RAL-94-030

Science and Engineering Research Council
Rutherford Appleton Laboratory
Chilton DIDCOT Oxon OX11 0QX

RAL-94-030

Hall Conductance as a Topological Invariant

G I Watson

March 1994

Hall Conductance as a Topological Invariant

Greg Watson
Rutherford Appleton Laboratory
Oxfordshire OX11 0QX, UK
(Email: giw@isise.rl.ac.uk)

January 1994

These lecture notes[1] form an informal review of work published around 1982–8 on the topological approach to the theory of the integer quantum Hall effect (IQHE). The scope is the "TKNN" paper (Thouless *et al.* 1982) and subsequent work inspired by it. I will give several versions of the reasoning and try to explain some of the mathematical jargon which sometimes appears.[2]

A popular explanation of the IQHE is the Laughlin (1981) argument, which considers the macroscopic response of a two-dimensional electron system to a change of external vector potential. The TKNN work is a rigorous formulation of his physical argument, in terms of a microscopic calculation using linear response theory. It proves, under very general assumptions, the following statement of the IQHE:

> When the Fermi energy of a two-dimensional electron gas lies in a gap and the system's ground state is nondegenerate, the (zero-temperature) Hall conductance is an integer multiple of e^2/h. (1)

Here e is the electron charge and h is Planck's constant.

Where does topology enter the theory of the integer quantum Hall effect? The basic idea can be understood by imagining a non-interacting electron system, in two dimensions, with no substrate disorder. In those circumstances, the single-particle Bloch functions are well-defined, and are parametrised by two real numbers, namely the components of the Bloch wavevector. Topologically, the parameter space is a torus, since the wavefunction is periodic (up to a phase) in each parameter. Then it turns out that the Kubo formula for the Hall conductance can be expressed as e^2/h times the integral of a kind of curvature over the torus—a mathematical quantity

[1] Expanded from a talk given at the University of Tennessee, Knoxville, November 1993.

[2] See also reviews by Thouless (1984, 1987), Morandi (1988), Harper (1991) and Stone (1992).

known as the first Chern class of a $U(1)$ bundle, a topological invariant. This means two things. First, it is necessarily an integer. Second, it is insensitive to the precise form of the wavefunctions and the substrate potential. The proof can even be extended to deal with the effects of substrate disorder, sample boundaries, electron-electron interactions or nonuniformities in the magnetic field.

1 Bloch electrons in a magnetic field

I begin by outlining the historical context in which the TKNN work appeared. The classical formula for the Hall conductivity of a gas of particles of charge e is

$$\sigma_H = \frac{nec}{B}, \tag{2}$$

where n is the area density of electrons and B is the magnetic field. Actually this gives the right answer also in the simplest quantum mechanical calculation—treating the electrons as noninteracting and moving in free space. In that case, the single-particle energy is quantized in units of the cyclotron frequency ω_c, and the energy spectrum is a regularly spaced series of *Landau levels*, each of which has degeneracy equal to the total flux through the system in units of the flux quantum $\Phi_0 = hc/e$. A noninteracting electron gas consisting of exactly j filled Landau levels has density

$$n = \frac{eBj}{hc},$$

and according to (2) its Hall conductance is j in units of e^2/h.

Next consider the effects of the periodic background potential in which the (still noninteracting) electrons are moving. For sufficiently large magnetic field, the periodic potential can be thought of as a perturbation—it broadens the Landau levels into bands and splits each one into a number of subbands (see Fig. 1). What is the Hall conductance if the Fermi energy lies in a gap between subbands? According to the classical formula, it should be some fraction of e^2/h, since the Landau level is partially filled. But that contradicts the Laughlin reasoning, which says that, owing to the gap at the Fermi energy, it should still be an integer. It was one of the aims of TKNN to resolve this paradox. Somewhat unexpectedly, their work turned out also to provide a topological theory of the quantum Hall effect.

In order to work out the details, we will need a specific Hamiltonian. Most of the time we will use

$$H = \frac{1}{2m} \sum_{i=x,y} \left(p_i - \frac{e}{c} A_i \right)^2 + U(x,y). \tag{3}$$

This is a single-particle Hamiltonian, so the electrons' mutual interaction is neglected. Each electron moves in the xy-plane and is subject to a uniform

perpendicular magnetic field $\mathbf{B}$, written as $\mathbf{B} = \nabla \times \mathbf{A}$, where $\mathbf{A}$ is the vector potential. The term $U(x, y)$ represents the periodic potential of a square lattice,

$$U(x, y) = U(x + a, y) = U(x, y + a),$$

with lattice spacing a.

Note that there is no disorder potential in (3). Actually there is a problem with this. If there really were no disorder, then the spectrum would consist purely of bands of extended states. With the Fermi energy in a gap, an infinitesimal increase in the filling factor would make it jump to the next band. The quantization of conductance would be observable only at isolated values of the filling. So disorder is needed to explain the plateaux of quantized conductance—with the Fermi energy in a gap, added electrons go into localized states which carry no current.

For the moment, let us keep things simple and exclude disorder from our Hamiltonian. But it must be kept in mind that it describes only the extended states (those responsible for conduction), and the localized states are somehow lurking in the background to provide the reservoir of electron states necessary for finite plateaux.

1.1 Magnetic translations

Having set aside the disorder, we can now take advantage of the translational symmetry properties of the system. It turns out that much can be deduced from symmetry alone, so it is worth going through this in some detail.

Everybody knows how to handle a Hamiltonian with a discrete translational symmetry: the translation operators, defined by[3]

$$T_x\psi(x, y) = \psi(x + a, y); \qquad T_y\psi(x, y) = \psi(x, y + a),$$

commute with the Hamiltonian and with each other, and therefore one can construct simultaneous eigenstates of H, T_x and T_y. These are the Bloch wavefunctions, and they are labelled completely by a wavevector in the Brillouin zone together with a discrete band index.

If we try to do the same thing for a system with a uniform magnetic field, we immediately run into a problem. With nonzero magnetic field, the translations do not commute with the Hamiltonian! In other words, the translated Hamiltonian $T_i^{-1}HT_i$ $(i = x, y)$ is not equal to H.

That sounds crazy. Translating the Hamiltonian amounts to choosing a different origin for the coordinate system, which should make no difference.

[3]Some authors write T_{ax} for translation by a distance a in the x-direction, but here we are interested only in translations by a fixed distance a, so we write just T_x.

The problem is that although the field is uniform, the vector potential is not. It transforms under translations:

$$T_i^{-1} H(\mathbf{A}) T_i = H(T_i^{-1}\mathbf{A}).$$

But this is nothing but a gauge transformation. Here is why. The transformed vector potential is written

$$\mathbf{A}' = T_i^{-1}\mathbf{A} = \mathbf{A} - a\frac{\partial}{\partial r_i}\mathbf{A},$$

where r_i are the components of the position vector. The second equality follows from the fact that the vector potential is a linear function of the coordinates. Thus

$$\mathbf{A}' = \mathbf{A} - a\nabla\Lambda_i,$$

where

$$\Lambda_i = \mathbf{r}\cdot\frac{\partial \mathbf{A}}{\partial r_i}.$$

The change in vector potential is the gradient of a scalar function, which represents a gauge transformation. So the system has a slightly unusual form of translational symmetry—it is invariant under the combined action of a translation and a gauge transformation.

We can use this to construct operators which commute with the Hamiltonian. The operator form of a gauge transformation corresponding to Λ_i is $\exp(-iae\Lambda_i/\hbar c)$, so we are led to define[4]

$$\tilde{T}_i = \exp\left[\frac{ia}{\hbar}\left(p_i - \frac{e}{c}\Lambda_i\right)\right].$$

These are the *magnetic translation operators*. The momentum operator generates the translation part, and Λ_i generates the gauge transformation part.

It is easy to check that $\tilde{T}_x$ and $\tilde{T}_y$ commute with the Hamiltonian, so next we want to check that they commute with each other. But they don't, and for good reason! In fact,

$$\tilde{T}_x\tilde{T}_y = e^{2\pi i\phi}\tilde{T}_y\tilde{T}_x, \tag{4}$$

where

$$\phi = \frac{ea^2B}{hc}.$$

The number ϕ is dimensionless and proportional to the magnetic field. It equals the number of flux quanta (hc/e) passing through each unit cell of the lattice.[5] In physical terms, Eq. (4) says that if we translate the system

[4] Actually the product of the exponentials does not equal the exponential of the sum if p_i and Λ_i do not commute. But they differ at most by a constant phase factor, since the commutator is a multiple of the identity.

[5] Another physical interpretation is as the ratio of two characteristic time periods: the period of motion of an electron with crystal momentum $2\pi\hbar/a$, and the reciprocal of the cyclotron frequency.

around a closed loop (here the boundary of a square unit cell) we gain a nontrivial phase equal to 2π times the number of flux quanta passing through the loop.

The failure of magnetic translations to commute has deep consequences for the problem. It also leads to some interesting mathematics. An important early contribution was made by Zak (1964a,b), who worked out some of the representation theory of the (nonabelian) group of magnetic translations.

Anyway, it looks like we did not get much mileage out of this construction. We do not have a full set of mutually commuting operators. For arbitrary magnetic field, the best we can do is to label the states by one of the magnetic translations, $\tilde{T}_x$ say, and although the action of the other, $\tilde{T}_y$, takes us between degenerate eigenstates, it does not yield a well-defined second quantum number. There is however a case in which we can do better than this. If the magnetic field happens to be such that ϕ is a rational number,

$$\phi = \frac{p}{q}, \tag{5}$$

with p and q integers, then

$$[\tilde{T}_x^q, \tilde{T}_y] = 0,$$

and we do have two mutually commuting symmetry operators.[6] Equation (5) is the condition of *rational flux* or *commensurate flux*. Physically, it says that a rectangle made of q adjacent lattice unit cells contains an integer number of flux quanta. If it holds, then we can simultaneously diagonalize H, $\tilde{T}_x^q$ and $\tilde{T}_y$, giving wavefunctions $\psi_{\mathbf{k}}$ labelled by $\mathbf{k} = (k_x, k_y)$ and satisfying

$$\begin{aligned} \psi_{\mathbf{k}}(x + qa, y) &= e^{-iqa(k_x - e\Lambda_x/\hbar c)}\psi_{\mathbf{k}}(x, y) \\ \psi_{\mathbf{k}}(x, y + a) &= e^{-ia(k_y - e\Lambda_y/\hbar c)}\psi_{\mathbf{k}}(x, y). \end{aligned} \tag{6}$$

These are the *magnetic Bloch functions*. The states are completely labelled by $\mathbf{k}$ together with a discrete band index. The vector $\mathbf{k}$ lies in the *magnetic Brillouin zone*:

$$-\frac{\pi}{aq} \le k_x < \frac{\pi}{aq}, \qquad -\frac{\pi}{a} \le k_y < \frac{\pi}{a}.$$

Values of k_x differing by $2\pi/aq$ are equivalent, as are values of k_y differing by $2\pi/a$, so the magnetic Brillouin zone is topologically a torus.

One feature of this approach is somewhat bizarre: the treatment of the problem depends on the denominator of a fractional representation of the magnetic field. Two almost equal rational fields may have altogether different denominators, and very close to any rational field is an irrational one at which the formalism breaks down completely. This seems to violate the principle that physical properties should vary smoothly with bulk parameters. However, this is actually an artifact of the mathematical representation of

[6] We could equally well have used $[\tilde{T}_x, \tilde{T}_y^q] = 0$.

the problem via Bloch's theorem. Physical properties really do vary continuously with field: for example, it can be proved rigorously that spectral gaps are continuous in ϕ. One way to avoid the embarrassment of dependence on the denominator is to go to a more general mathematical framework, involving the concept of C^*-algebras, in which rational and irrational fields can be treated simultaneously, and a natural continuity in ϕ emerges. However, here I will use only the easier (Bloch) approach, despite the restriction to rational fields, because that is the one I understand.

1.2 Harper's equation

It will be helpful to have a definite picture of what the magnetic subband structure might actually look like, so I am going to digress to discuss a well-studied solution for a particular model.

As mentioned, the effect of a periodic lattice potential on the discrete Landau level structure is to broaden each level, partially breaking the degeneracy, and to split it into a number of subbands. The actual splitting can be calculated in the large field limit. This was first done by Rauh, Wannier and Obermair (1974) and Rauh (1974, 1975). One begins by writing the wavefunctions in the Landau gauge $\mathbf{A} = (0, Bx, 0)$ as $\chi_{nk} = e^{iky} u_n(x - x_k)$, where n is the Landau level index, $\hbar k$ is the momentum in the y-direction, u_n denotes the nth harmonic oscillator wavefunction, and $x_k = \hbar k / m\omega_c$. The wavefunction is a plane wave in one direction, and a harmonic oscillator centred at x_k in the other. For simplicity, let the lattice potential be

$$U(x, y) = U_0[\cos(2\pi x/a) + \cos(2\pi y/a)].$$

The matrix elements of U between Landau states can be calculated exactly. Transitions between different Landau levels can be neglected if the magnetic field is strong enough to satisfy $\phi \gg 1$, in other words, if there are many flux quanta passing through each lattice cell. Note that this condition does not involve U_0—the validity of the approximation does not depend on how strong the periodic potential is. So inter-level transitions are neglected, but within each Landau level the calculation is non-perturbative.

Within this approximation, the effect of the periodic potential on a particular Landau level is to mix states with k differing by $2\pi/a$. If the wavefunction is expanded in the unperturbed basis as

$$\psi = \sum_j c_j \chi_n(k_0 + 2\pi j/a),$$

the coefficients are found to satisfy Harper's (1955) equation

$$c_{j-1} + c_{j+1} + 2\cos(2\pi\phi^{-1} j + \Delta) c_j = \epsilon c_j, \tag{7}$$

with $\Delta = 2\pi l^2 k_0/a$, where l is the magnetic length, defined by $l^2 = \hbar c/eB$. The eigenvalue ϵ is proportional to the energy shift produced by U.

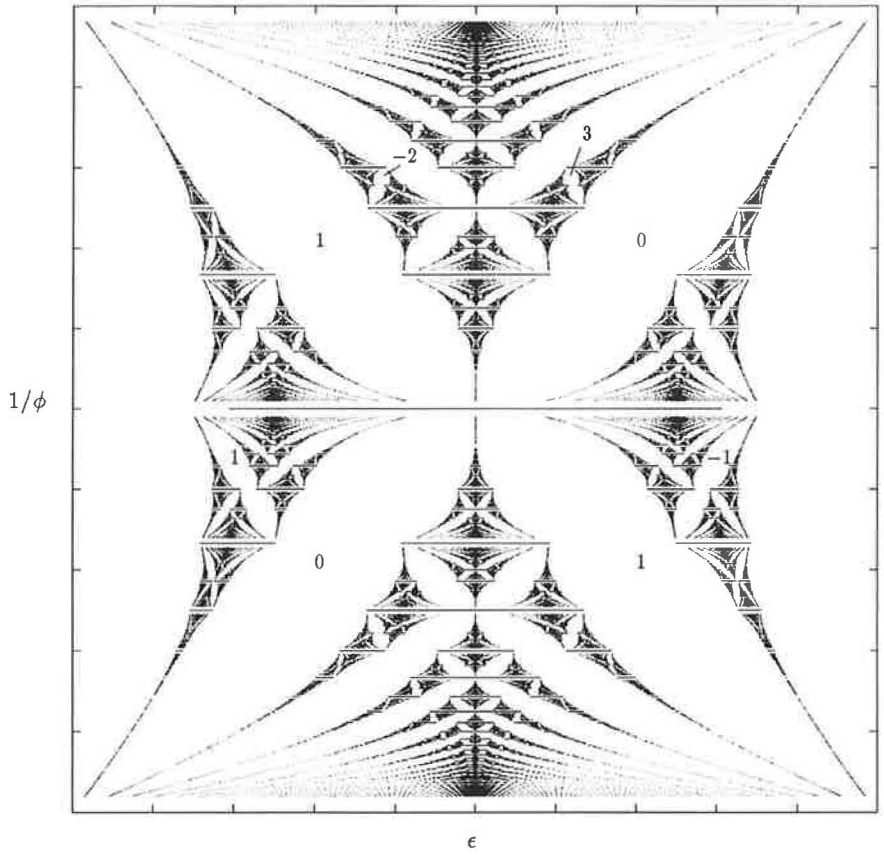

Figure 1: The Hofstadter (1976) butterfly spectrum. Energy bands for rational ϕ are drawn as horizontal lines, with $1/\phi$ (inversely proportional to magnetic field) running vertically from 0 to 1. A few of the gaps are labelled with their Hall conductance indices.

The problem of determining the broadening and splitting of the perturbed Landau level thus reduces to finding eigenvalues ϵ satisfying Harper's equation. The full spectrum is given by the union of the eigenvalues over all values of Δ. There is a large literature on this problem, reviewed by Sokoloff (1985) and Lovesey, Watson and Westhead (1991), amongst others. When ϕ is a rational number, say $1/\phi = p/q$, the equation is numerically tractable—it reduces to a $q \times q$ matrix eigenvalue problem, and there are some extra tricks for solving it efficiently. The spectrum consists of q separate bands, except for even q when two bands meet in the centre. The results are plotted in Fig. 1, the famous Hofstadter (1976) diagram.

It should be mentioned that there is an entirely independent way of arriving at Harper's equation in the opposite limit of weak field. In that treatment one considers the effect of the magnetic field as a perturbation

of the Bloch states. The usual approach is to include the magnetic field by the Peierls substitution—the replacement of the crystal momentum $\hbar\mathbf{k}$ in an effective Hamiltonian for a Bloch band by $\hbar\mathbf{k} - e\mathbf{A}/c$, where $\mathbf{A}$ is the vector potential. That leads again to Harper's equation, with the following differences of interpretation. First, ϕ is replaced by $1/\phi$. Second, the coefficients c_j now describe amplitudes of Wannier functions centred at site j. Third, the equation described the splitting of a Bloch band rather than the broadening of a Landau level. The fact that one obtains the same equation in opposite limits gives one some confidence that the picture is qualitatively correct.

Anyone who has not seen Fig. 1 before would be bound to admit that it is not what one would have expected. That is why I said it was a good idea to have a specific picture in mind. None of the rest of what I am going to say depends on the approximations used in this section.

2 Středa's argument

Now that we know what the bands look like, let us move on to our first explanation of quantization of the Hall conductance. It is very simple, and allows one to assign conductances to individual subbands. The starting point is a formula which relates the Hall conductance to the derivative of electron density with respect to magnetic field at constant Fermi energy:

$$\sigma_H = ec \left.\frac{\partial n}{\partial B}\right|_{E=E_F} = \frac{e^2}{h} \left.\frac{\partial \nu}{\partial \phi}\right|_{E=E_F}, \tag{8}$$

The second equality re-expresses it in terms of the electron density per unit cell ν, and the flux per unit cell ϕ. Středa (1982a,b) proved (8) from the Kubo formula.[7] It is valid whenever the longitudinal conductivity is zero.

So all we need is the electron density as a function of filling and field. That's easy. I claim that each subband contains exactly $1/q$ electron states per unit lattice cell. The proof is almost obvious and requires only the magnetic symmetry. Consider a finite rectangular system of dimensions L_x and L_y. As is customary we will apply periodic boundary conditions, and these should be compatible with the magnetic translation symmetry. Instead of requiring the wavefunction to be unchanged on translation by L_i, $(i = x, y)$, we require invariance under *magnetic* translation by L_i. But since the magnetic Bloch functions satisfy

$$\tilde{T}_x^q \psi_k = e^{iaqk_x} \psi_k; \qquad \tilde{T}_y \psi_k = e^{iak_y} \psi_k,$$

the boundary condition reads

$$e^{iL_x q k_x} = e^{iL_y k_y} = 1.$$

[7] There is also a thermodynamic derivation: see Widom (1983) and Středa and Smrčka (1983).

The total number of states in a subband is the number of values of $\mathbf{k}$ in the magnetic Brillouin zone which satisfy this equation, which is

$$\frac{L_x}{aq}\frac{L_y}{a}.$$

Since there are L_xL_y/a^2 unit cells in the system, the number of states per unit cell is $1/q$.

So far so good: if the Fermi energy lies in a gap between subbands, then $\nu = j/q$ with integer j. To express this in terms of ϕ so we can plug it into (8), we use an elementary fact: if p and q are relatively prime integers, then for any j there exist integers n and m such that $j = np + mq$, and hence

$$\nu = n\phi + m. \tag{9}$$

Now a non-elementary fact: the integers n and m are constant within each subband gap. In other words, if we change ϕ slightly at constant Fermi energy, n and m do not change. This is sometimes called the labelling theorem; it was first noticed by Wannier (1978) by empirical study of Fig. 1, and subsequently proved generally. Here it is all-important, as it allows us to differentiate (9) directly. From (8), this gives a Hall conductance of n in units of e^2/h. We have proved that the Hall conductance is quantized, using only the Středa formula and the translational symmetry of the pure noninteracting Hamiltonian. Eq. (9) is known as the Diophantine equation for the Hall conductance (Dana, Avron and Zak 1985).

To illustrate how the labelling theorem works, consider as an example the principal gap running from the top left corner to the centre of Fig. 1. We get the index from (9) by identifying ν and ϕ corresponding to that gap. For example, at $\phi = 3/2$ the gap lies between the first and second subbands. One subband filled, with $q = 2$, gives $\nu = 1/2$, and thus

$$1/2 = (3/2)n + m.$$

The same gap corresponds to two filled subbands at $\phi = 5/3$, giving

$$2/3 = (5/3)n + m.$$

We have two equations in unknowns m and n, which can be solved to give $n = 1$ and $m = -1$. The Hall conductance index is 1, as indicated on the figure.

There are plenty of other ϕ values we could have picked. For example, at $\phi = 4/3$ our gap lies at a filling of one subband, yielding

$$1/3 = (4/3)n + m.$$

The same values of n and m obtained above work in this equation. The same will be true at any value of ϕ for which the gap exists, and for any gap the values of n and m with this property are integers—that is the content of the labelling theorem.

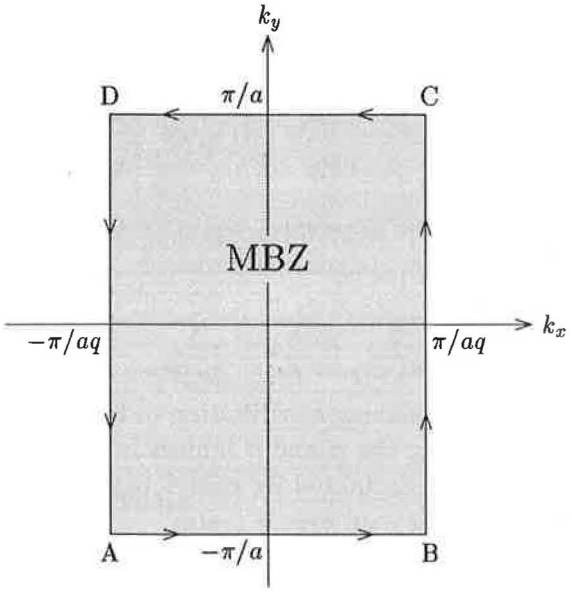

Figure 2: The magnetic Brillouin zone. The conductance of a subband is proportional to a line integral around the boundary of the zone as shown.

In the second line the integrand is written as the z-component of the curl of the quantity in parentheses. The third line uses Stokes' theorem to express the surface integral of the curl as a line integral around the boundary of the magnetic Brillouin zone. The integration contour is shown in Fig. 2.

The next step is to interpret this formula in terms of the change in phase of the wavefunction around a closed loop. Corresponding points on the boundary of the zone, such as points with k_y differing by $2\pi/a$, represent the same physical state. The wavefunctions can differ by at most a total phase factor:

$$u(k_x, \pi/a) = e^{i\Delta(\mathbf{k})} u(k_x, -\pi/a),$$

where the phase Δ is independent of $\mathbf{r}$. Thus we can write the contribution to σ_λ from the horizontal segments AB and CD as

$$\int_{\mathrm{A}}^{\mathrm{B}} dk_x \int d\mathbf{r} \left[u^* \frac{\partial u}{\partial k_x} - u^* \left(i \frac{d\Delta}{dk_x} u + \frac{\partial u}{\partial k_x} \right) \right] = i \int_{\mathrm{A}}^{\mathrm{B}} dk_x \frac{d\Delta}{dk_x}.$$

The $\mathbf{r}$ integral was set equal to unity using the normalization condition (11). The result is just the change in Δ from A to B. Suppose we assign the wavefunction $u_{\lambda\mathbf{k}}$ a phase θ, such that the phase factor $e^{i\theta}$ varies smoothly with $\mathbf{k}$ in going around the integration loop. Then the change in Δ along AB equals the change in θ along AB plus the change in θ along CD.

The line integral giving the conductance of subband λ has a similar contribution from the segments BC and DA. The total line integral is i times the change in wavefunction phase θ around the integration loop. That change must be an integer multiple of 2π since the wavefunction is nondegenerate. The $2\pi i$ cancels and we are left with the desired result that σ_λ is an integer.

That was the argument the way TKNN originally did it. Before we move on to the more topological version, let us pause for a moment and consider the generality of the reasoning. We began with a form of the Kubo formula valid when the ground state is nondegenerate; in the present (pure, non-interacting) system it is obtained by filling single-particle states up to the Fermi energy. The states are parametrized by the two components of the magnetic Bloch wavevector, and corresponding points on the boundary of the magnetic zone represent the same state. Very little information about the wavefunctions was needed. Even the magnetic boundary condition was not really essential—it was used to express the matrix elements as convenient integrals over a finite region, with the area dependence cancelling out. The crucial point is that the wavefunctions are labelled by two parameters of some sort, with a relationship between the phases on the boundary. The weakness of the assumptions suggests that the argument might be generalized to a case in which disorder destroys the magnetic translational symmetry and the Bloch wavevector is no longer available to label the states. That is exactly what Niu, Thouless and Wu (1985) did, and I will discuss that generalisation in Sec. 5.

A more fundamental assumption in the theory was the validity of the linear response formalism. That rests on regarding the electric field as a small perturbation, meaning smaller than the energy gap. Since experiments are done at some finite voltage, this limits the detail in Fig. 1 which might be resolved by the Hall conductance. However, that is perhaps academic at present, since not even the principal gap structure has, to my knowledge, been seen in experiments—the observed Hall conductance plateaux apparently correspond to filled entire Landau levels. Limitations in linear response theory may be more important in the fractional quantum Hall effect (Thouless 1989, Thouless and Gefen 1991).

4 Topological phantasmagoria

I have given two "proofs" of the IQHE. Here I will attempt to deal with the fact that published papers in this field invariably make some reference to fibre bundles and Chern classes. I am going to try to convey some feeling for what these concepts mean. Readers who don't care can skip this section. However, I think even a patchy understanding of Chern classes helps in appreciating the depth and generality of the topological approach to the IQHE.

Although at first seeming over-abstracted, the subject of fibre bundles is not particularly difficult, at least at the level required to appreciate the topological approach to the IQHE. It mainly amounts to learning some jargon. Actually, most physicists already know more than they think, since electromagnetism is a classic example of a $U(1)$ bundle. The best way of getting the hang of the mathematical lingo is to work through an example,

so in the next section I will describe the translation of the theory of Dirac monopoles into fibre bundle language. This approach is due to Wu and Yang (1975). The topological treatment of the IQHE also involves a $U(1)$ bundle, so the ideas used in the monopole example are directly applicable.

The following is a stripped down and intuitive account of an abstract mathematical topic. There is inevitably some loss of generality and in some cases accuracy. For the real story, I recommend the book by Nakahara (1990), although it is not easy reading. For an intermediate level introduction, see Schutz (1980) and Monastyrski (1987).

4.1 The Dirac monopole

One of Maxwell's equations says that the magnetic field is divergenceless, and hence there are no sources of magnetic flux. In the thirties, Dirac postulated the existence of (hitherto unobserved) particles with nonzero magnetic charge, called magnetic monopoles. He derived the now famous result that the magnetic charge must be quantized.

Consider an isolated monopole of magnetic charge g at the origin, defined by the introduction of a source term into one of Maxwell's equations:

$$\nabla \cdot \mathbf{B} = 4\pi g \delta(\mathbf{r}).$$

The magnetic field is divergenceless as usual except at the origin where there is a delta function source. The resulting magnetic field is directed radially outwards. Let S^2 be the surface of the unit sphere centred at the origin (the 2 denoting its two-dimensionality). The total flux outwards through the sphere is

$$\int_{S^2} \mathbf{B} \cdot \mathbf{dS} = \int \nabla \cdot \mathbf{B} = 4\pi g,$$

where the surface integral is written as a volume integral over the inside of the sphere by Gauss' divergence theorem. The magnetic field is uniform on the sphere, equalling g at any point.

This is a complete specification of the problem in classical physics. However, to do quantum mechanics with the monopole, we need to know the vector potential $\mathbf{A}$ associated with it. Here one immediately runs into a difficulty: it is impossible to find a vector potential which describes the required uniform magnetic field consistently over the whole sphere!

Proof: Let H_N be the northern hemisphere of S^2 and H_S the southern. Suppose we have a vector potential $\mathbf{A}$ describing the monopole. Its line integral around the equator can be written as a surface integral using Stokes' theorem in two different ways:

$$\oint \mathbf{A} \cdot \mathbf{dl} \;=\; \int_{H_N} \mathbf{B} \cdot \mathbf{dS} = 2\pi g$$

$$= -\int_{H_S} \mathbf{B} \cdot \mathbf{dS} = -2\pi g.$$

In the first line, the equator is the boundary of the H_N in a positive sense. In the second line, it is the boundary of H_S in a negative sense, giving a negative sign. It follows that $g = 0$ and there is no monopole. QED.

Here is another way of understanding the result. Basically we are trying to solve the differential equation $\mathbf{B} = \nabla \times \mathbf{A}$ to find $\mathbf{A}$ with a given $\mathbf{B}$ (uniformly outwards). Imagine starting at the north pole and using the differential equation to extend the solution outwards. If the surface were flat, we could extend this local solution indefinitely. Since the surface is a sphere, however, the local solutions extended in different directions eventually crowd in on each other at the opposite pole. We require that they all agree at the pole. But according to the above result, this is impossible (for nonzero monopole charge). The best we can do is find a vector potential that is well-defined except for a singularity at an isolated point.

There is a way out. Actually we do not require all of $\mathbf{A}$ to do quantum mechanics. It is a curious fact that the magnetic field does not contain enough information for a quantum mechanical description, but the vector potential contains *too much*. The second part is a consequence of gauge invariance: different vector potentials can describe the same physical system (via different gauges), and hence any particular vector potential contains redundant information. An equivalent statement is that the quantity of physical relevance is not the vector potential itself, but the *phase factor*,[8]

$$\exp\left(\frac{ie}{\hbar c}\oint \mathbf{A} \cdot \mathbf{dl}\right)$$

which a particle's wavefunction acquires when moving around some closed path. (Only closed paths appear because only phase *differences* are physically significant, and wavefunction phases can only be compared at the same point.) Two vector potentials which give identical phase factors around all closed paths are physically equivalent.

why ?
(see p17)

Returning to the monopole example, the identity obtained by integrating along the equator is modified to equality of the phase factors,

$$\exp\left(2\pi g\frac{ie}{\hbar c}\right) = \exp\left(-2\pi g\frac{ie}{\hbar c}\right),$$

implying

$$\frac{2eg}{\hbar c} = \text{integer}.$$

This is Dirac's quantization condition. Equivalently, the flux $4\pi g$ through the unit sphere must be an integer multiple of the flux quantum hc/e.

Fine, so quantum mechanics does not exclude magnetic monopoles altogether but it does require quantization of the flux source. But we are

[8]Here I adopt the (almost universal) convention of calling the exponential the phase factor and its argument the phase.

still left with the problem of how to make a vector potential to plug into Schrödinger's equation, when no such vector potential can exist throughout space. Dirac's fix was to use a vector potential with a point singularity (the "Dirac string"). That works, but has the drawback that the singularity is totally unphysical and simply a mathematical artifact of the description of the electromagnetic field in terms of a vector potential. The modern fix is to replace **A** by a more general concept—a connection on a fibre bundle.

4.2 Fibre bundles

Fibre bundles are tricky to explain because they are very general. I will attempt the usual kind of explanation, which is to give a few examples.[9] First, the mathematical lingo. Every fibre bundle has a *base space* and a *fibre space*. In a phrase like "$U(1)$ bundle over the torus," $U(1)$ is the fibre and the torus is the base space.

The idea is to attach a copy of the fibre at each point of the base space. So for most applications, think of the base space as physical space, and the fibre attached at a particular point as a space describing a degree of freedom, order parameter, or whatever other property particles or space might have at that point. For example, in a superconductor the physical space is three-dimensional Euclidean space $\mathbf{R}^3$, and the order parameter is a phase factor[10] —that is, an element of the group $U(1)$—so one can think of this as a $U(1)$ bundle over $\mathbf{R}^3$. At each point in the superconductor ($\mathbf{R}^3$), the attached fibre ($U(1)$) is the space of possible values of the order parameter at that point.

There is a second important ingredient in a fibre bundle, namely a specification of some global geometrical way of fitting the fibres together smoothly. A simple example is a line bundle over the circle. According to the recipe above, that bundle is obtained by attaching a line at each point on the circle, as shown in Fig. 3(a). The strip is the bundle, the dotted line is a circle (the base space), and at each point on it a perpendicular line (the fibre) is attached. The fibre twists smoothly as one goes around the base space. Fig. 3(b) is another way of making a line bundle over a circle. It has a half-twist, making a Möbius strip. The untwisted strip and the Möbius strip have the same base space and the same fibre, and hence are locally identical: if one cuts a small piece of the strip it is impossible to tell from it if the whole bundle was twisted or not. But these two bundles differ in their global geometry. They are two distinct line bundles over the circle.

Note that these two are the only distinct line bundles over the circle. What about strips with two or more half-twists? It turns out that these

[9] To keep things simple, I am not going to distinguish between coordinate bundles, fibre bundles and principal bundles.

[10] The order parameter space is $U(1)$ only because of the common neglect of variations in $|\psi|$.

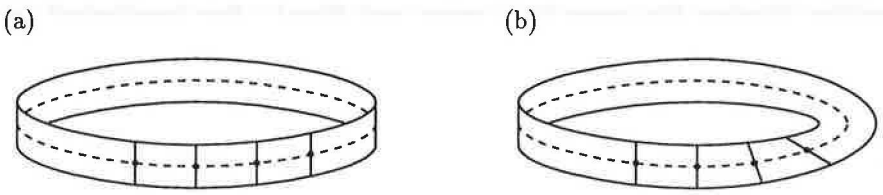

Figure 3: Two line bundles over the circle.

do not correspond to new bundles. Here is the reason. Since the fibre (the set of reals) has a characteristic direction (from negative to positive), one should imagine a little arrow on each fibre. As the fibre twists in going around the circle, the arrow comes back parallel or anti-parallel to its original orientation, depending on whether it went through an even or odd number of half-twists. As far as the bundle is concerned, all that matters is how the orientations match up, not how many twists the arrow went through to get there. The point is that that all strips with an even number of half-twists, for instance, are simply different embeddings, in three dimensions, of the same bundle. The fact that the strips look different to us is an artifact of the embedding—if we had more dimensions to work in (in this case four would be sufficient), we would be able to untwist any even number of half-twists "through" the extra dimensions.[11] So for smooth line bundles over the circle, we only have the two possibilities shown in Fig 3.

Only the simplest fibre bundles can be represented by drawings of three dimensional objects. Another example is a doughnut—a $U(1)$ bundle over the circle,[12] $U(1)$ being the unit circle in the complex plane. The Klein bottle is also a $U(1)$ bundle over the circle, one which cannot be embedded in three-dimensional space. The bundle of interest in the monopole problem is a $U(1)$ bundle over the sphere, and unfortunately I know of no way of visualizing this concretely. However, the twisted strip example gives us one of the important ideas of fibre bundles. Given a particular base space and fibre, there are geometrical restrictions on how they can be put together to make a fibre bundle. Usually there is only a discrete set of possibilities; in the strip example for instance there are only two. To emphasize: the discreteness here has a purely geometrical origin, depending only on the base and fibre spaces involved, and having nothing to do with the physical interpretation of those spaces in a given problem.

What does this have to do with monopoles? I will first outline the situation, then try to give some intuitive justification. A magnetic monopole with (quantized) magnetic charge g is described by a particular $U(1)$ bundle over the sphere. Different monopole charges correspond to geometrically distinct

[11] Just as a linked pair of rings can be unlinked in four dimensions. Think of the edges of the strip as the linked rings.

[12] One could say a $U(1)$ bundle over $U(1)$, but it is usual to give the fibre as a (Lie) group and the base space as a geometrical object.

bundles (having the same base space and fibre). For topological reasons there is only a discrete (but infinite) set of possible bundles, and that fact is equivalent to the statement that monopole charge must be quantized. Bundles are characterized by an integer topological invariant, called the Chern number—the term topological invariant meaning that equivalent bundles have the same number. The Chern number is proportional to the magnetic charge.

I will give two ways of approaching the Chern number for the monopole. The first uses some ideas from homotopy theory (familiar to many theoretical physicists from the theory of defects in homogeneous media) to clarify why there should exist an integer characterizing the fibre bundle. The second interprets the Chern number as the integral of a kind of curvature, which should appeal to anybody with a little familiarity with Riemannian geometry.

4.3 Chern number and homotopy

As mentioned, it is possible to construct a local vector potential in the vicinity of any point of the sphere, but it is impossible to patch local solutions together consistently over the whole sphere. Now suppose we have local vector potentials $\mathbf{A}_N$ and $\mathbf{A}_S$ satisfying $\mathbf{B} = \nabla \times \mathbf{A}$ in the northern and southern hemispheres separately. The two hemispheres meet at the equator, and we want the vector potentials to be consistent there where both are defined. Consistency means that the phase factor along any closed path is the same whether computed using $\mathbf{A}_N$ or $\mathbf{A}_S$. In particular, using the closed path going around the equator, consistency of wavefunction phases requires

$$\exp\left[\frac{ie}{\hbar c}\oint \mathbf{A}_N \cdot \mathbf{dl}\right] = \exp\left[\frac{ie}{\hbar c}\oint \mathbf{A}_S \cdot \mathbf{dl}\right],$$

and hence

$$\frac{ie}{\hbar c}\oint (\mathbf{A}_N - \mathbf{A}_S)\cdot \mathbf{dl} = 2\pi \times \text{integer}. \tag{13}$$

This is the monopole quantization condition again.

Now consider transporting an imaginary particle around the equator, starting at time $t = 0$. The phases ascribed to it by $\mathbf{A}_N$ and $\mathbf{A}_S$ will differ during the journey, but will agree when the particle arrives back at the starting point at $t = 1$. Define

$$\phi(t) = \exp\left[\frac{ie}{\hbar c}\int_0^{l(t)} (\mathbf{A}_N - \mathbf{A}_s)\cdot \mathbf{dl}\right],$$

which is the phase factor discrepancy between the two vector potentials up to time t. For each time t, $\phi(t)$ is a phase factor in $U(1)$. Obviously $\phi(0) = 1$, and from (13) also $\phi(1) = 1$. Hence as the particle goes around the equator, the function ϕ traces a loop in the space $U(1)$ which starts and finishes at $\phi = 1$.

The study of the topology of loops in a particular space is called homotopy theory. There are topologically distinct ways of forming loops in $U(1)$ and these are classified by the first or fundamental homotopy group of $U(1)$. That is easy to understand in this case. The space $U(1)$ is a circle. A point which moves in $U(1)$ and returns to its starting point must travel clockwise around the circle an integer number of times (with negative integers assigned to anticlockwise trajectories). Thus the paths of the point fall into classes characterized by the *winding number* of the path around $U(1)$, and the fundamental group of $U(1)$ is the set of integers.

It turns out that the winding number associated with ϕ is precisely the Chern number of the $U(1)$ bundle corresponding to a particular monopole charge. I hope that helps in understanding what the Chern number is. This view emphasizes once again that the Chern number expresses a discreteness arising from geometry. One warning, however: the connection with winding numbers is not very general. In this example it relied on being able to construct local vector potentials which meet in a boundary that is a closed loop. In general, there is no simple correspondence between homotopy and Chern numbers, but it continues to hold true that the Chern number is quantized in integer values.

4.4 Chern number and curvature

Mathematically, the Chern number is defined in terms of something called a connection. The precise definition is too abstruse to go into here. I have a vague physical way of looking at it, which I hope will help the reader. I am only going to sketch how the mathematical and physical ideas are related, without equations. The central ideas are those of parallel transport, holonomy, and curvature.

Recall that the only physically relevant quantities are phase factors acquired in going around closed loops, since it does not make sense, without additional information, to compare phase factors at different points in space. A connection is such a set of additional information. It specifies how to assign a reference phase factor to points along a curve, against which other phase factors can be measured. What does that mean physically? Well, what can give us a reference phase along a curve? The answer is a vector potential! Once we have some (local) vector potential, we can define the reference phase to be proportional to its line integral along the curve. So

$$\text{connection} \leftrightarrow \text{vector potential.}$$

The mathematical idea on the left corresponds to the physical idea on the right. There is a table of such correspondences in Wu and Yang's (1975) paper.

The idea of a connection also occurs in Riemannian geometry. Given a surface on which a particle can move, there is a tangent plane at each point

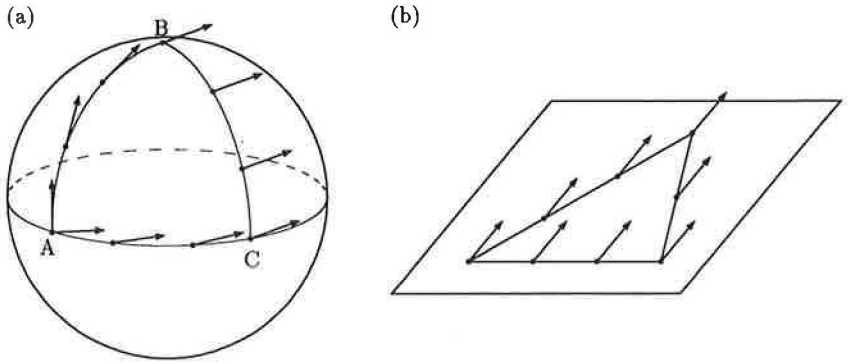

Figure 4: Parallel transport of a vector (a) on a sphere; (b) on a flat plane.

on the surface. The plane is the set of possible velocities of the particle as it goes through that point. But one cannot compare velocities at different points without going "outside" the surface, which depends on how the surface is embedded in a higher dimensional space. That is against the rules: the idea is define everything only within the space itself.[13] So one can only compare velocities at different points on a curve if one is given a canonical basis at each point on it. Equivalently, one needs a definition of *parallel transport* of a velocity vector along a curve. Starting with any basis at one point on the curve, one could then parallel transport the basis vectors along the curve, generating a set of basis vectors in each tangent plane.

So a connection is equivalent to a rule of parallel transport. This leads directly to the idea of curvature. Having forbidden talk of embedding our surface in three-dimensional space, I am now going to break that rule for purposes of illustration. The reason is that the embedding defines a simple connection: parallel transport a vector from A to an infinitesimally close point A′ by taking the projection of the vector at A onto the tangent plane at A′. Transport along a finite curve is then performed by continous projection onto a continuous series of infinitesimally varying tangent planes along the way. This is illustrated for a spherical surface in Fig. 4(a). A vector at point A on the equator is parallel transported to the pole B, then down to the equator at C, and finally back to A. When it arrives back it is no longer pointing in the same direction. It has rotated through a certain angle. The magnitude of the angle is important in quantifying the curvature of the manifold, and is called the *holonomy.* A similar thing for a flat surface is shown in Fig. 4(b). This time the vector arrives home unrotated—the holonomy is zero. Clearly, a nonzero holonomy on parallel transport around a closed loop is a direct consequence of curvature of the surface.

To quantify this idea, one may define the Riemannian curvature (of a

[13]For example, the formalism of general relativity does not depend on how space-time might be embedded in some unphysical higher dimensional space.

connection) as the holonomy per unit area of the loop as the area tends to zero. In other words, an infinitesimal loop enclosing an area dS around a given point gives a holonomy of dS times the Riemannian curvature at that point. It follows that for any finite loop, the holonomy is the surface integral of the curvature over the area bounded by the loop.

That takes care of Riemannian geometry, where we have an obvious geometrical meaning of a connection. But in the monopole example, the connection was something to do with assigning a phase factor (rather than a basis for the tangent plane) to each point on a curve. The trick is to define a curvature here also, motivated by Riemannian geometry. Before, the holonomy was the rotation of a vector in parallel transport (using the connection) around a loop; now we define the holonomy to be the change in phase (given by the connection) around the loop.[14] Recall that a connection is equivalent to a vector potential $\mathbf{A}$. The holonomy is then simply the usual line integral of $\mathbf{A}$ around a closed loop. By Stokes' theorem, it also equals the surface integral, over the area inside the loop, of the magnetic field. But the holonomy is the surface integral of the curvature, so we arrive at the correspondence

$$\text{curvature} \leftrightarrow \text{magnetic field}.$$

(This is somewhat oversimplified. More generally, there is a curvature tensor, which corresponds to the tensor field $F_{\mu\nu}$ including both magnetic and electric fields.)

Now for the Chern number. In the present context, it is simply defined as the (appropriately normalized) integral of the curvature over the entire base space. For the monopole,

$$\text{Chern number} = \frac{e}{\hbar c}\int \mathbf{B}\cdot\mathbf{dS} = \frac{2eg}{\hbar c}.$$

There is a theorem in topology that guarantees that the Chern number is always an integer. So we have arrived, by a very roundabout route, at a result we knew already—the magnetic monopole charge is quantized! From this point of view, the existence of an integer quantum number appears as a purely geometrical consequence of the fact that the coordinate space is a sphere and the gauge group is $U(1)$.

So what was the point of this digression to derive a result Dirac told us in 1931? It was, I hope, to de-mystify the concept of the Chern number and its quantization. A grasp of the physical basis for magnetic monopole quantization helps in appreciating the remarkable general result that the Chern number is an integer for bundles over a wide class of manifolds.

[14] Another application in physics of this notion of holonomy is Berry's quantum adiabatic phase (see Shapere and Wilczek 1989).

4.5 Back to the IQHE

Now we apply the topological ideas to the integer quantum Hall effect. The discovery that such an application exists was made by Simon (1983). Consider the dimensionless conductance of a single band in the form

$$\sigma_\lambda = \frac{1}{2\pi i} \int d\mathbf{k}\, \nabla_\mathbf{k} \times \left[\int d\mathbf{r}\, u^* \frac{\partial u}{\partial \mathbf{k}} \right]_z .$$

To make this resemble the monopole example, define a fictitious vector potential by

$$\mathbf{A}(\mathbf{k}) = \int d\mathbf{r}\, u^* \frac{\partial u}{\partial \mathbf{k}}. \tag{14}$$

Note that it is defined formally in terms of the electronic wavefunctions and has no direct connection with any real electromagnetic fields in the sample. The space over which $\mathbf{A}$ varies is the magnetic Brillouin zone of reciprocal space, rather than real space. The conductance is written as the surface integral of the curl:

$$\sigma_\lambda = \frac{1}{2\pi i} \int_{\text{MBZ}} (\nabla \times \mathbf{A}) \cdot \mathbf{dS}. \tag{15}$$

By Stokes' theorem, this equals the line integral of the vector potential around the boundary of the zone. But the magnetic Brillouin zone is topologically a torus, which has no boundary! Hence if the vector potential is defined smoothly over the entire torus, the Hall conductance of the band is zero.

Therefore, a necessary condition for a nontrivial Hall conductance is that is is impossible to find electron wavefunctions such that the resulting fictitious vector potential is smoothly defined on the torus. In topological language, this is expressed as the nontriviality of a $U(1)$ bundle. The fibre space here is $U(1)$ as for the monopole, because the fictitious vector potential has a gauge freedom—the wavefunctions are defined only up to a phase factor.

The following explicit construction is due to Kohmoto (1985).

The vector potential is completely defined by the electronic state as a function of $\mathbf{r}$ and $\mathbf{k}$. If the wavefunction is smooth over the $\mathbf{k}$-space torus so is $\mathbf{A}$, and then, for the reasons stated, the Hall conductance vanishes. Thus a nontrivial Hall conductance is possible only if there is a singularity in the wavefunction's dependence on the magnetic Bloch wavevector—for example, there might be a mismatch in the wavefunction phase along some boundary on the MBZ torus. Let us consider what goes wrong when one tries to define the phase smoothly over the entire zone.

Think of $u_\mathbf{k}(\mathbf{r})$ as a set of states—one state for each point $\mathbf{k}$ on the torus. We want to try to fix the phases of all the states, in such a way that the

ficticious vector potential $\mathbf{A}$, defined by (14), is a smooth function of $\mathbf{k}$. We can try to do this by selecting any point $\mathbf{r}_0$ in real space, and requiring that

$$u_{\mathbf{k}}(\mathbf{r}_0) = \text{real} \tag{16}$$

is satisfied separately for each $\mathbf{k}$. This condition is enough to nail down completely the phase freedom and fix the wavefunction uniquely—except it goes wrong if there happen to be values of $\mathbf{k}$ at which $u_{\mathbf{k}}(\mathbf{r}_0)$ vanishes. Suppose we don't have this problem, that is, we can find some r_0 for which $u_{\mathbf{k}}(\mathbf{r}_0)$ never vanishes for any state in the set. Then (16) fixes *all* the states, which yields a smooth fictitious vector potential over the whole torus, and the result is a vanishing Hall conductance. Hence, *if there exists an* $\mathbf{r}_0$ *at which the wavefunction does not have a zero for any* $\mathbf{k}$*, then the Hall conductance is zero.*

On the other hand, sometimes it is not possible to find a point $\mathbf{r}_0$ at which all the states on the torus are nonvanishing. In other worrds, for any specified point in the system we can find a state that has a node there. Then (16) is not capable of fixing the phase of that state, and the fictitious vector potential has a singularity at the corresponding value of $\mathbf{k}$. No matter what $\mathbf{r}_0$ we choose, we get a singularity somewhere or other. Does this remind you of the monopole? No matter how we try to define a vector potential, it ends up with a singularity. This is the signature of a nontrivial bundle, which means a nontrivial Chern number, which means a nonzero Hall conductance.

Thus we have an intimate connection between the Hall conductance and the zeros of the wavefunction as a function of wavevector. As will now be shown, there is a nontrivial contribution to σ_λ from each such zero. First look at a single isolated zero of $u_{\mathbf{k}}(\mathbf{r}_0)$ at some point $\mathbf{k}_0$ on the torus. The phase convention (16) defines a vector potential $\mathbf{A}$ with a singularity at $\mathbf{k}_0$. However, just as in the monopole example, the singularity is an artifact of the gauge freedom—it is not a sign of any pathological behaviour at that wavevector point. Furthermore, the particular point $\mathbf{k}_0$ is largely arbitrary. If we had chosen a different $\mathbf{r}_0$, say $\mathbf{r}_0'$, for which $u_{\mathbf{k}}(\mathbf{r}_0')$ does not vanish near $\mathbf{k}_0$, then this would define a new vector potential $\mathbf{A}'$ in a neighbourhood around $\mathbf{k}_0$. The two potentials are (gauge) equivalent in the region of common validity, meaning that they yield equivalent phase integrals around closed paths.

The situation is illustrated in Fig. 5. The vector potential $\mathbf{A}'$ is a smooth function defined on a small neighbourhood U' of $\mathbf{k}_0$, and $\mathbf{A}$ is smoothly defined on the remainder U. The two regions intersect in the directed loop C. The expression for the Hall conductance is now

$$\begin{aligned} \sigma_\lambda &= \frac{1}{2\pi i}\int_U (\nabla \times \mathbf{A})\cdot \mathbf{dS} + \frac{1}{2\pi i}\int_{U'} (\nabla \times \mathbf{A}')\cdot \mathbf{dS} \\ &= \frac{1}{2\pi i}\oint_C (\mathbf{A}' - \mathbf{A})\cdot \mathbf{dl}. \end{aligned}$$

The second equality follows from Stokes' theorem, with a relative minus sign coming from the fact that C bounds U and U' in opposite senses. How

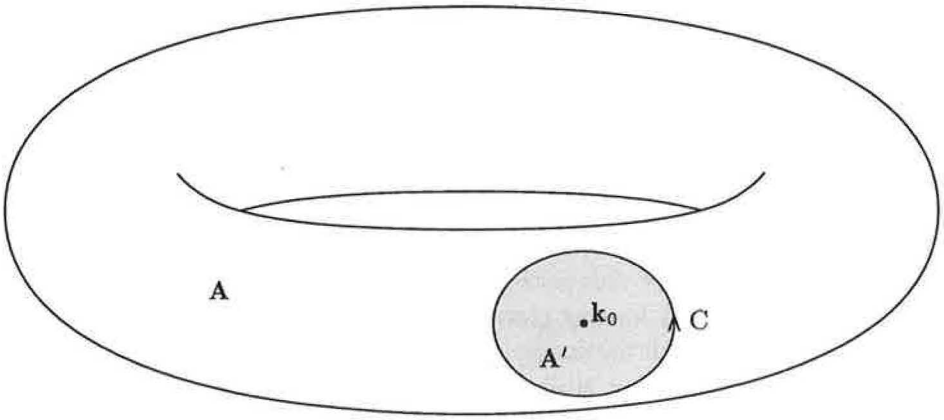

Figure 5: The magnetic Brillouin zone torus: a vector potential $\mathbf{A}(\mathbf{k})$ is patched with $\mathbf{A}'(\mathbf{k})$ in the region U' (shaded area) surrounding a singularity $\mathbf{k}_0$.

are the vector potentials related at points on C? They come from different choices of wavefunction phase:

$$u_{\mathbf{k}'} = e^{i\theta(\mathbf{k})} u_{\mathbf{k}}.$$

The primed and unprimed wavefunctions differ simply by an $\mathbf{r}$-independent phase factor. Hence

$$\mathbf{A}' = i\frac{d\theta}{d\mathbf{k}} + \mathbf{A},$$

which yields

$$\sigma_\lambda = \frac{1}{2\pi}[\text{change in } \theta \text{ around } C].$$

The phase θ is a single-valued function of wavevector, and consequently its change around a closed loop is an integer multiple of 2π. It follows that σ_λ is an integer.

The extension to more than one zero of $u_{\mathbf{k}}(\mathbf{r}_0)$ is obvious—the vector potential on the whole torus is patched together using local potentials defined in the neighbourhood of each zero by virtue of the phase freedom. The conductance becomes a sum of loop integrals along the boundaries of the neighbourhoods, each of which is an integer.

It should now be becoming clear why the Hall conductance integer is a topological quantum number, in fact the Chern number. The situation here is a slightly more complicated version of the monopole problem. Rather than go into the mathematical details of defining a connection on the torus (as Kohmoto 1985 does), I hope that the close analogy between the two problems makes matters clear. In both we have an electromagnetic vector potential whose definition involves a degree of arbitrariness corresponding to the gauge group $U(1)$—in other words, a $U(1)$ bundle. The fact that the potential is defined on a space which is not contractible to a point implies the

possibility of a nontrivial bundle. When the bundle is nontrivial, the vector potential cannot be defined over the whole space without singularities, and it is necessary to glue together functions defined in separate patches. The Chern number is a measure of the nontriviality of a bundle, defined as proportional to the integral over the whole space of the curvature, or magnetic field $\nabla \times \mathbf{A}$. By (15), the Chern number is precisely the dimensionless Hall conductance of a single electron band.

That concludes the discussion of topological considerations. Hopefully the reader can distil some useful ideas from this jumbled account. Apart from the generality and intuitive appeal of the mathematical concepts, I emphasize the physical content of the term "topological quantum number," which lies at the heart of the IQHE—a very precise quantization of a bulk quantity, insensitive to microscopic details and external perturbations.

5 General argument

Here we will sketch the argument, due to Niu, Thouless and Wu (1985), hereafter referred to as NTW, leading to the general statement (1) of the IQHE "theorem." The assumption of a pure noninteracting system is relaxed completely. All one needs is the nondegeneracy of the many-body ground state of the interacting electron system, and an energy gap at the Fermi energy, for whatever reason. The gap may be between Landau levels or magnetic subbands as previously considered, or it may be generated by the electron interactions.

The point about the gap and the nondegeneracy is that the physical zero-temperature state is separated by a finite energy from all other states higher in energy, and hence perturbation theory for the linear response to electromagnetic perturbations is valid. The result for the Hall conductance is analogous to that in the single-particle picture,

$$\sigma_H = \frac{ie^2\hbar}{A_0} \sum_{n \neq 0} \frac{v_x^{0n} v_y^{n0} - v_y^{0n} v_x^{n0}}{(E_0 - E_n)^2}.$$

The index n labels *many-body* electron states, 0 is the ground state, and v is the matrix element of the velocity.

Recall that the TKNN argument rests on the existence of well-defined quantum numbers k_x and k_y forming a torus. In the NTW version, their role is played by *macroscopic boundary conditions.*

NTW argue that the Hall conductance is insensitive to sample shape and surface effects and hence is a *bulk* property of the material. This view is supported by a simple calculation by NTW for a noninteracting system in the thermodynamic limit, which shows that the contribution of surface effects to the conductance is of the order of the localization length divided

by the size of the system. If this is correct, it does not matter what kind of boundary conditions are used in the calculation. However, it is not that simple. The Laughlin (1981) argument, which involves transfer of electrons from localized states at one edge to the other, suggests that the boundaries are important. In fact, Středa and Smrčka (1983) argue that all of the Hall conductance comes from surface currents, the bulk part being zero because of the excitation gap.

It seems to me important to clear up the role of bulk and surface currents before embarking on a calculation of the Hall conductance. However, I do not have time to do this. I will simply follow the NTW version, assuming that the conductance is a bulk property, independent of boundaries. Let us regard this as an experimental fact; then we choose to calculate the conductance for a system shaped as a torus. This has no boundaries, so there can be no surface currents. The calculation, in essence, measures the presence and nature of the delocalized electron states in the system.

Thus we are allowed to set up any boundary conditions we like. Following NTW, we choose "twisted" magnetic periodic boundary conditions,

$$\begin{aligned} \tilde{T}_{ix}(L_x)\psi &= e^{i\theta}\psi \\ \tilde{T}_{iy}(L_y)\psi &= e^{i\phi}\psi, \end{aligned}$$

where θ and ϕ are phases in the interval 0–2π. In other words, under magnetic translation of particle i through the length of the system, the state returns to itself multiplied by a phase factor. The phase is independent of the particle index as required to make the wavefunction totally antisymmetric.

Defining

$$u_n = \exp\left[-i\theta L_x^{-1}\sum x_i - i\phi L_y^{-1}\sum y_i\right]\psi_n,$$

one can perform algebraic manipulations of the Kubo expression entirely analogous to those leading from (10) to (12) in the noninteracting case. The result is

$$\sigma_H = \frac{ie^2}{h}\int d\mathbf{r}\left[\frac{\partial u_0^*}{\partial\theta}\frac{\partial u_0}{\partial\phi} - \frac{\partial u_0}{\partial\theta}\frac{\partial u_0^*}{\partial\phi}\right].$$

The integral is over the coordinates of all the particles and extends over the entire sample area.

Since we are calculating a bulk property, it should not depend on the boundary—the expression for σ_H, despite its superficial dependence on θ and ϕ, is actually independent of them. The next step is to average over macroscopic boundary conditions:

$$\sigma_H = \frac{1}{(2\pi)^2}\int\int_0^{2\pi} d\theta\, d\phi\, \sigma_H(\theta,\phi).$$

Obviously, since the conductance is independent of θ and ϕ, it does not matter if we average over them or not. But the averaged form is exactly

what we need to make the connection with the topological treatment: apart from the physical interpretation of some of the quantities, it is identical to the previous expression Eq. (12), with the magnetic Bloch wavevector replaced by the pair (θ, ϕ). Due to periodicity, the latter form a torus just like (k_x, k_y) did. We conclude without further ado that the Hall conductance is an integer times e^2/h.

That completes the actual "proof" of the quantization of Hall conductance. However, the reader might well feel uneasy about the unreasonable robustness and generality of the result. We need to explain why it does not flatly contradict the fractional quantum Hall effect, which is the quantization of conductance in *fractional* multiples of e^2/h. In the fractional effect the longitudinal conductance is observed to be zero, which suggests the presence of an excitation gap, and hence for the IQHE result to break down it is apparently necessary that the ground state be degenerate. In an infinite pure system the ground state is indeed degenerate as a result of the magnetic translation symmetry. However, for a finite system with disorder the origin of the degeneracy (if any) is, as far as I know, not settled. It is a rather curious situation the theorist finds himself in: the problem is to prove that the theory does not work!

6 Localization and conductance

In this final section we discuss in a little more detail the connection between wavefunction zeros and the Hall conductance, mentioned in Sec. 4.5.

Consider a gas of noninteracting electrons moving in a random background potential. The disorder is capable of localizing some or all of the single-electron states. We will now show that, as one would expect, the localized states do not contribute to the Hall conductance.

Recall the wavefunction phase convention, constructed in Sec. 4.5, using a reference point $\mathbf{r}_0$. The symbol $u_{\mathbf{k}}$ represents a collection of electron states, one for each point on the $\mathbf{k}$-space torus. We try to find a point $\mathbf{r}_0$ at which none of the wavefunctions have a node. If that is possible, the Chern number is zero and the set of states carries no Hall current; if is not possible, they carry a nonzero quantized amount of Hall current.

Let us rephrase this in the context of the NTW argument. It is appropriate to regard the magnetic Bloch wavevector as representing the macroscopic boundary condition. Then σ is nonzero if, by adjusting the boundary condition, we can make the wavefunction vanish at any given point in the sample—in other words, the state is very sensitive to the boundary condition. This is what one would call an extended (delocalized) state. Conversely, a localized state has appreciable magnitude only in some bounded region of space. Near the edges of the sample, the wavefunction amplitude

is so small that the conditions at the boundary have negligible effect on the state. In particular, we can choose some spot at which the wavefunction has appreciable amplitude, and then be sure that, no matter how we fool around with the boundary condition, the wavefunction can never be made to have a node there. With that choice of $\mathbf{r}_0$, $u_k(\mathbf{r}_0)$ is nonzero for all $\mathbf{k}$, and therefore the Hall current carried by the localized state is zero.

An interesting approach to the interplay of localisation properties and conductance was developed by Arovas *et al.* (1988). They considered non-interacting electrons in a finite system with a random substrate potential, with the motion projected onto the lowest Landau level—a good approximation in the limit of high magnetic field. They showed that the states have exactly N_s zeros, where N_s is the number of flux quanta passing through the sample, and that each state is completely determined by the location of its zeros. As one varies the boundary condition phases (θ, ϕ), the wavefunction zeros move around in characteristically different ways for localized and extended states. Extended states can be made to vanish at any specified point in the sample by a suitable choice of boundary conditions, and hence the union of wavefunction zeros over all boundary conditions covers the whole sample; localized states are "rigid" and the zeros are confined to a bounded region. I encourage the reader to look at the curious plots of the topology of wavefunction zeros given in the original paper, and to try to read off the Hall conductance from them.

One way of defining localization of electron states is in terms of the decay of the wavefunction amplitude at large distances. For the theorist solving specific models on a computer, however, this definition is somewhat impractical—it requires large system sizes to extract meaningful results. The work of Arovas *et al.* gives a much more useful definition of localization, in terms of a state's Chern number, which we have seen is directly related to the current-carrying capacity of the wavefunction. A state is defined to be localized if its Chern number is zero, and extended otherwise. The great practical benefit is that, even for small system sizes, it is rather simple to decide if a computed state is localized. An example of applying this idea is the calculation by Huo and Bhatt (1992) of the density of extended (nonzero Chern number) states in the lowest Landau level, including the effects of a disordered substrate potential. They found that in the thermodynamic limit, all the extended states are concentrated at a single energy, at the centre of the disorder-broadened level.

Acknowledgements

I would like to thank Geoff Canright for many discussions and critical comments, and Paul McCann and Stephen Lovesey for valuable suggestions on the text. I am grateful to John Baez for helping me learn some topology, and to Lizeng Zhang for encouraging me to give a talk on this topic.

References

Arovas D P, Bhatt R N, Haldane F D M, Littlewood P B and Rammal R 1988 *Phys. Rev. Lett.* **60** 619

Dana I, Avron Y and Zak J 1985 *J. Phys. C.: Solid State Phys.* **18** L67

Harper P G 1955 *Proc. Phys. Soc.* A **68** 874

——— 1991 *J. Phys.: Condens. Matter* **3** 3047

Hofstadter D R 1976 *Phys. Rev.* B **14** 2239

Huo Y and Bhatt R N 1992 *Phys. Rev. Lett.* **68** 1375

Kohmoto M 1985 *Ann. Phys.* **160** 343

Laughlin R B 1981 *Phys. Rev.* B **23** 5632

Lovesey S W, Watson G I and Westhead D R 1991 *Int. J. Mod. Phys.* B **5** 1313

Monastyrski M 1987 *Riemann, Topology and Physics* (Boston: Birkhäuser)

Morandi G 1988 *Quantum Hall Effect: Topological Problems in Condensed-Matter Physics* (Napoli: Bibliopolis)

Nakahara M 1990 *Geometry, Topology and Physics* (Bristol: Adam Hilger)

Niu Q, Thouless D J and Wu Y 1985 *Phys. Rev.* B **31** 3372

Rauh A 1974 *Phys. Stat. Sol.* B **65** K131

——— 1975 *Phys. Stat. Sol.* B **69** K9

Rauh A, Wannier G H and Obermair G 1974 *Phys. Stat. Sol.* B **63** 215

Schutz B 1980 *Geometrical Methods of Mathematical Physics* (Cambridge University Press)

Shapere A and Wilczek F (ed.) 1989 *Geometric Phases in Physics* (Singapore: World Scientific)

Simon B 1983 *Phys. Rev. Lett.* **51** 2167

Sokoloff J B 1985 *Phys. Rep.* **126** 189

Stone M 1992 *Quantum Hall Effect* (Singapore: World Scientific)

Středa P 1982a *J. Phys. C.: Solid State Phys.* **15** L717

——— 1982b *J. Phys. C.: Solid State Phys.* **15** L1299

Středa P and Smrčka L 1983 *J. Phys. C.: Solid State Phys.* **16** L895

Thouless D J 1984 *Phys. Rep.* **110** 279

——— 1987 in *The Quantum Hall Effect*, edited by R E Prange and S M Girvin (New York: Springer), Ch. 4

——— 1989 *Phys. Rev.* B **40** 12034

Thouless D J and Gefen Y 1991 *Phys. Rev. Lett.* **66** 806

Thouless D J, Kohmoto M, Nightingale M P and den Nijs M 1982 *Phys. Rev. Lett.* **49** 405

Wannier G H 1978 *Phys. Stat. Sol.* B **88** 757

Widom A 1982 *Phys. Lett.* **90A** 474

Wu T T and Yang C N 1975 *Phys. Rev.* D **12** 3845

Zak J 1964a *Phys. Rev.* **134** A1602

——— 1964b *Phys. Rev.* **134** A1607