3

Berry phases and curvatures

A "Berry phase" is a phase angle (i.e., running between 0 and 2π) that describes the global phase evolution of a complex vector as it is carried around a path in its vector space. It can also be referred to as a "geometric phase" or a "Pancharatnam phase," the latter after early work by Pancharatnam (1956). The concept was systematized and popularized in the 1980s by Sir Michael Berry, notably in a seminal paper (Berry, 1984), and by Berry and others in a series of subsequent publications that are well represented in the edited volume of Wilczek and Shapere (1989). A formal discussion of Berry phases and related concepts (fiber bundles, connections, Berry curvatures, etc.) can be found in modern texts on topological physics such as those by Frankel (1997), Nakahara (2003), and Eschrig (2011). Since then, Berry phases have found broad application in diverse realms of science including atomic and molecular physics, classical optics, wave mechanics, and condensed-matter physics.

Our goal here is to introduce the concept of the Berry phase and explain how it enters into the quantum-mechanical band theory of electrons in crystals. We will begin by introducing the Berry phase in its abstract mathematical form, and then discuss its application to the adiabatic dynamics of finite quantum systems. After these preliminaries, we will turn to the main theme of this book, where the complex vector in question is a Bloch wavevector, and the path lies in the space of wavevectors $\mathbf{k}$ within the Brillouin zone.

3.1 Berry phase, gauge freedom, and parallel transport

As indicated above, a Berry phase is a quantity that describes how a global phase accumulates as some complex vector is carried around a closed loop in a complex vector space. Since we are only interested in phases, we can take

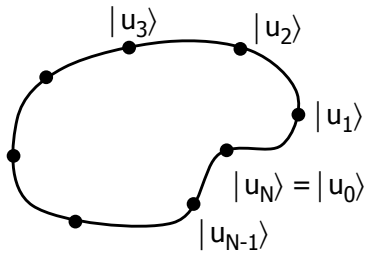

Figure 3.1 Illustration of the evolution of some complex unit vector $|u\rangle$ around a path in parameter space. The first and last points $|u_0\rangle$ and $|u_N\rangle$ are identical.

these complex vectors to be unit vectors, and we will typically identify them with the ground-state wavefunction of some quantum system. For example, the vector could represent the ground state of the electrons in a molecule with fixed nuclear coordinates, or that of a spinor in an external magnetic field. We then consider a gradual variation that returns the system to its starting point at the end of the loop. The situation is sketched in Fig. 3.1, where states $|u_0\rangle$, ..., $|u_7\rangle$ correspond to the $N = 8$ points around the loop (with $|u_8\rangle = |u_0\rangle$).

For a concrete example, consider the triatomic molecule shown in Fig. 3.2. It is almost equilateral, but a distortion has been introduced so as to shorten one of the three bonds slightly, as indicated by the "double bond" in the figure. We have in mind a continuous deformation path, with each of the three bonds being gradually shortened and lengthened in such a way that the illustrations in Fig. 3.2 represent snapshots along the way.[1] We then wish to consider the phase evolution of the ground state of this molecule as it is carried around the loop. Or, consider the evolution of the ground state of a spinor (e.g., an electron or proton) in an external magnetic field as the direction of this field varies around some closed loop on the unit sphere. In each case, the Berry phase will encode some information about the phase evolution of the ground state along the path in question.

3.1.1 Discrete formulation

Let's start with a discrete formulation, in which N representative vectors $|u_0\rangle$ to $|u_{N-1}\rangle$ are chosen around this loop, as for example with $N = 3$ in Fig. 3.2. Note that $|u_N\rangle$ and $|u_0\rangle$ are identical. The Berry phase ϕ is then

[1] Note that the molecule itself is not rotating; only the pattern of shortened bonds is rotating. This is often referred to as a "pseudorotation" and plays an important role in the physics of Jahn-Teller effects in molecules.

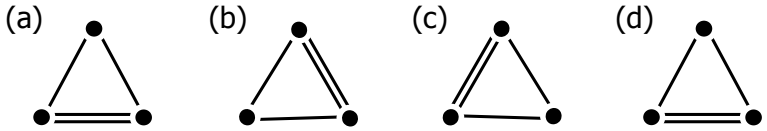

Figure 3.2 Triangular molecule going though a sequence of distortions in which first the bottom, then the upper-right, then the upper-left bond is the shortest and strongest of the three. The configurations in panels (a) and (d), representing the beginning and end of the loop, are identical.

defined to be

$$\phi = -\mathrm{Im}\ln\Big[\langle u_0|u_1\rangle\langle u_1|u_2\rangle ... \langle u_{N-1}|u_0\rangle\Big]. \tag{3.1}$$

Recall that for a complex number $z = |z|\,e^{i\varphi}$, the expression $\mathrm{Im}\ln z = \varphi$ just takes the complex phase and discards the magnitude. Thus, the Berry phase ϕ is minus the complex phase of the product of inner products of the state vectors at neighboring points around the loop. (Note that the sign convention is not universal, and some authors define the Berry phase without the minus sign in Eq. (3.1).)

Let's first consider a simple example based on the triatomic molecule of Fig. 3.2. Suppose that there are two degenerate states $|1\rangle$ and $|2\rangle$ for the equilateral triangle, but that the distortion causes a breaking of this degeneracy at all points along the loop. Following the lower-energy of the two states, we might find that these are

$$|u_a\rangle = |u_d\rangle = \tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}, \quad |u_b\rangle = \tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\e^{2\pi i/3}\end{pmatrix}, \quad |u_c\rangle = \tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\e^{4\pi i/3}\end{pmatrix}. \tag{3.2}$$

corresponding to the distorted states in Fig. 3.2, where the top and bottom elements in the column vector are the amplitudes on basis states $|1\rangle$ and $|2\rangle$ respectively. A trivial computation then shows that the corresponding Berry phase is

$$\phi = -\mathrm{Im}\ln\Big[\langle u_a|u_b\rangle\langle u_b|u_c\rangle\langle u_c|u_a\rangle\Big] = -\mathrm{Im}\ln\Big[\Big(\frac{e^{\pi i/3}}{2}\Big)^3\Big] = -\pi \tag{3.3}$$

(or equivalently $\phi = \pi$, since a phase is only well-defined modulo 2π.) Incidentally, you can see that the setting in a *complex* vector space is important; in the case of *real* vectors, the global product in Eq. (3.1) is always real, so the Berry phase is always 0 or π depending on the sign of that product.[2]

It is probably not yet obvious why the Berry phase defined in this way is

[2] For the states in Eq. (3.2) it happens that $\phi = \pi$, but this is an artifact of the special form of those states.

a useful quantity, but at least it is mathematically well-defined in the sense that it is *independent of the choices made for the phases of the individual* $|u_j\rangle$. That is, suppose we introduce a new set of N states

$$|\tilde{u}_j\rangle = e^{-i\beta_j}\,|u_j\rangle \tag{3.4}$$

(β_j is real) related to the old ones by a j-dependent phase rotation β_j, an operation that is known as a "gauge transformation" in the Berry-phase context.[3] Then the Berry phase ϕ is unaffected, since any given vector, such as $|u_2\rangle$, appears in Eq. (3.1) once in a ket and once in a bra, so that the phases $e^{\pm i\beta_j}$ cancel out. For example, we can replace $|u_c\rangle$ in Eq. (3.2) by the physically equivalent vector

$$|u_c\rangle = \tfrac{1}{\sqrt{2}}\begin{pmatrix} e^{2\pi i/3} \\ 1 \end{pmatrix} \tag{3.5}$$

and confirm that the final result of the computation in Eq. (3.3) is the same. The gauge-invariance of the Berry phase strongly hints that it may be connected with some physically observable phenomena.

We have passed over a subtlety above, namely the need to impose a branch choice on the definition of $\operatorname{Im}\ln z$, as by restricting it to the interval $-\pi < \phi \le \pi$. In this case, Eq. (3.1) always results in a Berry phase lying in this interval, while the nominally equivalent expression

$$\phi = -\sum_{j=0}^{N-1} \operatorname{Im}\ln\langle u_j|u_{j+1}\rangle \tag{3.6}$$

can yield a result that differs by an integer multiple of 2π. If we take the viewpoint that ϕ is just a shorthand for a phase angle, so that only $\cos\phi$ and $\sin\phi$ matter, then this distinction can be safely ignored. However, in any practical implementation the phase angles are normally mapped onto some interval on the real axis, and we can only claim that the Berry phase should be gauge-invariant modulo 2π in the context of an expression like that of Eq. (3.6).

You may be wondering about the magnitude information that has been discarded in Eqs. (3.1) and (3.6). Each inner product has a magnitude somewhat smaller than unity, so a partner function $-\operatorname{Re}\ln\prod_j\langle u_j|u_{j+1}\rangle$ would measure the extent to which the *character* of the states varies from point to point along the loop, whereas the Berry phase ϕ is instead related to the relative *phases* along the loop.

[3] The name is chosen in close analogy to the use of the same term in the theory of electromagnetism. A particular choice of gauge may influence the intermediate results of a calculation, but should not affect any physically meaningful prediction.

The concept of a Berry phase is also naturally described in terms of a notion of *parallel transport*, defined in the present context as follows. Suppose we have a chain of states $|u_0\rangle$, $|u_1\rangle$, ..., $|u_N\rangle$ with no special phase relations between them. We define a new set of "parallel transported" states $|\bar{u}_0\rangle$, $|\bar{u}_1\rangle$, ... to be the same as the previous set, except with their phases adjusted as follows. Set $|\bar{u}_0\rangle = |u_0\rangle$. Then choose $|\bar{u}_1\rangle$ to be $|u_1\rangle$ times a phase chosen such that $\langle\bar{u}_0|\bar{u}_1\rangle$ is real and positive. Similarly, choose $|\bar{u}_2\rangle$ such that $\langle\bar{u}_1|\bar{u}_2\rangle$ is also real and positive, and continue in this way around the loop, imposing the constraint

$$\mathrm{Im}\ln\langle\bar{u}_j|\bar{u}_{j+1}\rangle = 0 \tag{3.7}$$

on each link connecting neighboring points. Conclude by choosing $|\bar{u}_N\rangle$ such that its product with $\langle\bar{u}_{N-1}|$ is real and positive. This generates what is known as a *parallel transport gauge*.[4]

Assuming that the states form a closed loop as in Fig. 3.1, the two vectors $|u_N\rangle$ and $|u_0\rangle$ are identical. By contrast, while the two vectors $|\bar{u}_N\rangle$ and $|\bar{u}_0\rangle$ describe the same physical state, *they generally differ by a phase*. In fact, the phase mismatch between $|\bar{u}_0\rangle$ and $|\bar{u}_N\rangle$ is nothing other than the Berry phase! To see this, recall that Eq. (3.1) is gauge-invariant, so we can evaluate it using the parallel-transport gauge for the states $0, ..., N-1$, i.e., $\phi = -\mathrm{Im}\ln[\langle\bar{u}_0|\bar{u}_1\rangle...\langle\bar{u}_{N-1}|\bar{u}_0\rangle]$. Since $|\bar{u}_0\rangle$ and $|\bar{u}_N\rangle$ differ only by a phase, we can replace $|\bar{u}_0\rangle$ at the end of the product by $|\bar{u}_N\rangle\langle\bar{u}_N|\bar{u}_0\rangle]$ to get $\phi = -\mathrm{Im}\ln[\langle\bar{u}_0|\bar{u}_1\rangle...\langle\bar{u}_{N-1}|\bar{u}_N\rangle\langle\bar{u}_N|\bar{u}_0\rangle]$. Then all inner products are real and positive except the last, so that

$$\phi = -\mathrm{Im}\ln\langle\bar{u}_N|\bar{u}_0\rangle\,. \tag{3.8}$$

For the case of Eq. (3.2), for example, we get

$$|\tilde{u}_a\rangle = \begin{pmatrix}1\\1\end{pmatrix}, \quad |\tilde{u}_b\rangle = \begin{pmatrix}e^{-\pi i/3}\\e^{\pi i/3}\end{pmatrix}, \quad |\tilde{u}_c\rangle = \begin{pmatrix}e^{-2\pi i/3}\\e^{2\pi i/3}\end{pmatrix}, \quad |\tilde{u}_d\rangle = \begin{pmatrix}-1\\-1\end{pmatrix} \tag{3.9}$$

(where we have now dropped the irrelevant normalization prefactors), so that $\phi = -\mathrm{Im}\ln\langle\bar{u}_d|\bar{u}_0\rangle = \pi$ as before.

Note that the parallel transport-gauge is not quite unique, since there is still the freedom to choose the phase of the initial vector $|\bar{u}_0\rangle$. Since this choice of initial phase also propagates into $|\bar{u}_N\rangle$, however, it does not affect the value of ϕ coming from Eq. (3.8).

[4] The term "parallel transport" comes from differential geometry, where the basic idea is that one chooses a local orthonormal basis of vectors at each point along a path on a curved manifold in such a way that the basis is "as aligned as possible" with its neighbors everywhere along the path. Here, the phrase "as aligned as possible" is to be reinterpreted in terms of phase alignment.

For a closed loop of the kind that we are considering here, the parallel transport gauge is somewhat unsatisfying in that it has a discontinuity where the end of the loop rejoins the starting point. We can smooth out this discontinuity by constructing a "twisted parallel transport gauge" by starting from the parallel transport gauge and applying phase twists

$$|\tilde{u}_j\rangle = e^{-ij\phi/N}\,|\bar{u}_j\rangle\,. \tag{3.10}$$

The new gauge no longer has the discontinuity at the end of the loop. It has the property that $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ has the uniform value $-\phi/N$ at every point on the loop, which is manifestly consistent with Eq. (3.6). In other words, we have distributed the phase evolution uniformly along the loop in such a way as to iron out the gauge discontinuity that would otherwise occur at the end of the loop.

While the freedom in the choice of the twisted parallel-transport gauge is still strongly restricted, it is less restricted than for a true parallel-transport gauge for the following important reason. Now, in addition to rotating the phase of the starting state $|\tilde{u}_0\rangle$ (which amounts to a global rotation of all phases), we have the possibility of replacing ϕ by $\phi + 2\pi m$ (for some integer m) in Eq. (3.10). Taking $m=1$ for example, this changes all the $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ by $-2\pi/N$, which is still much less than 2π for large N. In other words, we are free to choose different ways of unwinding the phase discontinuity such that $\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle$ is identical for each pair of neighbors, and each of these is a different but equally valid twisted parallel transport gauge. We will usually choose the one such that $|\mathrm{Im}\ln\langle\tilde{u}_j|\tilde{u}_{j+1}\rangle|$ is minimum, but this is not a fundamental restriction. The gauge choice of Eq. (3.2) is an example of a twisted parallel-transport gauge.

3.1.2 Continuous formulation and Berry potential

Another hint that the Berry phase formula above may be physically meaningful arises from the fact that it has a well-defined continuum limit, shown in Fig. 3.3(c), obtained by increasing the density of points along the path as sketched in Fig. 3.3(a-b). In the continuum formulation, we can take the path to be parametrized by a real variable λ such that $|u_\lambda\rangle$ traverses the path as λ evolves from 0 to 1, with $|u_{\lambda=0}\rangle \equiv |u_{\lambda=1}\rangle$. (Such a convention should be familiar from Ch. 1.) We assume here that $|u_\lambda\rangle$ is a smooth and differentiable function of λ. To derive the continuum expression for the Berry phase

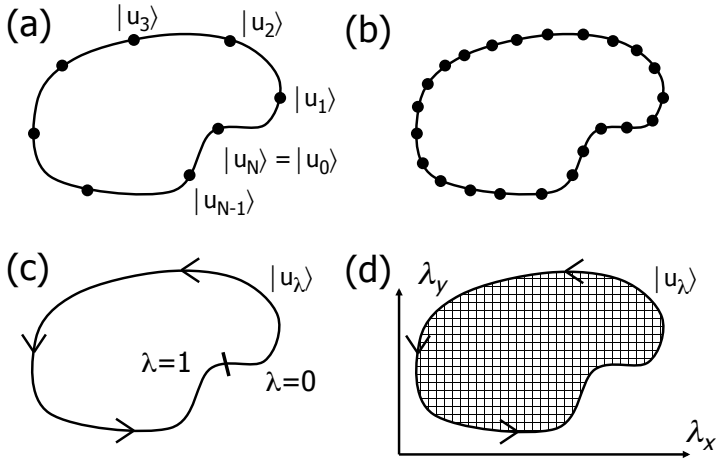

Figure 3.3 (a) Evolution of a state vector $|u\rangle$ in N discrete steps around a closed loop, as in Fig. 3.1. (b) Approach to the continuum limit by increasing the density of points around the loop. (c) Continuum limit, in which the parameter runs over $\lambda \in [0,1]$ with $|u_{\lambda=0}\rangle = |u_{\lambda=1}\rangle$. (d) Loop regarded as spanning a surface in a two-dimensional parameter space $\boldsymbol{\lambda} = (\lambda_\mu, \lambda_\nu)$.

that corresponds to Eq. (3.6) we note that

$$\begin{aligned}\ln \langle u_\lambda | u_{\lambda+d\lambda}\rangle &= \ln \langle u_\lambda | \left(|u_\lambda\rangle + d\lambda \frac{d|u_\lambda\rangle}{d\lambda} + \ldots \right) \\ &= \ln(1 + d\lambda \langle u_\lambda | \partial_\lambda u_\lambda \rangle + \ldots) \\ &= d\lambda \langle u_\lambda | \partial_\lambda u_\lambda \rangle + \ldots\end{aligned}$$

where ∂_λ is a shorthand for $d/d\lambda$ and '…' indicates terms of second order and higher in $d\lambda$. The latter can be discarded in taking the continuum limit of Eq. (3.6) and we obtain

$$\phi = -\mathrm{Im} \oint \langle u_\lambda | \partial_\lambda u_\lambda \rangle \, d\lambda \,. \tag{3.11}$$

In fact $\langle u_\lambda | \partial_\lambda u_\lambda \rangle$ is purely imaginary since

$$2\mathrm{Re} \langle u_\lambda | \partial_\lambda u_\lambda \rangle = \langle u_\lambda | \partial_\lambda u_\lambda \rangle + \langle \partial_\lambda u_\lambda | u_\lambda \rangle = \partial_\lambda \langle u_\lambda | u_\lambda \rangle = 0 \,,$$

so Eq. (3.11) can also be written as

$$\phi = \oint \langle u_\lambda | i \partial_\lambda u_\lambda \rangle \, d\lambda \,. \tag{3.12}$$

This is the famous expression for a Berry phase in the continuous formulation (Berry, 1984; Wilczek and Shapere, 1989).

The integrand on the right-hand side of Eq. (3.12) is known as the *Berry connection* or *Berry potential*,[5]

$$A(\lambda) = \langle u_\lambda | i\partial_\lambda u_\lambda \rangle = -\mathrm{Im}\,\langle u_\lambda | \partial_\lambda u_\lambda \rangle\,, \tag{3.13}$$

in terms of which the Berry phase is

$$\phi = \oint A(\lambda)\, d\lambda\,. \tag{3.14}$$

Let us understand how these quantities vary under a gauge transformation, which now takes the form

$$|\tilde{u}_\lambda\rangle = e^{-i\beta(\lambda)} |u_\lambda\rangle \tag{3.15}$$

where $\beta(\lambda)$ is some continuous real function of λ. We find

$$\widetilde{A}(\lambda) = \langle \tilde{u}_\lambda | i\partial_\lambda | \tilde{u}_\lambda \rangle = \langle u_\lambda | e^{i\beta(\lambda)}\, i\partial_\lambda\, e^{-i\beta(\lambda)} | u_\lambda \rangle = \langle u_\lambda | i\partial_\lambda | u_\lambda \rangle + \beta'(\lambda)$$

where $\beta'(\lambda) = d\beta/d\lambda$. Thus the Berry potential is *not gauge-invariant*; it is transformed under a gauge change according to

$$\widetilde{A}(\lambda) = A(\lambda) + \beta'(\lambda)\,. \tag{3.16}$$

But what about the Berry phase? Recall that since $\lambda = 0$ and $\lambda = 1$ label the same state, we must insist that $|\tilde{u}_{\lambda=1}\rangle = |\tilde{u}_{\lambda=0}\rangle$, just as was the case for $|u_\lambda\rangle$. But this implies that

$$\beta_{\lambda=1} = \beta_{\lambda=0} + 2\pi m \tag{3.17}$$

for some integer m. Then

$$\int_0^1 \beta'(\lambda)\, d\lambda = \beta_{\lambda=1} - \beta_{\lambda=0} = 2\pi m \tag{3.18}$$

so that replacing A by $\widetilde{A}$ in Eq. (3.14) and using Eq. (3.16) yields

$$\tilde{\phi} = \phi + 2\pi m\,. \tag{3.19}$$

That is, the Berry phase ϕ is gauge-invariant modulo 2π, or in other words, gauge-invariant when regarded as a phase angle!

Once again, we can think of the Berry phase as the phase that is "left over" after parallel transport around the loop. In the continuous case, a parallel-transport gauge is one in which the Berry connection $A(\lambda)$ vanishes:

$$\bar{A}(\lambda) = \langle \bar{u}_\lambda | i\partial_\lambda \bar{u}_\lambda \rangle = 0\,. \tag{3.20}$$

[5] These terms are generally used interchangeably. *"Connection"* is a term taken from differential geometry, while *"potential"* invokes an analogy with the vector potential of electromagnetism (see p. 75) and other gauge field theories.

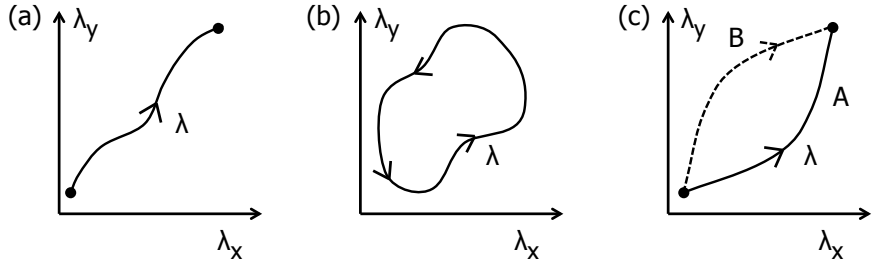

Figure 3.4 (a) Open path A in parameter space. (b) Closed path A in parameter space. (c) Pair of open paths A and B with common initial and final points, such that $A - B$ (that is, A followed by the reverse traversal of B) is a closed path.

If we impose such a gauge, then the Berry phase is just the phase mismatch at the end of the loop,

$$\phi = -\mathrm{Im} \ln \langle \bar{u}_{\lambda=1} | \bar{u}_{\lambda=0} \rangle \,, \tag{3.21}$$

exactly as in Eq. (3.8). We can also construct a twisted parallel transport gauge as $|\tilde{u}_\lambda\rangle = e^{-i\phi\lambda} |\bar{u}_\lambda\rangle$, in analogy with Eq. (3.10), with the result that $\widetilde{A}_\lambda$ is constant around the loop.

The fact that the Berry phase is gauge-invariant modulo 2π should not come as a surprise, reflecting as it does our experience with the discrete case, but its importance is profound. Because quantum probabilities are proportional to the norm squared of an amplitude, there is a tendency to think that "the phase doesn't matter." On the contrary, however, phases can lead to interference phenomena that are physically important. For example, if duplicate copies of a system are prepared, subjected to parallel transport along different paths in parameter space, and then recombined, the resulting phase difference can lead to physical and measurable interference effects.

We shall usually discuss Berry phases in the context of adiabatic evolution along closed paths, but it is useful to establish some terminology for open paths as well. For an open path such as that shown in Figure 3.4(a), we can define an open-path Berry phase

$$\phi = \int_{\mathrm{i}}^{\mathrm{f}} A(\lambda) \, d\lambda \,. \tag{3.22}$$

However, this kind of Berry phase is *not* gauge-invariant; a gauge transformation in the form of Eq. (3.15) changes ϕ by $\beta_\mathrm{f} - \beta_\mathrm{i}$. Only when the path is closed, as in Fig. 3.4(b), is the Berry phase gauge-invariant (modulo 2π). But Fig. 3.4(c) shows another interesting case: if a system is carried from λ_i to λ_f along two *different* paths A and B, the *relative* phase $\Delta\phi = \phi_B - \phi_A$ is again

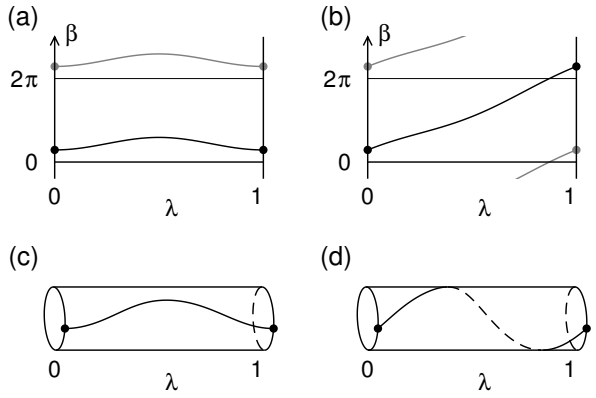

Figure 3.5 Possible behaviors of the function $\beta(\lambda)$ defining a gauge transformation through Eq. (3.15). (a-b) Conventional plots of "progressive" (a) and "radical" (b) gauge transformations, for which β returns to itself or is shifted by a multiple of 2π at the end of the loop, respectively. Shaded lines show 2π-shifted periodic images. (c-d) Same as (a-b) but plotted on the surface of a cylinder to emphasize the nontrivial winding of the radical gauge transformation in (b) and (d).

gauge-invariant. This follows trivially from the fact that $\Delta\phi$ is the Berry phase obtained by traversing path B, then path A in the reverse direction; this is equivalent to circulating around a closed path as in Fig. 3.4(b).

Returning now to closed paths, note that all possible gauge transformations given by Eq. (3.15) can be classified topologically according to the integer m appearing in Eq. (3.17), which is a "winding number" specifying how many times $e^{-i\beta}$ circulates around the unit circle in the complex plane as λ circulates around the loop. We shall refer to gauge changes characterized by $m=0$, illustrated in Fig. 3.5(a), as "progressive" gauge transformations. These have the special property that the gauge function $\beta(\lambda)$ can be smoothly deformed to the identity transformation ($\beta=0$ independent of λ). By contrast, we reserve the term "radical" for gauge changes with a nontrivial winding, as shown in Fig. 3.5(b) and (d).[6] (These are sometimes referred to as "small" and "large" gauge transformations respectively.)

[6] Note that the concept of a Berry phase does not impose any topological classification on adiabatic loops; the Berry phase itself is not quantized, and in the absence of special symmetries, its value can typically be adjusted by modifying the path of the loop or the Hamiltonian that determines the states along the loop. Instead, it is the set of *gauge transformations* on the loop that admits a topological classification.

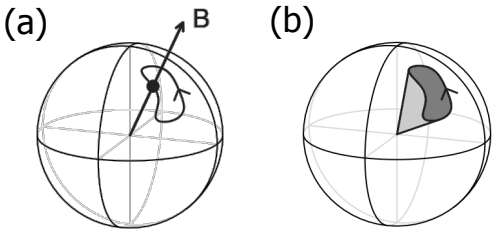

Figure 3.6 (a) Evolution of an applied magnetic field around a closed loop in **B** space. (b) Shaded region shows the solid angle swept out on the unit sphere in **B** space.

3.1.3 An example

So far, the discussion above has been entirely mathematical; we have been treating $|u_\lambda\rangle$ as a parametrized path in some complex vector space. For physical applications, we will usually be concerned with the case in which $|u_\lambda\rangle$ is the ground state of some quantum-mechanical Hamiltonian H_λ, with the ground state evolving smoothly as a consequence of the smooth evolution of H. For example, we might be concerned with the electronic ground state of a molecule as certain atomic structural coordinates are varied, or as external electric or magnetic fields are applied.

A simple and instructive example is the case of a spin-1/2 particle, such as an electron or a neutron, at rest in free space and subjected to a uniform magnetic field $\mathbf{B} = B\hat{\mathbf{n}}$ directed along $\hat{\mathbf{n}}$. Its Hamiltonian is just

$$H = -\gamma \mathbf{B}\cdot\mathbf{S} = -\left(\frac{\gamma\hbar B}{2}\right)\hat{\mathbf{n}}\cdot\boldsymbol{\sigma} \tag{3.23}$$

where γ is the gyromagnetic moment, $\mathbf{S} = \hbar\boldsymbol{\sigma}/2$ is the spin, and σ_j are the Pauli matrices. The ground state $|u_{\mathbf{B}}\rangle$ is a spin eigenstate of $\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}$, and is therefore completely independent of the *magnitude* of **B**. Thus it is natural to write it as $|u_{\hat{\mathbf{n}}}\rangle$, emphasizing that it depends only on the field direction $\hat{\mathbf{n}}$. We can then ask: What is the Berry phase of $|u_{\hat{\mathbf{n}}}\rangle$ as $\hat{\mathbf{n}}$ is carried around a loop in magnetic-field orientation space, as illustrated in Fig. 3.6(a)?

We shall see shortly that there is an elegant answer to this question, even for a curved loop such as that shown in Fig. 3.6(a), but let us first consider a simpler "triangular" loop in the discretized approximation. We let $\hat{\mathbf{n}}$ start along $\hat{\mathbf{z}}$, then rotate it to $\hat{\mathbf{x}}$, then to $\hat{\mathbf{y}}$, and then back to $\hat{\mathbf{z}}$, thereby tracing out one octant of the unit sphere. From Eq. (3.1), the Berry phase is

$$\phi = -\mathrm{Im}\ln\left[\langle\uparrow_z|\uparrow_x\rangle\langle\uparrow_x|\uparrow_y\rangle\langle\uparrow_y|\uparrow_z\rangle\right]$$

where $|\uparrow_n\rangle$ is the spinor that is "spin up in direction $\hat{\mathbf{n}}$." As shown in

standard quantum mechanics texts, such a spinor can be represented as

$$|\uparrow_{\hat{\mathbf{n}}}\rangle = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2)e^{i\varphi} \end{pmatrix}, \tag{3.24}$$

where (θ, φ) are the polar and azimuthal angles of $\hat{\mathbf{n}}$. Thus $|\uparrow_x\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $|\uparrow_y\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ i \end{pmatrix}$, and $|\uparrow_z\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. We can ignore the normalization factors when inserting into the expression for ϕ, obtaining $\phi = -\mathrm{Im}\ln[(1)(1+i)(1)] = -\pi/4$. As it happens, this result is exact; a more careful treatment using a dense mesh of intermediate points along each great-circle arc does not change this result. This will become clear when we obtain the general solution to the magnetic-field loop problem after introducing the concept of Berry curvature, which we do next.

Exercises

Exercise 3.1.1 Consider a path through the four spinor states

$$|u_0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |u_1\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad |u_2\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad |u_3\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ i \end{pmatrix} \tag{3.25}$$

which closes on itself with $|u_4\rangle = |u_0\rangle$. This corresponds to a path in which the spin points along $\hat{\mathbf{z}}$, $\hat{\mathbf{x}}$, $-\hat{\mathbf{z}}$, $\hat{\mathbf{y}}$, and then back to $\hat{\mathbf{z}}$. (a) Compute the discrete Berry phase for the path around this loop.

(b) Construct a parallel transport gauge for this path and check that the Berry phase computed from Eq. (3.8) agrees with your previous result..

(c) Construct a twisted parallel-transport gauge for this path.

Exercise 3.1.2 Consider the sequence of N spinor states described by Eq. (3.24) all with the same θ and with φ taking N equally spaced values from 0 to 2π.

(a) Show that the Berry phase is

$$\phi = -N\tan^{-1}\left[\frac{\sin^2(\theta/2)\sin(2\pi/N)}{\cos^2(\theta/2)+\sin^2(\theta/2)\cos(2\pi/N)}\right]. \tag{3.26}$$

(b) Find $\phi(\theta)$ in the limit that the discrete path becomes continuous (i.e., as $N \to \infty$).

(c) For $\theta = 45°$ compute ϕ numerically for $N = 3$, 4, 6, 12, and 100, and compare with the continuum limit.

Exercise 3.1.3

3.2 Berry curvature and the Chern theorem

3.2.1 Berry curvature

Consider a two-dimensional parameter space such as that illustrated in Fig. 3.3(d), so that we have vectors $|u_{\boldsymbol{\lambda}}\rangle$ as a function of $\boldsymbol{\lambda} = (\lambda_x, \lambda_y)$. Then the definition of the Berry potential in Eq. (3.13) naturally generalizes to that of a 2D vector $\mathbf{A}(\boldsymbol{\lambda}) = (A_x, A_y)$ via

$$A_\mu = \langle u_{\boldsymbol{\lambda}} | i\partial_\mu u_{\boldsymbol{\lambda}} \rangle \tag{3.27}$$

where $\partial_\mu = \partial/\partial\lambda_\mu$, and the Berry phase expression of Eq. (3.14) can be written as a line integral around the loop, i.e.,

$$\phi = \oint \mathbf{A} \cdot d\boldsymbol{\lambda} \,. \tag{3.28}$$

Then the *Berry curvature* $\Omega(\boldsymbol{\lambda})$ is simply defined as the Berry phase per unit area in (λ_x, λ_y) space. In a discretized context, as with the $\boldsymbol{\lambda}$ mesh shown in Fig. 3.3(d), Ω is identified with the Berry phase around one small plaquette[7] divided by the area of that plaquette. In a continuum framework, it becomes just the curl of the Berry potential,

$$\Omega(\boldsymbol{\lambda}) = \partial_x A_y - \partial_y A_x = -2\mathrm{Im}\,\langle \partial_x u | \partial_y u \rangle \,. \tag{3.29}$$

where the last equality follows from a cancellation of terms of the form $\langle u|\partial_x\partial_y u\rangle$ and noting that $\langle \partial_y u|\partial_x u\rangle^* = \langle \partial_x u|\partial_y u\rangle$.

Once Ω is defined as a curl, we can immediately write Stokes' theorem in the form

$$\phi = \oint_C \mathbf{A} \cdot d\boldsymbol{\lambda} = \int_S \Omega(\boldsymbol{\lambda})\, dS \tag{3.30}$$

where curve C traces the boundary of region S in the positive sense of circulation. In the discrete case, Stokes' theorem is just the statement that if we were to sum up the circulation of the Berry potential $\mathbf{A}$ around all the little plaquettes making up the region shown in Fig. 3.3(d), we would just obtain the Berry phase computed around its boundary, as is self-evident from the fact that the circulation computed along any one link is traversed once in each direction, leading to a cancellation.

A crucially important property of the Berry curvature is its gauge-invariance. That is, under a 2D gauge change $|\tilde{u}_{\boldsymbol{\lambda}}\rangle = e^{-i\beta(\boldsymbol{\lambda})}|u_{\boldsymbol{\lambda}}\rangle$, Eq. (3.16) is generalized to $\widetilde{\mathbf{A}} = \mathbf{A} + \nabla\beta$, and since the curl of a gradient is zero, the Berry curvature Ω in Eq. (3.29) is unchanged by the gauge transformation.

These concepts are easily generalized to a higher-dimensional parameter

[7] Because the plaquette is small, the magnitude of the Berry phase around the plaquette can safely be assumed to be $<< 2\pi$, so there is no ambiguity in its branch choice.

space $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$ by defining A_μ as an n-component vector following Eq. (3.27) and $\Omega_{\mu\nu}$ to be a (real) antisymmetric second-rank tensor

$$\Omega_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu = -2\,\mathrm{Im}\,\langle \partial_\mu u | \partial_\nu u \rangle\,. \tag{3.31}$$

Then Stokes' theorem becomes

$$\phi = \oint_C \mathbf{A}\cdot d\boldsymbol{\lambda} = \int_S \Omega_{\mu\nu}\, ds_\mu \wedge ds_\nu \tag{3.32}$$

where $ds_\mu \wedge ds_\nu$ is an area element on the surface S. This framework becomes most familiar for a 3D parameter space, where it is natural to use the pseudovector notation $\Omega_z \equiv \Omega_{xy} = -2\mathrm{Im}\,\langle \partial_x u | \partial_y u \rangle$ etc., which is sometimes written as $\boldsymbol{\Omega} = -\mathrm{Im}\,\langle \nabla_{\mathbf{k}} u | \times | \nabla_{\mathbf{k}} u \rangle$.[8] In this pseudovector notation, Stokes' theorem takes the familiar form

$$\phi = \oint_C \mathbf{A}\cdot d\boldsymbol{\lambda} = \int_S \boldsymbol{\Omega}\cdot\hat{\mathbf{n}}\, dS = \int_S \boldsymbol{\Omega}\cdot d\mathbf{S} \tag{3.33}$$

where $\hat{\mathbf{n}}$ is a unit vector normal to the surface element of area dS.

There is a close analogy connecting the real-space electromagnetic vector potential $\mathbf{A}(\mathbf{r})$ and its curl, the magnetic field $\mathbf{B}(\mathbf{r})$, with the parameter-space Berry potential $\mathbf{A}(\boldsymbol{\lambda})$ and its curl $\boldsymbol{\Omega}(\boldsymbol{\lambda})$. In both cases, the "potential" $\mathbf{A}$ is gauge-dependent, while the "field" $\mathbf{B}$ or $\boldsymbol{\Omega}$ is not. We shall have further opportunities to pursue this analogy later.

Let's apply this concept to the case of the spinor subjected to a magnetic field along $\hat{\mathbf{n}}$ as discussed earlier. We begin by calculating the Berry curvature Ω_{xy} at the "north pole" of the unit sphere in Fig. 3.6, i.e., at $\hat{\mathbf{n}} = (n_x, n_y, \sqrt{1-n_x^2-n_y^2})$. We again use the representation of Eq. (3.24), which we repeat here for convenience:

$$|\uparrow_{\hat{\mathbf{n}}}\rangle = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2)e^{i\varphi} \end{pmatrix}. \tag{3.34}$$

A gauge choice is implicit in this representation, which conveniently makes $|\uparrow_{\hat{\mathbf{n}}}\rangle$ smooth and continuous in the vicinity of $\theta = 0$. However, this comes at the expense of introducing a singularity at $\theta = \pi$, where the phase of $|\uparrow_{-\hat{\mathbf{z}}}\rangle$ depends on the azimuthal direction φ along which the limit $\theta \to \pi$ is taken.[9] Since the Berry-curvature formula of Eq. (3.31) only involves first derivatives of $|\uparrow_{\hat{\mathbf{n}}}\rangle$, we can expand to first order to get

$$|\uparrow_{\hat{\mathbf{n}}}\rangle \simeq \begin{pmatrix} 1 \\ (n_x + in_y)/2 \end{pmatrix}, \quad |\partial_x \uparrow_{\hat{\mathbf{n}}}\rangle = \frac{1}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad |\partial_y \uparrow_{\hat{\mathbf{n}}}\rangle = \frac{1}{2}\begin{pmatrix} 0 \\ i \end{pmatrix}, \tag{3.35}$$

so that Eq. (3.31) evaluates to $\Omega_{xy} = -1/2$.

[8] Note that this implies $\Omega_{\mu\nu} = \varepsilon_{\mu\nu\sigma}\Omega_\sigma$ but $\Omega_\mu = \frac{1}{2}\varepsilon_{\mu\nu\sigma}\Omega_{\nu\sigma}$, where $\varepsilon_{\mu\nu\sigma}$ is the antisymmetric tensor.

[9] An alternative would be to multiply the right side of Eq. (3.24) by $e^{-i\varphi}$, but this would leave a singularity at $\theta = 0$ where we want to compute the Berry curvature.

Using the pseudovector notation we can rewrite this as $\Omega_{\hat{\mathbf{n}}}$, which can be interpret as the Berry curvature per unit solid angle in $\hat{\mathbf{n}}$ orientation space. Having found that $\Omega_{\hat{\mathbf{n}}} = -1/2$ at $\theta = 0$, however, we can argue that it must take the same value everywhere else on the unit sphere. After all, the physics of a spinor in free space is intrinsically isotropic, and we are free to evaluate $\Omega_{\hat{\mathbf{n}}}$ using a coordinate system (x', y', z') with z' aligned with $\hat{\mathbf{n}}$. From this we conclude that $\Omega_{\hat{\mathbf{n}}} = -1/2$ everywhere on the unit sphere.

We can easily check the consistency of this result with the Berry phase that we computed on p. 72 for a path tracing one octant of the unit sphere. From Stokes' theorem, Eq. (3.30), we would expect that ϕ should be just $-1/2$ times the solid angle $4\pi/8$, or $\phi = -\pi/4$, which is precisely what we found above.

We now have the general answer to the problem posed in relation to Fig. 3.6. The Berry phase that results from the adiabatic evolution of a spinor around the loop in magnetic-field orientation space shown in Fig. 3.6(a) is simply $-1/2$ times the solid angle subtended by the loop, as sketched in Fig. 3.6(b). This is a beautiful and simple result of the mathematical physics of spinors. As a special case, the Berry phase obtained by rotating a spinor around a full great circle, as from $\hat{\mathbf{z}}$ to $\hat{\mathbf{x}}$ to $-\hat{\mathbf{z}}$ to $\hat{\mathbf{x}}$ and back to $\hat{\mathbf{z}}$ is just $-\pi$, since the solid angle of a hemisphere is 2π. This gives $e^{i\phi} = -1$, reflecting the well-known fact that the parallel transport of a spinor through a full 2π rotation results in a flip of the sign of the spinor wavefunction.

3.2.2 Chern theorem

You might be puzzled to note that the integral of the Berry curvature over the entire unit sphere does not vanish: it is -2π. At first sight this may seem impossible. For, imagine discretizing the surface of the unit sphere into small triangles or other polygonal plaquettes, and calculating the circulation of the Berry potential $\mathbf{A}$ (i.e., the Berry phase ϕ) around each plaquette. Shouldn't the sum of these vanish? One might think so, reasoning that each link between a pair of vertices is traversed once in each direction, leading to a cancellation. But we know that the integrated Berry curvature, which we identify with the total circulation of $\mathbf{A}$, should be -2π. Where have we gone wrong?

Let's look more closely, using a dodecahedral discretization of the sphere as shown in Fig. 3.7. The circulation of $\mathbf{A}$ on path C in panel (a), i.e., around a single pentagon, is $1/12$ of -2π or $-\pi/6$. Similarly, the total circulation around path C comprising six pentagons in panel (b) is $-\pi$, and in (c) it is $-11\pi/6$. On the other hand, path C in panel (c) traces the outline of

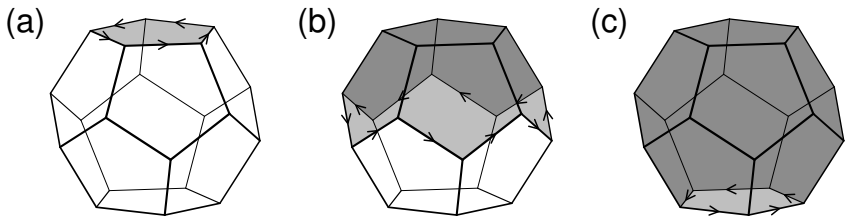

Figure 3.7 Application of Stokes' theorem to portions of the unit sphere in **B** space in a discretized approximation. The Berry phase, or circulation of **A**, around loop C must be equal to the sum of circulations of all the enclosed pentagons (shaded), for (a) the top pentagon only; (b) the top six pentagons; and (c) all but the bottom pentagon.

the bottom outward-directed pentagon backwards, so the circulation on this path should have been $+\pi/6$, not $-11\pi/6$. Is this a contradiction? No, for we are saved by the fact that a Berry phase is only well-defined modulo 2π, according to which $-11\pi/6$ and $+\pi/6$ are identical!

You can easily see that this argument generalizes: for any closed surface that is discretized into plaquettes, the total circulation must be 2π times an integer. In the continuum limit this becomes the famous Chern theorem, which states that the integral of the Berry curvature over any closed 2D manifold is quantized to be 2π times an integer.

Before deriving this theorem, it may first help to go back and clarify a potentially puzzling aspect of gauge invariance in the context of Stokes' theorem, Eq. (3.33). The right-hand side of this equation represents the Berry-curvature flux passing through surface patch S (i.e., the area integral of the surface-normal component of $\boldsymbol{\Omega}$ over S); since $\boldsymbol{\Omega}$ is fully gauge-invariant, the right-hand side is fully determined without any ambiguity. In contrast, the left-hand side is the Berry phase of the curve C that bounds S, and we know that a Berry phase is only well-defined modulo 2π. So which is it? Is there a 2π ambiguity, or not?

The answer is that if ϕ is to be determined using a knowledge of $|u_{\boldsymbol{\lambda}}\rangle$ *only on curve* C, then it is really only well-defined modulo 2π. In this case, we can rewrite Eq. (3.33) as

$$\int_S \boldsymbol{\Omega} \cdot d\mathbf{S} \;\coloneqq\; \oint_C \mathbf{A} \cdot d\boldsymbol{\lambda} \tag{3.36}$$

using a specialized notation that was introduced in Eq. (1.6). Recall that ':=' means that the unambiguously defined object on the left-hand side is equal to *one of the values* of the object on the right-hand side, which is ambiguous modulo a quantum. The meaning of Eq. (3.36), then, is that it is possible

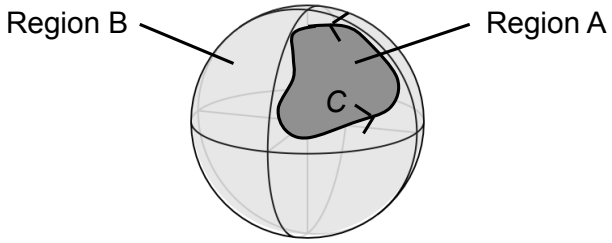

Figure 3.8 Proof of the Chern theorem for a manifold S having the topology of a sphere. Closed path C traces the boundaries of subregions A and B in the forward and reverse directions respectively. The uniqueness modulo 2π of the Berry phase around loop C is the key to the proof.

to make a choice of gauge for the phases of $|u_{\boldsymbol{\lambda}}\rangle$ around the loop C in such a way that this equation becomes an equality, while other choices will leave a mismatch of some integer multiple of 2π.

What kind of gauge gives the "correct" answer? Well, if we choose a gauge that is smooth and continuous *everywhere in* S, including on its boundary C, and use this gauge to evaluate the loop Berry phase, then Eq. (3.36) becomes an equality. For in this case, the logic leading to the derivation of Stokes' theorem – summing up circulations around small plaquettes to get the boundary circulation – is sound. While it is possible to make a radical gauge transformation that shifts ϕ_C by 2π when regarding $|u_{\boldsymbol{\lambda}}\rangle$ as a function defined on C only, such a gauge change cannot be smoothly continued into the interior S without creating a vortex-like singularity.

Now we are ready to prove the Chern theorem, which states that the integral of the Berry curvature over any closed 2D manifold is

$$\oint \boldsymbol{\Omega} \cdot \mathbf{dS} = 2\pi m \tag{3.37}$$

for some integer m. This integer is known as the "Chern number" or "Chern index" of the surface, and can be regarded as a "topological index" or "topological invariant" attached to the manifold of states $|u_{\boldsymbol{\lambda}}\rangle$ defined over the surface S.[10]

We prove this first for a surface having the topology of a simple sphere, as in Fig. 3.8, and divide the sphere into two regions A and B. The loop C forming the boundary between them traverses A in the forward direction and B in the reverse direction, so that applying Stokes' theorem to each of them,

[10] Another famous topological invariant is the Euler number χ, which is the integral of the Gaussian curvature over the surface S, and is related to the genus g via $\chi = 2 - 2g$. By contrast, the Chern index is not a characteristic of the surface itself, but of the manifold of states $|u_{\boldsymbol{\lambda}}\rangle$ defined over the surface.

we obtain $\int_A \mathbf{\Omega}\cdot d\mathbf{S} := \phi$ and $-\int_B \mathbf{\Omega}\cdot d\mathbf{S} := \phi$ where ϕ is the Berry phase around C. That is, the results of the two applications of Stokes' theorem must be consistent, but they only need to be consistent modulo 2π. Subtracting these two equations we get

$$\int_A \mathbf{\Omega}\cdot d\mathbf{S} + \int_B \mathbf{\Omega}\cdot d\mathbf{S} = \oint \mathbf{\Omega}\cdot \mathbf{dS} := 0 \tag{3.38}$$

which is equivalent to Eq. (3.37).

The same strategy applies to any orientable closed 2D surface, such as a torus. The general strategy is to decompose the surface S into an "atlas" composed of a series of "maps" (A and B above), such that a smooth and continuous gauge can be defined in each map. Then Stokes' theorem is applied to each map, and the results are summed. One side of the resulting equation is $\oint \mathbf{\Omega}\cdot\mathbf{dS}$ integrated over the entire surface S, while on the other is the sum of Berry phases along the boundaries, which must cancel modulo 2π. A more careful demonstration of the Chern theorem can be found in a number of topological physics texts such as those by Frankel (1997), Nakahara (2003), or Eschrig (2011), where a language of "fiber bundles" and ??? is typically introduced.

Looking back, we see that our result $\oint \mathbf{\Omega}\cdot\mathbf{dS} = -2\pi$ over the magnetic unit sphere in Sec. 3.2.1 is consistent with the Chern theorem with $m = -1$. In fact, if we had studied a spin-s particle with $s = 1$ or $s = 3/2$ instead of $s = 1/2$, we would have found $\Omega = -s$ and $m = -2s$. Since the latter has to be an integer, it follows that the only allowed spin representations are those with half-integral or integral s, a well-known fact which is usually derived in other ways in elementary quantum texts.

Note that when the Chern index is non-zero, it is impossible to construct a smooth and continuous gauge over the entire surface S. For, if there were such a gauge, then we could apply Stokes' theorem directly to the entire surface and conclude that the Chern number vanishes, in contradiction with the assumption. This is again well illustrated by the case of the spinor on the magnetic unit sphere. If we start from $\hat{\mathbf{n}} = +\hat{\mathbf{z}}$ and construct a gauge that is smooth in the vicinity of $\theta = 0$, and then extend this gauge as smoothly as possible with increasing θ, we get a gauge like that of Eq. (3.24). But while it is smooth in the "northern hemisphere," this gauge has a singularity ("vortex") at $\theta = \pi$, i.e., at the "south pole." This should remind you of the situation illustrated in Fig. 3.7, where a circulation of order 2π was left at the south pole at the end of the construction. If instead we start at $\hat{\mathbf{n}} = -\hat{\mathbf{z}}$ and work continuously towards the north pole, we can construct an equally

valid gauge described by

$$|\uparrow_{\hat{\mathbf{n}}}\rangle = \begin{pmatrix} \cos(\theta/2)e^{-i\varphi} \\ \sin(\theta/2) \end{pmatrix}. \tag{3.39}$$

But this gauge, while perfectly well-behaved in the southern hemisphere, has a vortex at the north pole. Indeed, there is no possible choice of gauge that is smooth and continuous everywhere on the unit sphere. In such a case, we say that the presence of a non-zero Chern index presents a "topological obstruction" to the construction of a globally smooth gauge.

Exercises

Exercise 3.2.1 Question...

3.3 Adiabatic dynamics

So far, we have been discussing the Berry phase as a property of the slow adiabatic evolution of a quantum system along a certain path in parameter space. It remains, however, to show how this relates to the actual quantum evolution of the system as described by the time-dependent Schrödinger equation.

Consider a Hamiltonian $H(\lambda)$ for some quantum system such as a molecule, with parameter $\lambda(t)$ being a slow function of t. (We shall quantify what is meant by "slow" shortly.) For a given λ the eigenstates of H are

$$H(\lambda)\,|n(\lambda)\rangle = E_n(\lambda)\,|n(\lambda)\rangle \tag{3.40}$$

where $n = 1, \ldots$ labels the eigenstates. We start the system in eigenstate n at time $t=0$ and then follow its subsequent time evolution.

If λ did not vary with t at all, the resulting wavefunction would evolve as $|\psi(t)\rangle = e^{-iE_n t/\hbar}|n\rangle$. In other words, the phase advances by an amount $e^{-iE_n\Delta t/\hbar}$ in an infinitesimal time interval Δt. Over a finite time the phase evolution is therefore

$$\prod e^{-iE_n\Delta t/\hbar} = e^{-i\sum E_n\Delta t/\hbar}\,.$$

In the continuum limit the sum turns into an integral, so we expect the phase evolution to be of the form $|\psi(t)\rangle = e^{i\gamma(t)}|n(t)\rangle$ with

$$\gamma(t) = -\frac{1}{\hbar}\int_0^t E_n(t')\,dt'\,. \tag{3.41}$$

This leads to the ansatz

$$|\psi(t)\rangle = c(t)\,e^{i\gamma(t)}\,|n(t)\rangle \tag{3.42}$$

where $|n(t)\rangle$ on the right-hand side is defined as $|n(\lambda(t))\rangle$, i.e., it is just the eigenstate $|n(\lambda)\rangle$ of the *time-independent* problem evaluated at $\lambda = \lambda(t)$. The factor $c(t)$ allows for the possibility that there may be some extra evolution beyond the guess based on $\gamma(t)$.

We shall see shortly that the ansatz (3.42) is only the zero-order term in a perturbation expansion in $\dot{\lambda} = d\lambda/dt$, but for now we plug this ansatz into the time-dependent Schrödinger equation

$$\Big[i\hbar\partial_t - H(t)\Big]|\psi(t)\rangle = 0 \tag{3.43}$$

(where $\partial_t = \partial/\partial t$) to find

$$0 = \dot{c}(t)\,|n(t)\rangle + c(t)\,\partial_t|n(t)\rangle\,. \tag{3.44}$$

To derive this equation, note that the time derivative ∂_t acts on all three terms in Eq. (3.42), but the term involving $\partial_t e^{i\gamma(t)}$ cancels against the $H(t)|n(t)\rangle = E_n(t)|n(t)\rangle$ term, leaving the two terms above. Acting with $\langle n(t)|$ on the left on both sides yields

$$\dot{c}(t) = ic(t)A_n(t) \tag{3.45}$$

where

$$A_n(t) = \langle n(t)|i\partial_t n(t)\rangle\,. \tag{3.46}$$

Comparing with Eq. (3.13) we note that $A_n(t)$ is a "Berry connection in time." The solution of Eq. (3.45) is just $c(t) = e^{i\phi(t)}$ with

$$\phi(t) = \int_0^t A_n(t')\,dt' \tag{3.47}$$

which we immediately recognize as a Berry phase.

Moreover, this Berry phase can be reexpressed in terms of λ. That is, since $|n(t)\rangle$ is defined as $|n(\lambda(t))\rangle$, application of the chain rule yields $\partial_t|n(t)\rangle = \dot{\lambda}\,\partial_\lambda|n(\lambda)\rangle$. It follows that $A_n(t) = \dot{\lambda}\,A_n(\lambda)$ where $A_n(\lambda) \equiv \langle n(\lambda)|i\partial_\lambda n(\lambda)\rangle$ is the Berry potential in parameter space. Substituting into Eq. (3.47) and using $d\lambda = \dot{\lambda}dt$ we then find

$$\phi(t) = \int_{\lambda(0)}^{\lambda(t)} A_n(\lambda)\,d\lambda\,. \tag{3.48}$$

This is a remarkable result; it says that the Berry phase entering into the time-evolving wavefunction is only a function of the path it has traced in parameter space, and is independent of the rate at which the path is traversed, so long as the parametric evolution is sufficiently slow.

Let's take stock. Our ansatz of Eq. (3.42) is successful only if the extra

Berry-phase term is included. The result is that, at leading order in adiabatic perturbation theory, the wavefunction evolves as

$$|\psi(t)\rangle = e^{i\phi(\lambda(t))}\, e^{i\gamma(t)}\, |n(t)\rangle \tag{3.49}$$

where the naively expected dynamical phase $e^{i\gamma}$ has to be augmented by the Berry phase $e^{i\phi}$ to find the correct phase evolution of the wavefunction.

As a special case, note that if we have chosen a parallel-transport gauge for $|n(\lambda)\rangle$, i.e., satisfying Eq. (3.20), the Berry-phase term is absent in Eq. (3.49). In other words, our derivation shows that, once the dynamical phase is factored out, the time evolution of the system is such that it follows a parallel-transport gauge.

There may be a tendency to think of the Berry-phase factor in Eq. (3.49) as "only a phase" with little in the way of physical consequences, since probabilities, not amplitudes, determine physical observations. However, as mentioned on p. 70, the Berry phase sometimes plays a crucial role by giving rise to interference phenomena. For example, if duplicate copies of a system are prepared and propagated along two paths on which they experience different Berry phases, this difference manifests itself when the systems are recombined. A simple example is discussed in Ex. 3.3.1.

We hinted earlier that higher-order terms might need to be added to Eq. (3.42) or (3.49) for some purposes. One situation where this is absolutely crucial is the discussion of *adiabatic charge transport*. This will play an important role for crystalline systems in Ch. 4, but for simplicity we consider it here only for a finite system such as an atom or molecule. Recall that the current density for an electron in state $|\psi\rangle$ is $(ie\hbar/2m)[\psi^*(\mathbf{r})\nabla\psi(\mathbf{r}) - \psi(\mathbf{r})\nabla\psi^*(\mathbf{r})]$, which vanishes identically if $\psi(\mathbf{r})$ is real. This will also be true of the wavefunction in Eq. (3.49) if $|n(\lambda)\rangle$ is real, since the phase factors in front are independent of $\mathbf{r}$. But in a typical case, such as for the ground electronic state of an H_2O molecule as one nucleus is gradually moved, $|n(\lambda)\rangle$ is indeed real. If we assumed (3.49), then, we would conclude that the motion of the nucleus induces no corresponding flow of electron charge. This is clearly nonsense, since the electronic charge density $\rho(\mathbf{r})$ changes with time, which it cannot do if there is no current flow.

The solution to this paradox is to carry the adiabatic perturbation theory to one higher power of $\dot{\lambda}$. We now expand Eq. (3.49) to become

$$|\psi(t)\rangle = e^{i\phi(\lambda(t))}\, e^{i\gamma(t)} \Big[\, |n(\lambda(t))\rangle + \dot{\lambda}\,|\delta n(t)\rangle \,\Big], \tag{3.50}$$

where the extra component $|\delta n(t)\rangle$ is to be determined. We already know that Eq. (3.50) solves the time-dependent Schrödinger equation to order zero

in $\dot{\lambda}$, but we now require that it should also do so at first order. For this purpose we can discard terms that go like $\ddot{\lambda}$ or $\dot{\lambda}^2$, including a $\dot{\lambda}\partial_t|\delta n\rangle$ term, and we find

$$(E_n - H_\lambda)\,|\delta n\rangle = -i\hbar\,(\partial_\lambda + iA_n)\,|n\rangle\,. \tag{3.51}$$

This is very similar to the inhomogeneous linear equation (2.75) for the perturbed wavefunction that arises in ordinary first-order perturbation theory, and has the formal solution

$$|\delta n\rangle = -i\hbar \sum_{m\neq n} \frac{\langle m|\partial_\lambda n\rangle}{E_n - E_m}\,|m\rangle \tag{3.52}$$

involving a sum over other eigenstates $|m(\lambda)\rangle$ at the same λ.[11] obtained in Using the notation of Sec. 2.3, this is just

$$|\delta n\rangle = -i\hbar T_n|\partial_\lambda n\rangle\,. \tag{3.53}$$

Note that time has again disappeared, and we can think in terms of evolution with respect to λ, except for the magnitude of $\dot{\lambda}$ appearing in Eq. (3.50).

We said that the adiabatic approximation should be valid if the Hamiltonian varies "slowly enough," and we are now in a position to quantify this. Namely, it should apply so long as $\dot{\lambda}\,|\delta n\rangle$ is small compared to $|n\rangle$. For an order-of-magnitude estimate we can replace $i\langle m|\partial_\lambda n\rangle$ by A_n and $E_n - E_m$ by a characteristic energy separation ΔE to find that the small dimensionless parameter describing the "slowness" of adiabatic evolution is $\hbar A_n\dot{\lambda}/\Delta E$.

An interesting feature of adiabatic perturbation theory is the fact that, leaving aside the phase information encoded in the Berry phase, the time-evolving wavefunction has only a short-term memory of the history of the path. Keeping terms to first order in $\dot{\lambda}$, for example, the state at time t depends only on $H_{\lambda(t')}$ for times t' at, and infinitesimally prior to, the current time t. This memory gets pushed back a little further as higher-order terms are included, but the overall picture is that the state vector rapidly "forgets" what happened earlier in its evolution along the path.

If we now use Eq. (3.50) when taking the expectation value of the current operator at order $\dot{\lambda}$, we can check that it does correctly describe the charge transport during the adiabatic evolution. This is carried through in Ex. 3.3.2.

Finally, we note that a common context for the application of adiabatic evolution is that of a system with "fast" and "slow" variables. The canonical example is that of electrons in molecules and solids, where the electron is many orders of magnitude lighter than then nuclei. From the viewpoint of

[11] A similar expression, but derived for the one-particle density matrix instead of the wavefunction, appears as Eq. (2.10) of Thouless (1983).

the time-evolving electron system, it is often an excellent approximation to treat the nuclear coordinates as classical variables that evolve slowly along a prescribed path. However, there is also a back-reaction on the system of nuclei, so that in a quantum treatment they experience a "gauge field" and a "gauge potential" arising from the electron system as it adiabatically follows the nuclear one. This is an important part of the theory of Berry phases as applied to molecular physics, and is outlined briefly in App. D.

Exercises

Exercise 3.3.1

Exercise

Exercise 3.3.2 Here we compute the current induced by an adiabatic change of the Hamiltonian and check that it correctly predicts the change in the electric dipole moment.

(a) Using Eq. (3.50), simplified to $|\psi(t)\rangle = e^{i\alpha}(|n\rangle + \dot{\lambda}\,|\delta n\rangle)$, show that the induced change in some arbitrary operator $\mathcal{O}$ is $\langle\mathcal{O}\rangle = 2\dot{\lambda}\mathrm{Re}\,\langle n|\mathcal{O}|\delta n\rangle$.

(b) Defining the current operator $\boldsymbol{\mathcal{J}} = -e\mathbf{v}$ in terms of the velocity operator $\mathbf{v}$ and using Eq. (3.52), show that

$$\langle\boldsymbol{\mathcal{J}}\rangle = -2e\hbar\dot{\lambda}\,\mathrm{Im}\sum_{m\neq n}\frac{\langle n|\mathbf{v}|m\rangle\langle m|\partial_\lambda n\rangle}{E_n - E_m}\,.$$

(c) Using Eq. (2.42), show that this becomes $\langle\boldsymbol{\mathcal{J}}\rangle = -2e\dot{\lambda}\,\mathrm{Re}\,\langle n|\mathbf{r}|\partial_\lambda n\rangle$. Hint: Note that $\langle n|\mathbf{r}|n\rangle\langle n|\partial_\lambda n\rangle$ is pure imaginary (why?).

(d) Noting that $\langle\boldsymbol{\mathcal{J}}\rangle$ has the interpretation of $d\mathbf{d}/dt$, where $\mathbf{d} = -e\mathbf{r}$ is the dipole operator, and cancelling the dt, show that this becomes $\partial_\lambda\langle\mathbf{d}\rangle = -2e\,\mathrm{Re}\,\langle n|\mathbf{r}|\partial_\lambda n\rangle = -e\partial_\lambda\langle n|\mathbf{r}|n\rangle$, which is self-evident. This shows that the calculation of the adiabatically induced current does correctly predict the change in electric dipole of the system.

3.4 Berryology of the Brillouin zone

Up until now in this chapter, we have considered Berry phases, connections and curvatures defined for some $|u_{\boldsymbol{\lambda}}\rangle$ in a generic parameter space $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ...)$. We now turn to the main theme of this book, where we specialize to the case that these parameters are the wavevector components k_j labeling Bloch states $|\psi_{n\mathbf{k}}\rangle$ of band n in the BZ, as described in Sec. 2.1.3.

We assume for the moment that band n is *isolated*, i.e., that it does not touch bands $n \pm 1$ anywhere in the BZ. This is a significant restriction, as

degeneracies between bands at high-symmetry points in the BZ are common in crystalline materials. When they occur, they typically introduce a non-analytic dependence of $|\psi_{n\mathbf{k}}\rangle$ on $\mathbf{k}$, which is problematic for the definitions of the Berry connection and curvature. We will lift this restriction in Sec. 3.6, but for now it allows us to assume that such singularities are not present, and the Berry formalism should apply.

However, we immediately encounter an important subtlety. Should we define the Berry phase and curvature in terms of the Bloch functions $|\psi_{n\mathbf{k}}\rangle$, or their cell-periodic versions $|u_{n\mathbf{k}}\rangle$? The answer is that we *must* use the latter. To see this, consider the case of a discretized Berry phase for a 1D crystal. If we were to substitute the $|u_j\rangle$ of Eq. (3.1) by the $|\psi_{nk}\rangle$, we would need to compute inner products $\langle\psi_{nk}|\psi_{n,k+b}\rangle$ which take a form like

$$\int_{-\infty}^{\infty} \psi_{nk}^*(x)\psi_{n,k+b}(x)\,dx = \int_{-\infty}^{\infty} e^{ibx}\,u_{nk}^*(x)u_{n,k+b}(x)\,dx \tag{3.54}$$

for some small but finite k-space separation b. However, the product of u factors on the right-hand side above is periodic with the unit cell, so that the phase factor e^{ibx} will average to zero when the integral is carried over all x. Thus, this inner product is ill-defined. Instead, the expression

$$\langle u_{nk}|u_{n,k+b}\rangle = \int_{\text{cell}} u_{nk}^*(x)u_{n,k+b}(x)\,dx \tag{3.55}$$

is perfectly well-behaved. (We adopt a a single-unit-cell normalization convention, i.e., $\int_{\text{cell}} |u_{nk}(x)|^2 = 1$.)

The essential observation is that all of the $|u_{nk}\rangle$ at different k have the same boundary conditions, and thus belong to the same Hilbert space. As a result, inner products between vectors at different k, or derivatives with respect to k, are well defined. This would not be the case if the formalism were based on the $|\psi_{nk}\rangle$ vectors. Note, however, that the k dependence reappears in a different guise, in that the $|u_{nk}\rangle$ are now solutions of a k-dependent Hamiltonian H_k as given by Eq. (2.39).

It is now straightforward to take the formalism of Sec. 3.1 over to the case of Bloch functions in the BZ. Returning to 3D, a Berry phase associated with band n takes the form

$$\phi_n = \oint \mathbf{A}_n(\mathbf{k}) \cdot d\mathbf{k} \tag{3.56}$$

where the Berry connection is

$$A_{n\mu}(\mathbf{k}) = \langle u_{n\mathbf{k}}|i\partial_\mu u_{n\mathbf{k}}\rangle \tag{3.57}$$

with $\partial_\mu = \partial/\partial k_\mu$, or equivalently, $\mathbf{A}_n(\mathbf{k}) = \langle u_{n\mathbf{k}}|i\nabla_{\mathbf{k}} u_{n\mathbf{k}}\rangle$. Similarly, the Berry

curvature is

$$\Omega_{n,\mu\nu}(\mathbf{k}) = \partial_\mu A_{n\nu}(\mathbf{k}) - \partial_\nu A_{n\mu}(\mathbf{k}) = -2\,\mathrm{Im}\,\langle \partial_\mu u_{n\mathbf{k}} | \partial_\nu u_{n\mathbf{k}} \rangle\,, \tag{3.58}$$

which in 3D can be reexpressed in pseudovector form as $\mathbf{\Omega}_n(\mathbf{k})$. And as before, we have a gauge freedom to transform the Bloch functions as

$$|\tilde{u}_{n\mathbf{k}}\rangle = e^{-i\beta(\mathbf{k})}|u_{n\mathbf{k}}\rangle \tag{3.59}$$

where $\beta(\mathbf{k})$ is some real function of $\mathbf{k}$. The Berry connection

$$\widetilde{\mathbf{A}}_n(\mathbf{k}) = \mathbf{A}_n(\mathbf{k}) + \nabla_{\mathbf{k}}\beta(\mathbf{k}) \tag{3.60}$$

is gauge-dependent; the curvature $\Omega_{n,\mu\nu}(\mathbf{k})$ is fully gauge-invariant; and the Berry phase of Eq. (3.56) is invariant modulo 2π.

It is, of course, equally straightforward to develop the theory in terms of the reduced wavevector $\boldsymbol{\kappa}$ of Eqs. (2.26-2.27). In this case, the derivatives entering the formalism are redefined as $\partial_\mu = \partial/\partial\kappa_\mu$.

In the discretized version of this theory, which will be used in any practical calculation, we have to take inner products of the form $\langle u_{n\mathbf{k}} | u_{n,\mathbf{k}+\mathbf{b}} \rangle$ between neighboring points $\mathbf{k}$ and $\mathbf{k}+\mathbf{b}$ in the BZ. This is quite unlike what we usually encounter when computing the expectation values of observables, Eq. (2.41), where the same wavevector $\mathbf{k}$ appears in both the bra and the ket. If the Berry phase and curvature are to have any physical significance, it will have to be in the context of a paradigm rather different from that of ordinary observables and their expectation values.

You may be wondering whether the Berry phases, connections, and curvatures defined above are actually nonzero in crystals of interest. This question requires better framing for the case of the Berry phase, where it depends on choice of path, and for the Berry potential, which depends on the gauge choice. But the Berry curvature $\mathbf{\Omega}_n(\mathbf{k})$ for band n is a uniquely defined function in the BZ. It is fairly straightforward to show that:

1. If the crystal has inversion (I) symmetry, then $\mathbf{\Omega}_n(\mathbf{k}) = \mathbf{\Omega}_n(-\mathbf{k})$.
2. If the crystal has time-reversal (TR) symmetry, then $\mathbf{\Omega}_n(\mathbf{k}) = -\mathbf{\Omega}_n(-\mathbf{k})$. Quantities involving an integral of $\mathbf{\Omega}_n$ over the BZ will vanish.
3. If the crystal has both I and TR symmetry,[12] then $\mathbf{\Omega}_n(\mathbf{k}) = 0$ identically.
4. If the crystal has some other spatial or magnetic symmetries, as described by the magnetic point group, then there are additional relations imposed on $\mathbf{\Omega}_n(\mathbf{k})$. For example, a simple 3-fold axis imposes the corresponding 3-fold rotational symmetry on the $\mathbf{\Omega}_n(\mathbf{k})$ field.

[12] Actually, it is enough if the crystal has I∗TR symmetry

It is easy to remember these rules if you note that they are the same as those governing the magnetic field $\mathbf{B}(\mathbf{r})$ of a molecule having some specified symmetries, but in reciprocal space rather than real space.

Items 2 and 3 suggest that the Berry curvature will be of most interest in magnetic systems, i.e., those with spontaneously broken TR symmetry. We shall see in Chs. 5 and 7-8 that this is indeed the case. Certainly the Berry curvature vanishes everywhere in the BZ for many nonmagnetic centrosymmetric materials such as Si (diamond structure), Cu (fcc structure), or Bi_2Se_3 (a van der Walls bonded layered structure). Nevertheless, it turns out that Berry-phase concepts also play a central role in aspects of these materials, as will be discussed in Chs. 4 and 6.

Since we have mentioned magnetic materials, we should briefly return to the subject of Sec. 2.1.2 and clarify the treatment of spin and spin-orbit coupling.

- If spin-orbit coupling is *not* included, then the systems of spin-up and spin-down electrons can be treated independently. If in addition the system is *nonmagnetic*, then spin up and down behave identically; we can think of the electrons as scalar particles and label bands by a single integer n, including a factor of two in sums and integrals such as in Eq. (2.41) to account for spin degeneracy. In a magnetic system, instead, we need to add a spin label s to the Bloch functions $|u_{ns\mathbf{k}}\rangle$ and carry the same label on Berry-related quantities such as ϕ_{ns}, $\mathbf{A}_{ns}$, and $\mathbf{\Omega}_{ns}$.
- If spin-orbit coupling is *included*, then the number of band labels n is doubled, and the Bloch functions $|u_{n\mathbf{k}}\rangle$ are spinor wavefunctions in the sense of Eq. (2.19). The Berry-related quantities carry only the label n, but the inner products in Eqs. (3.57-3.58) are taken between spinor wavefunctions.

We shall soon have to ask what physical interpretation can be given to the Berry phases and curvatures associated with the energy bands, but before we do, let us discuss what kinds of *paths* the Berry phases might be defined on. For a 1D crystal (of lattice constant a) there is really only one closed path of interest, namely, the one that circulates around the BZ. Recall from the discussion on pp. 39-40 that the BZ of a 1D system is best viewed as a closed loop or unit circle. This is illustrated in Fig. 3.9, where the panel at left shows the conventional view in which an energy band $E_n(k)$ is plotted as a periodic function of k with period $2\pi/a$. However, this picture is misleading, since a state at k is *one and the same* as a state at $k + 2\pi/a$. In other words, these two wavevectors are duplicate labels for the same state. The panel at right shows a more unconventional and yet more natural way of thinking about an energy band, in which the wavevector axis is wrapped into a circle

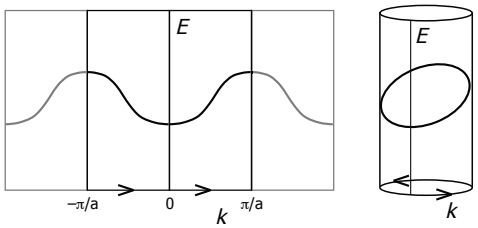

Figure 3.9 At left, conventional view of a 1D band structure, with the first Brillouin zone highlighted in black and the extended-zone scheme shown in gray. At right, a more topologically natural view in which the Brillouin zone is wrapped onto a circle and the band structure is plotted on a cylinder.

and the energy is plotted on the surface of the resulting cylinder. Then a possible object of interest is the Berry phase

$$\phi_n = \oint_{\mathrm{BZ}} A_n(k)\, dk = \oint_{\mathrm{BZ}} \langle u_{nk} | i\partial_k u_{nk} \rangle \tag{3.61}$$

defined on this loop. The notation $\oint_{\mathrm{BZ}}$ indicates an integral taken around the loop formed by the 1D BZ.

Such a Berry phase was first discussed by Zak (1989) and is sometimes called a "Zak phase." In the introductory paragraph of this paper, Zak listed some of the areas in which the concept of the Berry phase was making an impact in atomic, molecular, and nuclear physics, and then wrote

"It seems, however, that one important and natural system for the appearance of Berry's phase was left out. We have in mind the motion of an electron in a periodic solid."

This farsighted observation by Zak set the stage for many of the later developments that are discussed in this book. At the time, Zak was mainly concerned with symmetry properties and the relation of the Berry phase to the so called "band center," which we now identify with a Wannier center (see Sec. 3.5). It was to take a few years before the connection to the theory of electric polarization would be made, as will be discussed in the next chapter.

There is an important subtlety that arises when it comes to computing this Berry phase. For the Berry phase to be well-defined, we need $|u_{nk}\rangle$ to be a smooth function of k everywhere on the loop. If we represent the loop by letting k range from 0 to $2\pi/a$, for example,[13] then we also have to insure smoothness across the artificial boundary point where k crosses from $2\pi/a$

[13] Similar issues arise for any other choice of BZ, such as from $-\pi/a$ to π/a.

back to 0. That is, we must insist that

$$\psi_{n,k=2\pi/a}(x) = \psi_{n,k=0}(x)\,. \tag{3.62}$$

That is, the Bloch functions at the two ends of the interval $[0, 2\pi/a]$ must be equal not just up to a phase, but with the same phase. In an extended-zone context, where quantities such as E_{nk} are regarded as periodic functions of k, Eq. (3.62) is referred to as the "periodic gauge condition."

But recall that the Berry connection is defined not in terms of the Bloch functions $|\psi_{nk}\rangle$, but their cell-periodic partners $|u_{nk}\rangle$. Using Eq. (2.37), the condition of Eq. (3.62) translates to the condition

$$u_{n,k=2\pi/a}(x) = e^{-2\pi i x/a} u_{n,k=0}(x)\,. \tag{3.63}$$

So the vectors $|u_{n,k=2\pi/a}\rangle$ and $|u_{n,k=0}\rangle$ are not equal! And it is not just that they are unequal up to a global phase, since the phase factor in Eq. (3.63) is x-dependent. This is an essential feature of the cell-periodic Bloch functions; we have to learn to live with it.

To compute the Berry phase in practice, we discretize the BZ into N equal intervals and compute $|u_{nk_j}\rangle$ for $j = 0, ..., N-1$ with $k_j = 2\pi j/N$. This typically involves calling a matrix diagonalization routine that returns eigenvectors whose phase is not under our control. Thus, if we were to call this routine to compute $|u_{nk_0}\rangle$ and $|u_{nk_N}\rangle$ independently, we *cannot* assume that they will obey Eq. (3.63). Instead, we use Eq. (3.63) to construct $|u_{nk_N}\rangle$ from $|u_{nk_0}\rangle$, with the correct phase relation. That is, we compute the Berry phase of band n as

$$\phi_n = -\mathrm{Im}\ln\Big[\langle u_{nk_0}|u_{nk_1}\rangle\langle u_{nk_1}|u_{nk_2}\rangle ... \langle u_{nk_{N-1}}|e^{-2\pi i x/a}|u_{nk_0}\rangle\Big]. \tag{3.64}$$

It is important not to forget the phase factor in the last term when completing the loop, i.e., when k wraps from $k = 2\pi/a$ back to $k = 0$.

In two dimensions, there is more freedom in the choice of path. Some examples are shown in Fig. 3.10. This figure is drawn in terms of the reduced wavevectors κ_j of Eq. (2.30), each of which runs between 0 and 2π. Paths A and B are simple closed paths. Remember that the left and right edges of the BZ are identified, as are the top and bottom, with the BZ regarded as a closed torus. From this point of view, there is no intrinsic difference between a path like A that lies entirely inside the square BZ shown in the figure, and one like B that traverses the artificial boundary at $\kappa_1 = 2\pi$.[14]

On the other hand, paths C and D wind around the BZ in a nontrivial way, in analogy to the 1D path described earlier. Path C returns to itself

[14] For the latter, however, one would have to include phase factors like the one in Eq. (3.65) when crossing the conventional BZ boundaries.

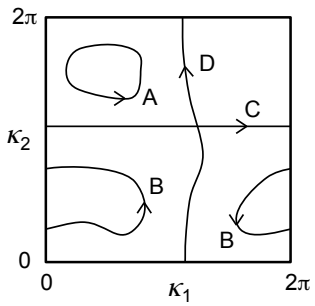

Figure 3.10 Sketch of four possible closed paths in the BZ of a 2D crystal. Paths A and B are trivially closed, while paths C and D wrap by reciprocal lattice vector $\mathbf{b}_1$ and $\mathbf{b}_2$ respectively.

after κ_1 has been incremented by 2π. This corresponds to translating $\mathbf{k}$ by $\mathbf{b}_1$, so when crossing the artificial boundary at $\kappa_1 = 2\pi$ an extra phase factor

$$\langle u_{n\mathbf{k}_{N-1}} | e^{-i\mathbf{b}_1\cdot\mathbf{r}} | u_{n\mathbf{k}_0} \rangle \tag{3.65}$$

has to be included, in analogy with Eq. (3.64). Similar considerations apply to path D, which winds in direction κ_2 instead.[15]

In 2D, we can also define a Chern number m_n, defined via

$$\oint_{\mathrm{BZ}} \Omega \, dS = 2\pi m_n \,, \tag{3.66}$$

associated with each band n. As in Eq. (3.61), the notation $\oint_{\mathrm{BZ}}$ denotes an integral over the BZ regarded as a closed manifold, but now it is a surface integral over a 2D manifold. In practice the Chern number is most easily computed by discretizing the 2D BZ on an $N \times N$ mesh $\mathbf{k}_{j_1 j_2} = (j_1/N)\mathbf{b}_1 + (j_2/N)\mathbf{b}_2$. One then computes the eigenvectors on the $(N+1)^2$ mesh points as the j_μ ($\mu = 1, 2$) run over $0, ..., N$, computes the Berry phase around each of the N^2 plaquettes [with branch choice $(-\pi, \pi)$], and sums these.[16] Following the argument given in Sec. 3.2.2, this must be an integer multiple of 2π, from which it is straightforward to extract the Chern index m_n.

In 2D, a *periodic gauge* is defined as one for which $|\psi_{n,\mathbf{k}+\mathbf{b}_\mu}\rangle = |\psi_{n\mathbf{k}}\rangle$ on the boundaries of the conventional BZ, so that it can be wrapped smoothly onto the 2-torus. A periodic gauge that is also smoothly defined on the interior of the BZ would thus be smooth and continuous everywhere on the 2-torus. But recall from the discussion on p. 79 that it is *impossible* to construct such a gauge when the Chern index m_n is nonzero. Such a topological obstruction

[15] Similar factors would have to be included twice for path B as well.

[16] Alternatively the states at $k_\mu = N$ can be obtained from those at $k_\mu = 0$ by multiplying by $e^{-i\mathbf{b}_\mu\cdot\mathbf{r}}$ following Eq. (3.63).

causes no problem for the calculation of particular Berry phases, such as those along paths A-D, or of the Chern number, which is an integral of a gauge-invariant quantity. But it will have other consequences, especially regarding the ability to construct Wannier functions, as we shall discuss in the next section.

For a 3D crystal the BZ forms a 3-torus. Concerning Berry phases, we can consider simple loops that close without winding around the 3-torus, or ones that wind in the $\mathbf{b}_1$, $\mathbf{b}_2$, or $\mathbf{b}_3$ direction (e.g., like path D in Fig. 3.10 but in 3D). Concerning Chern indices, we can compute these on any closed 2D manifold lying in the 3D BZ. For example, the Fermi surface of a metal is a suitable closed surface on which a Chern index can be computed. But we can also consider 2D manifolds that span the 3D BZ. For example, consider the Chern index $m_{1n}(k_1)$ defined on the κ_2-κ_3 "plane" lying at some κ_1. This "plane" is really a 2-torus when we recall that the "edges" at $\kappa_j = 0$ and 2π can be regarded as being seamlessly glued together. But if band n is isolated (i.e., it does not touch band $n-1$ or $n+1$ anywhere in the 3D BZ), then all properties of band n on this plane must evolve smoothly as κ_1 is varied. In particular, $m_{1n}(\kappa_1)$ must be a continuous function of κ_1. But it is also an integer-valued function, and a continuous inter-valued function must be completely constant. We are thus free to compute it at any κ_1 of our choice (say $\kappa_1 = 0$) and to drop the κ_1 argument, writing it simply as m_{1n}. Similarly, defining m_{2n} and m_{3n} to be the Chern indices for κ_3-κ_1 and κ_1-κ_2 planes respectively, we conclude that any fully isolated band n in a 3D crystal is characterized by a triplet of integer Chern indices (m_{n1}, m_{n2}, m_{n3}).

So far, all of this is very abstract. What, if any, physical interpretation can be attached to the Berry-phase quantities described above? This is the subject that will concern us in subsequent chapters. Before we finally turn to this part of the story, however, it is useful to cover two more details of a somewhat mathematical nature. These are the construction of the Wannier representation and the multiband treatment, which will be discussed in the two remaining sections of this chapter.

Exercises

Exercise 3.4.1 Exercises

3.5 Wannier functions

If we have an isolated band $E_n(\mathbf{k})$, i.e., one that never touches the band below or above it, then we have a right to expect that $E_n(\mathbf{k})$ is a smooth

and periodic function of $\mathbf{k}$ in 3D reciprocal space. It is then natural to consider its Fourier transform to real space, defined by

$$E_{n\mathbf{R}} = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}}\, E_{n\mathbf{k}}\, d^3k\,, \tag{3.67a}$$

$$\Updownarrow \text{ FT}$$

$$E_{n\mathbf{k}} = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}\, E_{n\mathbf{R}}\,. \tag{3.67b}$$

The second equation above is the inverse transform; the consistency between this pair of equations is associated with the two orthogonality identities

$$\int_{\text{BZ}} e^{i\mathbf{k}\cdot(\mathbf{R}-\mathbf{R}')}\, d^3k \;=\; \frac{(2\pi)^3}{V_{\text{cell}}}\, \delta_{\mathbf{R},\mathbf{R}'}\,, \tag{3.68}$$

which is relatively intuitive,[17] and

$$\sum_{\mathbf{R}} e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}} \;=\; \frac{(2\pi)^3}{V_{\text{cell}}}\, \delta^3(\mathbf{k}-\mathbf{k}'), \tag{3.69}$$

which is less so.[18] These and other Fourier transform conventions are summarized in App. B. Insofar as $E_n(\mathbf{k})$ is smooth in $\mathbf{k}$-space, we can expect $E_{n\mathbf{R}}$ to be large only for a few lattice vectors $\mathbf{R}$ near the origin, and to decay rapidly with increasing $|\mathbf{R}|$.

Now suppose we can choose a smooth and periodic gauge for the Bloch functions $|\psi_{n\mathbf{k}}\rangle$ associated with this band. Having done so, we should be able to Fourier transform these in a similar way, i.e., by defining

$$|w_{n\mathbf{R}}\rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}}\, |\psi_{n\mathbf{k}}\rangle\, d^3k\,, \tag{3.70a}$$

$$\Updownarrow \text{ FT}$$

$$|\psi_{n\mathbf{k}}\rangle = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}\, |w_{n\mathbf{R}}\rangle\,. \tag{3.70b}$$

The Fourier-transform partners to the Bloch functions defined in Eq. (3.70a) are known as the *Wannier functions* associated with band n; Eq. (3.70b) provides the inverse transform back from Wannier to Bloch functions. Again, the idea is that as long as $\psi_{n\mathbf{k}}(\mathbf{r})$ is a smooth function of $\mathbf{k}$, then $w_{n\mathbf{R}}(\mathbf{r})$ decays rapidly with $|\mathbf{R}|$ for a given $\mathbf{r}$. Actually, it turns out that each Wannier function $w_{n\mathbf{R}}(\mathbf{r})$ is a localized function centered near $\mathbf{R}$, so it is more natural to describe the situation by saying that $w_{n\mathbf{R}}(\mathbf{r})$ decays rapidly with $|\mathbf{r}-\mathbf{R}|$ for a given $\mathbf{R}$. Moreover, since the Fourier transform expressed by Eq. (3.70) is

[17] The phases cancel on the left unless $\mathbf{R}=\mathbf{R}'$, in which case the left side is the BZ volume.

[18] The wavevectors $\mathbf{k}$ on the right side should be interpreted as living on the 3-torus, i.e., $\mathbf{k}-\mathbf{k}'$ can be replaced by $\mathbf{k}-\mathbf{k}'+\mathbf{G}$ for any reciprocal lattice vector $\mathbf{G}$.

really just a special case of a unitary transformation, we can view the Bloch and Wannier functions as providing two different basis sets to describe the same manifold of states associated with the electron band in question.

A word of warning is in order concerning the inner-product and normalization conventions to be used here. Recall that the true Bloch function $|\psi_{\mathbf{k}}\rangle$ is related to the cell-periodic function $|u_{\mathbf{k}}\rangle$ by $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r})$. Also let $|\chi_{\mathbf{k}}\rangle$ and $|v_{\mathbf{k}}\rangle$ be another pair related in the same way. For the cell-periodic functions we adopt the inner-product convention

$$\langle u_{\mathbf{k}} | v_{\mathbf{k}'} \rangle \equiv \int_{V_{\text{cell}}} u^*_{\mathbf{k}}(\mathbf{r}) v_{\mathbf{k}'}(\mathbf{r})\, d^3r\,, \tag{3.71}$$

since the "natural domain" of such a function is a single unit cell. Moreover, we normalize $\psi_{n\mathbf{k}}(\mathbf{r})$ and $u_{n\mathbf{k}}(\mathbf{r})$ such that

$$\int_{V_{\text{cell}}} |\psi_{n\mathbf{k}}(\mathbf{r})|^2 = \int_{V_{\text{cell}}} |u_{n\mathbf{k}}(\mathbf{r})|^2 = \langle u_{n\mathbf{k}} | u_{n\mathbf{k}} \rangle = 1\,. \tag{3.72}$$

However, we use a very different inner-product convention for the Bloch functions themselves, namely

$$\langle \psi_{\mathbf{k}} | \chi_{\mathbf{k}'} \rangle \equiv \int \psi^*_{\mathbf{k}}(\mathbf{r}) \chi_{\mathbf{k}'}(\mathbf{r})\, d^3r \tag{3.73}$$

where the integral is over *all space*. This is a natural choice for the Bloch functions, which describe physical electrons that are delocalized throughout the crystal. It follows that $\langle \psi_{n\mathbf{k}} | \psi_{n\mathbf{k}} \rangle$ is not unity, but is instead infinite, within our conventions. A useful formula, somewhat analogous to Eq. (3.69), is

$$\langle \psi_{\mathbf{k}} | \chi_{\mathbf{k}'} \rangle = \frac{(2\pi)^3}{V_{\text{cell}}} \langle u_{\mathbf{k}} | v_{\mathbf{k}'} \rangle\, \delta^3(\mathbf{k} - \mathbf{k}')\,. \tag{3.74}$$

This formula is derived in App. B as Eq. (B.13). From this it follows that the Wannier functions obey the orthonormality condition $\langle w_{n\mathbf{R}} | w_{n'\mathbf{R}'} \rangle = \delta_{n,n'}\, \delta_{\mathbf{R},\mathbf{R}'}$, as will be shown in Ex. 3.5.1.

3.5.1 Properties of the Wannier functions

In standard solid state physics texts it is demonstrated that the Wannier functions have the following interesting and useful properties:

1. As hinted above, they are *localized* functions in real space. That is,

$$|w_{n\mathbf{R}}(\mathbf{r})| \to 0 \;\text{ as }\; |\mathbf{r} - \mathbf{R}| \quad \text{gets large.} \tag{3.75}$$

   We can think of $|w_{n\mathbf{R}}\rangle$ as being peaked in cell $\mathbf{R}$, even if its tails extend into neighboring unit cells.

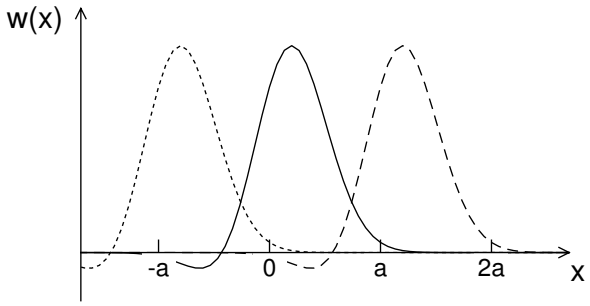

Figure 3.11 Sketch of three adjacent Wannier functions $w_{nR}(x)$ for band n in a 1D crystal of lattice constant a. The Wannier center assigned to the home unit cell $R = 0$ is shown as the full curve; dotted and dashed curves represent those in cells at $-a$ and a respectively. The Wannier functions are mutually orthonormal at the same time that they are translational images of one another.

2. The Wannier functions are translational images of one another, i.e.,

$$w_{n\mathbf{R}}(\mathbf{r}) = w_{n\mathbf{0}}(\mathbf{r} - \mathbf{R})\,. \tag{3.76}$$

   More formally, $|n\mathbf{R}\rangle = T_{\mathbf{R}}\,|n\mathbf{0}\rangle$ where, as in Sec. 2.1.3, $T_{\mathbf{R}}$ is the operator that translates the system by lattice vector $\mathbf{R}$.
3. The Wannier functions form an orthonormal set, i.e.,

$$\langle w_{n\mathbf{R}}|w_{n\mathbf{R}'}\rangle = \delta_{\mathbf{R}\mathbf{R}'}\,. \tag{3.77}$$

4. The Wannier functions span the same subspace of the Hilbert space as is spanned by the Bloch functions from which they are constructed. Defining $\mathcal{P}_n$ to be the projection operator onto band n, this property can be expressed as

$$\mathcal{P}_n \;=\; \frac{V_{\rm cell}}{(2\pi)^3}\int_{\rm BZ} |\psi_{n\mathbf{k}}\rangle\langle\psi_{n\mathbf{k}}|\;d^3k \;=\; \sum_{\mathbf{R}} |w_{n\mathbf{R}}\rangle\langle w_{n\mathbf{R}}|\,. \tag{3.78}$$

   From this it also follows that the total charge density $\rho_n(\mathbf{r})$ in band n,

$$\rho_n(\mathbf{r}) = -e\,\langle\mathbf{r}|\mathcal{P}_n|\mathbf{r}\rangle = -e\,\frac{V_{\rm cell}}{(2\pi)^3}\int_{\rm BZ} |\psi_{n\mathbf{k}}(\mathbf{r})|^2\;d^3k = -e\sum_{\mathbf{R}} |w_{n\mathbf{R}}(\mathbf{r})|^2\,,$$

   is the same when computed in either representation.

The first three properties above are illustrated in Fig. 3.11, where a possible set of Wannier functions are sketched for a 1D crystal. Each one is exponentially localized and normalized, and the neighboring Wannier functions are periodic images of one another. Moreover, the Wannier functions

are shown as having a negative lobe so that $\langle w_{n0}|w_{na}\rangle$ can plausibly vanish as a result of cancellation between contributions of opposite sign in the integral over x.

It is also possible to prove two remarkable properties about matrix elements of operators between Wannier functions:

5. The Hamiltonian matrix elements between Wannier functions are band-diagonal, and the diagonal elements are nothing other than the coefficients $E_{n\mathbf{R}}$ in the Fourier expansion of the band energy:

$$\langle w_{n\mathbf{0}}|H|w_{n\mathbf{R}}\rangle = E_{n\mathbf{R}}\,. \tag{3.79}$$

6. The position matrix elements between Wannier functions are

$$\langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{R}}\rangle = \mathbf{A}_{n\mathbf{R}}\,. \tag{3.80}$$

In this last equation the $\mathbf{A}_{n\mathbf{R}}$ are the Fourier transform coefficients of the Berry connection $\mathbf{A}_n(\mathbf{k})$ defined in Eq. (3.57). That is, they are related in complete analogy with Eq. (3.67) or (3.70) by

$$\mathbf{A}_{n\mathbf{R}} = \frac{V_{\text{cell}}}{(2\pi)^3}\int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}}\,\mathbf{A}_n(\mathbf{k})\,d^3k\,, \tag{3.81a}$$

$$\Updownarrow \text{FT}$$

$$\mathbf{A}_n(\mathbf{k}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}\,\mathbf{A}_{n\mathbf{R}}\,. \tag{3.81b}$$

We shall return shortly to provide the mathematical derivations of Eqs. (3.79) and (3.80), but before we do, let us comment on their significance. Equation (3.79) is a remarkable result, as it implies that the Wannier functions provide an *exact* tight-binding representation of the dispersion $E_{n\mathbf{k}}$ of band n. That is, we construct a TB model consisting of one orbital per cell having site energy $E_{n\mathbf{0}}$ and hoppings $E_{n\mathbf{R}}$ to its neighbors located at relative cell $\mathbf{R}$. In this context, Eq. (2.63) is exactly Eq. (3.67b), so that this Wannier-based TB model reproduces the exact band dispersion. Since the Wannier functions are localized, the hopping matrix elements fall off quickly with distance, so that only a small number of hoppings typically have to be retained.

Equation (3.80) is in some ways even more remarkable. It says that matrix elements of the position operator between Wannier functions have a form that is highly reminiscent of the Berry-phase formalism. An especially important quantity is the center of charge of a Wannier function,[19] defined

[19] This is analogous to the center of mass of an object, but defined based on charge rather than mass.

as the diagonal position-operator matrix element

$$\bar{\mathbf{r}}_n = \langle w_{n\mathbf{0}} | \mathbf{r} | w_{n\mathbf{0}} \rangle \,. \tag{3.82}$$

From Eq. (3.80) this is nothing other than $\mathbf{A}_{n\mathbf{0}}$, which Eq. (3.81a) tells us is just the BZ average of the Berry connection $\mathbf{A}_n(\mathbf{k})$, i.e.,

$$\begin{aligned} \bar{\mathbf{r}}_n &= \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} \mathbf{A}_n(\mathbf{k}) \, d^3k \\ &= \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} \langle u_{n\mathbf{k}} | i\nabla_{\mathbf{k}} u_{n\mathbf{k}} \rangle \, d^3k \,. \end{aligned} \tag{3.83}$$

In 1D this becomes $\bar{x}_n = (a/2\pi) \int_0^{2\pi/a} \langle u_{nk} | i\partial_k u_{nk} \rangle \, dk$ or

$$\bar{x}_n = a \, \frac{\phi_n}{2\pi} \,. \tag{3.84}$$

That is, the location of the Wannier center (in units of the lattice constant) is nothing other than the Berry phase associated with band n (in units of 2π)! This connection between Wannier centers and Berry phases is of fundamental importance, and lies at the heart of the developments to be presented in the next chapter.

We now return to the derivations of Eqs. (3.79) and (3.80), which were deferred earlier. For the Hamiltonian acting on a Wannier function we have

$$H \, |w_{n\mathbf{R}}\rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}} \, H |\psi_{n\mathbf{k}}\rangle \, d^3k \tag{3.85}$$

which we then multiply on the left by $\langle w_{n\mathbf{0}}|$. Since $H|\psi_{n\mathbf{k}}\rangle$ has the form of a Bloch function of wavevector $\mathbf{k}$,[20] we can use Eq. (3.74) to obtain

$$\langle w_{n\mathbf{0}} | H | w_{n\mathbf{R}} \rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}} \, \langle u_{n\mathbf{k}} | H_{\mathbf{k}} | u_{n\mathbf{k}} \rangle \, d^3k \tag{3.86}$$

which reduces to Eq. (3.79) using $E_{n\mathbf{k}} = \langle u_{n\mathbf{k}} | H_{\mathbf{k}} | u_{n\mathbf{k}} \rangle$ with Eq. (3.67).

The situation is illustrated for a band in a 2D crystal in Fig. 3.12. The position of the Wannier center in the home unit cell at $\bar{\mathbf{r}}_n$ is given by Eq. (3.82), or equivalently, Eq. (3.83). The Hamiltonian matrix elements between Wannier functions, defined in Eq. (3.86), are shown as the dashed hoppings in the figure. If all further neighbors are included, the tight-binding Hamiltonian defined in this way is guaranteed to reproduce the bandstructure of the crystal exactly. If only near-neighbor hoppings are retained, it will not longer be exact, but may still provide a very useful approximation to the full bandstructure.

[20] Actually it is just $E_{n\mathbf{k}}|\psi_{n\mathbf{k}}\rangle$ but we keep it in this form for the moment to make an analogy later with Eq. (3.89).

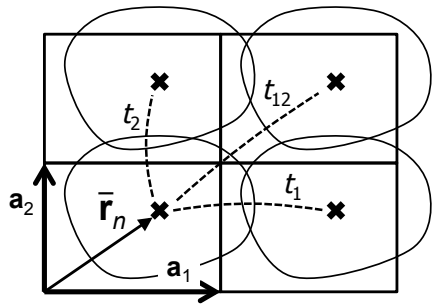

Figure 3.12 Sketch of four of the infinite lattice of Wannier functions (irregular blobs) for a single band in 2D with lattice vectors $\mathbf{a}_1$ and $\mathbf{a}_2$. Crosses mark the Wannier centers located at $\bar{\mathbf{r}}_n + \mathbf{R}$; dashed lines indicate the nearest-neighbor hoppings $t_1 = \langle w_{n\mathbf{0}}|H|w_{n\mathbf{a}_1}\rangle$, $t_2 = \langle w_{n\mathbf{0}}|H|w_{n\mathbf{a}_2}\rangle$, and $t_{12} = \langle w_{n\mathbf{0}}|H|w_{n,\mathbf{a}_1+\mathbf{a}_2}\rangle$ of the corresponding tight-binding model.

The equation that corresponds to Eq. (3.85), but for the position operator acting on $|w_{n\mathbf{R}}\rangle$ is a bit trickier to derive. We do it first in 1D, where we find that

$$\begin{aligned}(x-R)|w_{nR}\rangle &= \frac{a}{2\pi}\int_0^{2\pi/a}(x-R)\,e^{ik(x-R)}\,|u_{nk}\rangle\,dk \\ &= \frac{a}{2\pi}\int_0^{2\pi/a}\big(-i\partial_k e^{ik(x-R)}\big)\,|u_{nk}\rangle\,dk \\ &= \frac{a}{2\pi}\int_0^{2\pi/a} e^{ik(x-R)}\big(i\partial_k|u_{nk}\rangle\big)\,dk\,. \end{aligned} \tag{3.87}$$

In taking the last step above we have applied an integration by parts, making use of the fact that $e^{ik(x-R)}\,u_{nk}(x) = \psi_{nk}(x-R)$ has the same value at $k=0$ and $2\pi/a$. Generalizing this result to 3D and moving $\mathbf{R}$ to the other side, we find that the analogy to Eq. (3.85) is

$$\mathbf{r}\,|w_{n\mathbf{R}}\rangle = \frac{V_{\rm cell}}{(2\pi)^3}\int_{\rm BZ} e^{-i\mathbf{k}\cdot\mathbf{R}}\Big[e^{i\mathbf{k}\cdot\mathbf{r}}(\mathbf{R}+i\nabla_{\mathbf{k}})\,|u_{n\mathbf{k}}\rangle\Big]\,d^3k\,. \tag{3.88}$$

The object in brackets on the right-hand side has the form of a Bloch function of wavevector $\mathbf{k}$, so we can again multiply on the left by $\langle w_{n\mathbf{0}}|$ and use Eq. (3.74) to obtain

$$\langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{R}}\rangle = \frac{V_{\rm cell}}{(2\pi)^3}\int_{\rm BZ} e^{-i\mathbf{k}\cdot\mathbf{R}}\,\langle u_{n\mathbf{k}}|\mathbf{R}+i\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle\,d^3k\,. \tag{3.89}$$

The term involving $\mathbf{R}$ yields $\mathbf{R}\,\delta_{\mathbf{0},\mathbf{R}} = 0$ and can be discarded, yielding the important result

$$\langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{R}}\rangle = \frac{V_{\rm cell}}{(2\pi)^3}\int_{\rm BZ} e^{-i\mathbf{k}\cdot\mathbf{R}}\,\langle u_{n\mathbf{k}}|i\nabla_{\mathbf{k}}u_{n\mathbf{k}}\rangle\,d^3k\,. \tag{3.90}$$

We recognize the right-hand side as $\mathbf{A}_{n\mathbf{R}}$ from Eq. (3.81a), giving Eq. (3.80) as claimed.

3.5.2 Gauge freedom

We emphasized earlier that the phases of the Bloch functions are not unique; they can be twisted by a gauge change $e^{-i\beta(\mathbf{k})}$ as in Eq. (3.59). What effect does this have on Eqs. (3.79) and (3.80)? Because Eq. (3.86) is diagonal in $\mathbf{k}$ the phase factors $e^{i\beta(\mathbf{k})}$ cancel out, and the matrix elements $E_{n\mathbf{R}}$ are *exactly unchanged* by the gauge transformation. On one level this appears surprising, since the Wannier functions themselves may change shape, or become somewhat more or less localized, as a result of the gauge change. On the other hand, we showed that the $E_{n\mathbf{R}}$ are really just the Fourier-transform coefficients of the bandstructure $E_{n\mathbf{k}}$, which are unique, so they must indeed be gauge-invariant.

Before discussing the effect on the real-space matrix elements in Eq. (3.80), it is worthwhile to review the distinction between a *progressive* and a *radical* gauge transformation that was introduced in Sec. 3.1.2. In 1D, the idea is that each gauge transformation is characterized by a winding number m, which is the integer appearing in Eq. (3.19); the transformation is "progressive" if $m=0$ and "radical" otherwise. In 3D k-space the gauge twist $e^{i\beta(\mathbf{k})}$ is characterized by a triplet of integers that we denote by (n_1, n_2, n_3), such that $\beta(\mathbf{k}+\mathbf{b}_j) = \beta(\mathbf{k}) + 2\pi n_j$ where $\mathbf{b}_j$ is a primitive reciprocal lattice vector as defined above Eq. (2.29). Letting $\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$ be the real-space lattice vector defined by this triplet of integers, we find that each gauge transformation is characterized by a lattice vector $\mathbf{R}$ such that

$$\beta(\mathbf{k}+\mathbf{G}) = \beta(\mathbf{k}) + \mathbf{G}\cdot\mathbf{R}\,. \tag{3.91}$$

This condition is necessary and sufficient to insure that $e^{-i\beta(\mathbf{k})}$ is invariant under a translation by $\mathbf{G}$, i.e., that the transformation preserves the periodic-gauge property. A 3D gauge transformation is said to be "progressive" if $\mathbf{R}=\mathbf{0}$ and "radical" otherwise.

From the above, it is clear that any 3D gauge transformation can be decomposed into two steps: a radical part $\beta(\mathbf{k}) = \mathbf{k}\cdot\mathbf{R}$ for some $\mathbf{R}$, which we denote as a "shift," and a progressive part that obeys $\beta(\mathbf{k}+\mathbf{G}) = \beta(\mathbf{k})$. The effect of the shift is just to transform $\mathbf{A}_{n\mathbf{k}}$ via Eq. (3.60) into $\tilde{\mathbf{A}}_{n\mathbf{k}} = \mathbf{A}_{n\mathbf{k}} + \mathbf{R}$ so that $\tilde{\bar{\mathbf{r}}}_n = \bar{\mathbf{r}}_n + \mathbf{R}$, i.e., the Wannier center simply shifts by a lattice vector. In fact what has happened is that the Wannier functions themselves have all shifted by a lattice vector, i.e., $|\tilde{w}_{n\mathbf{R}'}\rangle = |w_{n,\mathbf{R}'+\mathbf{R}}\rangle$. This is really just a relabeling of the original Wannier functions by shifting all of them by the

same lattice vector $\mathbf{R}$, or equivalently, a change in the choice of the Wannier function assigned to the home unit cell $\mathbf{R}=\mathbf{0}$.

The effect of the progressive part can easily be worked out via Eq. (3.60), and we find that

$$\begin{aligned}\langle \tilde{w}_{n\mathbf{0}}|\mathbf{r}|\tilde{w}_{n\mathbf{R}}\rangle &= \langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{R}}\rangle + \frac{V_{\text{cell}}}{(2\pi)^3}\int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}}\,\nabla_{\mathbf{k}}\beta(\mathbf{k})\,d^3k \\ &= \langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{R}}\rangle + i\mathbf{R}\frac{V_{\text{cell}}}{(2\pi)^3}\int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}}\,\beta(\mathbf{k})\,d^3k \qquad (3.92)\end{aligned}$$

where an integration by parts has been applied.[21] For the diagonal element ($\mathbf{R}=\mathbf{0}$) the second term vanishes and we obtain the important result

$$\tilde{\bar{\mathbf{r}}}_n = \bar{\mathbf{r}}_n\,. \qquad (3.93)$$

This implies that the *centers of charge of the Wannier functions are gauge-invariant* in the progressive case. More generally $\tilde{\bar{\mathbf{r}}}_n = \bar{\mathbf{r}}_n + \mathbf{R}$ where $\mathbf{R}$ characterizes the shift part of the transformation, but it remains true that the *lattice of Wannier centers* remains invariant. In the 1D case, this is nothing other than the gauge-invariance of the Berry phase (modulo 2π) in Eq. (3.84).

To summarize this part, we have found that the band subspace associated with band n can be thought of equally well as spanned by the Bloch states $|\psi_{n\mathbf{k}}\rangle$ for $\mathbf{k}$ running over the BZ, or by a periodic lattice of localized Wannier functions $|w_{n\mathbf{R}}\rangle$ constructed from the $|\psi_{n\mathbf{k}}\rangle$ via a Fourier transform. Because of the gauge freedom in the phases of the $|\psi_{n\mathbf{k}}\rangle$, the Wannier functions are not unique, but the location of the Wannier center $\bar{\mathbf{r}}_n$ in the home unit cell is unique up to a lattice vector. This gauge-invariance property provides a strong hint that the Wannier charge centers may be related to some physically measurable property; we shall see in the next chapter that this is indeed the case.

Exercises

Exercise 3.5.1 Using Eq. (3.74), prove the orthonormality condition $\langle w_{n\mathbf{R}}|w_{n'\mathbf{R}'}\rangle = \delta_{n,n'}\,\delta_{\mathbf{R},\mathbf{R}'}$ for the Wannier functions. This implies that each Wannier function is automatically orthogonal to all of its periodic images for the same band, as well as to any Wannier function belonging to a different band, regardless of the choice of gauge.

Exercise 3.5.2

[21] The last term in Eq. (3.92) is just $i\mathbf{R}\,\beta(\mathbf{R})$ where $\beta(\mathbf{R})$ is the Fourier transform of $\beta(\mathbf{k})$.

(a) Show that the spatial second moment of a Wannier function for band n in 1D is given by

$$\langle w_n|x^2|w_n\rangle = \frac{a}{2\pi}\int_0^{2\pi/a}\langle\partial_k u_{nk}|\partial_k u_{nk}\rangle\,dk\,.$$

Here $|w_n\rangle = |w_{n0}\rangle$ is the Wannier function in the home unit cell.
(b) Using $1 = \mathcal{P}_{nk} + \mathcal{Q}_{nk}$ where $\mathcal{P}_{nk} = |u_{nk}\rangle\langle u_{nk}|$, show that

$$\langle w_n|x^2|w_n\rangle = \widetilde{\Omega}_n + \frac{a}{2\pi}\int_0^{2\pi/a} A_{nk}^2\,dk$$

where

$$\widetilde{\Omega}_n = \frac{a}{2\pi}\int_0^{2\pi/a}\langle\partial_k u_{nk}|\mathcal{Q}_{nk}|\partial_k u_{nk}\rangle\,dk\,.$$

(c) The quadradic spatial spread of the Wannier function, or mean square variation relative to its center, is

$$\Omega_n = \langle w_n|(x-\bar{x}_n)^2|w_n\rangle = \langle w_n|x^2|w_n\rangle - \bar{x}_n^2\,.$$

Show that this is also given by

$$\Omega_n = \widetilde{\Omega}_n + \frac{a}{2\pi}\int_0^{2\pi/a}(A_{nk}-\bar{A}_n)^2\,dk$$

where we have used $\bar{A}_n$, the BZ average of A_{nk}, as a synonym for $\bar{x}_n$.
(d) Show that $\widetilde{\Omega}_n$ is gauge-invariant.
(e) Now consider the change of Ω_n under a progressive gauge change (i.e., one for which $\beta_{\lambda=1} = \beta_{\lambda=0}$). Since $\bar{A}_n$ is also invariant in this context, conclude that the minimizing gauge is one that makes $A_{nk} = \bar{A}_n$ independent of k. (This is just the "twisted parallel-transport gauge" discussed on p. 70.)
Note: The generalization of these results to 2D and 3D is non-trivial, because the Berry connection cannot be flattened to be constant in all Cartesian directions simultaneously. See Marzari and Vanderbilt (1997) for details.

Exercise 3.5.3 Consider a translation of the entire crystal by an arbitrary vector $\mathbf{r}_0$ (or equivalently, the result of moving the origin to $-\mathbf{r}_0$), such that $\tilde{\psi}_{n\mathbf{k}}(\mathbf{r}) = \psi_{n\mathbf{k}}(\mathbf{r}-\mathbf{r}_0)$. Show that the Wannier centers move to $\tilde{\bar{\mathbf{r}}}_n = \bar{\mathbf{r}}_n + \mathbf{r}_0$.

3.6 Multiband formulation

In order to carry out the Wannier construction for band n above, we had to assume that $|\psi_{n\mathbf{k}}\rangle$ was smooth and periodic over the entire BZ. This is

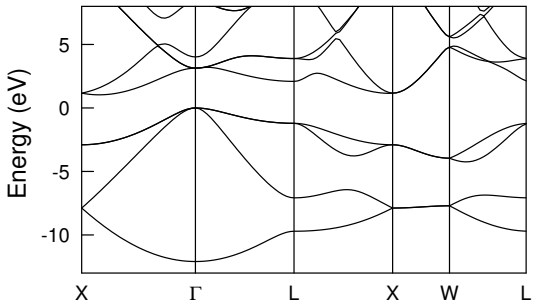

Figure 3.13 Band structure of silicon plotted along several high-symmetry lines in the BZ. The four bands at negative energies are the occupied valence bands. Note degeneracies among bands at symmetry points and along some symmetry lines.

usually not hard to arrange for an *isolated band*, i.e., one that remains separated by a finite energy gap from the next-lowest and next-highest band everywhere in the BZ. Unfortunately, that situation is actually rather uncommon. In most crystals, the occupied valence bands become degenerate at some high-symmetry points, and as a result the Bloch functions, defined as energy eigenstates, often have a singularity as a function of $\mathbf{k}$ in the vicinity of the degeneracy. This presents a problem not just for the formulation and construction of Wannier functions, but for other quantities to be discussed later, such as the electric polarization.

Fortunately, there is an elegant solution to these problems. One can consider a group of bands that are glued together by degeneracies in this way as comprising a "composite group," and develop methods for treating this group as a whole. This is the subject of the this last section of Ch. 3.

3.6.1 Multiband Wannier functions

Consider, for example, the bandstructure of Si, shown in Fig. 3.13. There are four occupied valence bands covering the region from about $-15\,\text{eV}$ to the valence-band maximum at $0\,\text{eV}$. Bands 1 and 2 are degenerate at the zone-boundary point X, while bands 2-4 are degenerate at the BZ center Γ. If we would try to construct a Wannier function for band $n=2$ using Eq. (3.70a), we would encounter a serious problem in that $|\psi_{2\mathbf{k}}\rangle$ has a singularity for $\mathbf{k}$ at Γ. This occurs because $\lim_{\mathbf{k}\to\mathbf{0}}|\psi_{2\mathbf{k}}\rangle$ depends on the direction along which the limit is taken; the states turn out to be of Si $3p_x$, $3p_y$, or $3p_z$ character when the approach is along the (100), (010), or (001) axis respectively. A similar non-analytic behavior occurs at the X point because of the crossing

with the lowest band. In the context of a Fourier transform, a sharp structure in **k**-space translates into a loss of localization in **r**-space, which is to say that the resulting Wannier function $w_{n\mathbf{R}}(\mathbf{r})$ would no longer be well localized.[22]

Does this mean that it is hopeless to construct well-localized Wannier functions for such a system? Luckily, the answer is no, as long as we are willing to abandon the notion that each Wannier function should be associated with one and only one energy band. Instead, we ask for a set of four Wannier functions $|w_{n\mathbf{R}}\rangle$ ($n = 1, ..., 4$) that spans the same subspace as the Bloch bands $|\psi_{n\mathbf{k}}\rangle$ considered as a group. That is, we construct a set of Wannier functions

$$|w_{n\mathbf{R}}\rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}} |\tilde{\psi}_{n\mathbf{k}}\rangle \, d^3k \tag{3.94}$$

out of a set of Bloch-like functions $|\tilde{\psi}_{n\mathbf{k}}\rangle$ that are smooth functions of **k** everywhere in the BZ, and that are related to the true (energy-eigenstate) Bloch functions via a unitary transformation of the form

$$|\tilde{\psi}_{n\mathbf{k}}\rangle = \sum_{m=1}^{4} U_{mn}(\mathbf{k}) \, |\psi_{m\mathbf{k}}\rangle \, . \tag{3.95}$$

Here $U_{mn}(\mathbf{k})$ is a manifold of 4×4 unitary matrices whose **k**-dependence near Γ and X has to be chosen in such a way as to "iron out" the nonanalytic behavior of $|\psi_{n\mathbf{k}}\rangle$, so that $|\tilde{\psi}_{n\mathbf{k}}\rangle$ is smooth everywhere. If this can be done, then the Wannier functions resulting from Eqs. (3.94-3.95) should be legitimate exponentially-localized Wannier functions similar to those generated from single isolated bands.

It is far from obvious that such an ironing-out procedure, resulting in globally smooth and periodic $|\tilde{\psi}_{n\mathbf{k}}\rangle$, is always possible. Clearly the manifold $U_{mn}(\mathbf{k})$ must itself have nonanalyticities that cancel out those of the underlying $|\psi_{m\mathbf{k}}\rangle$. It turns out that, under rather general conditions, this can always be done. We shall return to this point in Sec. 3.6.2, where we will see that the ironing-out can always be done locally in any small region within the BZ. The fact that it can be done globally follows from a deep mathematical analysis of Brouder et al. (2007). For now, let us just assume that this step is possible.

In general, we define an *isolated group of J bands* to be a set of J consecutive energy bands that do not become degenerate with any lower or higher band anywhere in the BZ. (In insulators this is normally taken to coincide with set of occupied valence bands, but other choices are often possible.)

[22] Typically, $w_{n\mathbf{R}}(\mathbf{r})$ acquires power-law tails, such that matrix elements of operators such as H and **r** are no longer well-defined.

Noting that the $|\psi_{n\mathbf{k}}\rangle$ and $|u_{n\mathbf{k}}\rangle$ transform in the same way, we can rewrite and generalize Eq. (3.95) as

$$|\tilde{u}_{n\mathbf{k}}\rangle = \sum_{m=1}^{J} U_{mn}(\mathbf{k})\,|u_{m\mathbf{k}}\rangle\,, \tag{3.96}$$

which we refer to as a *multiband* or *non-Abelian*[23] gauge transformation. Equation (3.96) is the natural generalization of Eq. (3.59) to the multiband case, since the special case of diagonal matrices $U_{mn}(\mathbf{k}) = \delta_{mn}\, e^{-i\beta_n(\mathbf{k})}$ corresponds to the application of a simple phase twist $\beta_n(\mathbf{k})$ to each band individually.

Note that the $|\tilde{\psi}_{n\mathbf{k}}\rangle$ are no longer eigenstates of H, since we have mixed different energy bands to construct them. Nevertheless, because the transformations in Eqs. (3.94) and (3.95) are both unitary, we can equally well use the $|\psi_{n\mathbf{k}}\rangle$, the $|\tilde{\psi}_{n\mathbf{k}}\rangle$, or the Wannier functions $|w_{n\mathbf{R}}\rangle$ constructed from the latter, as a representation of the occupied band subspace. Thus, total charge densities constructed from any one of these sets of functions must be identical, as are the expectation values of other operators when expressed as traces over the band subspace.

Once again, the Wannier functions are non-unique, because there are many choices of multiband gauge transformations $U_{mn}(\mathbf{k})$ that can successfully iron out the nonanalyticities of the Bloch energy eigenstates. If fact, as soon as we have one such gauge transformation, we can always construct another one by following with second multiband gauge transformation having the form of Eq. (3.96), but this time with a $U(\mathbf{k})$ that itself is smooth and periodic everywhere in the BZ. As a result, the Wannier functions may change shape, becoming somewhat more or less localized.

In spite of this gauge-dependence, the six properties of single-band Wannier functions enumerated in Sec. 3.5.1 all have their counterparts in the multiband case. Equations (3.75) and (3.76) remain true, while Eq. (3.77) is generalized to a multiband orthonormality condition $\langle w_{m\mathbf{R}}|w_{n\mathbf{R}'}\rangle = \delta_{mn}\,\delta_{\mathbf{R}\mathbf{R}'}$. As for Eq. (3.78), the total band projection $\mathcal{P} = \sum_n \mathcal{P}_n$ is now invariant.

The fact that the Wannier functions provide an exact TB-like representation remains true, with the modification that instead of just computing the $E_{n\mathbf{R}}$ as in Eq. (3.79), we have to construct a $J \times J$ *Hamiltonian matrix* $H_{mn}(\mathbf{R}) = \langle w_{m\mathbf{0}}|H|w_{n\mathbf{R}}\rangle$. This is then Fourier transformed in the manner of Eq. (3.81b) to obtain $H_{mn}(\mathbf{k})$, i.e.,

$$H_{mn}(\mathbf{k}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}\, H_{mn}(\mathbf{R})\,, \tag{3.97}$$

[23] The name refers to the fact that $J \times J$ matrices generally do not commute, and to the connection with non-Abelian gauge theories.

from which it is easily shown that $H_{mn}(\mathbf{k}) = \langle\tilde{\psi}_{n\mathbf{k}}|H|\tilde{\psi}_{n\mathbf{k}}\rangle$. The eigenvalues $E_{n\mathbf{k}}$ are then the eigenvalues of the matrix secular equation $\det[H(\mathbf{k}) - E_{n\mathbf{k}}] = 0$.

As for the position matrix elements of Eq. (3.80), it is again most interesting to focus on the diagonal elements, i.e., the Wannier centers $\bar{\mathbf{r}}_n$. Under a multiband gauge transformation these may shift in position, but as we shall see in Ex. 3.6.1, the sum of Wannier-center vectors

$$\bar{\mathbf{r}}_{\rm tot} = \sum_{n=1}^{J} \bar{\mathbf{r}}_n \tag{3.98}$$

remains multiband-gauge-invariant, modulo a lattice vector, just as for the single-band Wannier center as expressed in Eq. (3.93).[24] This fact will prove important later on.

If the Wannier functions are not unique, how can we calculate them and plot them in practice for materials of interest? An answer to this question was provided by Marzari and Vanderbilt (1997), where we suggested to focus on the maximally-localized Wannier functions (MLWF), defined as those chosen according to the criterion of minimizing the sum of quadratic spreads of the Wannier functions in one unit cell,[25]

$$\Omega_{\rm spread} = \sum_{n=1}^{J} \Big[\langle w_{n\mathbf{0}}|r^2|w_{n\mathbf{0}}\rangle - |\bar{\mathbf{r}}_n|^2\Big], \tag{3.99}$$

where the quantity in brackets represents the mean square variation of the Wannier electron density away from its mean position $\bar{\mathbf{r}}_n = \langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{0}}\rangle$. Such a construction is by now a standard procedure in the computational electronic structure community, as reviewed by Marzari et al. (2012) and implemented in the open-source WANNIER90 code package (Mostofi et al., 2008, 2014).

As an example, Fig. 3.14 shows a MLWF constructed for Si in this way. The procedure results in four equivalent Wannier functions, each of which is located on one of the four nearest-neighbor Si–Si bonds in the unit cell, and each one having the character of a σ bond orbital, i.e., a bonding combination of two inward-directed sp^3 hybrids from the two neighboring atoms. We emphasize that there is no one-to-one correspondence between the four Wannier functions in the unit cell and the four bands in Fig. 3.13; instead,

[24] On p. 98 we introduced the distinction between radical and progressive gauge transformations in the single-band case, depending on whether or not $e^{i\beta(\mathbf{k})}$ winds by a nonzero multiple of 2π as $\mathbf{k}$ loops around the BZ. For the multiband gauge transformation of Eq. (3.96), the corresponding question is whether $\det[U(\mathbf{k})]$ has a corresponding winding. For a progressive gauge transformation there is no such winding and $\bar{\mathbf{r}}_{\rm tot}$ is fully gauge-invariant.

[25] Despite a similarity in notation, $\Omega_{\rm spread}$ has no relation to a Berry curvature.

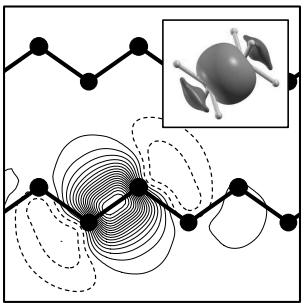

Figure 3.14 One of the four maximally localized Wannier functions in Si, shown as an amplitude contour plot on a (110) plane containing the chains of Si atoms (filled circles connected by bonds). Inset: 3D view showing isosurfaces of the Wannier function at a positive (central region) and a negative (two end caps) contour level.

the four Wannier functions taken together span the same space as is spanned by the four lowest bands in Fig. 3.13.

It is satisfying to see that the MLWfs correspond to our usual understanding of the chemical bonding in Si. This turns out to be a rather general feature of the MLWF construction (Marzari and Vanderbilt, 1997; Marzari et al., 2012), and is responsible in part for its popularity. For the purposes of this book, however, the MLWF criterion is not particularly important. When speaking of Wannier functions in the multiband case, we shall only require that these are reasonably well localized, such that Hamiltonian and position matrix elements between Wannier functions are well defined.

3.6.2 Multiband parallel transport

In Sec. 3.1.1 we introduced a notion of parallel transport for the case of a single state $|u_\lambda\rangle$ for discretized λ. Taken over to the case of a single band transported along a path in $\mathbf{k}$-space in the BZ of a 3D crystal, in which the phase of the first state $|\bar{u}_{n\mathbf{k}_0}\rangle$ has already been chosen, the procedure was to choose the phase of the next state $|\bar{u}_{n\mathbf{k}_1}\rangle$ such that $\langle\bar{u}_{n\mathbf{k}_0}|\bar{u}_{n\mathbf{k}_1}\rangle$ is real and positive, and then similarly for $\langle\bar{u}_{n\mathbf{k}_1}|\bar{u}_{n\mathbf{k}_2}\rangle$ and all subsequent points. The criterion that $\langle\bar{u}_{n\mathbf{k}_j}|\bar{u}_{n\mathbf{k}_{j+1}}\rangle$ should be real and positive is equivalent to requiring that this inner product should be as close to unity as possible, intuitively corresponding to a notion of "optimal alignment" of the states. In the case that this string of points is a closed loop, the physical states must be identical at the final and initial points, but the phases do not necessarily match. Translating Eq. (3.21) to this context, the Berry phase of the loop

is then just the phase mismatch

$$\phi_n = -\mathrm{Im}\ln\langle\bar{\psi}_{n\mathbf{k}_N}|\bar{\psi}_{n\mathbf{k}_0}\rangle = -\mathrm{Im}\ln\langle\bar{u}_{n\mathbf{k}_N}|e^{-i\mathbf{b}_\mu\cdot\mathbf{r}}|\bar{u}_{n\mathbf{k}_0}\rangle\,, \tag{3.100}$$

where the $e^{-i\mathbf{b}_\mu\cdot\mathbf{r}}$ factor is needed only if the loop winds around the BZ by a primitive reciprocal lattice vector $\mathbf{b}_\mu$.

To generalize this to the multiband case, imagine that we start with a set of J orthonormal states $|\bar{u}_{n\mathbf{k}_0}\rangle$ ($n = 1, ..., J$); these could be the cell-periodic energy-band states themselves at $\mathbf{k}_0$, or some unitarily mixed combination of them. We then want to carry out a $J \times J$ unitary transformation on the states $|u_{n\mathbf{k}_1}\rangle$ to get new states $|\bar{u}_{n\mathbf{k}_1}\rangle$ having "optimal alignment" with those at $\mathbf{k}_0$. This means that the matrix

$$\bar{M}_{mn}^{(\mathbf{k}_0,\mathbf{k}_1)} = \langle\bar{u}_{m\mathbf{k}_0}|\bar{u}_{n\mathbf{k}_1}\rangle \tag{3.101}$$

should be as close to the unit matrix as possible.

To accomplish this, we first construct the overlap matrix

$$M_{mn}^{(\mathbf{k}_0,\mathbf{k}_1)} = \langle\bar{u}_{m\mathbf{k}_0}|u_{n\mathbf{k}_1}\rangle \tag{3.102}$$

between the chosen states at $\mathbf{k}_0$ and the initial states at $\mathbf{k}_1$. We then subject $M^{(\mathbf{k}_0,\mathbf{k}_1)}$ to a *singular-value decomposition* (SVD)

$$M^{(\mathbf{k}_0,\mathbf{k}_1)} = V\Sigma W^\dagger \tag{3.103}$$

where V and W are $J \times J$ unitary matrices, and $\Sigma_{mn} = s_n\,\delta_{mn}$ is a diagonal matrix whose elements, the "singular values" s_n, are all real and positive. The SVD is a standard operation that is implemented in virtually all linear algebra software packages. Since we are considering the overlap of states at two nearby points along the path, there should be some choice of unitary rotations that would make the states very similar, so that we can expect all the eigevalues s_n to be close to unity.[26] We then construct the unitary matrix

$$\mathcal{M}^{(\mathbf{k}_0,\mathbf{k}_1)} = VW^\dagger \tag{3.104}$$

which can be regarded as the best unitary approximation to $M^{(\mathbf{k}_0,\mathbf{k}_1)}$, i.e., the unitary matrix that tells us approximately how the states at $\mathbf{k}_0$ got rotated in going to $\mathbf{k}_1$. To undo this rotation, we let the new states at $\mathbf{k}_1$ be

$$|\bar{u}_{n\mathbf{k}_1}\rangle = \sum_m (\mathcal{M}^\dagger)_{mn}|u_{m\mathbf{k}_1}\rangle\,. \tag{3.105}$$

Note that if we now construct the overlap matrix via Eq. (3.101) the result is $\bar{M} = V\Sigma V^\dagger$. This is a Hermitian matrix with near-unit eigenvalues, i.e.,

[26] All the s_n must be ≤ 1, since two nomalized vectors can never have an inner product that exeeds unity.

nearly the identity matrix. The $|\bar{u}_{n\mathbf{k}_1}\rangle$ can thus be regarded as the states that are "maximally aligned" to the $|\bar{u}_{n\mathbf{k}_0}\rangle$. Repeating the same procedure to chose states $|\bar{u}_{n\mathbf{k}_2}\rangle$ that are optimally aligned with the $|\bar{u}_{n\mathbf{k}_1}\rangle$, and so on, provides a parallel-transport gauge $|\bar{u}_{n\mathbf{k}_0}\rangle, ..., |\bar{u}_{n\mathbf{k}_N}\rangle$ for the states along the chain.

As an aside, recall that in Sec. 3.6.1 we encountered the problem of constructing a smooth gauge in the vicinity of some reference point $\mathbf{k}_0$, as for example for the four valence bands at Γ in Si. We can now give a simple prescription for doing so. At $\mathbf{k}_0$, make some arbitrary choice of J orthonormal functions spanning the band subspace there; for Si we can choose the lowest-energy state as $j=1$, and then arbitrarily pick three orthonormal states to represent the next three degenerate states at the valence band maximum. Then for each nearby $\mathbf{k}$, calculate $M_{mn}^{(\mathbf{k}_0,\mathbf{k})} = \langle u_{m\mathbf{k}_0}|u_{n\mathbf{k}}\rangle$, obtain the SVD, and rotate the states at $\mathbf{k}$ according to Eq. (3.105). This insures maximal alignment of the states at each point in the neighboring region to the states at $\mathbf{k}_0$, and consequently generates a smooth manifold in this region.

Let us return to the case of a chain of N $\mathbf{k}$-points along a path, but suppose, as in Sec. 3.1.1, that the path is a loop such that $\mathbf{k}_0$ and $\mathbf{k}_N$ represent the same Hamiltonian. After the sequential parallel-transport procedure is applied around this loop, the states at $\mathbf{k}_N$ will have acquired an overall global unitary rotation characterized by the unitary matrix

$$\mathfrak{U}_{mn} = \langle u_{m\mathbf{k}_N}|u_{n\mathbf{k}_0}\rangle\,, \tag{3.106}$$

where $\mathfrak{U}$ is a $J\times J$ unitary matrix.[27] We now define the *total Berry phase* associated with this path to be

$$\phi_{\rm tot} = -{\rm Im}\ln\det\mathfrak{U}\,. \tag{3.107}$$

Equations (3.106) and (3.107) constitute the multiband generalization of Eq. (3.8).

We can do more and extract a set of individual Berry phases associated with the multiband parallel transport. Since $\mathfrak{U}$ is unitary, its eigenvalues λ_m are unimodular, i.e., $\lambda_m = e^{-i\phi_m}$ for real ϕ_m. Thus, Eq. (3.107) can also be written as

$$\phi_{\rm tot} = -{\rm Im}\ln\prod_{m=1}^{J}\lambda_m = -\sum_{m=1}^{J}{\rm Im}\ln\lambda_m = \sum_{m=1}^{J}\phi_m\,. \tag{3.108}$$

The individual phases ϕ_m are known as "multiband Berry phases" or "Wil-

[27] If the path involves a winding by reciprocal lattice vector $\mathbf{b}_\mu$, then as in Eq. (3.100), a factor of of $e^{-i\mathbf{b}_\mu\cdot\mathbf{r}}$ has to be inserted into the inner product in Eq. (3.106).

son loop eigenvalues" (Wilson, 1974), and correspond to the multiband generalization of the single-band Berry phase of Eq. (3.8).

One way to attach a meaning to these multiband Berry phases is to consider what happens if we use the eigenvectors of $\mathfrak{U}$ to rotate the states $|u_{n\mathbf{k}_0}\rangle$ into new states $|\tilde{u}_{m\mathbf{k}_0}\rangle$, and then use these as the starting point for the parallel-transport gauge construction. Then the matrix $\tilde{\mathfrak{U}}$ obtained at the end of this construction will be diagonal, with the diagonal elements being just the $e^{-i\phi_m}$. That is, each state $|\tilde{u}_{m\mathbf{k}}\rangle$ will return precisely to itself at the end of the loop, having acquired the Berry phase ϕ_m.

Note that we can also determine the total Berry phase ϕ, or the individual multiband Berry phases ϕ_j, without explicitly constructing the parallel-transport gauge at all. We start with states the $|u_{n\mathbf{k}_0}\rangle, ..., |u_{n\mathbf{k}_{N-1}}\rangle$ in any arbitrary gauge, and for each neighboring pair of points, construct the overlap matrix $M_{mn}^{(\mathbf{k}_j,\mathbf{k}_{j+1})} = \langle u_{m\mathbf{k}_j}|u_{n\mathbf{k}_{j+1}}\rangle$. We then replace each $M^{(\mathbf{k}_j,\mathbf{k}_{j+1})}$ by its unitary approximation $\mathcal{M}^{(\mathbf{k}_j,\mathbf{k}_{j+1})}$ as given by Eq. (3.104), and compute the product $\mathfrak{U} = \prod_{j=0}^{N-1} \mathcal{M}^{(\mathbf{k}_j,\mathbf{k}_{j+1})}$. This is easily shown to be identical to the $\mathfrak{U}$ of Eq. (3.106), and the ϕ_j are extracted from its eigenvalues as above. The total Berry phase is just the sum of these eigenvalues, and can also be written as

$$\phi_{\rm tot} = -{\rm Im}\ln\det\prod_{j=0}^{N-1}\mathcal{M}^{(\mathbf{k}_j,\mathbf{k}_{j+1})}\,. \tag{3.109}$$

Often $\phi_{\rm tot}$ is approximated in practice by

$$\phi_{\rm tot} = -{\rm Im}\ln\det\prod_{j=0}^{N-1}M^{(\mathbf{k}_j,\mathbf{k}_{j+1})} \tag{3.110}$$

since this gives the same result in the limit of increasing mesh density and is easier to calculate in practice.

In the case of a 1D multiband system in which the path winds around the BZ, one can establish a more provocative interpretation of the ϕ_m in terms of Wannier centers. Recall that in the single-band case discussed in Sec. 3.5.1, we saw that the position of the Wannier center in the unit cell is just proportional to the single-band Berry phase as expressed via Eq. (3.84). This connection generalizes to the multiband case, where it turns out that in 1D the charge centers of the MLWFs are again given in terms of the multiband Berry phases by an identical equation,

$$\bar{x}_m = a\,\frac{\phi_m}{2\pi}\,. \tag{3.111}$$

The only difference is that these Wannier functions no longer have a one-to-

one relation to individual energy bands, but are extracted from the entire set of bands treated as a group. The demonstration of Eq. (3.111) is given in Sec. IV.C.1 of Marzari and Vanderbilt (1997).

3.6.3 Multiband Berry potentials and curvatures

The discussion above has motivated the idea that an isolated set of J adjacent bands can profitably be treated as a unified group. We have already seen how the concept of the Berry phase can be generalized to this case. In fact, the entire formalism of Berry potentials and curvatures introduced earlier in this chapter can be generalized to the multiband case, where objects that were previously scalars and now replaced by $J \times J$ matrices. The resulting formalism is often referred to as *nonabelian* because matrices generally do not commute. The formalism can be developed in any space of parameters λ_j as in the early parts of this chapter, but here we shall use the notation of Bloch states of a 3D crystal for consistency with the preceding discussion. In subsequent chapters, we will often assume isolated bands for simplicity, so that it is possible to skim or omit this subsection on a first reading. In any practical implementation, however, the use of methods that treat groups of occupied bands as a whole are highly advantageous.

We assume that some definite choice has been made for the cell-periodic Bloch-like states $|u_{n\mathbf{k}}\rangle$ such that these are smooth functions of $\mathbf{k}$ that are unitarily related to the true (energy-eigenstate) Bloch functions of the band group in question. You can imagine that these are the result of some smoothing procedure as discussed above. Then the multiband Berry potential and curvature are defined as

$$A_{mn,\mu}(\mathbf{k}) = \langle u_{m\mathbf{k}} | i\partial_\mu u_{n\mathbf{k}} \rangle \tag{3.112}$$

and

$$\begin{aligned} \Omega_{mn,\mu\nu}(\mathbf{k}) &= \partial_\mu A_{mn,\nu}(\mathbf{k}) - \partial_\nu A_{mn,\mu}(\mathbf{k}) \\ &= i\langle \partial_\mu u_{m\mathbf{k}} | \partial_\nu u_{n\mathbf{k}} \rangle - i\langle \partial_\nu u_{m\mathbf{k}} | \partial_\mu u_{n\mathbf{k}} \rangle \end{aligned} \tag{3.113}$$

As before, we sometimes hide the Cartesian indices by writing these as vectors $\mathbf{A}_{mn}$ and pseudovectors $\mathbf{\Omega}_{mn}$.

It is also useful to define the matrix traces

$$A^{\mathrm{tr}}_\mu(\mathbf{k}) = \sum_n A_{nn,\mu}(\mathbf{k}) \tag{3.114}$$

and

$$\Omega^{\mathrm{tr}}_{\mu\nu}(\mathbf{k}) = \sum_n \Omega_{nn,\mu\nu}(\mathbf{k}) \tag{3.115}$$

where the sum is over the J bands in the group. These traced quantities behave very much like the single-band Berry quantities of Sec. 3.4. In particular, for any closed path C in $\mathbf{k}$-space we can define a total Berry phase

$$\phi_{\rm tot} = \oint_C \mathbf{A}^{\rm tr}(\mathbf{k}) \cdot d\mathbf{k} = \int_S \mathbf{\Omega}^{\rm tr}(\mathbf{k}) \cdot d\mathbf{S} \tag{3.116}$$

which precisely corresponds to the discretized total Berry phase of Eqs. (3.107-3.108). In the second form above, Stokes' theorem has been invoked to convert the line integral into a surface integral over a patch of surface S whose boundary is C. In the special case that the individual bands are nowhere degenerate and the $|u_{n\mathbf{k}}\rangle$ are Bloch energy eigenstates, we can identify the diagonal matrix elements $A_{nn\mu}$ and $\Omega_{nn,\mu\nu}$ in the current notation with $A_{n\mu}$ and $\Omega_{n,\mu\nu}$ of Sec. 3.4, and the total Berry phase $\phi_{\rm tot}$ around a closed path is just the sum $\sum_n \phi_n$ of the individual-band Berry phases.

Let us see how these Berry connection and curvature matrices transform under a multiband gauge change, Eq. (3.96), which we now write as

$$|\tilde{u}_{n\mathbf{k}}\rangle = \sum_m U_{mn} \, |u_{m\mathbf{k}}\rangle \, . \tag{3.117}$$

(We have dropped the explicit $\mathbf{k}$-dependence of U_{mn} and it is understood that all sums are over the J states.) Any J-component vector that transforms according to Eq. (3.117) is denoted as a "gauge-covariant vector." A matrix

$$\mathcal{O}_{mn} = \langle u_{m\mathbf{k}} | \hat{\mathcal{O}} | u_{n\mathbf{k}} \rangle \tag{3.118}$$

constructed from matrix elements of an ordinary single-particle operator $\hat{\mathcal{O}}$ is also referred to as "gauge covariant," in which case we mean that it transforms as

$$\widetilde{\mathcal{O}}_{mn} = (U^\dagger \mathcal{O} U)_{mn} \, , \tag{3.119}$$

where $J \times J$ matrix products are implied on the right-hand side. This is easily checked by inserting Eq. (3.117) into $\widetilde{\mathcal{O}}_{mn} = \langle \tilde{u}_{m\mathbf{k}} | \hat{\mathcal{O}} | \tilde{u}_{n\mathbf{k}} \rangle$.

Unfortunately, the vector

$$|\partial_\mu \tilde{u}_{n\mathbf{k}}\rangle = \sum_m U_{mn} \, |\partial_\mu u_{m\mathbf{k}}\rangle + \sum_m (\partial_\mu U_{mn}) \, |u_{m\mathbf{k}}\rangle \tag{3.120}$$

is not gauge-covariant, in view of the appearance of the second term above. By the same token, the Berry potential matrix

$$\widetilde{A}_{\mu,mn} = (U^\dagger A_\mu U)_{mn} + (U^\dagger i\partial_\mu U)_{mn} \tag{3.121}$$

is not gauge-covariant, again because of the extra term on the right. A similar extra term appears in the transformed Berry curvature tensor $\widetilde{\Omega}_{\mu\nu,mn}$. For some purposes this is inconvenient. For example, any physical observable

should be gauge-invariant, and it turns out to be easy to construct invariants out of gauge-covariant matrices.[28] We should not be surprised that the Berry potential is not gauge-covariant, since neither was it gauge-invariant in the single-band case. However, the Berry curvature *was* gauge-invariant in the single-band case, and it is unsatisfactory that we have lost the corresponding property here.

This can be repaired as follows. First, at each $\mathbf{k}$ define the projection operator onto the group of bands

$$\mathcal{P}_{\mathbf{k}} = \sum_n |u_{n\mathbf{k}}\rangle\langle u_{n\mathbf{k}}| \tag{3.122}$$

where the sum is over states in the group, and its complement $\mathcal{Q}_{\mathbf{k}} = 1 - \mathcal{P}_{\mathbf{k}}$. (In the case that the group is the full set of occupied states of an insulator, $\mathcal{P}_{\mathbf{k}}$ and $\mathcal{Q}_{\mathbf{k}}$ are just the projections onto valence and conduction states respectively.) Then we can define a "gauge-covariant derivative" of $|u_{n\mathbf{k}}\rangle$ as

$$|\check{\partial}_\mu u_{n\mathbf{k}}\rangle = \mathcal{Q}_{\mathbf{k}} |\partial_\mu u_{n\mathbf{k}}\rangle \tag{3.123}$$

which deserves this name because it satisfies

$$|\check{\partial}_\mu \tilde{u}_{n\mathbf{k}}\rangle = \sum_m U_{mn} |\check{\partial}_\mu u_{m\mathbf{k}}\rangle \tag{3.124}$$

since the $\mathcal{Q}_{\mathbf{k}}$ term yields zero when acting on the $|u_{m\mathbf{k}}\rangle$ in the last term of Eq. (3.120). There is no such thing as a gauge-covariant Berry connection tensor, since $\langle u_{n\mathbf{k}}|\check{\partial}_\mu u_{m\mathbf{k}}\rangle$ vanishes, but we can define a gauge-covariant Berry curvature tensor

$$\begin{aligned}\check{\Omega}_{mn,\mu\nu} &= i\langle\check{\partial}_\mu u_{m\mathbf{k}}|\check{\partial}_\nu u_{n\mathbf{k}}\rangle - i\langle\check{\partial}_\nu u_{m\mathbf{k}}|\check{\partial}_\mu u_{n\mathbf{k}}\rangle \\ &= i\langle\partial_\mu u_{m\mathbf{k}}|\mathcal{Q}_{\mathbf{k}}|\partial_\nu u_{n\mathbf{k}}\rangle - i\langle\partial_\nu u_{m\mathbf{k}}|\mathcal{Q}_{\mathbf{k}}|\partial_\mu u_{n\mathbf{k}}\rangle\end{aligned} \tag{3.125}$$

satisfying an equation like Eq. (3.119). We can write this covariant curvature more concisely using

$$\begin{aligned}|\check{\partial}_\mu u_{n\mathbf{k}}\rangle &= |\partial_\mu u_{n\mathbf{k}}\rangle - \sum_m |u_{m\mathbf{k}}\rangle \langle u_{m\mathbf{k}}|\partial_\mu u_{n\mathbf{k}}\rangle \\ &= |\partial_\mu u_{n\mathbf{k}}\rangle + i\sum_m A_{mn,\mu} |u_{m\mathbf{k}}\rangle .\end{aligned} \tag{3.126}$$

from which it follows that

$$\check{\Omega}_{mn,\mu\nu} = \Omega_{mn,\mu\nu} - i[A_\mu, A_\nu]_{mn} \tag{3.127}$$

[28] For example, the trace of any gauge-covariant matrix is automatically gauge-invariant. More generally, if $A_1 \ldots A_M$ are gauge-covariant matrices, then the trace of their product is gauge-invariant.

Note that the band-index traces of $\check{\Omega}_{\mu\nu}$ and $\Omega_{\mu\nu}$ are identical, with both yielding Eq. (3.115), since the trace of the commutator appearing in Eq. (3.127) vanishes.

Exercises

Exercise 3.6.1 We want to show that the vector sum $\bar{\mathbf{r}}_{\rm tot}$ of Wannier centers in Eq. (3.98) is gauge-invariant, modulo a lattice vector, under a multiband gauge transformation as in Eq. (3.96).

(a) Starting from Eq. (3.121), show that the traced Berry potential of Eq. (3.114) transforms as

$$\mathrm{Tr}\,[\widetilde{A}_\mu(\mathbf{k})] = \mathrm{Tr}\,[A_\mu(\mathbf{k})] + \mathrm{Tr}\,[U^\dagger(\mathbf{k}) i\partial_\mu U(\mathbf{k})]\,.$$

(b) Show that this can be rewritten as

$$\mathrm{Tr}\,[\widetilde{A}_\mu] = \mathrm{Tr}\,[A_\mu] + \partial_\mu \beta$$

where $\beta(\mathbf{k}) = -\mathrm{Im}\ln\det U(\mathbf{k})$. Note the similarity with Eq. (3.60).

Hint: Recall that a trace can be evaluated in any representation. Here it is convenient to use the one that diagonalizes $U(\mathbf{k})$; that is, assume $U(\mathbf{k}) = V(\mathbf{k})B(\mathbf{k})V^\dagger(\mathbf{k})$ where V is unitary and B is unimodular diagonal, $B_{mn} = e^{-i\beta_n}\delta_{mn}$ for real β_n.

(c) Assuming that a global branch choice has been made for $\beta(\mathbf{k})$ such that it is smooth and periodic over the 3D BZ, prove the invariance of $\bar{\mathbf{r}}_{\rm tot}$, starting from Eq. (3.83). (For this part and the next, you can assume an orthorhombic $a\times b\times c$ unit cell and let $\mu = x$.)

(d) The gauge transformation in (c) is multiband "progressive" in the sense of the discussion on p. 98. It is also possible for $\beta(\mathbf{k})$ to increase by $2\pi n_j$ as $\mathbf{k} \to \mathbf{k} + \mathbf{b}_j$, where n_j is an integer associated with primitive reciprocal lattice vector $\mathbf{b}_j$, and still be consistent with a $U(\mathbf{k})$ that is smooth and periodic over the 3D BZ. In this case of a "radical" gauge transformation, show that $\bar{\mathbf{r}}_{\rm tot}$ shifts by a lattice vector.