B^1(M)}. \tag{2.89}$$

If you wonder why the prefix "co-" appears in cohomology, there is a dual theory of linear combinations of curves, surfaces, and so forth, called homology, in which the differential operator in de Rham cohomology is replaced by the boundary operator. However, while arguably more basic mathematically, homology seems to crop up less frequently in physics.

An even simpler object is the zeroth de Rham cohomology group. To understand this, realize that a closed 0-form is one whose gradient is zero, that is, one that is constant on each connected component of U. There are no (-1)-forms and hence no exact 0-forms. So the zeroth group is just $\mathbb{R}^n$, where n is the number of connected components.

Note that there are many different ways to compute the winding number of a path, for example either continuously, by the integral of the 1-form, or discretely, by counting the signed number of crossings of the positive x-axis. Many of the more sophisticated topological invariants appearing later in this book likewise can be

computed either continuously through an integral or discretely through a counting process, but often the integral is more fundamental in the sense that it is what arises directly from the microscopic expression for an observable quantity.

We can show that $H^1 = \mathbb{R}$ for the unit circle S^1 using the angle form f in Eq. 2.87, by showing that this form (more precisely, its equivalence class up to exact forms) provides a basis for H^1. Given some other form $\tilde{f}$, we use the unit circle path, parameterized by an angle θ going from zero to 2π, to define

$$c = \frac{\int_0^{2\pi} \tilde{f}}{\int_0^{2\pi} f}. \tag{2.90}$$

Now $\tilde{f} - cf$ integrates to zero. We can define a function g via

$$g(\theta) = \int_0^{\theta} (\tilde{f} - cf). \tag{2.91}$$

Now g is well defined and periodic because of how we defined c, and $\tilde{f} = cf + dg$, which means that $\tilde{f}$ and cf are in the same equivalence class as their difference dg is an exact form. We say that $\tilde{f}$ and cf are cohomologous because they differ by an exact form. So cf, $c \in \mathbb{R}$, generates H^1, and $H^1(S^1)$ is isomorphic to $\mathbb{R}$. With a little more work, one can show that $\mathbb{R}^2 - (0, 0)$ also has $H^1 = \mathbb{R}$.

We close this section with a few more remarks on cohomology that are not essential on a first reading. Actually we can connect the results of this section to the previous one: a general expression for the Euler characteristic is

$$\chi(M) = \sum_i (-1)^i \dim H^i(M) = \sum_i (-1)^i \dim \frac{Z^i(M)}{B_i(M)}. \tag{2.92}$$

The dimension of the ith cohomology group is called the ith Betti number (to be pedantic, the Betti numbers are defined for homology rather than cohomology, but these are related by a simple duality). There is a compact way to express the idea of cohomology and homology that lets us introduce some notation and terminology that comes in useful later. If Ω_r is the vector space of r-forms, and C_r is the dual space of r-chains (here duality means that an r-chain can be viewed as an r-dimensional path on which an r-form gives a number), then the action of the boundary operator ∂ and the differential d (i.e., exterior derivative) is as follows:

$$\begin{aligned} &\longleftarrow C_r \xleftarrow[\partial_{r+1}]{} C_{r+1} \xleftarrow[\partial_{r+2}]{} C_{r+2} \longleftarrow \\ &\longrightarrow \Omega_r \xrightarrow[d_{r+1}]{} \Omega_{r+1} \xrightarrow[d_{r+2}]{} \Omega_{r+2} \longrightarrow . \end{aligned} \tag{2.93}$$

The kernel of a linear map like the differential is defined as the set of elements in the initial (source) space taken to zero, and the image is the set of elements in the

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

final (target) space that are images under the map of some point in the initial space. Thus our definitions earlier in this section can be summarized by saying that exact forms are the image of the differential map, and closed forms are the kernel. The rth cohomology group is the quotient $\ker d_{r+1}/\mathrm{im}\, d_r$, and the rth homology group is $\ker \partial_r/\mathrm{im}\, \partial_{r+1}$.

The duality relationship is provided by Stokes's theorem. Recall that this theorem relates the integral of a form over a boundary to the integral of the differential of the form over the interior. In terms of the linear operator (f, c) that evaluates the form f on the chain c, the theorem has the compact expression

$$(f, \partial c) = (df, c). \tag{2.94}$$

Now we move on to a different type of topology that is perhaps more intuitive and will be useful for our first physics challenge: how to classify defects in ordered systems.

2.7.3 Winding Numbers and Homotopy

What if we did not want to deal with smooth functions and calculus? An even more basic type of topology, homotopy theory, can be defined without reference to calculus, differential forms, and so on (although in physics the assumption of differentiability is usually applicable). Suppose that we are given a continuous map from $[0, 1]$ to a manifold M such that 0 and 1 get mapped to the same point; we can think of this as a closed curve on M. We say that two such curves γ_1, γ_2 are homotopic if there is a continuous function (a homotopy) f from $[0, 1] \times [0, 1]$ to M that satisfies

$$f(x, 0) = \gamma_1(x), \quad f(x, 1) = \gamma_2(x). \tag{2.95}$$

Intuitively, f describes how to smoothly distort γ_1 to γ_2. Now homotopy is an equivalence relation and hence defines equivalence classes: $[\gamma_1]$ is the set of all paths homotopic to γ_1. Furthermore, concatenation of paths (i.e., tracing one after the other) defines a natural group structure on these equivalence classes: the inverse of any path can be obtained by tracing it in the opposite direction. (To be precise, one should define homotopy with reference to a particular point where paths start and end; for a symmetric space where all points are basically equivalent, which is the usual case in physics, this is unnecessary.) We conclude that the equivalence classes of closed paths form a group $\pi_1(M)$, called the fundamental group or first homotopy group. Higher homotopy groups $\pi_n(M)$ are obtained by considering mappings from the n-sphere S^n to M in the same way.

The homotopy groups of a manifold are not totally independent of the cohomology groups: for example, if $\pi_1(M)$ is trivial, then so is the first de Rham

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

cohomology group. The manifolds used as examples above, $\mathbb{R}^2 - (0,0)$ and S^1, both have $\pi_1(M) = \mathbb{Z}$: thus there is an integer that we can use to classify (equivalence classes of) paths. In fact this integer should be thought of just as winding number and it can be computed by the angle form given above; as in this case, it is frequently useful to use a cohomological integral to compute which homotopy class includes a given path.

So our two-dimensional examples already contains the two types of topology that occur most frequently in physics: de Rham cohomology and homotopy. A powerful relationship between them is given by the Hurewicz theorem: the first (co)homology group is the Abelianization of the first homotopy group for a connected space.[8] We will use homotopy in more detail in the following section, when we explain how it can be used to classify topological defects such as vortices in symmetry-breaking phases. Higher homotopy groups π_n will be defined there and used to classify other kinds of topological defects and configurations, for example, in the case of skyrmions in the quantum Hall effect (Section 3.7).

2.8 Topological Defects in Symmetry-Breaking Phases

Topological defects in an ordered phase can be classified using mappings from spheres in real space to the order parameter manifold M, that is, the homotopy groups $\pi_n(M)$. We will explain this result and see a number of examples. Another reasons the manifold M is useful in practice is that moving from one point in M to another is naively a soft or massless fluctuation, while changing the magnitude of the order parameter is a hard or massive fluctuation. A field-theory description that involves only the degrees of freedom in M, known for historical reasons as a nonlinear σ-model, is frequently useful at low energies for this reason.

Continuous spins also allow a number of important phenomena related to topological defects: the simplest example of this idea is a vortex in an XY model or, equivalently, a neutral superfluid. We will discuss this kind of vortex in 2D systems in the most detail, as the vortex is easily visualized and also controls the phase diagram (Box 2.3). In general, a topological defect refers to a configuration of a continuous model that cannot smoothly relax to a uniform configuration because of a nontrivial topological invariant (usually a winding number or some generalization thereof).

Vortices are the simplest example of topological defects. The local spin variable in the 2D XY model is a unit vector on the circle. Suppose that we are at low temperature so that the spin tilts only slightly from one site to the next. Then, in going around a large circle, we can ask how many times the spin winds around

[8] This theorem also gives an isomorphism between higher homology and homotopy groups if the lower homotopy groups are trivial.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

the unit circle, and define this as the winding number $n \in \mathbb{Z}$. Note that if the winding number is nonzero, then the continuum limit must break down at some point within the circle, as otherwise we would have the same angular rotation $2\pi n$ around circles of smaller and smaller radius, implying larger and larger gradients and hence infinite energy density, since the energy density is proportional to the squared gradient of the order parameter. In fact the total energy is weakly divergent as computed in Box 2.3.

Many types of topological defects are now known in various condensed matter and particle physics systems. Vortices in the XY model have integer charge or winding number, but frequently topological defects have charge taking values in a finite group like $\mathbb{Z}_2$ (the group with two elements ± 1). Finally, in addition to the thermodynamics of such defects, one can consider their dynamics: many observed properties of superfluid helium are controlled by the motion of vortex loops.

The mathematical classification of topological defects has been carried out for a variety of systems. Vortex-like defects (defects that can be circled by a loop) are related to the group $\pi_1(M)$, where M is the manifold of degenerate values of the order parameter once its magnitude has been set (for example, S^1 for XY and S^2 for Heisenberg, where S^d is the unit sphere in $d+1$ dimensions). $\pi_1(M)$ is known as the first homotopy group and is the group of equivalence classes of mappings from S^1 to M: for example, the mappings from S^1 to S^1 are characterized by an integer winding number $n \in \mathbb{Z}$, so $\pi_1(S^1) = \mathbb{Z}$, while $\pi_1(S^2) = 0$ (the group with one element) as any loop on the sphere is contractible to a point.[9]

In other words, $\pi_1(M)$ gives the set of equivalence classes up to smooth deformations of closed paths on M. The group operation on equivalence classes in the group is defined by concatenation of paths. An example of what this means physically can be understood from Figure 2.4. The order parameter winds clockwise (counterclockwise) on circling the positive (negative) vortex clockwise. Going around the rectangular path shown would wind and unwind around the order parameter manifold only slightly, with zero net winding, and this path can be continuously deformed to the path far away, which does not wind at all. The second homotopy group $\pi_2(M)$ classifies mappings from S^2 to M, and describes defects circled by a sphere, such as pointlike defects in 3D. For example, $\pi_2(S^2)$ is nonzero, and there are stable point defect configurations in 3D of Heisenberg spins (known descriptively as hedgehogs) but not of XY spins.

There are also topological configurations that do not involve defects; an example is the skyrmion of Heisenberg spins in $d = 2$, which we discuss in more detail in the context of the quantum Hall effect in Section 3.7. Not only does the skyrmion not have a core as it can be everywhere continuous; it also can be uniform at infinity. The use of $\pi_2(S^2) = \mathbb{Z}$ in two dimensions is that the skyrmion is a nontrivial map

[9] Intuitively, imagine a basketball with an unknotted rubber band stretched to form a loop on its surface. Then move the rubber band continuously to a small circle around the North Pole.

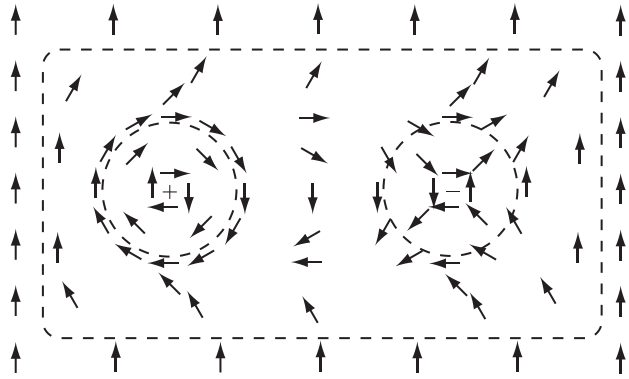

Fig. 2.4 The direction of an XY-type order parameter can be represented as a vector on the unit circle. On going around either dashed circle, the order parameter winds once around a circle, but it winds in opposite directions in going around the positive (+) and negative (−) vortices. The interaction energy between a vortex–antivortex pair is finite as the pair can be embedded in a configuration that is uniform at long distances from the center of the pair, because the topological charges of the two defects cancel: there is no nontrivial winding of the order parameter on going around the dashed rectangle enclosing both vortex cores. This illustrates what it means to say that there is an addition (group operation) in the first homotopy group. Note also that the velocity of a superfluid is given by the gradient of the order parameter, and hence the circulation is opposite around positive and negative vortices.

from the whole plane, compactified at infinity,[10] to the order parameter manifold, rather than just a map from the circle at infinity. Shankar's monopoles and other defect-free configurations in 3D of order parameter manifolds such as SO(3) are related to the group π_3 (cf. Nakahara, 1998).

Vortices in symmetry-broken phases become richer when the order parameter carries electromagnetic charge, as in the case of superconductors (Section 8.5). The most important difference will be that the superconducting vortex is not an infinitely large object like the superfluid vortex. Instead, the magnetic field through the center of the vortex is screened by supercurrents, and ultimately the strength of these currents goes to zero exponentially with radius beyond a certain scale known as the London penetration depth. In the superfluid vortex the currents fall off slowly as a power law in radius, unless an antivortex intervenes (Figure 2.4).

When we move beyond the treatment of just the order parameter and think about the excitations of the system, additional physics results as particles can be required to be present in the vortex core (see Section 8.7), for topological reasons. The richest topological defects of all are in states combining symmetry breaking and topological order, such as topological superconductors (Chapter 9), where

[10] Compactified means that we only consider configurations or functions on the plane that take the same value for all points at spatial infinity. Then the plane becomes topologically equivalent to the sphere, by stereographic projection.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

one focus of current research is to find exotic Majorana zero mode quasiparticles trapped in vortices; such quasiparticles offer promise for a topological approach to quantum computation.

Box 2.3 The Berezinskii–Kosterlitz–Thouless Transition

One of the most remarkable examples of a collective effect arising from many topological defects is the superfluid transition in two spatial dimensions. We sketch a theoretical prediction by Kosterlitz and Thouless, anticipated in part in work of Berezinskii, that received spectacular experimental confirmation in work of Bishop and Reppy on ^{4}He films. While the full analysis of Kosterlitz and Thouless is a tour de force of renormalization-group methods, which we will not cover here, the essentials can be obtained from simple physical arguments about topological defects.

Our starting point is the two-dimensional XY model: the local spin variable on each site of a lattice is a unit vector on the circle, with the lattice Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \mathbf{s}_i \cdot \mathbf{s}_j = -J \sum_{\langle ij \rangle} \cos(\theta_i - \theta_j), \tag{2.96}$$

where J is an energy and the sum is over nearest-neighbor pairs. In the second equality we have introduced an angle θ via $s_x + i s_y = e^{i\theta}$.

This model has the same symmetry as the superfluid transition in a film of atoms with no low-temperature internal degrees of freedom, by the following argument: the Bose condensation transition means that one quantum state has a macroscopic number of atoms, and the wavefunction of this state $\psi(\mathbf{r})$ can be taken loosely as the order parameter in the ordered phase.[a]

Suppose that we are at low temperature so that the spin moves only slightly from one site to the next. Then, in going around a large circle, we can ask how many times the spin winds around the unit circle, and define this as the winding number $n \in \mathbb{Z}$. As mentioned above, if the winding number is nonzero, then the continuum limit must break down at some point within the circle or else the energy would diverge; in a moment, we will calculate the energy of a vortex and a vortex-antivortex pair in Eq. 2.105 and see this divergence.

This will also let us see from a fairly simple calculation how there can be continuously varying exponents in the power law correlations of the 2D XY model at low temperature. The assumption we'll need to make is that vortices are unimportant at sufficiently low temperature, so that the 2π periodicity of the phase can be ignored (a vortex in 2D is a point where the magnitude of the order parameter vanishes, around which the phase of the order parameter changes by a multiple of 2π). This calculation can be justified by looking at the renormalization group flow in a space of two parameters: the temperature, and the vortex fugacity (essentially a parameter controlling how many vortices there are). An excellent reference for this RG flow is the original paper by Kosterlitz and Thouless (1973). If there are no vortices and the

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

magnitude of the order parameter is constant, then the effective partition function is

$$Z = \int D\theta(r)\, e^{-\frac{K}{2}\int (\nabla\theta)^2 d^2r}, \tag{2.97}$$

where θ is no longer restricted to be periodic. Here K is a dimensionless coupling incorporating temperature that, in a lattice model such as (2.96), can be obtained by linearization. We can define a superfluid stiffness with units of energy $\rho_s = (k_B T)K$ that measures the energy required to create a twist in the superfluid phase. One way to look at this nonlinear sigma model is as describing slow variations of the ordered configuration at low temperature: the magnitude is fixed because fluctuations in magnitude are energetically expensive, but since there are degenerate states with the same magnitude but different θ, slow variation of θ costs little energy.

In the model with no vortices, that is, with θ treated as a real-valued rather than periodic field, the spin correlation function, which we will find goes as a power law, is

$$\langle \mathbf{s}(0) \cdot \mathbf{s}(r)\rangle = \mathrm{Re}\langle e^{i\theta(0)} e^{-i\theta(r)}\rangle. \tag{2.98}$$

(Actually taking the real part is superfluous if we define the correlator of an odd number of θ fields to be zero.) We choose not to rescale the θ field to make K equal to unity, since such a rescaling would modify the periodicity constraint $\theta = \theta + 2\pi$ in the model once vortices are restored. To obtain this correlation function, we first compute the correlation of the θ fields $G(r) = \langle \theta(0)\theta(r)\rangle$, which we will need to regularize by subtracting the infinite constant $G(0) = \langle \theta(0)^2\rangle$.

From previous experience with Gaussian theories, we know to write $G(r)$ as an integral over Fourier components:

$$G(r) = \frac{1}{(2\pi)^2 K} \int^{a^{-1}} \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{k^2}\, d\mathbf{r}. \tag{2.99}$$

However, this integral is divergent at $k = 0$ (this is called infrared divergent; by contrast, ultraviolet divergent means a divergence at short length scales, at $k = \infty$). In condensed matter we expect that short length scales are cut off by some microscopic length a, as in the above integral, while long length scales are more interesting. We can regularize the short-distance divergence by subtracting out the formally infinite quantity $G(0)$, and then using $\tilde{G}(r) = G(r) - G(0)$ to calculate physical quantities:

$$\tilde{G}(r) = -\frac{1}{(2\pi)^2 K} \int^{a^{-1}} \frac{1 - e^{i\mathbf{k}\cdot\mathbf{r}}}{k^2}\, d\mathbf{r}. \tag{2.100}$$

Now the integral can only be a function of the ratio r/a (you can check this by changing variables). As $a \to 0$, the leading term in the integration is proportional to

$$\tilde{G}(r) = \frac{1}{(2\pi)^2 K} \int^{a^{-1}} \frac{dk}{k} = -\frac{\log(r/a)}{2\pi K} + \ldots, \tag{2.101}$$

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

where one factor of 2π was picked up by the angular integration. Note that at large r, $\tilde{G}$ is divergent, which makes sense since θ is unbounded.

We now want to calculate the resulting spin-spin correlator. To do this we need to use a fact about Gaussian integrals. The Gaussian average

$$\langle e^{iJ\phi}\rangle = (A/\sqrt{2\pi})\int_{-\infty}^{\infty} d\phi\, e^{iJ\phi} e^{-\frac{1}{2}A\phi^2} = e^{-\frac{1}{2}A^{-1}J^2} = e^{-\frac{1}{2}\langle J^2\phi^2\rangle} \tag{2.102}$$

generalizes to the continuum limit in the following way:

$$\langle e^{-i\theta(r)+i\theta(0)}\rangle = e^{-\frac{1}{2}\langle(\theta(r)-\theta(0))^2\rangle} = e^{G(r)-G(0)}. \tag{2.103}$$

So finally,

$$\langle \mathbf{s}(0)\cdot\mathbf{s}(r)\rangle = \mathrm{Re}\langle e^{i\theta(r)-i\theta(0)}\rangle = \frac{1}{(r/a)^{1/2\pi K}}. \tag{2.104}$$

We expect on physical grounds that this power law correlation function (algebraic long-range order) cannot survive up to arbitrarily high temperatures; above some maximum temperature, there should be a disordered phase with exponentially decaying correlations.

To understand how our physical expectation of exponentially short correlations at high temperature is met, we give a simple picture due to Kosterlitz and Thouless (which is supported by a more serious RG calculation). The picture is that the phase transition results from an unbinding of logarithmically bound vortex-antivortex pairs, which can be viewed as the plasma-gas transition of a two-dimensional Coulomb plasma. The phase with unbound vortices is a plasma phase since it is has unbound positive and negative charges (vortices) as in a plasma. Note, of course, that the logarithmic interaction between vortex charges in 2D is just like that between Coulomb charges in 2D. A more serious calculation constructs an RG flow in terms of vortex fugacity to show that below a critical temperature vortices are irrelevant (their fugacity scales to 0), while above that temperature vortices are relevant (their fugacity increases upon rescaling). Vortex-antivortex pairs (see Figure 2.4) are logarithmically bound because the energy of a single vortex of winding number n goes as, from integrating the energy density $(k_BT)K(\nabla\theta)^2/2$ from (2.97),

$$E = \frac{1}{2}K(k_BT)\int_a^L (n/r)^2\, d^2r \sim \pi n^2 K\log(L/a). \tag{2.105}$$

Here L is the long-distance cutoff (e.g., system size) and a is the short-distance cutoff (e.g., vortex core size). Although the energy of a single vortex in the infinite system diverges, the interaction energy of a vortex-antivortex pair does not; each vortex has energy given by (2.105), but with the system size replaced by the intervortex spacing. Note that changes of order unity in the definition of this spacing or the core size will add constants to the energy but not change the coefficient of the logarithm.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

We would like to compare the free energy of two phases: one in which vortices are bound in pairs, and essentially do not modify the Gaussian model, and one in which vortices are numerous and essentially free, although the system is still charge-neutral (total winding number 0). Suppose the vortices in the free phase have typical separation L_0. Then each vortex can be distributed over a region of size ${L_0}^2$, and the entropic contribution to the free energy per vortex is $-TS = -T \log \Omega = -2T \log(L_0/a)$, where $\Omega = (L_0/a)^2$ is the number of distinguishable states. The energy cost is $E = \pi n^2 J \log L_0/a$, so there should be a phase transition somewhere near $T_{KT} = \pi J/2k_B$, where we have written $K = J/k_B T$ in order to define a coupling energy scale J.

A bit more work shows that this coupling scale, as we have defined it, is exactly the superfluid stiffness ρ_s that measures the energy induced by a twist in the superfluid phase. More precisely, the Berezinskii–Kosterlitz–Thouless transition occurs when the asymptotic long-distance stiffness ρ_s^∞, including renormalization by bound vortex pairs, satisfies

$$\rho_s^\infty = \frac{2k_B T_{KT}}{\pi}. \tag{2.106}$$

This prediction of a universal jump at T_{KT} in the superfluid stiffness was beautifully confirmed in experiments (Bishop and Reppy, 1978). Another way to state this result is that a 2D superfluid that starts at short distances with stiffness less than this value allows vortices to proliferate and reduce the long-distance superfluid stiffness to zero. This behavior is rather different from that in higher dimensions, where the superfluid stiffness flows to zero smoothly.

It is remarkable that such relatively straightforward considerations can uncover the KT transition. Equally remarkably, the renormalization group can predict phase diagrams in considerable detail. As a rich illustrative example, we return to the case of the triangular Ising magnet and consider it as a quantum magnet in a transverse field of strength Γ:

$$\mathcal{H} = -J \sum_{\langle ij \rangle} s_i^z s_j^z - \Gamma \sum_i s_i^x \tag{2.107}$$

Here, the Ising spins s_i of Eq. 2.62 have been replaced by spin-1/2 operators s^α. Via a Trotter–Suzuki transformation this can be transformed into a stacked classical magnet, with the same exchange interactions in the plane but with ferromagnetic couplings between adjacent planes. In this mapping, the temperature of the quantum magnet is encoded by the extent of the additional, third dimension, which is infinite at $T = 0$ but finite otherwise. For details of this mapping, see Box 6.1.

The coupling in the third direction adds a term proportional to $\cos q_z$, with a negative prefactor, to the Fourier transform of the couplings (Eq. 2.67), which remains unchanged otherwise. The candidate ordering thus takes place at $(\pm 4\pi/3, 0, 0)$ now, and the Landau–Ginzburg functional (Eq. 2.68) remains essentially unchanged except for the integration over the additional dimension (and extra gradient terms, which we have suppressed).

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press

The analysis of the XY-model with a clock term was analyzed in great detail by José et al. (1977) in a milestone paper demonstrating the power of the renormalization group in its relatively early days. There, it was found that in $d = 2$ dimensions, the KT transition out of the paramagnet is not changed by the addition of the clock term b_6 (this term is thus said to be irrelevant at the transition). Upon lowering the temperature further, the KT phase with its drifting exponent is encountered, which terminates at another transition where the clock term finally becomes relevant. The order parameter then locks into one of the clock directions. This corresponds to a three-sublattice ordering at wavevector $\pm\mathbf{K}$, as expected.

This contrasts to the quantum phase transition at $T = 0$ in the original quantum model, which occurs when the transverse field strength Γ is tuned. This corresponds to an infinite extent in the third dimension of the effective classical model, and hence a $d = 3$ RG analysis is needed. Here, it is found that the clock term is still irrelevant at the transition, which therefore is in the XY universality class in $d = 3$. However, it immediately becomes relevant inside the ordered phase; such terms are known as dangerously irrelevant, and again lead to locking into one of six symmetry-breaking phases.

The resulting phase diagram is shown in Figure 2.5. The ordered phase is labeled as bond-ordered, on account of its appearance in the dual quantum dimer model on the honeycomb lattice (see Box 6.1). The details of the conventional ordered phases are unimportant in the context of our interest in topological phases; however, looking ahead to Chapter 6, the robust ordering tendency of quantum dimer models visible in this phase diagram is common to all quantum dimer models on bipartite lattices such as the honeycomb, and also responsible for the absence of the desired topologically ordered resonating valence bond liquids on the square lattice.

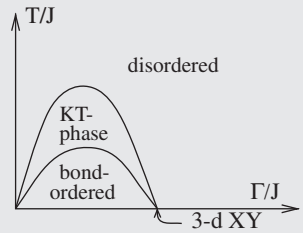

Fig. 2.5 Phase diagram of the transverse field Ising model on the triangular lattice. Upon lowering transverse field and/or temperature, a KT transition to a phase with algebraic correlations takes place. The exponent of the algebraic correlations drifts until long-range order sets in at another KT transition. At $T = 0$, these collapse to a single quantum phase transition in the 3D XY universality class. From Moessner and Sondhi (2001a).

[a] More precisely, as we will use in the discussion of superconductivity and the Josephson effect in Chapter 8, the order parameter comes from an expectation value of an operator changing the particle number, and its phase is not an overall wavefunction phase.

https://doi.org/10.1017/9781316226308.003 Published online by Cambridge University Press