NOTES FOR PHYSICS 223A

LEON BALENTS

Leon Balents[*]

UCSB

CONTENTS

[*]balents@spinsandelectrons.com (scribe)

LIST OF FIGURES

1 BASIC NOTIONS AND THE STABILITY OF MATTER

Lecture 1 (1h 15mins)
26th September 2022

This is the first quarter of a course on condensed matter physics. This quarter covers what is traditionally called solid state physics, focusing on the elementary description of crystalline solids. We will cover the mathematical description of periodic structures, the band theory of solids, transport, and lattice dynamics. Along the way, we will discuss aspects of topology in band theory, quantum Hall effects, and a variety of experimental tools and the theory behind them.

One can think of condensed matter physics in several ways. On the one hand, it is about the physics of materials. It addresses the question of why a given material has the properties it does, e.g. why is gold shiny and conducting

and malleable, why is ceramic brittle, why is iron magnetic while aluminum is not, why does $La_{2-x}Sr_xCuO_4$ become a high temperature superconductor, why is mercury liquid at room temperature? We can ask for the phase diagram of collection of molecules versus temperature a pressure. One can envision an infinity of such questions, or varying degrees of importance.

Another goal of condensed matter physics is to learn to manipulate matter. Semiconducting devices, which are the "brains" of all modern computing and electrics, are the outcome of decades of study of condensed matter physics. Today researchers are using condensed matter physics to try to develop quantum devices for quantum computing. Under the umbrella of this goal is the invention of new tools, such as scanning tunneling microscopes, atomic force microscopes, nuclear magnetic resonance, photoemission spectroscopy, etc. These tools use condensed matter physics but also serve to help us interrogate condensed matter further, and are useful both for new insights and for applications.

Yet a third view is that condensed matter physics is about the fundamental problem of emergence. It aims to understand the mechanisms by which large scale – macroscopic or "mesoscopic" – behaviors result from the microscopic laws governing many constituents (e.g. particles) and how to describe these large scale behaviors. Emergence includes ordering – the development of magnetic, superconducting, electric dipole, etc. order in a system – the phenomena of spontaneous symmetry breaking and associated universal low energy excitations. It also includes topological phenomena, which emerge in many condensed matter contexts. The apparently classical nature of the macroscopic world is something that emerges, and how it does is a problem that fits, at least in part, into condensed matter physics.

1.1 *Why condensed matter is condensed*

The field of "condensed matter" is ultimately about the matter around us. In this class, we will describe matter as the combination of positively charged nuclei and negatively charged electrons. We don't need to concern ourselves with the forces that hold the nucleus together (the strong force), or that govern its decay (the weak force). Particles apart from these two types play basically no role in condensed matter, though neutrons and other particles are often used as probes to study it, and sometimes photons can be trapped long enough to interact strongly with matter. Gravity is so weak that it is irrelevant (at least in terrestrial settings) to the structure and properties of matter. What is left is electromagnetism, which is really the sole force that we will care about, microscopically. The other important ingredient is quantum mechanics.

I would like to start by discussing the defining property of *condensed* matter, which is that it holds itself together. From the above discussion, it is clear that what holds matter together is the electrostatic Coulomb force. Let's see if we can understand a bit better how this occurs.

1.1.1 The fundamental Hamiltonian

The problem is very concrete. Indeed, we can write one single Hamiltonian which describes the vast majority of physics of condensed matter.

$$H = \sum_{i=1}^{N_n} \frac{P_i^2}{2M_i} + \sum_{i=1}^{N_e} \frac{p_i^2}{2m_e} - \sum_{i=1}^{N_n}\sum_{j=1}^{N_e} \frac{Z_i e^2}{|\mathbf{X}_i - \boldsymbol{x}_j|} + \sum_{i<j=1}^{N_n} \frac{Z_i Z_j e^2}{|\mathbf{X}_i - \mathbf{X}_j|} + \sum_{i<j=1}^{N_e} \frac{e^2}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}. \tag{1}$$

Here lower case and capital letters refer to coordinates/momenta of electrons and nuclei, respectively. $M_i \sim 2Z_i m_n$ is the mass of the i^{th} nucleus, Z_i is its atomic number, m_n is of order the proton or neutron mass, and m_e is the electron mass. This is to be treated using quantum mechanics, e.g. eigenstates are found via

$$H(\{\boldsymbol{p}_i, \boldsymbol{x}_i, \mathbf{P}_i, \mathbf{X}_i\})\Psi(\boldsymbol{x}_1, \cdots \boldsymbol{x}_{N_e}; \mathbf{X}_1, \cdots \mathbf{X}_{N_n}) = E\Psi(\boldsymbol{x}_1, \cdots \boldsymbol{x}_{N_e}; \mathbf{X}_1, \cdots \mathbf{X}_{N_n}), \tag{2}$$

where $\mathbf{P}_i = -i\hbar\nabla_{\mathbf{X}_i}$, $\boldsymbol{p}_i = -i\hbar\nabla_{\boldsymbol{x}_i}$ in the coordinate representation. I have suppressed spin indices, but in general Ψ should also be a function of each electron and nuclear spin (without spin-orbit coupling, this only imposes some degeneracies and symmetry constraints on the orbital part of the wavefunction, since spin does not explicitly enter Eq. (2)). This is a complete description except in the presence of external and time-dependent electromagnetic fields, and neglecting relativistic effects, i.e. spin-orbit coupling. These are not major modifications, and Eq. (1) contains most of what will be important in this course, though spin-orbit coupling will play a role. Eq. (1) defines the generic many-body problem for condensed matter. It describes atoms, molecules, solids, and liquids.

1.1.2 Born-Oppenheimer approximation

A critical fact about this Hamiltonian is that the nuclear mass is much larger than the electronic one: the ratio of the electron to proton mass is about 1/1800. This is the only generic small parameter in the description of ordinary matter. It is a very important one. It implies that motions of the nuclei are much slower than those of the electrons. This justifies the *Born-Oppenheimer approximation*, which says that the quantum description of the electrons can be separated from that of the nuclei, and that, to lowest order, it consists of neglecting the dynamics of the nuclei completely. Indeed, taking $M_i \to \infty$ in Eq. (1), we obtain a Hamiltonian in which the $\mathbf{X}_i$ are classical variables, and constants of the motion:

$$H_{BO}(\{\boldsymbol{p}_i, \boldsymbol{x}_i; \mathbf{X}_i\}) = \sum_{i=1}^{N_e} \frac{p_i^2}{2m_e} - \sum_{i=1}^{N_n}\sum_{j=1}^{N_e} \frac{Z_i e^2}{|\mathbf{X}_i - \boldsymbol{x}_j|} + \sum_{i<j=1}^{N_e} \frac{e^2}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} + \sum_{i<j=1}^{N_n} \frac{Z_i Z_j e^2}{|\mathbf{X}_i - \mathbf{X}_j|}, \tag{3}$$

where the last term is simply constant for fixed atomic positions. We can view the nuclear coordinates as defining a family of quantum Hamiltonians for the electrons. Each such Hamiltonian can be solved separately, for example to compute the eigenstates and energy eigenvalues, which are then functions of the classical nuclear coordinates. This is obviously a huge simplification.

Born-Oppenheimer approximation: Due to the large nuclear mass, it is an excellent approximation to treat the atomic coordinates as fixed classical numbers, and solve the quantum mechanical problem for fixed atomic positions. At zero temperature this consists of finding the ground state energy $E_0(\{\mathbf{X}_i\})$ of Eq. (3). The structure is determined by the lowest minima over these coordinates of $E_0(\{\mathbf{X}_i\})$.

1.2 *One atom*

With this understanding, let us return to the question of cohesion of matter. First let us recall the situation of a single atom, where Eq. (1) reduces to

$$H = \sum_{i=1}^{Z} \left(\frac{p_i^2}{2m_r} - \frac{Ze^2}{|\boldsymbol{x}_i|} \right) + \sum_{i<j=1}^{Z} \frac{e^2}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}, \tag{4}$$

where we separated and eliminated the center of mass coordinate, and $m_r \approx m_e$ is the reduced mass (in the Born-Oppenheimer approximation, $m_r = m_e$, which is seen to be very accurate).

1.2.1 Hydrogen atom: length and energy scales

Everyone has solved the problem for $Z = 1$, the hydrogen atom. You may have forgotten that there is anything surprising about this problem, but it is good to remember that, classically, even the hydrogen atom with $Z = 1$ is unstable, because an electron can spiral closer and closer to the nucleus and continually lower its energy. The collapse of an atom is halted by quantum uncertainty: approaching the nucleus very closely requires reduction of the electron's positional uncertainty, and consequently an increase in the quantum fluctuations of its momentum. When the latter becomes too large, the kinetic energy increases and this increase exceeds the energy gain due to the Coulomb potential, halting the inward motion of the electron. The Bohr radius of the hydrogen atom is determined by this balance of the kinetic energy determined through the uncertainty relation, $p^2/2m_e \sim \hbar^2/m_e a^2$, where a is the radius of localization, which should be balanced against the Coulomb energy $\sim e^2/a$. This gives $a = a_0 = 1/m_e e^2$ (in cgs units), which is just the Bohr radius, equal to about half an angstrom. This sets the basic length scale for atoms in condensed matter. Typical inter-atomic distances are a few angstroms. The basic quantum level spacing in the atomic problem is also familiar. The binding energy of the hydrogen atom defines the Rydberg constant, $R_\infty = e^2/2a_0 = m_e e^4/2\hbar^2$ in

cgs units or $R_\infty = m_e e^4/(2(4\pi\epsilon_0)^2\hbar^2)$ in SI units, and it is useful to remember $R_\infty \approx 13eV$. One should remember that typical energy scales for electronic motion in solids are electron volts. It is a good idea to also remember the conversion between eV and K: 1eV = 1.4×10^4K.

Scales in condensed matter: The fundamental length and energy scales in solids are set by the competition of Coulomb potential and electronic kinetic energy, and consequently are of the same order as the Bohr radius and binding energy (Rydberg) of the hydrogen atom. Thus typical length scales (electron wavelengths, atomic spacings) in solids are of order an angstrom, and typical energy scales are of order electron volts. Since 1eV $\approx 1.4\times10^4$K, even room temperature is "low" for electrons, i.e. the electrons are close to their ground state even at room temperature.

Hence typical electronic energy scales in condensed matter are of order 10,000K. This is significantly larger than room temperature (300K), i.e. $R_\infty \gg k_B T$, where k_B is Boltzmann's constant and $T \leq 300$K. Hence in most of the matter around us, electrons can be expected to be close to their ground state.

1.2.2 Atomic collapse for fictitious bosonic electrons

The next step is the many-electron atom. This raises a new fundamental question: how does the size of matter change as you increase the number of charges? We are familiar with conventional matter, which is *extensive*: the volume of material is proportional to the number of electrons and protons in it, i.e. to its mass. How does this end up being the case? If one proton attracts one electron, more protons will attract more electrons more strongly, so why would matter not *decrease* in size as we add more electrons and protons? What prevents such a many-particle *collapse* of matter?

In the context of a multi-electron atom, the question becomes: how does the atomic radius depend upon the atomic number Z? In fact, it is easy to show that avoiding many-particle collapse relies not only on uncertainty but also on Fermi-Dirac statistics and the Pauli exclusion principle. If electrons were bosons, then actually collapse would indeed occur! We can see this by constructing a bosonic variational wavefunction for Eq. (4) which puts every electron in the same, spherically symmetric, state: $\Psi = \prod_i \psi(r_i)$, where $r_i = |\boldsymbol{x}_i|$.

Then we can evaluate the variational energy

$$\begin{aligned}\langle\Psi|H|\Psi\rangle =& \frac{4\pi Z}{2m_r}\int_0^\infty dr r^2\left(\frac{d\psi}{dr}\right)^2 - 4\pi Z^2 e^2\int_0^\infty dr r|\psi(r)|^2 \\ &+ 8\pi^2 e^2\frac{Z(Z-1)}{2}\int_0^\infty dr_1 dr_2 r_1^2 r_2^2|\psi(r_1)|^2|\psi(r_2)|^2\int_{-1}^1 \frac{d\cos\theta}{\sqrt{r_1^2+r_2^2-2r_1r_2\cos\theta}} \\ =& \frac{4\pi Z}{2m_r}\int_0^\infty dr r^2\left(\frac{d\psi}{dr}\right)^2 - 4\pi Z^2 e^2\int_0^\infty dr r|\psi(r)|^2 \\ &+ 16\pi^2 e^2\frac{Z(Z-1)}{2}\int_0^\infty dr_1 dr_2\frac{r_1^2 r_2^2}{\max(r_1,r_2)}|\psi(r_1)|^2|\psi(r_2)|^2. \end{aligned} \tag{5}$$

Now we choose a specific form, which is just that of the ground state of the hydrogen atom:

$$\psi(r) = \frac{1}{\sqrt{\pi a^3}}e^{-\frac{r}{a}}, \tag{6}$$

where a will determine the "size" of the atom and is a variational parameter. This is a normalized wavefunction so that $\int d^3\boldsymbol{x}|\psi(|\boldsymbol{x}|)|^2 = 4\pi\int_0^\infty dr r^2|\psi(r)|^2 = 1$. We obtain then

$$\langle\Psi|H|\Psi\rangle = \frac{Z}{2m_r a^2} - \frac{Z^2e^2}{a} + \frac{5Z(Z-1)e^2}{16a}. \tag{7}$$

Here the first term is the kinetic energy, the second term is the attraction of the electrons to the nuclei, and the third is the electron-electron repulsion. The crucial fact is that for large $Z \gg 1$, the Coulomb terms become dominant and scale as Z^2, and moreover the coefficient of the Z^2 term from the attraction, -1, is larger in magnitude than the coefficient of the Z^2 term from the repulsion, $5/16 < 1$, so that attraction dominates (in fact the net Coulomb effect is attractive for all $Z \geq 1$). Optimizing the energy over a, we find the atomic radius

$$a = \frac{16}{11Z+5}a_0 \sim_{Z\gg1} \frac{16}{11Z}a_0, \tag{8}$$

where $a_0 = 1/me^2$ is the Bohr radius. We see that the size of the atom decreases linearly with the atomic number. *Real atoms do not do this!!!!* In real atoms, the atomic radius actually has a complicated evolution and is much larger than Eq. (8) for large Z.

1.2.3 Fermi statistics to the rescue: Thomas-Fermi theory of the atom

In this discussion, we introduce some basic concepts of the electron gas, which are important well beyond Thomas-Fermi theory.

The difference from Eq. (8) is of course due to Fermi statistics. We cannot solve a large atom exactly, but in fact there is a theoretical description that applies asymptotically for large Z. This is Thomas-Fermi theory. The Thomas-Fermi approximation is a semi-classical one, basically relying upon a large electron density. The idea is that when the density is large (as it is in most of space when we have a large atom), we can break up space into many small regions, such that the electron density is approximately constant in each region. If the density itself is large, then the number of electrons in each region is large ($\gg 1$), and the law of large numbers applies. The density of electrons in each region, i.e. the local density, becomes a classical variable $n(\boldsymbol{x})$. This has two implications. First, the kinetic energy in this region can be approximated by the kinetic energy of a gas of uniform density (equal to the average in this region) multiplied by the volume of the region. Second, the potential energy – i.e. the Coulomb interaction of the electrons with the nucleus and amongst one other – can be approximated by the corresponding electrostatic energy of a continuous charge density.

Let us formulate the Thomas-Fermi theory of the atom now explicitly. First we need an expression for the kinetic energy as a function of density. To obtain this, consider a cubic box of linear size L containing N electrons, with periodic boundary conditions – the result is actually independent of the shape and boundary conditions of the box. The assumption is that the electrostatic potential is approximately constant over the box, so the electrons behave as free particles, and the potential does not play any role in the kinetic energy. The electron wavefunctions are just plane waves,

$$\psi_{\boldsymbol{k}}(\boldsymbol{x}) = \frac{1}{\mathrm{L}^{3/2}} e^{i\boldsymbol{k}\cdot\boldsymbol{x}}, \qquad \boldsymbol{k} = \frac{2\pi}{\mathrm{L}}\left(m_1\hat{\boldsymbol{x}} + m_2\hat{\boldsymbol{y}} + m_3\hat{\boldsymbol{z}}\right), \tag{9}$$

where m_μ are integers. Taking into account the two spin states of each electron, the many-electron ground state is obtained consistent with the Pauli principle by putting two electrons sequentially in states beginning with $\boldsymbol{k} = 0$ and then into those with increasing magnitude of momentum, up to some maximum which is known as the *Fermi momentum* k_F. We are interested in the limit $\mathrm{N} \to \infty$, $\mathrm{L} \to \infty$, with $n = \mathrm{N}/\mathrm{L}^3$ fixed. Then the filled states form a "Fermi sphere" with radius k_F, whose volume is $4\pi k_\mathrm{F}^3/3$. The number of discrete momenta contained in this sphere is this volume, divided by the phase space volume $(2\pi/\mathrm{L})^3$. The number of electrons is twice this, hence we see that $\mathrm{N} = 2 \times 4\pi k_\mathrm{F}^3/3 \times (\mathrm{L}/2\pi)^3$ so that

$$n = \frac{k_\mathrm{F}^3}{3\pi^2}, \qquad k_\mathrm{F} = \left(3\pi^2 n\right)^{1/3}. \tag{10}$$

The kinetic energy density $\mathrm{T}(n) = 1/\mathrm{L}^3 \sum_i k_i^2/2m$ is obtained by adding $k^2/(2m)$

up for every electron, which gives

$$\mathrm{T}(n) = 2\int_0^{k_\mathrm{F}} dk k^2 4\pi \frac{k^2}{(2\pi)^3}\frac{k^2}{2m} = \frac{2k_\mathrm{F}^5}{20\pi^2 m} = \frac{3}{10}\left(3\pi^2\right)^{2/3}\frac{n^{5/3}}{m}. \tag{11}$$

The Thomas-Fermi expression for the total kinetic energy is just the integral of this,

$$\mathrm{T}_{\mathrm{TF}}[n(\boldsymbol{x})] = \int d^3\boldsymbol{x}\frac{3(3\pi^2)^{2/3}}{10m}[n(\boldsymbol{x})]^{5/3}. \tag{12}$$

This is a classical *functional* of the density. Note that if density is increased, the kinetic energy density increases as a power of the density larger than one, hence it is "expensive" to accumulate many fermions in one place. This is a reflection of Fermi statistics and the Pauli principle: fermions cannot get too close together. Keep in mind this is a purely statistical repulsion, not an electrostatic one, since the result was obtained without using any Coulomb energy. This result is enough to understand intuitively how collapse is avoided (see the box).

Interpretation of the Thomas-Fermi result and why it avoids collapse: The Thomas-Fermi result has a simple interpretation in terms of the *energy per particle*. To get this, we simply divide the energy density (integrand of Eq. (12)) by the density, which scales as the density to the 2/3 power. Now if the typical spacing between electrons is ℓ, the density is ℓ^{-3}, to the Thomas-Fermi result can be re-phrased to say that *the kinetic energy per particle of a collection of fermions is at least as large as a constant times* $1/(m\ell^2)$ (it is a lower bound because it came from an estimate of the ground state energy of the electron gas). In this form it can be anticipated by dimensional analysis because ($\hbar = 1$) the kinetic energy operator in quantum mechanics is $-\frac{1}{2m}\nabla^2$. This requires Fermi rather than Bose statistics because only for fermions does the Pauli principle force the typical wavenumber to be comparable to the inverse inter-particle spacing: for bosons it can be much smaller and the kinetic energy need not rise as ℓ decreases.
To understand why fermions evade collapse, you can compare this energy cost as ℓ decreases with the possible gain from Coulomb attraction of electrons and nuclei. Because of Coulomb's law, we expect that the most the energy can be lowered by attraction is an amount proportional to $-e^2/\ell$ per electron. Collapse would mean $\ell \to 0$, but in this limit, the Thomas-Fermi kinetic energy per particle increase will overwhelm the Coulomb gain because $1/(m\ell^2) \gg e^2/\ell$ when ℓ is sufficiently small.

Now let's add the kinetic and potential energy to produce the Thomas-

Fermi (TF) energy functional for an atom,

$$E_{TF}[n] = T_{TF}[n] - Ze^2 \int d^3\boldsymbol{x}\, \frac{1}{|\boldsymbol{x}|} n(\boldsymbol{x}) + \frac{e^2}{2} \int d^3\boldsymbol{x} d^3\boldsymbol{x}'\, \frac{n(\boldsymbol{x})n(\boldsymbol{x})}{|\boldsymbol{x}-\boldsymbol{x}'|}. \tag{13}$$

This should be minimized subject to the constraints $n(\boldsymbol{x}) \geq 0$ and $\int d^3\boldsymbol{x}\, n(\boldsymbol{x}) = N_e = Z$ (actually we could study ionized atoms $N_e \neq Z$ but we seek the neutral case here). the total charge density via a Lagrange multiplier, we write

$$L[n] = E_{TF}[n] - \mu\left(\int d^3\boldsymbol{x}\, n(\boldsymbol{x}) - Z\right), \tag{14}$$

and the variational equation becomes

$$\frac{\delta L}{\delta n(\boldsymbol{x})} = C\,[n(\boldsymbol{x})]^{2/3} - \frac{Ze^2}{|\boldsymbol{x}|} + \int d^3\boldsymbol{x}'\, \frac{e^2}{|\boldsymbol{x}-\boldsymbol{x}'|} n(\boldsymbol{x}') - \mu = 0, \tag{15}$$

where $C = (3\pi^2)^{2/3}/2m$. It is convenient to rewrite Eq. (15) as

$$C\,[n(\boldsymbol{x})]^{2/3} = \frac{Ze^2}{|\boldsymbol{x}|} - \int d^3\boldsymbol{x}'\, \frac{e^2}{|\boldsymbol{x}-\boldsymbol{x}'|} n(\boldsymbol{x}') + \mu \equiv \phi(\boldsymbol{x}), \tag{16}$$

where $\phi(\boldsymbol{x})$ is an effective potential. Using $\nabla^2 \frac{1}{|\boldsymbol{x}|} = -4\pi\delta(\boldsymbol{x})$ as you learned in electromagnetism, we can simplify Eq. (16) by taking the Laplacian of the last two expressions:

$$\nabla^2\phi(\boldsymbol{x}) = 4\pi e^2\,(n(\boldsymbol{x}) - Z\delta(\boldsymbol{x}))\,. \tag{17}$$

Now using the equality between the first and third terms in Eq. (17), we can express this equation entirely in terms of ϕ:

$$\nabla^2\phi(\boldsymbol{x}) = 4\pi e^2 \left[\left(\frac{\phi(\boldsymbol{x})}{C}\right)^{3/2} - Z\delta(\boldsymbol{x})\right]. \tag{18}$$

This is known as the Thomas-Fermi equation, and it is simple electrostatics: just Poisson's equation using the Thomas-Fermi relation $\phi = Cn^{2/3}$. We can use spherical symmetry to write $n(\boldsymbol{x}) = n(r)$, and obtain

$$\frac{1}{r}\frac{d^2}{dr^2}(r\phi) = 4\pi e^2 \left(\frac{\phi(r)}{C}\right)^{3/2}, \qquad r > 0. \tag{19}$$

The condition that the total number of electrons is Z implies charge neutrality, which means that at infinity the potential decays faster than $1/r$, i.e. $\lim_{r\to\infty} r\phi(r) = 0$. Near the origin, the bare nuclear charge dominates and $\lim_{r\to 0} r\phi(r) = Ze^2$. Hence let us define $\psi(r) = r\phi(r)/(Ze^2)$. It obeys

$$\frac{d^2\psi}{dr^2} = \frac{4\pi Z^{1/2} e^3}{\sqrt{r}} \left(\frac{\psi(r)}{C} \right)^{3/2}, \tag{20}$$

and $\lim_{r\to\infty} \psi(r) = 0$, $\lim_{r\to 0} \psi(r) = 1$. Inspection of Eq. (20) shows that the dependence on Z, e, and C can be made explicit by taking $\psi(r) = \tilde{\psi}(r/R)$, with

$$1/R^{3/2} = \frac{Z^{1/2} e^3}{C^{3/2}} \qquad \Rightarrow \qquad R = (\frac{Z^{1/2} e^3}{C^{3/2}})^{-2/3} = \frac{C}{Z^{1/3} e^2}, \tag{21}$$

and

$$\frac{d^2\tilde{\psi}}{d\tilde{r}^2} = \frac{4\pi}{\sqrt{\tilde{r}}} \left(\tilde{\psi}(\tilde{r}) \right)^{3/2}, \tag{22}$$

where $\tilde{r} = r/R$. We can go back to express the density as

$$n(r) = (\phi/C)^{3/2} = (Ze^2\psi/(Cr))^{3/2} = \frac{Z^2 e^6}{C^2} \left(\frac{\tilde{\psi}(r/R)}{r/R} \right)^{2/3}. \tag{23}$$

Eq. (23) shows that the size of the atom (R) decreases as $1/Z^{1/3}$ in Thomas-Fermi theory, much more slowly than $1/Z$ as it would for bosons. Note however that the density is still large for $Z \gg 1$ so the Thomas-Fermi approximation is justified. Eq. (22) is still non-trivial to solve and Thomas-Fermi theory of the atom has many more interesting aspects, but they would take us too far afield and we will not explore the solution further. I would like to comment that while the Thomas-Fermi approximation is crude, and has many deficiencies, it is a useful concept in many ways. It is the inspiration for density functional theory, which is the basis for modern calculations of electronic structure and the vast majority of computational studies of materials (and widely used in chemistry as well). It forms the basis for the theory of screening, which is very important for understanding properties of metals. Finally, the Thomas-Fermi approximation can actually be shown, with small modifications, to provide a rigorous lower bound on the energy[2].

1.3 *Cohesion and structure of macroscopic matter*

The Thomas-Fermi treatment shows how Fermi statistics prevents collapse at the atomic scale. It is a separate but related problem to *prove* that the same mechanism leads to stability of macroscopic matter, i.e. that in a system with N atoms, the volume grows proportionally to N when N is large, so that a material has an intrinsic density. It is equivalent to say that the ground state energy E of the system of N_n nuclei is proportional to N_n when $N_n \to \infty$. The basic idea is again to show that collapse is avoided, i.e. that atoms do not bind increasingly tightly as more of them are put together. The many-body Hamiltonian itself contains terms which are not bounded below, and it is

conceivable that the energy could be made increasingly negative by pulling electrons increasingly close to increasingly many nuclei at large N_n. This is exactly the regime in which Thomas-Fermi theory can be applied to the problem, since if this occurs, the electron density becomes large. We will not discuss this in detail, but just state some results. For a general ionic potential $V(\boldsymbol{x})$, the TF energy functional is

$$E_{TF}[n, V] = T_{TF}[n] + \int d^3\boldsymbol{x}\, V(\boldsymbol{x}) n(\boldsymbol{x}) + \frac{e^2}{2} \int d^3\boldsymbol{x} d^3\boldsymbol{x}' \, \frac{n(\boldsymbol{x}) n(\boldsymbol{x})}{|\boldsymbol{x} - \boldsymbol{x}'|}. \tag{24}$$

Because there is no binding in Thomas-Fermi theory, cohesion of solids necessarily goes beyond it! Thomas-Fermi theory is good for the dense "cores" of atoms, but all the binding occurs in the outer "halos" of atoms where electrons are not so dense and Thomas-Fermi theory does not apply. Modern density functional theory follows in the spirit of the Thomas-Fermi approximation but is able to capture binding, and is the standard way to find E_0.

For a general set of nuclei, we have from Eq. (1)

$$V(\boldsymbol{x}) = -\sum_{i=1}^{N_n} \frac{Z_i e^2}{|\mathbf{X}_i - \boldsymbol{x}|}. \tag{25}$$

This energy functional provides a good approximation to the energy if the electron density is large, so it can be used to describe the possibility of collapse. More specifically, it provides a proof by contradiction that such collapse does not occur. To summarize the logic: if collapse occurs (i.e. the atoms minimize their energy by coming very close together as their number increases), then Thomas-Fermi theory is a good approximation to the energy. However, Edward Teller proved in 1962 that *there is no binding within Thomas-Fermi theory*. In fact, the Thomas-Fermi energy of any assembly of atoms is always minimized by bringing the atoms infinitely far from one another. Thus collapse does not occur. This and other interesting things are described (and rigorously proven) in the beautiful paper by Lieb[2].

Let us consider the consequences of these deep statements for more practical things. How are the density and in more detail the structure of a material determined? We return to the many-body problem of Eq. (1) in the Born-Oppenheimer approximation. For a given set of atomic (nuclear) coordinates $\{\mathbf{X}_i\}$, we may approximately consider the electrons in their ground state, which defines some function $E_0(\mathbf{X}_1, \cdots, \mathbf{X}_{N_n})$.

In general $E_0(\mathbf{X}_1, \cdots, \mathbf{X}_{N_n})$ is a very complicated function, which can at best be approximately computed numerically for specific values of the $\mathbf{X}_i$. The proof by Lieb shows that it is bounded below, as an energy should be, so that some absolute minima exist. However, even simple functions can have many local minima, and generally this is expected to be the case for the true energy functions. A physicist's or chemist's intuition might come from considering some limits. For example, when the nuclei are well separated, the low energy electronic states become close to those of isolated atoms. The dependence of the energy upon the separation results from perturbations around this limit due to induced multipoles and their interactions. This is the famous van der Waals interaction, and generally leads to a weak attraction between atoms at long distances, i.e. the energy reduces as atoms are brought together

from infinity. Conversely, if two nuclei approach one another very closely, the energy will grow due to the repulsion between their positive charges. Putting these two tendencies together, we can deduce that the minimum energy configurations generally have some finite and non-zero separation between nuclei. It is natural to try to encapsulate these two tendencies by postulating that the energy is a sum of two-body interactions. However, in general this is not the case, and a priori the minimum energy structures occur precisely at intermediate distances between nuclei where there are no simple approximations. Fortunately, for most purposes, we do not really need to know how to express the full energy function, but only that it exists and so do local minima. Such local minima comprise locally stable *atomic structures*.

The focus on the minima of the energy function presumes that the thermal energy is small compared to the depth of the minima. This is generally true due to the considerations of energy scales we discussed earlier. In particular, we expect that varying a nuclear coordinate $\mathbf{X}_i$ by an amount of order a Bohr radius will change the energy E_0 by an amount of order R_∞. This is a large energy compared to $k_\mathrm{B}\mathrm{T}$, hence we can expect that in equilibrium, the nuclei will be predominantly found in configurations in which E_0 is close to its minimum. Therefore it is natural to focus on the structural ground states, determined in principle by minimizing this function E_0. There are certainly instances where this is insufficient, for example in considering the motions of small molecules, or for very light atoms. More generally in the consideration of liquids, atoms are constantly in motion and occupying at least a large manifold of low-energy states.

In this class, we will be concerned with solids, and in the thermodynamic limit, when $\mathrm{N}_e, \mathrm{N}_n \to \infty$, in which case we expect to find structures that have finite density. The simplest examples of finite density structures are *periodic crystals*, which can be locally stable and may be the global minimum energy solutions in the thermodynamic limit. In this class we will concentrate on such structures. We will soon discuss the description of periodic solids in some detail. For now we simply summarize these as arrangements of nuclei which are repeated to tile space, in such a way that when the solid is displaced by multiples of certain displacements, which define the periodicity, the structure is unchanged. Periodic solids are spatially uniform on large scales. Their nature means that local properties do not depend upon where they are measured, at least from one repeating unit to another. They comprise *materials*, whose intrinsic properties are well defined and do not depend upon the boundary, etc. Periodicity means that only a very small amount of data is needed to specify the locations of all the nuclei in a system of arbitrary size, and we can indeed take a sensible thermodynamic limit in which the volume of the system goes to infinity.

Once one assumes periodicity, modern theory is actually rather good at predicting the structure of many solids, using computational techniques. However, most of condensed matter physics really starts once the structure is determined, either by theory or by experiment. Then one has reduced the

infinite family of many body problems for the electrons to a single one, given by Eq. (1) (at $M_i = \infty$) and the now-determined values of the nuclear coordinates $\mathbf{X}_i$. This is still a very formidable quantum problem, with extremely rich physics. It will comprise most of this class.

2 PERIODIC STRUCTURES

Lecture 2 (1h 15mins)
28th September 2022

2.1 *Crystal lattices*

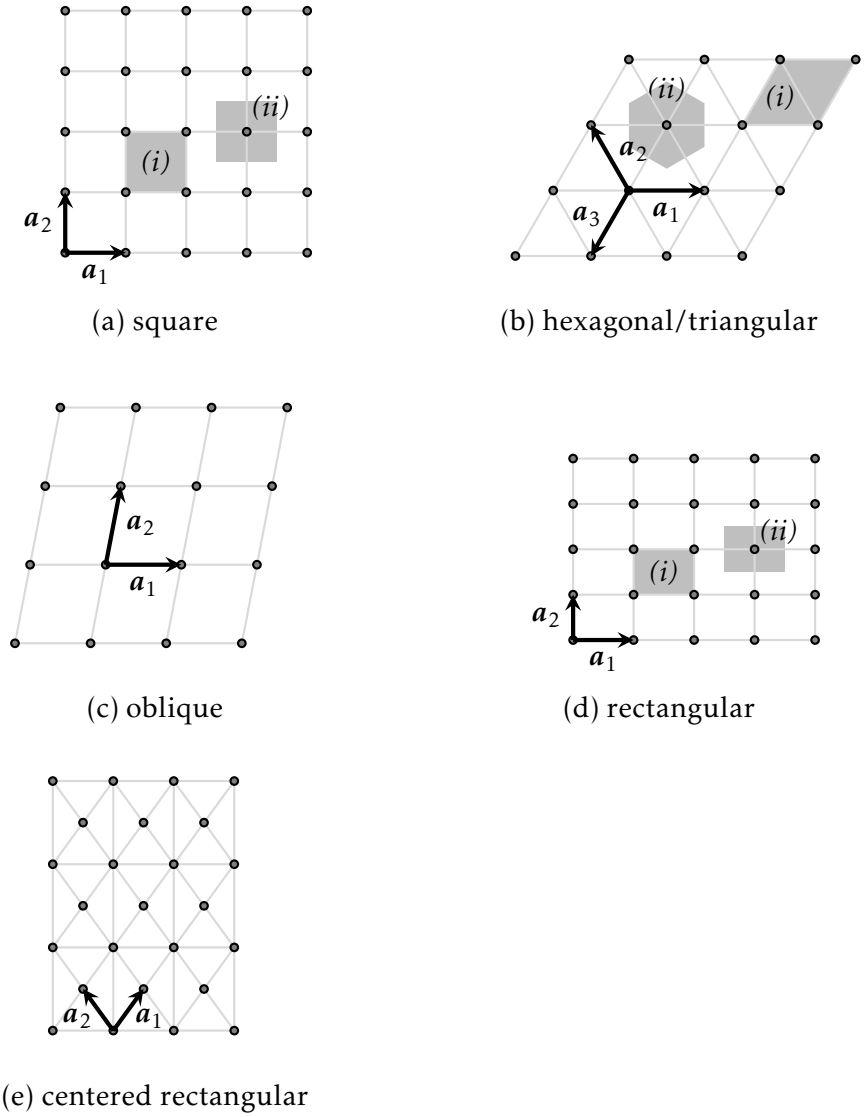
We now turn to the mathematical description of periodic structures and crystals in particular. A periodic solid, also called a crystal, is an infinite arrangement of nuclei which is generated by first taking a finite set of nuclei in fixed positions, and then translating this set by all linear combinations of d linearly independent vectors in d dimensions (bulk materials have $d = 3$, but we will encounter $d = 1, 2$ here). The finite set of atoms is called a *basis*, and the translation vectors are called *Bravais lattice vectors*. To specify a crystal structure, one needs to give the Bravais lattice vectors, $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_d$, and the locations of the atoms in the basis $\boldsymbol{d}_1, \cdots, \boldsymbol{d}_{n_b}$, if n_b atoms are in the basis. Here we should label by Z_i the atomic number of the atomic at site i in the basis. Then the nuclei are located at positions given by

$$\boldsymbol{x}_{i;n_1,\cdots,n_d} = \sum_{\mu=1}^{d} n_\mu \boldsymbol{a}_\mu + \boldsymbol{d}_i, \tag{26}$$

where $n_1 \cdots n_d \in \mathbb{Z}$ and $i = 1 \cdots n_b$. This data is not necessarily unique: different sets of Bravais lattice vectors and basis vectors may describe the same solid. The set of $\boldsymbol{a}_\mu$ are known as *primitive* lattice vectors if they describe the structure with the smallest set of basis vectors possible, i.e. with the minimal possible n_d.

The simplest crystals are those with only one atom in the basis. Then we may by choice of origin take $\boldsymbol{d}_1 = \mathbf{0}$ and every atom is specified by a set of d integers. The set of these points comprises what is called a *Bravais lattice*. In a Bravais lattice all points are equivalent and connected by a lattice translation. The more general solid may be referred to as a Bravais lattice with a basis. For any crystal structure, we can associate a Bravais lattice in this way; i.e. it is the lattice defined by the $\boldsymbol{a}_\mu$ only. One may regard a Bravais lattice also more abstractly as defining the set of translations which leave the crystal invariant, and this is the same association of a Bravais lattice with the crystal structure. If the atoms are at the positions defined by Eq. (26), then a translation by a vector in the Bravais lattice, i.e. an integer linear comination of the $\boldsymbol{a}_\mu$, leaves the set of atomic positions invariant, because it can be compensated by a re-labeling of the integers n_μ.

The unit cell is a useful concept. Thinking of the Bravais lattice as a set of translations that leave a crystal invariant, one can define a unit cell as a compact connected volume (in 3d) or area (in 2d) which, when translated by



(a) square

(b) hexagonal/triangular

(c) oblique

(d) rectangular

(e) centered rectangular

Figure 1: The five two dimensional Bravais lattices. In (b) any two of $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ may serve as primitive lattice vectors. In several panels, two unit cells are shaded and labeled as (i) and (ii).

all vectors in the Bravais lattice, covers all of space, without any overlaps or gaps (except possibly for the boundaries of the cell). A unit cell is not unique, but there are often conventional ones. One simple way to define a unit cell is to introduce coordinates y_μ along the primitive lattice vectors: $\boldsymbol{x} = \sum_\mu y_\mu \boldsymbol{a}_\mu$. One unit cell for this Bravais lattice is defined by the volume with $0 \leq y_\mu < 1$. The shape of this unit cell is a parallelogram in two dimensions or a parallelepiped in three dimensions (marked *(i)* in Fig. 1). In a similar way, one may associate the atoms in crystal structure with a basis with "fractional coordinates", i.e. one can rewrite Eq. (26) in the form

$$\boldsymbol{x}_{i;n_1,\cdots,n_d} = \sum_{\mu=1}^{d} \left(n_\mu + x_{i,\mu}\right) \boldsymbol{a}_\mu, \tag{27}$$

where $x_{i,\mu}$ are fractional coordinates for atom i in the basis, and we can take $0 \leq x_{i,\mu} < 1$ which we can think of as specifying each atom within one unit cell in this form. This is typically how crystal structures are represented in crystallographic databases and in journal papers. From the formula for the volume of a parallelepiped, one can obtain the volume of the unit cell as

$$\mathrm{V}_{\mathrm{u.c.}} = \left|\det\left(\boldsymbol{a}_1 | \cdots | \boldsymbol{a}_d\right)\right|. \tag{28}$$

Another popular choice of unit cell is the Wigner-Seitz cell. This is defined as the set of points which are closer to the origin than to any other point in the Bravais lattice. Except at boundaries, every point in space is closest to one Bravais lattice site, so clearly translating the Wigner-Seitz cell reproduces all of space, and there are no overlaps between these cells. The Wigner-Seitz cell can be constructed geometrically by drawing lines (2d) or planes (3d) which perpendicularly bisect segments connecting the origin to all other Bravais lattice sites (in practice, only a finite number of sites near the origin need be considered), and assigning every which can be accessed from the origin without crossing one of these lines/planes to the Wigner-Seitz cell. The Wigner-Seitz cell is useful because it shows more explicitly the symmetries of the Bravais lattice. For the hexagonal lattice, it is a hexagon, as indicated by *(ii)* in Fig. 1(b).

A *primitive unit cell* is a unit cell with the smallest possible volume, making it a minimal one to describe a periodic solid. Here are a few facts about primitive unit cells. A primitive unit cell is generated from the above procedures if the $\boldsymbol{a}_\mu$ are primitive lattice vectors. Translation of any point within a primitive unit cell by a Bravais lattice vector generates a point which is outside the primitive unit cell. A primitive unit cell contains exactly one exemplar for each atom in the basis.

It is often conventional to describe crystals by non-primitive unit cells. The most common examples are for the face-centered cubic (fcc) and body-centered cubic (bcc) Bravais lattices. As Bravais lattices, the primitive unit cell for each of these structures contains just one site. However, it is conventional to describe these by cubic unit cells, which contain 4 and 2 sites, for the fcc and bcc cases, respectively.

2.2 *Bragg scattering and reciprocal space*

2.2.1 Bragg scattering

How do we know that materials actually have periodic crystal structures? The best evidence comes from x-ray scattering. The periodic structure of a crystal

lattice diffracts electromagnetic radiation at x-ray wavelengths similarly to the way a grating diffracts visible light. The phenomena was discovered by William Henry Bragg and his son Lawrence Bragg in 1913. They won the Nobel prize in physics for the discovery in 1915, and are the only father son team to do so. We now know that any wave-like object can undergo Bragg scattering from a periodic solid, but x-rays predominate because they are readily available and their wavelength is the same order as the inter-atomic distances in solids. The Braggs discussed their finding by interpreting a solid as exhibiting sets of parallel planes upon which x-rays reflect and different reflections interfere with one another constructively for certain angles of reflection. A periodic crystal has in fact an infinite variety of such "Bragg planes", and it is easier and more modern to describe the phenomena mathematically using the concept of reciprocal space, which we will develop below. This formulation of scattering is due to Laue.

The basic idea is as follows, and applies to any type of wave which weakly scatters from the atoms. A plane wave is incident on the sample with a wavevector $\boldsymbol{k}_0$, with amplitude

$$\psi_{\text{incident}}(\boldsymbol{x}) = \mathrm{A}_0 e^{i\boldsymbol{k}_0\cdot\boldsymbol{x}} \tag{29}$$

at position $\boldsymbol{x}$. It scatters elastically, i.e. conserving its energy, off at atom at position $\boldsymbol{x}_i$ to produce an outgoing spherical wave, which is detected far way at position $\boldsymbol{y}$. At long distances the scattered component from this atom behaves as

$$\psi_{\text{scattered from } \boldsymbol{x}_i}(\boldsymbol{y}) = \mathrm{A}_{s,a_i} e^{i\boldsymbol{k}_0\cdot\boldsymbol{x}_i} e^{ik|\boldsymbol{y}-\boldsymbol{x}_i|}, \tag{30}$$

where $k = |\boldsymbol{k}_0|$ because the scattering is elastic (we do not write a power-law prefactor which does not affect the result). Here the first exponential represents the amplitude of the incident wave at the location of the scatterer, and the second is the outgoing spherical wave. The scattered wave has an amplitude A_{s,a_i} which is the same for all atoms of the same type (indicated by a_i) in the crystal. It is convenient to define the outgoing wavevector measured at the detector position by defining the direction $\hat{\boldsymbol{k}}_f = (\boldsymbol{y} - \boldsymbol{x}_i)/|\boldsymbol{y} - \boldsymbol{x}_i|$. When the detector is far away from the sample, this direction $\hat{\boldsymbol{k}}_f$ is approximately the same for all atoms. Then we have

$$\psi_{\text{scattered from } \boldsymbol{x}_i}(\boldsymbol{y}) = \mathrm{A}_{s,a_i} e^{i\boldsymbol{k}_0\cdot\boldsymbol{x}_i} e^{i\boldsymbol{k}_f\cdot(\boldsymbol{y}-\boldsymbol{x}_i)}, \tag{31}$$

where $\boldsymbol{k}_f = k\hat{\boldsymbol{k}}_f$. Now we can add up the contributions from all the atoms to the scattered wave

$$\psi_{\text{scattered}}(\boldsymbol{y}) = \left(\sum_i \mathrm{A}_{s,a_i}\, e^{i(\boldsymbol{k}_0-\boldsymbol{k}_f)\cdot\boldsymbol{x}_i}\right) e^{i\boldsymbol{k}_f\cdot\boldsymbol{y}}. \tag{32}$$

The sum in the parenthesis reflects wave interference, and should be taken over all atoms in the crystal. We can define it as the scattering amplitude

$$\mathrm{A}(\boldsymbol{k}) = \sum_i \mathrm{A}_{s,a_i} e^{i\boldsymbol{k}\cdot\boldsymbol{x}_i}, \tag{33}$$

where the wavevector $\boldsymbol{k} = \boldsymbol{k}_0 - \boldsymbol{k}_f$ is called the scattering wavevector (the final exponential in Eq. (32) outside the parenthesis is an overall phase that does not depend upon the atomic positions and will become trivial in the intensity). The scattering amplitude in Eq. (445) is a general expression for an arbitrary arrangement of scatterers, e.g. in liquids as well as solids, or for non-periodic solids. One can study the resulting scattered intensity using statistical mechanics to average over thermal and quantum fluctuations according to

$$\mathrm{I}(\boldsymbol{k}) \propto \left\langle |\mathrm{A}(\boldsymbol{k})|^2 \right\rangle = \sum_{i,j} \mathrm{A}_{s,a_i} \mathrm{A}^*_{s,a_j} \left\langle e^{i\boldsymbol{k}\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)} \right\rangle. \tag{34}$$

Here the angular brackets denote thermal/quantum averages. This is the *equal time structure factor* for scattering, and expressions like this appear in many contexts.

For now we specialize to ordered atomic arrangements, so the atomic positions may be regarded as fixed and classical, and we do not need to do any averaging. In that case we just consider A_s directly. For each atom in the basis, there is a macroscopic sum of terms involving copies of that atom generated by translation to all other unit cells, which generates interference. Writing for atom of type a that $\boldsymbol{x}_i = \sum_\mu n_\mu \boldsymbol{a}_\mu + \boldsymbol{d}_a$, we have

$$\mathrm{A} = \sum_{i=1}^{n_b} \mathrm{A}_{s,a(i)} \sum_{n_\mu \in \mathbb{Z}} e^{i\boldsymbol{k}\cdot(\sum_\mu n_\mu \boldsymbol{a}_\mu) + \boldsymbol{d}_i)} = \left(\sum_{i=1}^{n_b} \mathrm{A}_{s,a(i)} e^{i\boldsymbol{k}\cdot\boldsymbol{d}_i} \right) \left(\sum_{n_\mu \in \mathbb{Z}} e^{i\boldsymbol{k}\cdot(\sum_\mu n_\mu \boldsymbol{a}_\mu)} \right). \tag{35}$$

Apologies for the notation: in this equation i just denotes a finite sum of the different atoms in the basis (as opposed to in e.g. Eq. (445) where it sums over all atoms in the crystal). Note that the amplitude factors into the finite sum, which is a smooth function of wavevector, and a sum which is infinite in the thermodynamic limit. The latter leads to the Bragg/Laue condition on scattering and very sharp (singular in the thermodynamic limit) dependence of the scattering on wavevector. Specifically, this latter sum contains a huge number of oscillating terms that will destructing interfere unless $\boldsymbol{k}\cdot(\sum_\mu n_\mu \boldsymbol{a}_\mu)$ is an integer multiple of 2π. Since the n_μ are arbitrary integers, this is equivalent to the condition that

$$\boldsymbol{k}\cdot\boldsymbol{a}_\mu \in 2\pi\mathbb{Z}, \tag{36}$$

for each μ. Eq. (36) is called the Laue condition. It can be shown to be

equivalent to Bragg's condition. Quantum mechanically, $\hbar\boldsymbol{k}$ represents the momentum transferred from the light to the sample. The Laue condition represents in three dimensions three linear equations for three unknown components of $\boldsymbol{k}$, for a given choice of integers on the right hand side. So for each choice of integers the solution is just a single point in wavevector/momentum space (taking $\hbar = 1$, we do not distinguish these), often called *reciprocal space*. Repeating for all the integers, one obtains an infinite set of points. They form a lattice in momentum space known as the reciprocal lattice (below we will show it forms a Bravais lattice). When the scattering wavevector coincides with a point of the reciprocal lattice, strong scattering is possible, and one may see a peak in scattered intensity of a detector oriented to collect such scattered x-rays.

However, one should note that we have not actually used any information on the crystal structure beyond the Bravais lattice in obtaining Eq. (36). This is because it came entirely from the second factor in Eq. (35). Hence, the Laue condition only tells us when the scattering from each atom in the basis adds constructively. It is still possible for contributions from different atoms in the unit cell to interfere with one another destructively, so that sometimes the intensity of scattered waves at a reciprocal lattice point vanishes. This is called an "extinction". This arises mathematically from the first factor in Eq. (35):

$$\mathrm{A}^{\mathrm{geom}}_{\boldsymbol{k}} = \sum_{i=1}^{n_b} \mathrm{A}_{s,a(i)} e^{i\boldsymbol{k}\cdot\boldsymbol{d}_i}, \tag{37}$$

which is called the "geometrical structure factor". For a non-Bravais lattice this smooth function may vanish at some reciprocal lattice vectors. In general, the intensity of a Bragg peak is determined by the absolute value squared of this factor.

To summarize: scattered waves appear only at reciprocal lattice vectors, and the relative intensity of different Bragg peaks, i.e. different reciprocal lattice vectors, is calculated through a finite sum of atomic amplitudes. This intensity may vanish for some reciprocal lattice vectors.

2.2.2 Reciprocal lattice

A full set of solutions of the Laue conditions can be obtained as follows. Define a set of d linearly independent basis vectors $\boldsymbol{b}_\mu$ in reciprocal space such that

$$\boldsymbol{b}_\mu \cdot \boldsymbol{a}_\nu = 2\pi\delta_{\mu\nu}. \tag{38}$$

The factor of 2π is one common convention. This completely determines the $\boldsymbol{b}_\mu$ from the $\boldsymbol{a}_\mu$. Indeed one can solve this equation explicitly to obtain, in three dimensions:

$$\boldsymbol{b}_1 = 2\pi \frac{\boldsymbol{a}_2 \times \boldsymbol{a}_3}{\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 \times \boldsymbol{a}_3}, \qquad + \text{cyclic permutations.} \tag{39}$$

Now consider a scattering wavevector which is an integer linear combination of the new vectors,

$$\boldsymbol{k} = \sum_{\mu} m_{\mu} \boldsymbol{b}_{\mu}, \qquad m_{\mu} \in \mathbb{Z}. \tag{40}$$

Then one sees that

$$\boldsymbol{k} \cdot \boldsymbol{a}_{\nu} = \sum_{\mu} m_{\mu} \boldsymbol{b}_{\mu} \cdot \boldsymbol{a}_{\nu} = 2\pi m_{\nu}, \tag{41}$$

which satisfies the Laue condition for every choice of m_{μ}. Hence every wavevector of the form of Eq. (40) exhibits strong constructive scattering. Evidently *the scattering wavevectors form a Bravais lattce in momentum space*, known as the *reciprocal lattice*. Every point in the reciprocal lattice defines a momentum which can be transferred to or from a wave scattering elastically from the crystal lattice. While we have discussed x-rays, this is also true for neutrons, and even electrons. The idea of Bragg scattering of electrons will lead us soon to band theory.

What do you need to know about the reciprocal lattice?

- The reciprocal lattice is a Bravais lattice, so we can apply all the concepts we developed above to it.
- The scale of reciprocal and real Bravais lattices are reversed: if the real space unit cell gets larger, the unit cell of the reciprocal lattice gets smaller. Specifically, using Eq. (444) and the expression in Eq. (28), the volume of the unit cell of the reciprocal lattice times the volume of the (primitive) unit cell of the Bravais lattice is $(2\pi)^3$ (it generalizes to $(2\pi)^d$ in d dimensions):
$$V_{\mathrm{BZ}} V_{\mathrm{p.u.c.}} = (2\pi)^d. \tag{42}$$
- The reciprocal lattice has the same symmetries as the Bravais lattice. We can just as well identify point symmetry groups, i.e. Wyckoff positions, with locations within the unit cell of the reciprocal lattice. These points are conventionally indicated by capital letters (including Greek ones). Unlike in real space (there is no natural origin in real space because we can always just translate the whole crystal), there is a natural origin in reciprocal space: the point $\boldsymbol{k} = 0$, or zero momentum, has the highest symmetry, and is often called the Γ point.
- The Wigner-Seitz unit cell of the reciprocal lattice, centered on the Γ point, is known as the **Brillouin zone**, or sometimes the first Brillouin zone. It will play a central role in the discussion of electron states.

2.3 *Symmetries of crystals*

2.3.1 Symmetry operations and their composition

Bravais lattices, and crystal structures in general, can be categorized using the powerful theoretical apparatus of symmetry. This is the subject of crystallography. Symmetry considerations are very useful in solids because they can also be applied to the electrons and to dynamical excitations of the lattice. Recall the basic structure of symmetries in physics. A symmetry g is one which leaves the system invariant. For a crystal structure, it consists of a *space group operation*, which is a linear coordinate transformation which keeps the metric invariant and preserves the configurations of the atoms. One way to view this is that one starts with the full set of symmetries of free space, which are Euclidean transformations, and selects the subset which leave the crystal invariant. The most general space group operation g is the combination of a translation and an O(d) orthogonal transformation, i.e. a map

$$g : \boldsymbol{x} \to \mathsf{O}_g \boldsymbol{x} + \boldsymbol{t}_g, \tag{43}$$

where $\boldsymbol{t}$ is a translation vector, and O is a d×d orthogonal matrix, $\mathsf{O}^{\mathrm{T}}\mathsf{O} = 1_d$ in d dimensions. The orthogonal matrix can describe a rigid rotation, or a reflection, or an inversion. Any operation of the form of Eq. (43) is a symmetry of free space, and some subset of these, i.e. choices of O, $\boldsymbol{t}$, are symmetries of a given crystal. An operation g with $\mathsf{O}_g, \boldsymbol{t}_g$ is a symmetry of a crystal if it maps the *entire set* of atomic positions of a given atomic number back to itself.

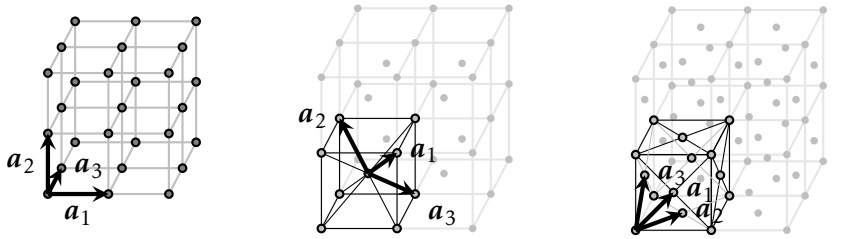
Symmetries have a group structure. Two operations may be composed, and we write symbolically $g = g_2 g_1$ to represent the operation g obtained by first performing g_1 and then g_2. We see that under g:

$$g : \boldsymbol{x} \to_{g_1} \mathsf{O}_{g_1}\boldsymbol{x} + \boldsymbol{t}_{g_1} \to_{g_2} \mathsf{O}_{g_2}\left(\mathsf{O}_{g_1}\boldsymbol{x} + \boldsymbol{t}_{g_1}\right) + \boldsymbol{t}_{g_2} = \mathsf{O}_{g_2}\mathsf{O}_{g_1}\boldsymbol{x} + \mathsf{O}_{g_2}\boldsymbol{t}_1 + \boldsymbol{t}_2, \tag{44}$$

which has the same form as Eq. (44), showing that a group structure exists, and one can then read off the specific way in which the elements compose:

$$\mathsf{O}_g = \mathsf{O}_{g_2}\mathsf{O}_{g_1}, \qquad \boldsymbol{t}_g = \mathsf{O}_{g_2}\boldsymbol{t}_1 + \boldsymbol{t}_2. \tag{45}$$

For a periodic lattice, by definition an infinite set of pure translations, i.e. g with $\mathsf{O}_g = 1_d$, are symmetries, where the translation vectors $\boldsymbol{t}_g$ are integer linear combinations of the three primitive vectors $\boldsymbol{a}_\mu$ in Eq. (26). There are generically also pure "point group" operations in which $\boldsymbol{t}_g = \boldsymbol{0}$. Finally there are operations in which both O_g and $\boldsymbol{t}_g$ are non-trivial. These can exist because of the group composition law, i.e. one can simply apply a translation and a point group operation in sequence, if both separately are symmetries. There are also sometimes operations which cannot be built in that way, i.e. for which the O_g and $\boldsymbol{t}_g$ are not symmetries when applied separately as pure point or translation operations. Examples are "screw axes" and "glide planes".



(a) cubic (b) body-centered cubic (c) face-centered cubic

Figure 2: The three dimensional Bravais lattices in the cubic family

2.3.2 Classification of Bravais lattices

One can use symmetries to categorize many objects in crystallography. Perhaps the simplest is to discuss Bravais lattices. A general Bravais lattice in d dimensions is just parametrized by a set of d linearly independent d-component vectors $\boldsymbol{a}_\mu$, which is equivalent to a d-dimensional matrix with non-vanishing determinant, and moreover we can take this determinant to be always positive by taking $n_1 \to -n_1$ if necessary. Obviously this forms a continuous and completely connected set, so in this sense all Bravais lattices are deformable to one another. However, we can break them into symmetry classes, by declaring two Bravais lattices equivalent if and only if they can be deformed into one another smoothly without changing their symmetry. With this standard definition, it turns out that there are 5 Bravais lattices in two dimensions: oblique, rectangular, centered-rectangular, hexagonal, and square. A useful (but partial) understanding of the symmetries of the Bravais lattices is through their point symmetry: the set of all orthogonal coordinate transformations that leave a site, e.g. the origin, invariant, forms the *point group* of the Bravais lattice. Point groups can be defined more generally for any point in space within a crystal. All such crystallographic point groups are finite groups, and there are finitely many of them. For the 5 Bravais lattices listed above, the corresponding point groups are C_2, D_2, D_2, D_6, and D_4, respectively, in so-called Schönflies notation. The group C_n is the group of all n-fold in-plane rotations, i.e. rotations by $2\pi/n$ radians. The group D_n contains these rotations *and* n two-fold rotations about axis in the plane (within the plane these appear like mirror reflections across lines). Notice that the rectangular and centered rectangular Bravais lattices share the same point group, but they are distinct structures and cannot be transformed into one another without modifying the symmetry. The centered rectangular lattice also contains some non-point symmetries like glide planes which are not present in the rectangular structure, so point symmetry is not the full story. Sometimes people talk about "crystal classes" or "crystal families", in which several Bravais lattices are unified; I do not quite understand the nomenclature, but according to wikipedia there are 4 cystal families in 2d, because the rectangular and centered-rectangular Bravais lattices are placed in the same "orthorhombic" family, presumably because they share the same point group D_2.

The hexagonal and square lattices have the highest symmetries, but they are distinct; for example, the hexagonal lattice has a C_6 symmetry, while the square lattice has a C_4 symmetry. Stretching the hexagonal lattice along one direction converts it to centered rectangular, which has lowered symmetry compared to the hexagonal lattice, for example, the centered rectangular lattice has only a C_2 axis. One can also deform the square lattice into a centered rectangular one by shearing it, but during the process of shearing the symmetry will fall generally into the oblique case.

In three dimensions, there are 14 Bravais lattices, and 7 lattice systems, i.e. 7 different point groups that arise. I will not list them all but you can find a list on wikipedia. For example, there are three Bravais lattices in the cubic family/crystal system, which are illustrated in Fig. 2. They are the simple cubic, the face-centered cubic (fcc) and body-centered cubic (bcc) lattices. The fcc lattice is actually a quite common structure for pure elemental solids.

2.3.3 Space groups

Most materials have structures with a basis, which have symmetries that are different from the Bravais lattices. A complete description of their symmetries is the space group. This is the full group of transformations in Eq. (43) discussed above. Because the set of allowed translations is infinite, a space group necessarily has an infinite number of group elements. It is however a countably infinite and discrete group. Intuitively, it is good to abstract these symmetry groups a bit from the explicit formulation in terms of O_g and $\boldsymbol{t}_g$ parameters. This is because the latter are not independent of trivial factors such as the choice of coordinate origin or axes, or of trivial rescaling like a change of the overall lattice constant. Such changes do indeed not change the group multiplication relations, so do not affect the space group as an abstract group. Hence, we should regard two different parametrizations obtained by different such choices as defining the same space group. With this convention (please see original references like the international tables of crystallography for a more precise definition) all the crystallographic space groups have been found and their properties tabulated. There are 17 two dimensional space groups, also known as *wallpaper groups*, and 230 space groups in three dimensions. There are many useful online resources for space group data. The traditional reference are the International Tables for Crystallography, which may be found online. Another very useful reference is the Bilbao Crystallographic Server. Wikipedia also has articles on space groups etc.

One could spend forever on space groups. What is actually useful to know? For a given material, one can usually find an experimentally or computationally determined crystal structure, which identifies the space group by its number, gives the lattice constant(s), and the locations of all the atoms in the unit cell. For each space group, there are conventional choices of the Bravais lattice vectors, and conventional definitions of the lattice constants that define the length scale(s). Then the atomic locations are defined by "atomic posi-

tions", which are given by fractional coordinates, i.e. the $x_{i,\mu}$ in Eq. (27). When the atomic positions are given, atoms which are present in multiple places within a unit cell which are related by symmetries are specified only once. This is done in a standardized way, and you can find all this data tabulated in papers reporting crystal structures. There is also a standard file format known as a ".cif" file (crystallographic information file) which can be read by various software programs to display or otherwise use crystal structures. I recommend the free software vesta, which is useful for visualizing crystal structures.

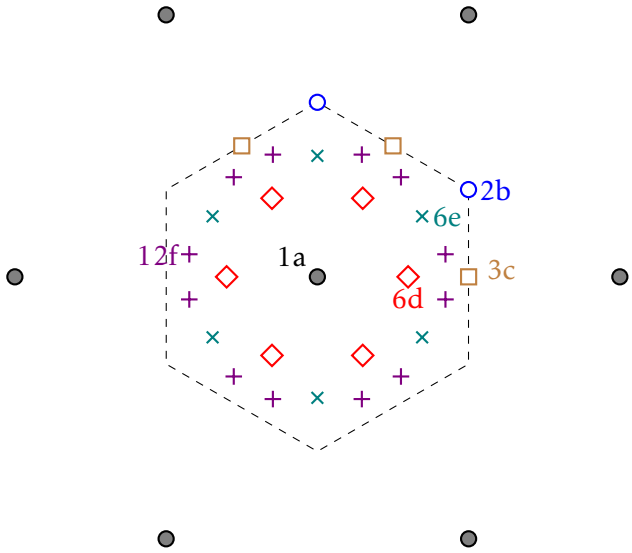


Figure 3: A Kaleidoscope of Wyckoff positions for wallpaper group 17, P6mm, which is the symmetry group of the triangular/hexagonal lattice. Some of the Bravais lattice points of the latter are shown as gray filled circles. There are 6 sets of equivalent Wyckoff positions 1a,2b,3c,6d,6e, and 12f, which are shown within the Wigner-Seitz cell *indicated by the dashed hexagon. Note that the 6d and 6e positions are free to move radially, and the 12f position has full freedom of movement, so long as it does not become one of the other positions. Each set of equivalent positions is found by taking one single point and acting on it with* C_6 *rotations about the origin, and reflections across lines passing through the origin at angles that are multiples of 30 degrees from the x axis.*

To understand the way atomic positions are presented in such tables or .cif files, as well as for many symmetry analyses, it is useful to introduce the concept of a *Wyckoff position*. This works as follows. Pick a point inside the unit cell. Under each symmetry operation of the crystal, this point may either remain fixed, or transform to another point, which by suitable translation can be brought back into the original unit cell. For any given starting point, a finite number of points within the unit cell are generated; the number of such points generated is called the degeneracy of the point. These points are equivalent positions, and if there is an atom at one such position we know the same atom must be present at all of them, hence we need specify

only one. Furthermore, all points within the unit cell may be separated into different families according to the set of symmetry operations which leave that point invariant. This is a subgroup of the space group and for a given point is called the site symmetry group of that point. All points within the unit cell can be divided into classes, within which each point has the same site symmetry group and the same degeneracy, and every set of points within each class can be smoothly transformed into every other set. Each class is known as a Wyckoff position. There are a finite set of such Wyckoff positions for each space group, and you can find them tabulated as well in the resources listed above. They have standardized names generally specified by a letter. Those with the highest symmetry have the lowest degeneracy, and they may be fixed at isolated locations in the unit cell. Positions with lower symmetry have higher degeneracy, and these points can "slide" within the unit cell with changing their Wyckoff class. The most general position has no special symmetry, i.e. a trivial point group, and has the largest degeneracy. Please see Figure 3 for an example. We will use the concept of special symmetry positions later in the class.

3 PHONONS

We understand from the previous section that in the Born-Oppenheimer approximation, the (free) energy of a solid is in principle expressed as a function of the coordinates of all the atoms, and that the minimum (global, or at least a deep local one) of this function is often a periodic crystal. To a first approximation, we can regard the atoms as fixed at this minimum energy configuration. But what happens beyond this? The atoms can of course move away from these positions, and this can and does occur for many reasons. It occurs in equilibrium via thermal fluctuations. It occurs even at absolute zero temperature due to quantum zero point motion. Excitation of the electrons away from their ground state will generally affect the atomic positions. The solid may also be distorted as a response to external or internal stresses.

In this section, we will content ourselves with *small* deviations of the atoms from their minimum energy positions. Such deviations in a periodic crystal give rise to quantized modes of lattice vibrations called *phonons*.

3.1 *A one dimensional chain*

As a warm up, let us consider a one dimensional chain of atoms. We assume the atoms are all identical and occupy some equilibrium positions $X_n = na$, with integer n, and their actual position x_n can be slightly displaced from this by an amount u_n:

$$x_n = X_n + u_n. \tag{46}$$

Now we imagine the atoms are connected by springs with equilibrium length a so the Hamiltonian becomes

$$\mathrm{H} = \sum_{n=0}^{\mathrm{N}} \left[\frac{p_n^2}{2\mathrm{M}} + \frac{\mathrm{K}}{2} (u_{n+1} - u_n)^2 \right]. \tag{47}$$

Here because u_n differs from x_n just by a constant shift, we have $[u_n, p_{n'}] = i\delta_{n,n'}$. The spring potential energy depends only on the difference of the two neighboring displacements, because of course a uniform shift of all the atoms does not change the energy. For simplicity we can put our atoms on a ring, and apply periodic boundary conditions, $u_{\mathrm{N}} = u_0$, $p_{\mathrm{N}} = p_0$. To diagonalize Eq. (47) we first make make a Fourier transformation,

$$\begin{aligned} u_n &= \frac{1}{\sqrt{\mathrm{N}}} \sum_k e^{ik\mathrm{X}_n} \tilde{u}_k, \\ p_n &= \frac{1}{\sqrt{\mathrm{N}}} \sum_k e^{ik\mathrm{X}_n} \tilde{p}_k, \end{aligned} \tag{48}$$

where here we sum over $k = 0, \frac{2\pi}{\mathrm{L}}, \cdots, \frac{2\pi}{\mathrm{L}}(\mathrm{N}-1)$, with $\mathrm{L} = \mathrm{N}a$. The inverse Fourier transform is

$$\begin{aligned} \tilde{u}_k &= \frac{1}{\sqrt{\mathrm{N}}} \sum_n e^{-ik\mathrm{X}_n} u_n, \\ \tilde{p}_k &= \frac{1}{\sqrt{\mathrm{N}}} \sum_n e^{-ik\mathrm{X}_n} p_n, \end{aligned} \tag{49}$$

from which one deduces the commutation relations

$$[\tilde{u}_k, \tilde{p}_{k'}] = i\delta_{k,-k'}, \tag{50}$$

where the delta-function is to be understood as periodic in $k \to k + 2\pi/a$. The Hamiltonian becomes

$$\mathrm{H} = \sum_k \left[\frac{\tilde{p}_k \tilde{p}_{-k}}{2\mathrm{M}} + \mathrm{K}(1 - \cos ka)\, \tilde{u}_k \tilde{u}_{-k} \right]. \tag{51}$$

Now we have decoupled the system into pairs of $k, -k$ which remain coupled. This is true except for $k = 0$ where the potential vanishes and we just have a free mode $\mathrm{H}_0 = \tilde{p}_0^2/(2\mathrm{M})$, which describes the motion of the center of mass (note $\tilde{p}_0 = \frac{1}{\sqrt{\mathrm{N}}} \sum_n p_n = p_{\mathrm{tot}}/\sqrt{\mathrm{N}}$). This is close enough to a simple harmonic oscillator we can guess how to finish diagonalizing it, defining (for $k \neq 0$)

$$a_k = \frac{1}{\sqrt{2}} \left(f(k)\tilde{u}_k + \frac{i}{f(k)} \tilde{p}_k \right), \tag{52}$$

where $f(k) = (2\mathrm{MK}(1 - \cos ka))^{1/4}$. This is canonical, so that using Eq. (50) we obtain

$$\left[a_k, a_{k'}^\dagger\right] = \delta_{k,k'}. \tag{53}$$

Using $\tilde{u}_k^\dagger = \tilde{u}_{-k}$ and $\tilde{p}_k^\dagger = \tilde{p}_{-k}$, one can invert this to obtain

$$\tilde{u}_k = \frac{a_k + a_{-k}^\dagger}{\sqrt{2} f(k)}, \qquad \tilde{p}_k = f(k)\frac{a_k - a_{-k}^\dagger}{\sqrt{2} i}. \tag{54}$$

Then inserting into the Hamiltonian and collecting terms, we obtain

$$\mathrm{H} = \frac{\tilde{p}_0^2}{2\mathrm{M}} + \sum_{k\neq 0} \omega(k)\left(a_k^\dagger a_k + \frac{1}{2}\right), \tag{55}$$

where

$$\omega(k) = \sqrt{\frac{2\mathrm{K}(1-\cos ka)}{\mathrm{M}}}. \tag{56}$$

We see that there are N − 1 normal modes for the N atoms, plus the center of mass degree of freedom (which is very heavy and can be usually neglected). The modes have the same dispersion as a classical chain. There are a few features to note. The maximum frequency $\omega_{\max} = 2\sqrt{\mathrm{K}/\mathrm{M}}$. For small wavevector, $ka \ll 1$, the dispersion becomes linear, $\omega(k) \approx v|k|$, with $v = \sqrt{\mathrm{K}/\mathrm{M}}a$. This represents a "sound" wave of compression along the chain, and because the dispersion $\omega(k)$ approaches zero in this way for small k it is called an "acoustic" mode.

In this example, we obtained a single branch of phonon modes because we had just one degree of freedom, the displacement of the one atom along the chain, per unit cell. In systems with more atoms per unit cell, each atom can move independently in each dimension, so in general there are $n_b \times d$ branches of modes. We will see that in general d of them are acoustic, and the remainder have frequencies which remain non-zero as $k \to 0$. Those are called "optical" modes.

3.2 *Energy scales for phonons*

It's good to have an idea of the quantitative scales that appear for typical phonons. To understand that, we need estimates for K, M and a. The mass M is easy, just the mass of an atom, which of course varies but would be estimated as $\mathrm{M} \approx 2Zm_p$, where m_p is the proton mass and Z is the atomic number. The length scale a is typically of order an Ångstrom for most solids. The spring constant K is the trickier one. It should be though to arise from the potential energy of the solid, which in turn is just the Born-Oppenheimer energy of the electronic problem, $\mathrm{E}_0(\{\mathbf{X}_i\})$. In that problem, the typical energy scales are those of the hydrogen atom, i.e. Rydbergs, and the typical length scales are of order the Bohr radius a_0. So we can expect crudely on dimensional grounds that $\mathrm{K} \sim \mathrm{R}_\infty/a_0^2$. Hence we can estimate the maximum phonon energy as

$$\omega_{\max} \sim \sqrt{\frac{K}{M}} \sim \sqrt{\frac{R_\infty}{2Zm_p a_0^2}} \sim \sqrt{\frac{m_e}{2Zm_p}} R_\infty, \tag{57}$$

using the fact that $R_\infty \sim 1/(m_e a_0^2)$. Now the proton is about 2000 times heavier than the electron, and for typical solids we can take $Z \sim 20$ or so. Then the factor $\sqrt{\frac{m_e}{2Zm_p}} \sim 1/300$ or so. This will give a value $\omega_{\max} \sim 600$K. The estimate of a Rydberg is a bit high: more often the basic scale is just a few eV rather than 13eV (probably this is due to screening). Then a typical frequency is a bit lower than our estimate, in the range of 200K or so. Clearly the estimates are crude, and one can envision that a more careful and complete treatment will lead to a range of numbers that depend upon the details of the structure. Regardless, this sort of number is a good rule of thumb: the typical range of energies of phonons corresponds to temperatures comparable to room temperature. This is much smaller than the typical energies of electronic excitations, which are as we already indicated in the eV range. Usually this characteristic phonon scale is referred to as the *Debye temperature* (or Debye frequency, Debye energy, depending upon the units we choose). See the discussion in Sec. 3.7 about phonons and thermodynamics.

Using this estimate for the Debye frequency, one can estimate a typical acoustic velocity $v \sim a\omega_{\max} \sim 1\text{Å} \times (200\text{K} \times \frac{k_B}{\hbar}) \sim 2600 m/s$. This is actually in the correct range for sound speed in solids (look it up!). It is much faster than in air.

3.3 *Atomic displacements*

Now we'll present the general treatment. Let us assume the minimum of the configurational energy is realized with a periodic structure consisting of n_b atoms per primitive unit cell. We consider a small displacement of each atom in an arbitrary direction, modifying Eq. (26) to

$$\boldsymbol{x}_{i;n_1,\cdots,n_d} = \sum_{\mu=1}^{d} n_\mu \boldsymbol{a}_\mu + \boldsymbol{d}_i + \boldsymbol{u}_{i,n_1,\cdots,n_d}, \tag{58}$$

where $\boldsymbol{u}_{i,n_1,\cdots,n_d}$ is the vector displacement of the i^{th} atom in the unit cell indexed by $n_1, \cdots, n_d$. It is convenient to trade all these unit cell indices for just the location of the center $\mathbf{R} = \sum_{\mu=1}^{d} n_\mu \boldsymbol{a}_\mu$ of each unit cell. Then we have

$$\boldsymbol{x}_i(\mathbf{R}) = \mathbf{R} + \boldsymbol{d}_i + \boldsymbol{u}_i(\mathbf{R}). \tag{59}$$

A correct count of the number of atoms and their displacements is obtained by including each distinct periodic Bravais lattice position $\mathbf{R}$ once, and again $i = 1, \ldots, n_b$. Note that the total number of degrees of freedom (i.e. real

variables needed to specify the position of all the atoms) is the dimensionality d times the number of atoms in the basis n_b times the number of unit cells.

3.4 *Expansion of the energy*

In the Born-Oppenheimer approximation, the energy is a function of all the atomic coordinates, and by assumption it is minimized when $\boldsymbol{u}_i(\mathbf{R}) = 0$. Assuming further that it is analytic around this minimum (a generally correct assumption), a Taylor expansion of the BO energy begins at quadratic order,

$$\begin{aligned} \mathrm{E}_{\mathrm{BO}}[\{\boldsymbol{u}_i(\mathbf{R})\}] &\equiv \mathrm{E}_{\mathrm{BO}}[\{\mathbf{R} + \boldsymbol{d}_i + \boldsymbol{u}_i(\mathbf{R})\}] - \mathrm{E}_{\mathrm{BO}}[\{\mathbf{R} + \boldsymbol{d}_i\}] \\ &= \frac{1}{2} \sum_{ij} \sum_{\mathbf{R},\mathbf{R}'} \sum_{\mu\nu} \mathrm{V}_{ij;\mu\nu}(\mathbf{R}, \mathbf{R}') u_i^{\mu}(\mathbf{R}) u_j^{\nu}(\mathbf{R}') + \mathrm{V}_{\mathrm{ah}}[\{\boldsymbol{u}_i(\mathbf{R})\}], \end{aligned} \tag{60}$$

where $\mathrm{V}_{\mathrm{ah}}[\{\boldsymbol{u}_i(\mathbf{R})\}]$ contains anharmonic terms beginning at third order.

For many purposes the quadratic term in the BO energy suffices. It is parametrized by the tensor $\mathrm{V}_{ij;\mu\nu}(\mathbf{R}, \mathbf{R}')$, which in general takes $d^2 n_b^2 \mathrm{N}_{\mathrm{u.c.}}^2$ values (including all values of $i, j, \mu, \nu, \mathbf{R}, \mathbf{R}'$). We can simplify this for the case of crystals by using the translational invariance of the solid. In particular, if all atoms in the crystal are shifted over to the neigboring unit cell, the energy must be unchanged. This means that one can translate the labels $\mathbf{R} \to \mathbf{R} + \boldsymbol{a}, \mathbf{R}' \to \mathbf{R}' + \boldsymbol{a}$, where $\boldsymbol{a}$ is an arbitrary Bravais lattice vector. In turn it means that $\mathrm{V}_{ij;\mu\nu}(\mathbf{R}, \mathbf{R}') = \mathrm{V}_{ij;\mu\nu}(\mathbf{R} - \mathbf{R}')$ is a function of $\mathbf{R} - \mathbf{R}'$ only. This reduces the number of parameters in the quadratic energy function by a factor of $\mathrm{N}_{\mathrm{u.c.}}$. Hence

$$\mathrm{E}_{\mathrm{BO}}[\{\boldsymbol{u}_i(\mathbf{R})\}] = \frac{1}{2} \sum_{ij} \sum_{\mathbf{R},\mathbf{R}'} \sum_{\mu\nu} \mathrm{V}_{ij;\mu\nu}(\mathbf{R} - \mathbf{R}') u_i^{\mu}(\mathbf{R}) u_j^{\nu}(\mathbf{R}') + \mathrm{V}_{\mathrm{ah}}[\{\boldsymbol{u}_i(\mathbf{R})\}]. \tag{61}$$

The starting point for the treatment of lattice vibrations is to truncate Eq. (61) to quadratic order and include the kinetic energy of the atoms,

$$\begin{aligned} \mathrm{H}_{\mathrm{latt}} &= \hat{\mathrm{T}}_{\mathrm{atoms}} + \mathrm{E}_{\mathrm{BO}}^{(2)}[\{\boldsymbol{u}_i(\mathbf{R})\}] \\ &= \sum_i \sum_{\mathbf{R}} \frac{\mathbf{P}_i^2(\mathbf{R})}{2\mathrm{M}_i} + \frac{1}{2} \sum_{ij} \sum_{\mathbf{R},\mathbf{R}'} \sum_{\mu\nu} \mathrm{V}_{ij;\mu\nu}(\mathbf{R} - \mathbf{R}') u_i^{\mu}(\mathbf{R}) u_j^{\nu}(\mathbf{R}'). \end{aligned} \tag{62}$$

Notice that since we label the atoms by the discrete index i and the unit cell label $\mathbf{R}$, the momenta also have the same labels.

There is an important constraint on the couplings $\mathrm{V}_{ij;\mu\nu}(\mathbf{R})$ due to translational invariance. This might seem strange: didn't we already use translational symmetry to conclude V was a function of $\mathbf{R} - \mathbf{R}'$? It's a bit subtle, but what we used so far was invariance under discrete translations of $\mathbf{R}$. This does not actually translate the solid, but rather permutes atoms. Instead we can make a *continuous* translation of the solid by letting $u_i^{\mu}(\mathbf{R}) \to u_i^{\mu}(\mathbf{R}) + \delta^{\mu}$, where $\boldsymbol{\delta}$ is

the translation vector. Note that for a rigid translation of the solid, the vector δ is the same for all atoms. For such an atomic motion, there is no restoring force. Therefore we expect that there must be a mode that has zero oscillation frequency. We will indeed find this is the case.

Mathematically, for the energy to be invariant under a translation by an arbitrary δ, we need to have

$$\sum_{\mathbf{R}} \sum_{j} \mathrm{V}_{ij;\mu\nu}(\mathbf{R}) = 0. \tag{63}$$

The energy should also be invariant under a rigid rotation of the solid. An infinitesimal rotation by angle θ around the α axis takes $x_\mu \to x_\mu + \theta\,\epsilon_{\alpha\mu\nu}x_\nu$. Applying this to the solid corresponds to taking $u_i^\mu(\mathbf{R}) \to u_i^\mu(\mathbf{R}) + \theta\,\epsilon_{\alpha\mu\nu}(\mathrm{R}_\nu + d_{i,\nu})$. Independence of the energy for arbitrary small θ implies that

$$\sum_{\mathbf{R}} \sum_{j} \mathrm{V}_{ij;\mu\nu}(\mathbf{R})\epsilon_{\alpha\nu\gamma}(d_{j\gamma} - \mathrm{R}_\gamma) = 0. \tag{64}$$

This must hold for arbitrary α, μ, ν and i. The conditions in Eqs. (63,64) will be important.

3.5 *Normal modes*

The Hamiltonian in Eq. (62) is quadratic, so it must be equivalent to a set of harmonic oscillators. A relatively quick way to get the frequencies of these oscillators is to look at the classical equations of motion. These equations are valid also in the quantum case by Ehrenfest's theorem. Hamilton's equations, recalling that $\boldsymbol{u}_i(\mathbf{R})$ and $\mathbf{P}_i(\mathbf{R})$ are conjugate variables, are

$$\partial_t u_i^\mu(\mathbf{R}) = \frac{\partial \mathrm{H}}{\partial \mathrm{P}_i^\mu(\mathbf{R})} = \frac{1}{\mathrm{M}_i}\mathrm{P}_i^\mu(\mathbf{R}), \tag{65}$$

$$\partial_t \mathrm{P}_i^\mu(\mathbf{R}) = -\frac{\partial \mathrm{H}}{\partial u_i^\mu(\mathbf{R})} = -\sum_{j,\nu}\sum_{\mathbf{R}'} \mathrm{V}_{ij;\mu\nu}(\mathbf{R}-\mathbf{R}')u_j^\nu(\mathbf{R}'). \tag{66}$$

Taking another time derivative of the first equation and multiplying by M_i, one can insert the second equation to obtain

$$\mathrm{M}_i\partial_t^2 u_i^\mu(\mathbf{R}) = -\sum_{j,\nu}\sum_{\mathbf{R}'} \mathrm{V}_{ij;\mu\nu}(\mathbf{R}-\mathbf{R}')u_j^\nu(\mathbf{R}'). \tag{67}$$

Now we can seek solutions of the plane wave form

$$u_i^\mu(\mathbf{R},t) = \tilde{u}_i^\mu e^{i\boldsymbol{k}\cdot\mathbf{R}-i\omega t}, \tag{68}$$

which is a solution provided

$$\mathrm{M}_i \omega^2 \tilde{u}_i^\mu = \sum_{j,\nu} \tilde{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) \tilde{u}_j^\nu, \tag{69}$$

with

$$\tilde{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) = \sum_{\mathbf{R}} \mathrm{V}_{ij;\mu\nu}(\mathbf{R}) e^{-i\boldsymbol{k}\cdot\mathbf{R}}. \tag{70}$$

Eq. (69) defines a system of $n_b \times d$ homogeneous linear equations (for each $\boldsymbol{k}$) which only have solutions for special values of the frequencies, the normal mode frequencies $\omega_n(\boldsymbol{k})$, with $n = 1 \ldots d n_b$. The number of branches of normal modes corresponds to the number of translational degrees of freedom in a unit cell.

If we want to get detailed, Eq. (69) doesn't quite look like a usual eigenvalue problem because of the M_i factors. To bring it to standard form, let $\tilde{u}_i^\mu = \varepsilon_{i\mu}/\sqrt{\mathrm{M}_i}$. Then one obtains

$$\sum_{j,\nu} \hat{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) \varepsilon_{j\nu}^{(n)} = \omega_n^2 \varepsilon_{i\mu}^{(n)}, \tag{71}$$

where

$$\hat{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) = \frac{1}{\sqrt{\mathrm{M}_i}} \tilde{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) \frac{1}{\sqrt{\mathrm{M}_j}}. \tag{72}$$

We included in Eq. (71) the index $n = 1 \ldots d n_b$.

Note that from Eq. (62) the potential can be chosen symmetric, $\mathrm{V}_{ij;\mu\nu}(\mathbf{R}) = \mathrm{V}_{ji;\nu\mu}(-\mathbf{R})$ which leads to $\tilde{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) = \left[\tilde{\mathrm{V}}_{ji;\nu\mu}(\boldsymbol{k})\right]^*$ and also $\hat{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k}) = \left[\hat{\mathrm{V}}_{ji;\nu\mu}(\boldsymbol{k})\right]^*$. The latter means that, regarded as matrices, $\tilde{\mathrm{V}}(\boldsymbol{k})$ and $\hat{\mathrm{V}}(\boldsymbol{k})$ are Hermitian. This means the eigenvalue problem in Eq. (71) has real eigenvalues ω_n^2, and that the eigenvectors can be taken orthonormal,

$$\sum_{i\mu} \bar{\varepsilon}_{i\mu}^{(m)}(\boldsymbol{k}) \varepsilon_{i\mu}^{(n)}(\boldsymbol{k}) = \delta_{mn}, \tag{73}$$

where $\bar{\varepsilon}_{i\mu}^{(m)}(\boldsymbol{k}) = \left(\varepsilon_{i\mu}^{(n)}(\boldsymbol{k})\right)^*$. They also obey the completeness relation (resolution of the identity),

$$\sum_{n} \bar{\varepsilon}_{j\nu}^{(n)}(\boldsymbol{k}) \varepsilon_{i\mu}^{(n)}(\boldsymbol{k}) = \delta_{ij}\delta_{\mu\nu}. \tag{74}$$

The positivity of the eigenvalues of $\hat{\mathrm{V}}$ is actually a stability requirement, i.e. all the eigenvalues of $\hat{\mathrm{V}}$ should be positive semi-definite (greater than or equal to zero) if the undeformed lattice is a local minimum.

What can we say about these modes? One important constraint comes from

Eq. (63). Comparing to Eq. (70), we can see that implies that $\sum_j \tilde{V}_{ij;\mu\nu}(\boldsymbol{k} = \mathbf{0}) = 0$. One can see then that

$$\tilde{u}_i^\mu = v^\mu, \qquad \leftrightarrow \qquad \varepsilon_{i\mu} = \frac{1}{\sqrt{M_i}} v^\mu, \tag{75}$$

solves the eigenvalue problem at $\boldsymbol{k} = \mathbf{0}$ with zero frequency, $\omega = 0$, for any vector v^μ. There are d linearly independent components of v^μ, which implies that there are d branches of modes whose frequency vanishes as the wavevector vanishes. Generally for these modes the frequency tends to zero linearly in $|k|$ (though it can depend upon direction), i.e. $\omega_n \sim v(\hat{\boldsymbol{k}})|k|$, so these are called *acoustic modes*, by analogy with the dispersion relation of sound. The remaining $(n_b - 1)d$ modes have frequencies which are non-zero at $\boldsymbol{k} = 0$, and these are known as *optical modes*. This is because it is these phonons that are most easily observed by optical spectroscopy, as we'll explain later.

3.6 *Quantization*

We can expect that each normal mode at each wavevector $\boldsymbol{k}$ describes a harmonic oscillator, and should be promoted to a quantum harmonic oscillator, i.e. corresponds to a set of equally spaced levels with energy spacing $\hbar\omega_n(\boldsymbol{k})$. Seeing how this happens is a standard exercise in quantum mechanics. Here we'll jump to the answer and the reader is welcome to check that it works.

In the quantum theory, we implement canonical commutation relations

$$\left[u_i^\mu(\mathbf{R}), P_j^\nu(\mathbf{R}')\right] = i\delta_{ij}\delta_{\mu\nu}\delta_{\mathbf{R},\mathbf{R}'}. \tag{76}$$

We then define the Fourier mode expansions

$$\begin{aligned} u_i^\mu(\mathbf{R}) &= \frac{1}{\sqrt{M_i}} \frac{1}{\sqrt{2N}} \sum_{n,\boldsymbol{k}} \frac{1}{\sqrt{\omega_n(\boldsymbol{k})}} \left(a_{n\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\mathbf{R}} \varepsilon_{i\mu}^{(n)}(\boldsymbol{k}) + a_{n\boldsymbol{k}}^\dagger e^{-i\boldsymbol{k}\cdot\mathbf{R}} \bar{\varepsilon}_{i\mu}^{(n)}(\boldsymbol{k}) \right), \\ P_i^\mu(\mathbf{R}) &= -i\sqrt{M_i} \frac{1}{\sqrt{2N}} \sum_{n,\boldsymbol{k}} \sqrt{\omega_n(\boldsymbol{k})} \left(a_{n\boldsymbol{k}} e^{i\boldsymbol{k}\cdot\mathbf{R}} \varepsilon_{i\mu}^{(n)}(\boldsymbol{k}) - a_{n\boldsymbol{k}}^\dagger e^{-i\boldsymbol{k}\cdot\mathbf{R}} \bar{\varepsilon}_{i\mu}^{(n)}(\boldsymbol{k}) \right), \end{aligned} \tag{77}$$

where N is the number of unit cells.

3.6.1 Periodic boundary conditions and state counting

Strictly speaking the plane wave form holds for periodic boundary conditions (PBCs), and to understand the sum we should look slightly more carefully. We define the PBCs by

$$\mathbf{R} \equiv \mathbf{R} + \mathbf{L}_i, \tag{78}$$

where $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$ are three linearly independent directions defining a parallelpiped shape for the system. Since the $\mathbf{R}$ are Bravais lattice vectors, so must

be the $\mathbf{L}_i$. Imposing that Eqs. (77) are independent of these translations implies that $\boldsymbol{k} \cdot \mathbf{L}_i \in 2\pi\mathbb{Z}$. This condition looks an awful lot like the one which defines the reciprocal lattice, Eq. (36), but with the primitive vectors $\boldsymbol{a}_i$ replaced by the $\mathbf{L}_i$. Consequently, PBCs implies that the wavevectors in our phonon sum should be taken from the reciprocal lattice defined by the $\mathbf{L}_i$. That is, we can define

$$\mathbf{B}_i \cdot \mathbf{L}_j = 2\pi\delta_{ij}, \tag{79}$$

and the $\mathbf{B}_i$ are the basis vectors of this new reciprocal lattice, i.e. the allowed wavevectors are

$$\boldsymbol{k} = \sum_{i=1}^{d} n_i \mathbf{B}_i. \tag{80}$$

Different choices of the integers n_i give different wavevectors. Naïvely there is an infinite choice of the d integers, and so an infinite set of these wavevectors. This would seem bizarre since for our finite crystal there are a finite number of unit cells and so a finite number of modes. The problem is resolved by realizing that any two wavevectors whose difference is a reciprocal lattice vector of the microscopic lattice, i.e. with $\boldsymbol{k} - \boldsymbol{k}' = \mathbf{Q}$ with $\mathbf{Q} \cdot \mathbf{R} \in 2\pi\mathbb{Z}$, describe the same mode. This is because the plane wave factors in Eq. (77) are identical for two such wavevectors. The equivalence of wavevectors differing by a reciprocal lattice vectors is something we will run into over and over again. It arises generally for waves moving in a periodic system. At a more mathematical level, the wavevector itself is not uniquely defined, so we might call it a "quasi-wavevector". In the quantum theory, wavevectors and momenta are identified, so this same notion will apply to momentum, becoming "quasi-momentum" inside the crystal. We will return to this in Sec. 5.1 in more depth when we discuss electronic wavefunctions.

Here, the equivalence means that if we include the infinite set of n_i in Eq. (80), we will count every physical mode an infinite number of times! So to do this properly, we should count each inequivalent mode only once. That is, when we sum over $\boldsymbol{k}$, or equivalently integers n_i in Eq. (80), we should pick a set of these wavevectors so that translating any wavevector by a reciprocal lattice vector $\mathbf{Q}$ is always outside the set. There are a priori an infinite number of different ways to do this. Any particular choice is just a convention. A common choice is to choose our $\boldsymbol{k}$ to be those points which are closer to the origin than to any other reciprocal lattice vector. This defines a set which is actually the Wigner-Seitz cell of the reciprocal lattice, that we discussed in Sec. 2.2.2, and is called the 1st Brillouin zone.

So we conclude that a good definition of the momentum sums in Eqs. (77)

is, more explicitly

$$\sum_{\boldsymbol{k}} = \sum_{\substack{\boldsymbol{k}\in \mathrm{BZ} \\ \boldsymbol{k}\cdot\mathbf{L}\in 2\pi\mathbb{Z}}} = \sum_{\substack{n_1\cdots n_d \\ |\sum_i n_i\mathbf{B}_i|<|\sum_i n_i\mathbf{B}_i-\mathbf{Q}|\forall\mathbf{Q}\neq 0}} . \tag{81}$$

This is obviously a bit nasty to write which is why we usually do not do it explicitly.

If one wants something more explicit and also simple, we can consider a special case. For example, suppose we take $\mathrm{L}_i = \mathrm{N}_i \boldsymbol{a}_i$ with integers N_i, where $\boldsymbol{a}_i$ are the primitive lattice vectors. Then we have just $\mathbf{B}_i = \boldsymbol{b}_i/\mathrm{N}_i$. A unique set of inequivalent wavevectors is then found by taking $0 \leq n_i < \mathrm{N}_i$, or, for N_i odd, more symmetrically, $-\mathrm{N}_i/2 < n_i < \mathrm{N}_i/2$ (for N_i even we should use instead $-\mathrm{N}_i/2 \leq n_i < \mathrm{N}_i/2$).

Regardless of the choice of wavevectors, the *number* of terms in the wavevector sum should be fixed, and equal to the number of unit cells in the solid. This correctly counts the number of degrees of freedom in the crystal. One can check the above logic by recalling that the volume of wavevector space corresponding to a unit cell in reciprocal space is just $(2\pi)^d/v_{\mathrm{u.c.}}$, where $v_{\mathrm{u.c.}}$ is the volume of the real space unit cell. The volume per allowed wavevector defined via Eq. (80) is the volume of the parallelpiped defined by $\mathbf{B}_1 \cdots \mathbf{B}_d$. This is just the volume of the fictitious reciprocal space defined by the $\mathbf{L}_i$, and so in turn the volume per allowed wavevector is $(2\pi)^d/\mathrm{V}$, where V is the volume of the crystal defined by the $\mathbf{L}_i$. Dividing the volume of the reciprocal space unit cell by the latter volume gives the number of inequivalent wavevectors, which is just $\mathrm{V}/v_{\mathrm{u.c.}} = \mathrm{N}$, the number of unit cells.

I want to emphasize that most of the time we are interested in large samples containing many unit cells, and so only in those properties which are associated with the bulk, and not sensitive to details of the crystal shape. If one takes the dimensions $\mathbf{L}_i$ large, the details of the shape drop out and one achieves the thermodynamic limit as they become infinite. In fact, for a real finite crystal, we of course do not have periodic boundary conditions! Please take it as a matter of faith that bulk properties are not affected by the choice of boundary conditions. Apologies for the apparent detour into numerology, but it's useful to know this stuff, and it will come up again when we talk about electron bands.

3.6.2 Back to physics

Eq. (77) generalizes the undergraduate expression for the position and momentum operators in terms of ladder operators in the simple harmonic oscillator. One can check that with these definitions, the $a, a^\dagger$ operators are canonical, i.e. hermitian conjugates of one another, and satisfy

$$\left[a_{n\boldsymbol{k}}, a^\dagger_{n'\boldsymbol{k}'}\right] = \delta_{nn'}\delta_{\boldsymbol{k}\boldsymbol{k}'}, \qquad \left[a_{n\boldsymbol{k}}, a_{n'\boldsymbol{k}'}\right] = \left[a^\dagger_{n\boldsymbol{k}}, a^\dagger_{n'\boldsymbol{k}'}\right] = 0. \tag{82}$$

One can now insert Eqs. (77) into Eq. (62) and carry out the sums over $\mathbf{R}$ and i, j, μ, ν. One finds eventually the standard result

$$\mathrm{H}_{\mathrm{latt}} = \sum_{n\boldsymbol{k}} \hbar\omega_n(\boldsymbol{k}) \left(a^\dagger_{n\boldsymbol{k}} a_{n\boldsymbol{k}} + \frac{1}{2} \right). \tag{83}$$

Here we restored the factor $\hbar = 1$ for appearance's sake.

A word of interpretation: in the theory of the quantum harmonic oscillator, the operator $\mathrm{N}_{n\boldsymbol{k}} \equiv a^\dagger_{n\boldsymbol{k}} a_{n\boldsymbol{k}}$ is found to have non-negative integer eigenvalues $\mathrm{N}_{n\boldsymbol{k}} = 0, 1, 2, \ldots$. Consequently it can be considered a "number" operator. Each term in Eq. (83) is just the zero point energy of that mode plus the energy $\hbar\omega_n(\boldsymbol{k})$ times the number operator for that mode.

This energy has a natural particle interpretation. We can think of the total energy as that of a system of $\mathrm{N}_{n\boldsymbol{k}}$ particles of type n and wavevector $\boldsymbol{k}$ (plus a zero point contribution which is the energy in the absence of any particles). The ground state $|0\rangle$ is the eigenstate with $\mathrm{N}_{n\boldsymbol{k}} = 0$ for all $n, \boldsymbol{k}$, and is called the *vacuum*. In the particle picture it is the state with no particles. Acting on the vacuum with raising operators $a^\dagger_{n\boldsymbol{k}}$ "creates" particles. For example,

$$|\psi\rangle = a^\dagger_{n\boldsymbol{k}} a^\dagger_{n'\boldsymbol{k}'} |0\rangle, \tag{84}$$

is a state with two particles, with discrete quantum numbers n, n' and wavevectors/momenta $\boldsymbol{k}, \boldsymbol{k}'$. These particles are quantum mechanically *identical*, because there is a just a single quantum state with $\mathrm{N}_{n\boldsymbol{k}} = k$ particles when $k > 1$, for example when $n' = n$, $\boldsymbol{k}' = \boldsymbol{k}$ above this is a unique state. Moreover, these particles are bosons, because the creation operators *commute*, which means that (1) permuting their order does not matter and (2) one can have an arbitrarily large number of particles in each state. This is nothing but the "second quantized" Hamiltonian for identical bosons. We call these quantum particles associated with lattice vibrations *phonons*.

3.6.3 Continuum elasticity

It is useful to keep in mind a particular case of a deformed crystal, in which the displacements are not only small but also *slowly varying in space*. These correspond to *strains*, and describe the way in which a solid responds to relatively small forces exerted over long distances. For example, if pressure is applied to the top and bottom surfaces of a cube of solid, it will typically compress in the vertical direction and expand horizontally. Such slowly varying deformations are the subject of *continuum elasticity*.

In this case, one can regard $\boldsymbol{u}_i(\mathbf{R})$ as a continuum function of position $\mathbf{R}$, which is a smooth interpolation of the discrete values at which it is really defined. Moreover for such smooth deformations all the atoms in the unit cell move together, so we can take $\boldsymbol{u}_i(\mathbf{R}) = \boldsymbol{u}(\mathbf{R} + \boldsymbol{d}_i)$, with a single function $\boldsymbol{u}(\boldsymbol{r})$ rather than a separate one for each atom in the unit cell. In turn one can

define derivatives of this smooth displacement function,

$$u_j^\nu(\mathbf{R}') \approx u^\nu(\mathbf{R}' + \boldsymbol{d}_j) \approx u^\nu(\mathbf{R}) + (\mathrm{R}'_\gamma - \mathrm{R}_\gamma + d_j^\gamma)\partial_\gamma u^\nu(\mathbf{R}) + \frac{1}{2}(\mathrm{R}'_\gamma - \mathrm{R}_\gamma + d_j^\gamma)(\mathrm{R}'_\lambda - \mathrm{R}_\lambda + d_j^\lambda)\partial_\gamma\partial_\lambda u^\nu(\mathbf{R}) + \mathcal{O}(\partial^3 u), \tag{85}$$

when $|\mathbf{R}' - \mathbf{R}|$ is not large.

Now we can insert this into the expression for the lattice potential energy

$$\mathrm{E}^{(2)}_{\mathrm{BO}}[\{\boldsymbol{u}_i(\mathbf{R})\}] = \frac{1}{2}\sum_{ij}\sum_{\mathbf{R},\mathbf{R}'}\sum_{\mu\nu}\mathrm{V}_{ij;\mu\nu}(\mathbf{R}-\mathbf{R}')u_\mu(\mathbf{R})\Big[u_\nu(\mathbf{R}) + (\mathrm{R}'_\gamma - \mathrm{R}_\gamma + d_j^\gamma)\partial_\gamma u_\nu(\mathbf{R}) + \frac{1}{2}(\mathrm{R}'_\gamma - \mathrm{R}_\gamma + d_j^\gamma)(\mathrm{R}'_\lambda - \mathrm{R}_\lambda + d_j^\lambda)\partial_\gamma\partial_\lambda u_\nu(\mathbf{R})\Big]. \tag{86}$$

Now change summation variables $\mathbf{R}' \to \mathbf{R} - \mathbf{R}'$. One obtains

$$\begin{aligned}\mathrm{E}^{(2)}_{\mathrm{BO}}[\{\boldsymbol{u}_i(\mathbf{R})\}] = \frac{1}{2}\sum_{\mu\nu}\sum_{\mathbf{R}}\Bigg[&\left(\sum_{ij}\sum_{\mathbf{R}'}\mathrm{V}_{ij;\mu\nu}(\mathbf{R}')\right)u_\mu(\mathbf{R})u_\nu(\mathbf{R}) \\ &+ \sum_\gamma\left(\sum_{ij}\sum_{\mathbf{R}'}\mathrm{V}_{ij;\mu\nu}(\mathbf{R}')\left(d_j^\gamma - \mathrm{R}'_\gamma\right)\right)u_\mu(\mathbf{R})\partial_\gamma u_\nu(\mathbf{R}) \\ &+ \frac{1}{2}\sum_{\gamma\lambda}\left(\sum_{ij}\sum_{\mathbf{R}'}\mathrm{V}_{ij;\mu\nu}(\mathbf{R}')\left(d_j^\gamma - \mathrm{R}'_\gamma\right)\left(d_j^\lambda - \mathrm{R}'_\lambda\right)\right)u_\mu(\mathbf{R})\partial_\gamma\partial_\lambda u_\nu(\mathbf{R})\Bigg].\end{aligned} \tag{87}$$

The first line above vanishes due to the condition of translational invariance, Eq. (63). In the remaining terms, since the displacement fields are presumed slowly varying, we can reply the summation over $\mathbf{R}$ by an integral,

$$\mathrm{E}^{(2)}_{\mathrm{BO}}[\{\boldsymbol{u}_i(\mathbf{R})\}] \approx \frac{1}{2v_{\mathrm{u.c.}}}\int d^d\mathbf{R}\left[\sum_{\mu\nu\gamma} v_{\mu\nu\gamma}u_\mu\partial_\gamma u_\nu + \sum_{\mu\nu\gamma\lambda} v_{\mu\nu\gamma\lambda}u_\mu\partial_\gamma\partial_\lambda u_\nu\right], \tag{88}$$

where

$$\begin{aligned}v_{\mu\nu\gamma} &= \sum_{ij}\sum_{\mathbf{R}'}\mathrm{V}_{ij;\mu\nu}(\mathbf{R}')\left(d_j^\gamma - \mathrm{R}'_\gamma\right) = \sum_{ij}\left(d_j^\gamma - i\frac{\partial}{\partial k_\gamma}\right)\tilde{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k})\Big|_{\boldsymbol{k}=0}, \\ v_{\mu\nu\gamma\lambda} &= \frac{1}{2}\sum_{ij}\left(d_j^\gamma - i\frac{\partial}{\partial k_\gamma}\right)\left(d_j^\lambda - i\frac{\partial}{\partial k_\lambda}\right)\tilde{\mathrm{V}}_{ij;\mu\nu}(\boldsymbol{k})\Big|_{\boldsymbol{k}=0},\end{aligned} \tag{89}$$

These two terms can be simplified by using the invariance of the crystal energy under rigid rotations of the solid, Eq. (64). I invite the student to verify that this implies first that the term proportional to $v_{\mu\nu\gamma}$ vanishes. By integration

by parts, the last term in Eq. (88) can be written in terms of gradients only,

$$\begin{aligned} \mathrm{E}^{(2)}_{\mathrm{BO}}[\{\boldsymbol{u}_i(\mathbf{R})\}] &\approx -\sum_{\mu\nu\gamma\lambda} \frac{1}{2v_{\mathrm{u.c.}}} \int d^d\mathbf{R}\, v_{\mu\nu\gamma\lambda}\partial_\gamma u_\mu \partial_\lambda u_\nu \\ &= \sum_{\mu\nu\gamma\lambda} \frac{1}{2} \int d^d\mathbf{R}\, c_{\mu\nu\gamma\lambda} u_{\gamma\mu} u_{\lambda\nu}, \end{aligned} \tag{90}$$

where

$$u_{\mu\nu} = \frac{1}{2}\left(\partial_\mu u_\nu + \partial_\nu u_\mu\right) \tag{91}$$

is the symmetric *strain tensor*. The passage from the first to the second line requires imposing rotational invariance. Indeed, invariance under rotations forces the energy to depend on displacements only through the strain tensor in the elastic limit. This is a well-known result from classical elasticity. It is easily understood, by realizing that an infinitesimal rotation corresponds to a configuration

$$u^{\mathrm{rigidrot}}_\mu = \theta\, \epsilon_{\gamma\mu\nu} x_\nu, \tag{92}$$

for a rotation by angle $\theta \ll 1$ around the γ axis. The symmetrized strain tensor is invariant under stuch rotations. We do not give a formula for the elastic tensor $c_{\mu\nu\gamma\lambda}$.

The final form in Eq. (90) gives the elastic potential energy as a quadratic form in the strain tensor. Because each strain tensor is linear in derivatives, the elastic energy in Fourier space is quadratic in wavevector. This evinces explicitly the linear in wavevector behavior of the acoustic mode frequencies (recall that for a SHO the spring constant of the quadratic potential = $\mathrm{M}\omega^2$). The elastic description captures the acoustic modes but not the optical ones.

The general elastic tensor $c_{\mu\nu\gamma\lambda}$ is a material property which is constrained by crystal symmetries: the lower the symmetry of the crystal, the larger the number of independent values of the these coefficients, called *elastic moduli*. Details can be easily figured out or found in books. It is probably noteworthy that the elastic form of the potential energy holds even for non-crystalline solids, i.e. amorphous solids, rubber, glasses, etc., in which the atoms do not occupy a regular periodic array. Perhaps surprisingly, an amorphous solid actually has *less* independent elastic moduli – only two are needed to describe a completely amorphous solid. This is because typical non-crystalline solids are *statistically isotropic*: because of their irregularity, there are no preferred directions in space if one considers a large volume of the solid. We won't explore this further here.

3.7 *Thermodynamics*

Phonons make an important contribution to the heat capacity of solids. The heat capacity is a thermodynamic quantity, derived from the free energy, or partition function, so it depends only upon the energies of the levels and not on any other quantum numbers like wavenumber/momentum. Hence it is convenient to introduce the *density of states* (DOS) $\mathrm{G}(\epsilon)$, which counts the density of phonon levels in an infinitesimal interval around the energy ϵ:

$$\mathrm{G}(\epsilon) = \sum_n \sum_{\boldsymbol{k}} \delta(\epsilon - \hbar\omega_n(\boldsymbol{k})). \tag{93}$$

The integral $\int_{\epsilon_1}^{\epsilon_2} \mathrm{G}(\epsilon) d\epsilon$ gives the total number of states with energy between ϵ_1 and ϵ_2. In the large volume limit, the discrete wavevectors become very tightly spaced, and the sum above can be converted to an integral

$$\mathrm{G}(\epsilon) \to \mathrm{V} g(\epsilon), \tag{94}$$

$$g(\epsilon) = \sum_n \int_{\mathrm{BZ}} \frac{d^d \boldsymbol{k}}{(2\pi)^d} \, \delta(\epsilon - \hbar\omega_n(\boldsymbol{k})), \tag{95}$$

using the fact, obtained in Sec. 3.6.1, that the volume per allowed wavevector is $(2\pi)^d/\mathrm{V}$. The quantity $g(\epsilon)$ gives the density of states per unit volume, and is an intrinsic quantity.

What does $g(\epsilon)$ look like? We know that it is by definition positive semi-definite. It is zero for $\epsilon < 0$, and also when $\epsilon > \epsilon_{\max} = \max_{n\boldsymbol{k}} \epsilon_n(\boldsymbol{k})$, i.e. above the maximum of the phonon bands. When the energy is small but not zero, $\epsilon \ll \epsilon_{\max}$, the contributions to the density of states come only from the *acoustic* phonons, which are the ones which persist to low energy. In this range, for the d acoustic modes, we can substitute $\omega_n(\boldsymbol{k}) = v_n(\hat{k})|k|$, which gives

$$\begin{aligned} g(\epsilon) &\sim \sum_{n=1}^{d} \int_{\mathrm{BZ}} \frac{d^d \boldsymbol{k}}{(2\pi)^d} \, \delta(\epsilon - \hbar v_n(\hat{k})|k|) \\ &\sim \sum_{n=1}^{d} \int_0^\infty \frac{dk k^{d-1} \int d\boldsymbol{\Omega}}{(2\pi)^d} \delta(\epsilon - \hbar v_n(\boldsymbol{\Omega})k) = \sum_{n=1}^{d} \frac{\epsilon^{d-1}}{(2\pi)^d} \int \frac{d\boldsymbol{\Omega}}{(\hbar v_n(\boldsymbol{\Omega}))^d}. \end{aligned} \tag{96}$$

In the second line we changed to spherical coordinates and collapsed the delta function. We see that the result is proportional to ϵ^{d-1}, so the density of states vanishes as a power law on approaching zero energy (in $d > 1$). In a simple model with mode-independent and angle-independent velocity, $v_n(\boldsymbol{k}) = v$, one has

$$g_{Debye}(\epsilon) = \frac{d\mathrm{S}_d}{(2\pi\hbar v)^d} \epsilon^{d-1}, \tag{97}$$

where S_d is the surface area of the d-dimensional sphere. A crude approximation to the full DOS is to simply keep the power law form of Eq. (97) up to some maximum frequency, and then set $g(\epsilon) = 0$ above that. This model is called the "Debye" model, and the maximum frequency the Debye frequency, ω_D. More generally, one loosely speaks of a Debye frequency as giving the scale for the width of the phonon energy spectrum.

One should also note that the total integral of the density of states is fixed, because the total number of phonon states is just the total number of translational degrees of freedom, i.e. $\int d\epsilon G(\epsilon) = N_{u.c.} dn_b$, where $N_{u.c.}$ is the number of unit cells in the solid. This means that

$$\int d\epsilon g(\epsilon) = \frac{dn_b}{v_{u.c.}}, \tag{98}$$

where $v_{u.c.}$ is the volume of a primitive unit cell.

The above properties of the phonon DOS are general. Specific crystals will have other notable features such as peaks and edges in $g(\epsilon)$.

Armed with the DOS, we can readily obtain the internal energy U and internal energy density u of the phonon system using Bose-Einstein statistics:

$$u = \frac{U}{V} = \frac{\langle H \rangle}{V} = \int d\epsilon \, g(\epsilon) \epsilon n_B(\epsilon), \tag{99}$$

where

$$n_B(\epsilon) = \frac{1}{e^{\beta\epsilon} - 1} \tag{100}$$

is the Bose distribution function, with $\beta = 1/(k_B T)$. Note that because there is no reason to impose conservation of the *number* of phonons, there is no chemical potential for phonons.

Taking the temperature derivative, we obtain

$$c_v = \frac{\partial u}{\partial T} = \int d\epsilon \, g(\epsilon) \frac{\epsilon^2}{4 k_B T^2 \sinh^2 \frac{\beta\epsilon}{2}}. \tag{101}$$

At low temperatures, When $k_B T \ll \hbar\omega_D$, we can approximate $g(\epsilon) \approx A\epsilon^{d-1}$, with A from Eq. (96) or Eq. (97). Then we can change variables to $x = \beta\epsilon$ to obtain

$$c_v \sim \frac{A}{4} k_B^{d+1} T^d \int_0^\infty dx \frac{x^{d+1}}{\sinh^2 \frac{x}{2}}. \tag{102}$$

The integral is just a number, so we see that the low temperature specific heat is proportional to T^d. For the Debye model in three dimensions, one has

$A = \frac{3}{2\pi^2}\frac{1}{(\hbar v)^3}$ and $\int_0^\infty dx \frac{x^4}{\sinh^2 \frac{x}{2}} = 16\pi^4/15$, so we find

$$c_v \sim \frac{2\pi^2}{5\hbar^3 v^3} k_B^4 T^3. \tag{103}$$

The T^3 behavior of the phonon heat capacity is a robust property of crystalline solids at temperatures below the Debye scale.

3.8 *Other phenomena involving phonons*

Phonons are involved in many other physical effects in solids. Some are obvious, because they explicitly require the lattice to move:

- **Thermal expansion:** As temperature varies, a solid expands or contracts: in most cases solids grow in volume with increasing temperature, but not always. This is a consequence of anharmonic (i.e. cubic and higher) terms in the energy of lattice deformations, and the competition of energy versus entropy.
- **Elasticity:** We already saw in Sec. 3.6.3 that the long wavelength acoustic modes are just the elastic degrees of freedom.
- **Ferroelectricity:** Sometimes a crystal undergoes, as a function of temperature, a phase transition in which the crystal structure changes. If in this transition an electronic dipole moment develops, it is called ferroelectricity. Ferroelectricity often occurs by a particular mode of lattice distortion becoming on average non-zero.

Another important role of phonons is closely related to their thermodynamics:

- **Thermal conductivity:** A propagating phonon is a particle that moves at some velocity and carries some energy, so it contributes to energy currents. Indeed, usually phonons dominate the thermal conductivity, which is defined as the coefficient of proportionality between a heat current and (minus) an imposed temperature gradient. Specifically $\boldsymbol{j}_Q$ the thermal current density is given as $\boldsymbol{j}_Q = -\underline{\kappa}\boldsymbol{\nabla} T$, which actually defines a matrix of thermal conductivities $\underline{\kappa}$.

Phonons can be measured by various techniques. Traditionally, scattering measurements determine their dispersion relations:

- **Infra-red absorption:** Infrared radiation has an energy comparable to typical phonons, but such light has very long wavelength. Thus energy and momentum conservation require that an infrared photon can be (sometimes) absorbed to create an *optical* phonon (not an acoustic phonon, which has zero energy in the long-wavelength limit). This is why optical phonons are called optical. Only *infrared active* phonons couple efficiently to light in this way, which places some symmetry constraints.

- **Light scattering:** Electromagnetic radiation with frequencies in the optical range are too high energy to be absorbed by creating a single phonon, but they can scatter off a solid and exchange some energy with phonons. This is called Brillouin and Raman scattering.

- **Neutron scattering:** Neutrons are readily available with energies and momenta that match phonons reasonably well. So a neutron scattering off a solid can exchange momentum and energy with it and create or absorb a phonon, which allows one to deduce the phonon dispersion relation if you track carefully the changes of energy of the neutron.

Phonons also play an important role sometimes in electronic physics. They can interact with electrons, which is an important source of electronic scattering (see Sec. 7.2.2), and can induce new physics such as superconductivity and charge density waves.

4 FROM MANY ELECTRONS TO ONE

Now that we've discussed the structure of crystals, we are ready to come back to understand the electrons within them. If the nucleii are in a fixed periodic arrangement, then the quantum problem for the electrons from Eq. (3) reduces to

$$\mathrm{H}_e = \sum_{i=1}^{\mathrm{N}_e} \left[\frac{|\boldsymbol{p}_i|^2}{2m} + \mathrm{V}(\boldsymbol{x}_i) \right] + \sum_{i<j} \mathrm{U}(\boldsymbol{x}_i - \boldsymbol{x}_j), \tag{104}$$

where $\mathrm{U}(\boldsymbol{x}) = e^2/|\boldsymbol{x}|$ is the repulsive electron-electron interaction, and $\mathrm{V}(\boldsymbol{x})$ is the (attractive) one-electron potential due to all all the ions,

$$\mathrm{V}(\boldsymbol{x}) = \sum_{j=1}^{\mathrm{N}_n} -\frac{\mathrm{Z}_j e^2}{|\boldsymbol{x} - \mathbf{X}_j|}. \tag{105}$$

It will be helpful at times to define the three terms in the Hamiltonian as operators, i.e. to write Eq. (104) as

$$\hat{\mathrm{H}} = \hat{\mathrm{T}} + \hat{\mathrm{V}} + \hat{\mathrm{U}}, \tag{106}$$

where

$$\hat{T} = \sum_i \frac{|\boldsymbol{p}_i|^2}{2m}, \tag{107}$$

$$\hat{V} = \sum_i V(\boldsymbol{x}_i), \tag{108}$$

$$\hat{U} = \sum_{i<j} U(\boldsymbol{x}_i - \boldsymbol{x}_j). \tag{109}$$

It is particularly useful to note that the 1-electron potential term can be expressed in terms of the density operator,

$$\hat{V} = \int d^3\boldsymbol{x}\, V(\boldsymbol{x})\hat{n}(\boldsymbol{x}), \tag{110}$$

where

$$\hat{n}(\boldsymbol{x}) = \sum_i \delta(\boldsymbol{x} - \boldsymbol{x}_i). \tag{111}$$

The one-electron potential has the important property that it shares the periodicity of the crystal, to wit

$$V(\boldsymbol{x} + \mathbf{R}) = V(\boldsymbol{x}), \qquad \mathbf{R} \in \text{B.L.}, \tag{112}$$

where $\mathbf{R}$ is any vector in the Bravais lattice. This is equivalent to the three linearly independent conditions

$$V(\boldsymbol{x} + \boldsymbol{a}_\mu) = V(\boldsymbol{x}). \tag{113}$$

The full Hamiltonian in Eq. (104) shares this periodicity, because the electron-electron interaction is invariant under any translation.

In this section, we will motivate the replacement of the full Hamiltonian in Eq. (104) by an approximate one with the same form, but *without interactions between electrons*, i.e. with $U(\boldsymbol{x}) = 0$, and with some "renormalized" $V(\boldsymbol{x})$, which is not simply the sum of pure atomic $1/r$ Coulomb attractions. Indeed, the astute among you probably realized that for an infinite crystal, the sum in Eq. (105), which involves all negative terms, diverges, because the $1/r$ potential does not decay rapidly enough at long distances. The full Hamiltonian, including electron-electron repulsion, is well-behaved because the negatively charged electrons screen the positive nuclear charges, and averaged over a volume larger than the unit cell, the full system is approximately neutral. Hence to capture this physics, and restore the finite nature of the local electrostatic potential, we must include the potential due to the electrons themselves.

The challenge of doing this is that the electrons in our treatment are

fully quantum-mechanical, and so their potential is not fixed, because their positions are not fixed. Hence the full electrostatic potential is not a classical variable, and has quantum fluctuations and quantum uncertainty. To deal with this, we must make some approximations. The premise, on which much of solid state physics is based, is that in some sense replacing this fluctuating potential by its average is a good starting point. This is an example of a "mean field theory", in which a fluctuating field is replaced by its average (a word of caution: the term "mean field theory" is used in many different ways in physics, and I am using it loosely here). I'd like to outline here two mean field approaches.

4.1 *Hartree-Fock theory*

This subsection is optional. Hartree-Fock is a very common theoretical method, but not essential to understanding the flow of the course.

The Hartree-Fock theory dates back to very early days of quantum mechanics. Normally the Hartree-Fock approximation is formulated as a variational one for the ground state. One takes as a variational wavefunction a Slater determinant of single-particle states $\psi_a(\chi)$, where a labels the single-electron states, and we defined a combined space/spin coordinate $\chi = (\boldsymbol{x}, \sigma)$, with $\sigma = \pm\frac{1}{2} = \uparrow / \downarrow$ is a spin-1/2 index:

$$\begin{aligned}\Psi(\chi_1, \cdots, \chi_N) &= \frac{1}{\sqrt{N!}} \det \begin{pmatrix} \psi_1(\chi_1) & \psi_1(\chi_2) & \cdots & \psi_1(\chi_N) \\ \psi_2(\chi_1) & \psi_2(\chi_2) & \cdots & \psi_2(\chi_N) \\ \vdots & \vdots & & \\ \psi_N(\chi_1) & \psi_N(\chi_2) & \cdots & \psi_N(\chi_N) \end{pmatrix} \\ &= \frac{1}{\sqrt{N!}} \sum_{a_1 \cdots a_N} \epsilon_{a_1 \cdots a_N} \psi_{a_1}(\chi_1) \cdots \psi_{a_N}(\chi_N), \end{aligned} \tag{114}$$

where here ϵ is the fully antisymmetric Levi-Civita symbol. The wavefunctions $\psi_a(\chi)$, which should form an orthonormal set (for Ψ to be normalized) can be considered variational parameters. Alternatively, we can regard Ψ as the ground state of an at this point arbitrary single-electron Hamiltonian, and it is this single-electron Hamiltonian which is to be varied, i.e. we seek the best possible single-electron Hamiltonian which approximates the full one. The two views can be seen to be equivalent as follows. Consider the variational Lagrangian for the ground state energy,

$$L = \langle \Psi | H_e | \Psi \rangle - \sum_a \epsilon_a (\langle \psi_a | \psi_a \rangle - 1), \tag{115}$$

with Lagrange multipliers ϵ_a introduced to enforce normalization of the states. First we can evaluate the one-electron terms in the variational energy as

$$\begin{aligned}\langle\Psi|\sum_i\left[\frac{|\boldsymbol{p}_i|^2}{2m}+V(\boldsymbol{x}_i)\right]|\Psi\rangle &= N\langle\Psi|\left[\frac{|\boldsymbol{p}_1|^2}{2m}+V(\boldsymbol{x}_1)\right]|\Psi\rangle \\ &= \sum_a\langle\psi_a|\frac{|\boldsymbol{p}|^2}{2m}+V(\boldsymbol{x})|\psi_a\rangle. \qquad (116)\end{aligned}$$

In the first line, we use the antisymmetry of the bra and ket (which means the full expression is symmetric in permutations) to note that every term in the i sum gives an equal contribution. Then the equality in the second line follows by inserting Eq. (114) for both the bra and ket in Eq. (116), and noting that the only non-vanishing terms are those in which the permutation of indices of ψ_a for the bra and ket match, because otherwise the orthogonality of the single-particle states gives zero. There are N! such terms, which can be divided into $(N-1)!$ terms in which the first particle (with argument χ_1) is in state 1, another $(N-1)!$ terms in which the first particle is in state 2, etc. This gives each of the terms in the sum on the second line, with coefficient $N\times(N-1)!/N! = 1$ for each term.

For the interaction term, we have

$$\begin{aligned}U_{HF} &\equiv \langle\Psi|\sum_{i<j}U(\boldsymbol{x}_i-\boldsymbol{x}_j)|\Psi\rangle = \frac{N(N-1)}{2}\langle\Psi|U(\boldsymbol{x}_1-\boldsymbol{x}_2)|\Psi\rangle \qquad (117)\\ &= \sum_{a<b}[\langle\psi_a\psi_b|U(\boldsymbol{x}_1-\boldsymbol{x}_2)|\psi_a\psi_b\rangle-\langle\psi_a\psi_b|U(\boldsymbol{x}_1-\boldsymbol{x}_2)|\psi_b\psi_a\rangle]\\ &= \frac{1}{2}\sum_{a,b}[\langle\psi_a\psi_b|U(\boldsymbol{x}_1-\boldsymbol{x}_2)|\psi_a\psi_b\rangle-\langle\psi_a\psi_b|U(\boldsymbol{x}_1-\boldsymbol{x}_2)|\psi_b\psi_a\rangle].\end{aligned}$$

Here we applied the same line of argument as before, except that when including the expansion of the two Slater determinants, only those single-particle states for coördinates $3\cdots N$ need to match between the bra and ket. This means that the two first states can be the same in the bra and ket, or they may be exchanged, and in the latter case there is a relative minus sign. In going from the second to the third line we use the fact that the expression inside the sum is symmetric and vanishes if $a = b$. The notation with two $|\psi_a\psi_b\rangle$ means that the first state in the ket (here a) has argument χ_1 and the second state in the ket (here b) has argument χ_2. So explicitly

$$\begin{aligned}&\langle\psi_a\psi_b|U(\boldsymbol{x}_1-\boldsymbol{x}_2)|\psi_c\psi_d\rangle \qquad (118)\\ &\quad=\sum_{\sigma,\sigma'}\int d^3\boldsymbol{x}_1 d^3\boldsymbol{x}_2\,\psi_a^*(\boldsymbol{x}_1,\sigma)\psi_b^*(\boldsymbol{x}_2,\sigma')U(\boldsymbol{x}_1-\boldsymbol{x}_2)\psi_c(\boldsymbol{x}_1,\sigma)\psi_d(\boldsymbol{x}_2,\sigma').\end{aligned}$$

The presence of the two terms in Eq. (117) is a signature of Fermi-Dirac

statistics, with a tell-tale minus sign. These terms are important enough that they have names: the first, positive, term is called the Hartree, or direct term, while the second, negative, term is called the Fock term, or exchange term.

Putting all the terms together, we get the Langrangian

$$\mathrm{L} = \sum_a \langle\psi_a|\frac{|\boldsymbol{p}|^2}{2m} + \mathrm{V}(\boldsymbol{x})|\psi_a\rangle + \frac{1}{2}\sum_{a,b}\langle\psi_a\psi_b|\mathrm{U}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\left(|\psi_a\psi_b\rangle - |\psi_b\psi_a\rangle\right) - \sum_a \epsilon_a(\langle\psi_a|\psi_a\rangle - 1). \tag{119}$$

Now we can (functionally) differentiate this with respect to $\langle\psi_a|$ to obtain

$$\left(\frac{|\boldsymbol{p}|^2}{2m} + \mathrm{V}(\boldsymbol{x}) - \epsilon_a\right)|\psi_a\rangle + \sum_b \langle\psi_b(2)|\mathrm{U}(1-2)\left(|\psi_a\psi_b\rangle - |\psi_b\psi_a\rangle\right) = 0. \tag{120}$$

The final terms may be a bit abstract, since the overlap integrals define the matrix element there is not written. Let us write out the form more explicitly:

$$\left(\frac{|\boldsymbol{p}|^2}{2m} + \mathrm{V}(\boldsymbol{x}) - \epsilon_a\right)|\psi_a(\boldsymbol{x},\sigma)\rangle + \sum_b\sum_{\sigma'}\int d^d\boldsymbol{x}'\,\psi_b^*(\boldsymbol{x}',\sigma')\mathrm{U}(\boldsymbol{x}-\boldsymbol{x}')\left(\psi_a(\boldsymbol{x},\sigma)\psi_b(\boldsymbol{x}',\sigma') - \psi_b(\boldsymbol{x},\sigma)\psi_a(\boldsymbol{x}',\sigma')\right) = 0. \tag{121}$$

It helps to separate the direct and exchange terms. One can do this by defining the direct and exchange potentials:

$$\mathrm{U}_d(\boldsymbol{x},\sigma) = \sum_b\sum_{\sigma'}\int d^d\boldsymbol{x}'\,\mathrm{U}(\boldsymbol{x}-\boldsymbol{x}')|\psi_b(\boldsymbol{x}',\sigma')|^2, \tag{122}$$

$$\mathrm{U}_{ex}(\boldsymbol{x}',\boldsymbol{x},\sigma',\sigma) = -\sum_b \mathrm{U}(\boldsymbol{x}-\boldsymbol{x}')\psi_b^*(\boldsymbol{x}',\sigma')\psi_b(\boldsymbol{x},\sigma). \tag{123}$$

Then Eq. (121) becomes

$$\left(-\frac{\boldsymbol{\nabla}^2}{2m} + \mathrm{V}(\boldsymbol{x}) + \mathrm{U}_d(\boldsymbol{x},\sigma)\right)\psi_a(\boldsymbol{x},\sigma) + \sum_{\sigma'}\int d^d\boldsymbol{x}'\,\mathrm{U}_{ex}(\boldsymbol{x}',\boldsymbol{x},\sigma',\sigma)\psi_a(\boldsymbol{x}',\sigma') = \epsilon_a\psi_a(\boldsymbol{x},\sigma). \tag{124}$$

You can see this appears like a single-particle Schrödinger equation for the states ψ_a, in which the interaction induces the additional direct and exchange "potentials". The Lagrange multipliers have taken on the role analogous to single-particle energies. The direct/Hartree potential has a rather simple interpretation: it is the electrostatic potential that would be induced on an electron by the electronic charge density, regarding that latter as a cloud with (number) density $n(\boldsymbol{x}) = \sum_{b,\sigma'}|\psi_b(\boldsymbol{x}',\sigma')|^2$. So the Hartree term realizes the first

goal of including the electronic contribution to the electrostatic potential.

The Fock/exchange potential goes beyond this. It can be of the opposite sign to the direct potential, and has the complication of being non-local. Both the direct and exchange potentials are implicitly dependent (through Eqs. (122)) on the full set of Hartree-Fock states. So generally, if they are to be solved, the Hartree-Fock equations are usually solved iteratively. One guesses a set of wavefunctions, computes the direct and exchange potentials, then solves Eq. (124) to obtain a new set of wavefunctions, and repeats until convergence is achieved. In principle, one should try this with different starting points, as the Hartree-Fock equations have multiple solutions, and one wants to find the one with the lowest total energy. We can find a simple expression for the total energy:

$$E_{HF} = \langle\Psi|H|\Psi\rangle = \sum_a \epsilon_a - \frac{1}{2}\sum_{a,b}\langle\psi_a\psi_b|U(\boldsymbol{x}_1 - \boldsymbol{x}_2)\left(|\psi_a\psi_b\rangle - |\psi_b\psi_a\rangle\right). \quad (125)$$

The first term in the total energy is what it would be if the system was really non-interacting, and the parameters ϵ_a were true single-particle energies. There is however a correction due to interactions, which is, a little surprisingly, of opposite sign to the original term in Eq. (117) (basically this is because we have already over-counted the interaction energy in the single-particle levels). I leave the derivation as an exercise to the reader.

At this point, instead of pursuing Hartree-Fock theory further, let us assume we have solved the equations, so that the direct and exchange potentials are known. Importantly, if the crystal is periodic, then we expect that a likely outcome is that the many-body wavefunction has the same periodicity, and in particular so does the charge density. Then we expect that the direct and exchange potentials we find are also periodic. This means that Eq. (124) defines an effective quantum mechanics problem of a particle in a (non-local) periodic potential. That is the problem we will turn to soon.

4.2 *Density functional theory*

By far the most common approach to actually deal with the interacting electron problem in materials is density functional theory. It is based on an idea of Hohenberg and Kohn, which we now explain. In a way it is a very formal result, but it led to the development of practical algorithms for computing properties of realistic materials. These methods have their limitations, but definitely are useful.

4.2.1 Hohenberg-Kohn theorems

The first idea of Hohenberg and Kohn is to try to work in terms of the electron density, which is easier to think about physically. If we look at Eq. (104), we can see that the differences between different types of matter, e.g. different materials, as well as different molecules, are all determined by specifying the

potential $V(\boldsymbol{x})$. For example, the difference between lead and gold, for example, is just the choice of potential. Given a potential, one can in principle solve for the ground state wavefunction $|\Psi\rangle$, which of course very complicated, but it is determined by V. Once one has this wavefunction, one can then calculate the density, i.e. the expectation value $n(\boldsymbol{x}) = \langle\Psi|\sum_i \delta(\boldsymbol{x} - \boldsymbol{x}_i)|\Psi\rangle$. Hence $n(\boldsymbol{x})$ is determined from $V(\boldsymbol{x})$. What Hohenberg and Kohn showed, first of all, is that this relation is true in the other direction: that a given density profile determines the potential $V(\boldsymbol{x})$. Since the potential determines the ground state wavefunction, we conclude that the density determines the ground state wavefunction. This can be shown via proof by contradiction. Suppose there are two different potentials, V_1 and V_2 (not trivially different by a constant), which have therefore two distinct ground states, but which give the same density $n(\boldsymbol{x})$. Call the ground state wavefunction in the two cases $|\Psi_1\rangle, |\Psi_2\rangle$, and the Hamiltonians H_1 and H_2. By the variational principle,

$$\begin{aligned} E_1 &< \langle\Psi_2|H_1|\Psi_2\rangle \\ &= \langle\Psi_2|H_2|\Psi_2\rangle + \langle\Psi_2|H_1 - H_2|\Psi_2\rangle \\ &= E_2 + \int d^3\boldsymbol{x}(V_1(\boldsymbol{x}) - V_2(\boldsymbol{x}))n(\boldsymbol{x}). \end{aligned} \tag{126}$$

where E_1 and E_2 are the two ground state energies. The last line holds because the two states have the same density. This is a strict inequality because by assumption the two states are different, and we assume that the ground states of the two Hamiltonians are non-degenerate. Reversing the roles of states 1 and 2 gives

$$E_2 < E_1 - \int d^3\boldsymbol{x}(V_1(\boldsymbol{x}) - V_2(\boldsymbol{x}))n(\boldsymbol{x}). \tag{127}$$

Adding the two equations gives $E_1 + E_2 < E_1 + E_2$, which is obviously false. This completes the proof.

Now that we know that the density determines the potential and hence the ground state, we can think of the expressing the energy as a function of the density. The reason this might be useful is that the expectation value of the potential term in any state is just the integral of the potential times the density of that state. It is the rest of the Hamiltonian that is more complicated. Therefore we separate out the potential and define

$$\hat{F} = \sum_{i=1}^{N_e}\left[\frac{|\boldsymbol{p}_i|^2}{2m} + \sum_{i<j} U(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] = \hat{T} + \hat{U}, \tag{128}$$

so that for a given potential, $H = \hat{F} + \hat{V}$. We would like to assign an energy to just $\hat{F}$, which is just determined by the density. One way to do it is to define

$$F[n] = \min_{|\Psi\rangle \text{ s.t. } \langle\Psi|\hat{n}|\Psi\rangle = n}\langle\Psi|\hat{F}|\Psi\rangle, \tag{129}$$

i.e. $F[n]$ is given by searching all states which have the density $n(\boldsymbol{x})$ and finding the minimum of the expectation value of $\hat{F}$ amongst those states. Note that in this definition, we do not require that these states are ground states corresponding to any potential V. This makes $F[n]$ defined for nearly all densities, as long as they are not too non-analytic (for example they cannot be discontinuous). Now we define a variational energy by just adding the energy of the potential,

$$E_V[n] = F[n] + \int d^3\boldsymbol{x}\, V(\boldsymbol{x}) n(\boldsymbol{x}). \tag{130}$$

Because $F[n]$ itself is the expectation value of $\hat{F}$ in some state $|\Psi_n\rangle$ which minimizes the latter and has the density n, we see that

$$E_V[n] = \langle \Psi_n | \hat{F} + \hat{V} | \Psi_n \rangle \geq E_0, \tag{131}$$

where E_0 is the ground state energy with the potential V. It is also clear that if we take $n(\boldsymbol{x}) = n_0(\boldsymbol{x})$, where n_0 is the density in the actual ground state with potential V, then $E_V[n_0] = E_0$. So we have established that $E_V[n]$ obeys a variational principle, and moreover, the non-trivial part of it, $F[n]$, is "universal", i.e. independent of the potential and the same for all materials. This is the second part of the "deep" theory of density function theory.

4.2.2 Kohn-Sham formulation

The problem with density functional theory is that of course we do not know the exact $F[n]$. Functionals are extremely complex objects, and an explicit representation of it is rather far-fetched to imagine. We can hope to approximate it, but even that, i.e. an approximation which works well for all choices of n, is very challenging. One can view the Thomas-Fermi energy funtional (removing the V term) in Eq. (24) as an approximation to it, but it is not a good enough one (as we saw it does not even predict stable molecules or solids!).

Kohn and Sham proposed an approximate functional which is practical in the sense that it can be calculated relatively quickly, and which is physically intuitive. They proposed to write the functional as

$$F[n] = T_f[n] + \frac{e^2}{2} \int d^3\boldsymbol{x} d^3\boldsymbol{x}' \, \frac{n(\boldsymbol{x}) n(\boldsymbol{x})}{|\boldsymbol{x} - \boldsymbol{x}'|} + V_{xc}[n], \tag{132}$$

where $T_f[n]$ is defined to be the kinetic energy of a free electron gas with the density n, i.e. with a potential $V_{KS}(\boldsymbol{x})$ chosen to give the electron density $n(\boldsymbol{x})$. That is, $T_f[n] = \langle \Psi_{KS} | \hat{T} | \Psi_{KS} \rangle$, where $|\Psi_{KS}\rangle$ is the ground state of the non-interacting "Kohn-Sham" Hamiltonian,

$$H_{KS} = \hat{T} + \hat{V}_{KS} = \hat{T} + \int d^3\boldsymbol{x}\, V_{KS}(\boldsymbol{x}) \hat{n}(\boldsymbol{x}), \tag{133}$$

and $V_{KS}(\boldsymbol{x})$ determined by the condition

$$\langle \Psi_{KS} | \hat{n}(\boldsymbol{x}) | \Psi_{KS} \rangle = n(\boldsymbol{x}). \tag{134}$$

By the first Hohenberg-Kohn theorem (uniqueness), this completely determines V_{KS} and hence $T_f[n]$ is completely specified. We can see that Eq. (132) therefore just trades the unknown functional $F[n]$ for a new unknown one $V_{xc}[n]$. This object is called the "exchange-correlation functional". The idea is that this functional should correct for two types of errors. First, the kinetic energy in an interacting state is not the same as the kinetic energy of a free gas. Second, the interaction energy is not just the classical interaction energy – comparing to the Hartree-Fock approach we can see that the classical interaction energy is missing the exchange term. So $V_{xc}[n]$ is meant to account for these mistakes. The hope is that these corrections are relatively small. The Hohenberg-Kohn theorems tell us that such a functional exists at least. In practical DFT, some simple approximations to the exchange-correlation energy are made.

Although we argued that $\hat{T}_f[n]$ is completely specified, it may not be so clear how to calculate it or to use this functional. It turns out to be simplest to look directly for minima of the Kohn-Sham functional. We wish to minimize $E_V[n]$ in Eq. (130) for a fixed number of electrons, and so consider the Lagrange functional

$$L[n] = F[n] + \int d^3\boldsymbol{x} V(\boldsymbol{x}) n(\boldsymbol{x}) - \mu \left(\int d^3\boldsymbol{x} n(\boldsymbol{x}) - N \right). \tag{135}$$

We apply the condition $\delta L / \delta n(\boldsymbol{x}) = 0$, which is a functional derivative. This gives

$$\frac{\delta T_f}{\delta n(\boldsymbol{x})} + V(\boldsymbol{x}) + e^2 \int d^3\boldsymbol{x}' \frac{n(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} + \frac{\delta V_{xc}}{\delta n(\boldsymbol{x})} - \mu = 0. \tag{136}$$

This involves two terms which need clarification: the functional derivative of the kinetic energy, and the derivative of the exchange-correlation energy. The latter will depend upon the approximation for this unknown functional. The former we can simplify. Note that from the definition, we can write

$$T_f[n] = E_0^{KS} - \langle \Psi_{KS} | \hat{V}_{KS} | \Psi_{KS} \rangle = E_0^{KS} - \int d^3\boldsymbol{x}' V_{KS}(\boldsymbol{x}') n(\boldsymbol{x}'), \tag{137}$$

where E_0^{KS} is the ground state energy of H_{KS}. So we have

$$\frac{\delta T_f}{\delta n(\boldsymbol{x})} = \frac{\delta E_0^{KS}}{\delta n(\boldsymbol{x})} - \int d^3\boldsymbol{x}' \frac{\delta V_{KS}(\boldsymbol{x}')}{\delta n(\boldsymbol{x})} n(\boldsymbol{x}') - V_{KS}(\boldsymbol{x}). \tag{138}$$

To proceed, we need to make the first term more explicit. To do so, remember that the way in which we actually vary $n(\boldsymbol{x})$ is by varying $V_{KS}(\boldsymbol{x})$. So we can

write

$$\frac{\delta E_0^{KS}}{\delta n(\boldsymbol{x})} = \int d^3\boldsymbol{x}' \frac{\delta E_0^{KS}}{\delta V_{KS}(\boldsymbol{x}')} \frac{\delta V_{KS}(\boldsymbol{x}')}{\delta n(\boldsymbol{x})}. \tag{139}$$

The variation of the energy E_0^{KS} with the potential V_{KS} is just the standard problem of perturbation theory of the ground state energy: for a small change in the Hamiltonian, the ground state energy shifts by the expectation value of that change. In particular, from Eq. (133), we see that

$$\frac{\delta E_0^{KS}}{\delta V_{KS}(\boldsymbol{x}')} = n(\boldsymbol{x}'). \tag{140}$$

Therefore Eq. (139) becomes

$$\frac{\delta E_0^{KS}}{\delta n(\boldsymbol{x})} = \int d^3\boldsymbol{x}' n(\boldsymbol{x}') \frac{\delta V_{KS}(\boldsymbol{x}')}{\delta n(\boldsymbol{x})}. \tag{141}$$

Using Eq. (141) in Eq. (138), we see that the first and second terms actually cancel, and we find simply that

$$\frac{\delta T_f}{\delta n(\boldsymbol{x})} = -V_{KS}(\boldsymbol{x}). \tag{142}$$

This simple result can be viewed as a form of the Legendre transform: the functional $T_f[n]$ is the Legendre transform (via Eq. (137)) of the functional $E_0[V_{KS}]$. In any case, Eq. (142) means that the stationarity condition of Eq. (136) implies just that

$$V_{KS}(\boldsymbol{x}) = V(\boldsymbol{x}) + e^2 \int d^3\boldsymbol{x}' \frac{n(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} + \frac{\delta V_{xc}}{\delta n(\boldsymbol{x})} - \mu. \tag{143}$$

This is the simple and useful final result. The Kohn-Sham potential is explicitly determined from the nuclear potential plus an explicit functional of the density. The latter terms can be thought of as giving a "screened" potential which describes self-consistently the total potential the electrons feel from both the nucleii and the other electrons.

The practical algorithm is then iterative. One starts with some initial guess at the density, and then calculates V_{KS}. This defines the non-interacting H_{KS}, which one can then solve, since it is just a free particle Hamiltonian, i.e. we find the one-particle eigenstates $\phi_n(\boldsymbol{x})$, which satisfy

$$\left(\frac{|\boldsymbol{p}|^2}{2m} + V_{KS}(\boldsymbol{x})\right)\phi_n(\boldsymbol{x}) = \epsilon_n \phi_n(\boldsymbol{x}), \tag{144}$$

and then we can recompute the density by summing over the occupied states:

$$n(\boldsymbol{x}) = 2 \sum_{n|\epsilon_n<\epsilon_F} |\phi_n(\boldsymbol{x})|^2. \tag{145}$$

From the new density, we repeat the same procedure with the new Kohn-Sham potential, and continue iterating in this way until convergence is achieved.

We should probably say a word about the exchange-correlation functional. There are different approximations in use. The original work by Kohn and Sham suggested the "local density approximation" (LDA), in which

$$V_{xc}^{\mathrm{LDA}}[n] = \int d^3\boldsymbol{x}\, \varepsilon_{xc}(n(\boldsymbol{x})), \tag{146}$$

where $\varepsilon_{xc}(n)$ is just a function of n (typically it is written as $\varepsilon_{xc}(n) = \epsilon_{xc}(n)n$). It is taken so that one obtains the correct energy versus density for the case of "Jellium", in which the density is constant in space. This is the simplest model of interacting electrons, in which the nuclei are replaced by a uniform positive charged background whose only role is to make the system electrically neutral. This model has been extensively studied using advanced many body techniques (much more computationally and conceptually sophisticated than DFT) and its ground state energy is very well known, so that the exchange-correlation energy may be extracted. It is of course a major approximation to assume that the only dependence of V_{xc} on variations of the density is in the local form of Eq. (146). However, the LDA seems to work surprisingly well. More modern calculations tend to use improved functionals beyond the LDA, in which the integrand in Eq. (146) is replaced by something which depends not only on the local density but on its gradients, e.g. the "generalized gradient approximation", or GGA. Regardless of the choice, the exchange-correlation part of the Kohn-Sham potential in Eq. (143) is just determined by a calculation from the density, so it is relatively simple.

These steps are incorporated into a large number of free and commercially available density functional software applications. To carry out such calculations one does not need to really understand the formalism, or even to write any numerical code, but just provide some inputs to one of these programs. Such "ab initio" programs have become extremely widespread in condensed matter physics, chemistry, and engineering. In the former context, which is the subject of this class, one will often see "band structures" that are produced from these codes (we will soon discover what these "bands" are in much more detail), which are representations of the Kohn-Sham eigenvalues ϵ_n in Eq. (144) using the converged Kohn-Sham potential. It is probably good to know what it is these software packages really calculate.

It is interesting to compare the Kohn-Sham potential to the potential which appears in Hartree-Fock theory, i.e. in Eqs. (121-124). One can see that the direct potential, i.e. the Hartree term, U_d, is identical to the classical electrostatic energy in the Kohn-Sham potential. The difference between the two is that the non-local exchange/Fock term in Hartree-Fock theory is replaced by the exchange-correlation potential in the Kohn-Sham formulation. The latter is *much* simpler since it involves only the classical density.

4.3 *A cautionary note*

Density functional theory is pervasive and by far the dominant method used in practice to do "realistic" calculations on solids, but it has limitations. There is the obvious one: we do not have an exact exchange-correlation functional. But beyond the obvious is a more general concern. DFT is formulated as a theory in which the density determines the ground state energy, and hence, to the extent that it achieves these goals, it is "reliable" for the density profile and the ground state energy. Even for these quantities it is not exact, and it is not really obvious how to improve it. But more seriously, DFT is routinely used to calculate not only the total ground state energy but the energy and wavefunctions of Kohn-Sham states. The former are usually plotted as the "band structure" calculated from DFT codes (we will discuss bands in the next section). However, there is nothing in the fundamental theory that relates the eigenvalues of the Kohn-Sham Hamiltonian to physical energies. Generally speaking, DFT is not designed to calculated excited states. It is a loose, and sometimes deceptive, interpretation to view the Kohn-Sham eigenvalues and wavefunctions as representative of real excitations. There are many ways to improve on this, but they go beyond the subject of this course. What you should remember is that DFT is far from a complete solution of the physics of electrons in solids, and sometimes it can fail dramatically.

Nevertheless, for now, we will adopt the above naïve and loose interpretation, and assume there is some effective one-particle potential that describes a solid. This situation is still remarkably rich and we can learn a lot from it.

5 THE ONE PARTICLE PROBLEM

5.1 *Bloch's theorem and bands*

Now we are going to take seriously the idea motivated in the previous section, that we can approximately describe a solid by a one-electron Hamiltonian, which we take (for now) to be

$$\mathrm{H} = \frac{\boldsymbol{p}^2}{2m} + \mathrm{V}(\boldsymbol{x}). \tag{147}$$

We will *not* take $\mathrm{V}(\boldsymbol{x})$ in the form of Eq. (105), as it should include at least the direct/Hartree contribution, and this is not easily known (indeed it requires a solution of the charge density). So we will regard it as an unknown potential except for the property of periodicity given in Eqs.112,113. We can get surprisingly far just with this picture in mind. For more specificity, we can try to just think physically about $\mathrm{V}(\boldsymbol{x})$: close to a nucleus it predominantly reflects the nuclear attraction, but as we move outward it gets progressively screened, and once we move to distances "halfway" between atoms it will be smooth and far from atomic.

For the moment, let us just discuss the general case. The Bloch Hamiltonian is periodic, and periodicity is a symmetry under discrete translations. Like

any symmetry in quantum mechanics, it is associated with a unitary operator that enacts the symmetry transformation on states. In quantum class, you also learned that translations are generated by momentum. So indeed we can write the translation operator

$$\hat{T}_{\mathbf{R}} = e^{i\boldsymbol{p}\cdot\mathbf{R}}. \tag{148}$$

Acting on a function of position, we have

$$\hat{T}_{\mathbf{R}} f(\boldsymbol{x}) = f(\boldsymbol{x} + \mathbf{R}). \tag{149}$$

The translation operator is unitary $\hat{T}_{\mathbf{R}}^{\dagger} = \hat{T}_{\mathbf{R}}^{-1} = \hat{T}_{-\mathbf{R}}$ and two translations compose

$$\hat{T}_{\mathbf{R}}\hat{T}_{\mathbf{R}'} = \hat{T}_{\mathbf{R}+\mathbf{R}'}. \tag{150}$$

This in turn implies translations commute with one another. They also obviously commute with momentum. So we have

$$\hat{T}_{\mathbf{R}}^{\dagger} H \hat{T}_{\mathbf{R}} = H, \qquad \mathbf{R} \in \mathrm{BL}. \tag{151}$$

This is equivalent to

$$\left[H, \hat{T}_{\mathbf{R}}\right] = 0, \qquad \mathbf{R} \in \mathrm{BL}. \tag{152}$$

As you learned in quantum mechanics, you can simultaneously diagonalize any set of commuting observables. We see that H and all the translation operators constitute such a set. That is, the energy eigenstates can be chosen as eigenstates of discrete translations. From Eq. (150), the translation operators are not independent; essentially, we need only consider translation operators by linearly independent translations, e.g. by the primitive vectors $\boldsymbol{a}_\mu$. So we can take our states to be eigenstates of $\hat{T}_\mu = \hat{T}_{\boldsymbol{a}_\mu}$. Since these are unitary operators, their eigenvalues are complex numbers with absolute value one, i.e. we have states

$$\hat{T}_\mu |\theta_1 \cdots \theta_d\rangle = e^{i\theta_\mu} |\theta_1 \cdots \theta_d\rangle, \tag{153}$$

where θ_μ are phases defined modulo 2π. For a general Bravais lattice vector,

$$\mathbf{R} = \sum_\mu n_\mu \boldsymbol{a}_\mu, \tag{154}$$

we have

$$\hat{T}_{\mathrm{R}} |\theta_1 \cdots \theta_d\rangle = e^{i\sum_\mu n_\mu \theta_\mu} ||\theta_1 \cdots \theta_d\rangle. \tag{155}$$

Now we can use the inner product of the $\boldsymbol{b}_\mu$ vectors from Eq. (38) to obtain the

integers n_μ from Eq. (154) as $n_\mu = \boldsymbol{b}_\mu \cdot \mathbf{R}/(2\pi)$. This then gives the expression

$$\hat{T}_\mathbf{R}|\theta_1 \cdots \theta_d\rangle = e^{i\sum_\mu \frac{\theta_\mu}{2\pi}\boldsymbol{b}_\mu\cdot\mathbf{R}}|\theta_1 \cdots \theta_d\rangle, \tag{156}$$

which we can write as

$$\hat{T}_\mathbf{R}|\boldsymbol{k}\rangle = e^{i\boldsymbol{k}\cdot\mathbf{R}}|\boldsymbol{k}\rangle, \tag{157}$$

with

$$\boldsymbol{k} = \sum_\mu \frac{\theta_\mu}{2\pi}\boldsymbol{b}_\mu. \tag{158}$$

Here $\boldsymbol{k}$ appears like a momentum, but because θ_μ are phases, it is defined only up to the addition of an integer linear combination of the $\boldsymbol{b}_\mu$ vectors, i.e. $\boldsymbol{k}$ is defined up to a reciprocal lattice vector. For this reason we call it the *quasi-momentum* or *crystal momentum*. Let's apply Eq. (157) to the wave function. It states that

$$\psi(\boldsymbol{x}+\mathbf{R}) = e^{i\boldsymbol{k}\cdot\mathbf{R}}\psi(\boldsymbol{x}). \tag{159}$$

We can arrive at a standard form by writing defining $u(\boldsymbol{x}) = \psi(\boldsymbol{x})e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}$, or

$$\psi(\boldsymbol{x}) = e^{i\boldsymbol{k}\cdot\boldsymbol{x}}u(\boldsymbol{x}). \tag{160}$$

Then Eq. (159) implies that

$$u(\boldsymbol{x}+\mathbf{R}) = u(\boldsymbol{x}), \tag{161}$$

which simply states that u has the periodicity of the Bravais lattice. In turn, then Eq. (160) means that the eigenstates of quasi-momentum are of the form of a plane wave multiplied by a periodic function. This is a remarkable result: the energy eigenstates of Eq. (147) are generically plane waves with an amplitude that is modulated just within the unit cell. This might be a surprising result: all these electron eigenstates are extended, despite the fact that they may feel strong potentials from the ions, which might be expected to form bound states. This fact is crucial to the existence of metals: the extended nature of the states means that electrons can propagate over long distances, and hence carry current.

It is customary to label the eigenstates explicitly by their quasi-momentum, i.e. to write Eq. (160) as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{x}) = e^{i\boldsymbol{k}\cdot\boldsymbol{x}}u_{n\boldsymbol{k}}(\boldsymbol{x}), \tag{162}$$

where we introduce an additional index n to represent any other quantum

numbers, i.e. to label multiple states with the same quasi-momentum. We will soon see these indeed exist. Eq. (162) is known as the Bloch form, and that this is true is called Bloch's theorem. It can be important that Eq. (162) has some ambiguity. First, it has the usual phase arbitrariness, which can be chosen separately for every eigenstate, i.e. for every $\boldsymbol{k}$ and n. There is a further ambiguity, however, due to the fact that $\boldsymbol{k}$ is defined only modulo a RLV. The plane wave factor in Eq. (162), however, is not independent of a shift of the $\boldsymbol{k}$ by a RLV. Rather, such a shift can be absorbed in a redefinition of $u_{n\boldsymbol{k}}(\boldsymbol{x})$, that is, $\psi_{n\boldsymbol{k}}(\boldsymbol{x})$ is unchanged under

$$\boldsymbol{k} \to \boldsymbol{k} + \mathbf{Q}, \qquad u_{n\boldsymbol{k}}(\boldsymbol{x}) \to e^{-i\mathbf{Q}\cdot\boldsymbol{x}} u_{n\boldsymbol{k}}(\boldsymbol{x}), \tag{163}$$

valid for any RLV $\mathbf{Q}$.

There are many ways to understand the Bloch form. A simple way follows from Fourier analysis: any periodic function can be written as a Fourier series in harmonics. For a function with the periodicity of the Bravais lattice, these harmonics are just the set of reciprocal lattice vectors. Hence one can write

$$u_{n\boldsymbol{k}}(\boldsymbol{x}) = \sum_{\mathbf{Q}\in\mathrm{RL}} \tilde{u}_{n\boldsymbol{k};\mathbf{Q}} e^{i\mathbf{Q}\cdot\boldsymbol{x}}, \tag{164}$$

where $\tilde{u}_{n\boldsymbol{k};\mathbf{Q}}$ are Fourier coefficients. Inserting this into Eq. (162) we have

$$\psi_{n\boldsymbol{k}}(\boldsymbol{x}) = \sum_{\mathbf{Q}\in\mathrm{RL}} \tilde{u}_{n\boldsymbol{k};\mathbf{Q}} e^{i(\boldsymbol{k}+\mathbf{Q})\cdot\boldsymbol{x}}. \tag{165}$$

Such a plane wave expansion can be a useful way to solve the Schrödinger equation. Conceptually, Eq. (172) has a simple interpretation in terms of scattering. We can imagine injecting an electron into the solid with momentum $\boldsymbol{k}$. Since in quantum mechanics, an electron behaves as a wave, the Bragg scattering theory applies to it. According to Sec. 2.2.1, there is an amplitude for the electron to scatter from the crystal lattice with any scattering wavevector in the reciprocal lattice. The term $\tilde{u}_{n\boldsymbol{k};\mathbf{Q}}$ is just the amplitude for this wave to scatter by momentum $\mathbf{Q}$. Since eventually all the RLVs are present in the wavefunction, one can view n as specifying the "initial" momentum out of all this set.

One can insert the Bloch form into the Schrödinger equation,

$$\left(\frac{\boldsymbol{p}^2}{2m} + \mathrm{V}(\boldsymbol{x})\right)\psi_{n\boldsymbol{k}}(\boldsymbol{x}) = \epsilon_{n\boldsymbol{k}}\psi_{n\boldsymbol{k}}(\boldsymbol{x}) \Rightarrow$$

$$\mathcal{H}_{\boldsymbol{k}} u_{n\boldsymbol{k}} \equiv \left(\frac{(\boldsymbol{p}+\boldsymbol{k})^2}{2m} + \mathrm{V}(\boldsymbol{x})\right) u_{n\boldsymbol{k}}(\boldsymbol{x}) = \epsilon_{n\boldsymbol{k}} u_{n\boldsymbol{k}}(\boldsymbol{x}). \tag{166}$$

Eq. (166), in combination with periodic boundary conditions on $u_{n\boldsymbol{k}}(\boldsymbol{x})$ due to Eq. (161), defines a quantum mechanics problem for a finite system, i.e.

entirely within a unit cell, like a particle in a box with periodic boundary conditions. The associated operator $\mathcal{H}_{\boldsymbol{k}}$ is called the Bloch Hamiltonian. In this problem, the quasimomentum appears just as a parameter. Consequently, the energy levels defined by Eq. (166) for fixed $\boldsymbol{k}$ form a discrete but infinite series – hence the n index. The solutions must be periodic – up to the ambiguities already mentioned – in $\boldsymbol{k}$ with the periodicity of the reciprocal lattice. We can choose to define n so that the states are consecutive in energy, i.e. $\epsilon_{1\boldsymbol{k}} \leq \epsilon_{2\boldsymbol{k}} \cdots$.

The solutions are called bands. The name comes from the fact that the energy $\epsilon_{n\boldsymbol{k}}$, for fixed n, is a continuous function of $\boldsymbol{k}$ and periodic, and hence extends just over some range of energy which is bounded both below and above – a "band" of energy. We will return to think more carefully about the parametric dependence of eigenstates and eigenvalues on $\boldsymbol{k}$ soon.

There are many ways to represent bands. The most compact is to eliminate the ambiguity in the quasimomentum by choosing a unit cell in momentum space. Generally one chooses the (1st) Brillouin zone. This choice is known as the reduced zone scheme. Recall that any momentum can be translated back into the Brillouin zone by shifting by a RLV. This means that while we can label states by any $\boldsymbol{k}$, if we count states with quasimomentum $\boldsymbol{k}$ and $\boldsymbol{k} + \mathbf{Q}$ separately (with $\mathbf{Q}$ a RLV), then we are counting the same state twice. By choosing the reduced zone scheme, we count every state exactly once.

5.2 *Nearly free electron bands*

Nothing we have done excludes the trivial case in which the periodic potential vanishes, $V(\boldsymbol{x}) = 0$, i.e. free electrons. Then obviously the states with *momentum* $\boldsymbol{k}$ have energy $\epsilon(\boldsymbol{k}) = k^2/2m$. This describes the eigenstates in terms of true momentum, not quasimomentum, and there is one state (per spin) for each value of the momentum. We are, however, allowed to follow the Bloch conventions and view the spectrum in terms of quasi-momentum. This is useful because the latter description becomes necessary when the potential is not zero. Then we can adopt the reduced zone scheme. For any given true momentum, we can find the unique quasi-momentum to which it corresponds by translating it back by a RLV into the Brillouin zone. This is called "folding" the bands. In this picture, one might imagine tiling all of momentum space with translated Brillouin zones, and every one of these copies gives rise to a band in the reduced zone scheme. Obviously there are an infinite number of such bands. We illustrate this for free electrons in one dimension in the left panel of Fig. 4.

Note that for free electrons, there are many places in which bands cross. These crossings occur because there are momenta with the same single particle energy (i.e. same magnitude of the momenta) which differ by a reciprocal lattice vector. The free electron states at these different momenta cannot mix because momentum is a good quantum number for free electrons. However, in quantum mechanics, generally levels that do not have different quantum numbers avoid being degenerate: this is the phenomena of *level repulsion*. If we

turn on the periodic potential, typically these pairs of degenerate levels, which differ in momentum but *not* in quasi-momentum, will be able to mix and the levels will repel. Consequently such crossings typically become "avoided crossings" when the periodic potential is included. This means that instead of two bands crossing, these bands become separated by an *energy gap*. This occurs unless there is some other symmetry (besides momentum conservation) which can protect the crossing of these levels. *Nearly* free electron bands in one dimension are shown in the right panel of Fig. 4.

One thing to take from this simple analysis is that *if* the periodic potential is weak, the bands are quite close to their free electron forms except near the band crossings. This is true in all dimensions. Consequently there are circumstances, e.g. for the alkali metals in particular and to some extend for the noble metals like gold, that just a free electron gas without any periodic potential is a reasonable approximation, at least for the bands which lie at the Fermi energy.

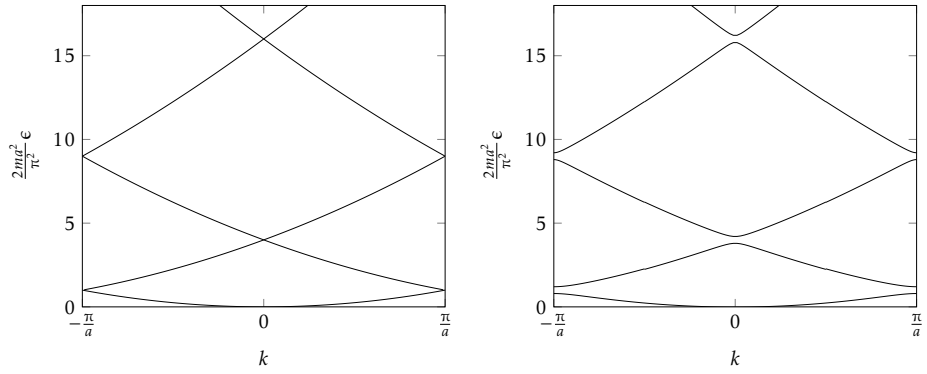


Figure 4: Left: one dimensional free electron bands in the reduced zone scheme, for a lattice constant a. Right: nearly *free electron bands for the same structure, showing gaps due to level repulsion at Bragg planes.*

5.3 *Tight binding bands*

When the periodic potential is strong rather than weak, the situation is opposite to the previous subsection. This limit is described by what is called a *tight-binding model*. Instead of starting from plane waves, we start from localized atomic orbitals.

The idea is to project the full Hamiltonian into the basis of some set of localized orbitals at different atomic sites. These may be thought of obtained by expanding around the minima of the potential $V(\boldsymbol{x})$. In the general tight binding model, we introduce some basis of localized states $|i\rangle$, where i may labels sites and orbitals. Then we write the Hamiltonian as

$$H = \sum_{ij} h_{ij} |i\rangle\langle j|, \tag{167}$$

which is simply a matrix in this space. When the two indices coincide, h_{ii} represents the energy of the orbital i. When $i \neq j$, the term h_{ij}, for $i \neq j$, describes hopping of an electron from an orbital j to another orbital i. In time-reversal symmetric systems without spin-orbit coupling, h_{ij} is real and is most often *negative*, which reflects the fact that electrons lower their kinetic energy by delocalizing. Note however that the sign (and indeed phase) of h_{ij} depends upon a sign/phase convention for the orbitals. The tight-binding description is useful when the overlap between different orbitals is negligible when the orbitals are far apart. This is typically true for atomic wavefunctions due to their exponential decay, and in many cases we can be content with only a few orbitals per unit cell and only overlaps h_{ij} between those localized on very nearby atoms.

We will illustrate the tight-binding model using the by now very popular and standard description of graphene. One includes just a single π (p^z) orbital on each site of a honeycomb lattice. You can find this discussed in many many places, for example this Reviews of Modern Physics article.

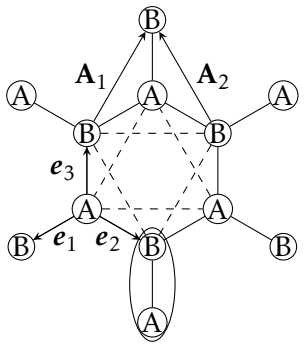


Figure 5: A hexagon of the honeycomb lattice, including all nearest-neighbors of the sites on the hexagon. Representative second *neighbor bonds are shown with dashed lines. Two linearly independent Bravais lattice (translation) vectors* $\mathbf{A}_1$, $\mathbf{A}_2$ *are shown, as are the three nearest-neighbor vectors* $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$. *A unit cell consists of a pair of A and B sites, one of which is enclosed by an ellipse.*

The geometry is shown in Fig. 5. The lattice sites are divided into A and B sublattices, connected by nearest-neighbor bonds shown as solid lines. We define a unit cell containing two sites on a vertical bond, for example the pair in the ellipse drawn in the figure. A site is indexed by the coordinate of its unit cell, which we take to be the coordinate of the A site within that cell, and the sublattice $s = 1, 2 =$ A, B. The A sites are then located at the sites of the triangular Bravais lattice, for which we may take $\mathbf{A}_1$ and $\mathbf{A}_2$ shown in the figure as primitive lattice vectors. We define also the three nearest-neighbor vectors $\boldsymbol{e}_i$, $i = 1, 2, 3$ as shown. One can see that $\mathbf{A}_1 = \boldsymbol{e}_3 - \boldsymbol{e}_1 = -2\boldsymbol{e}_1 - \boldsymbol{e}_2$ and $\mathbf{A}_2 = \boldsymbol{e}_3 - \boldsymbol{e}_2 = -2\boldsymbol{e}_2 - \boldsymbol{e}_1$. The basis vectors $\mathbf{B}_1$ and $\mathbf{B}_2$ of the reciprocal lattice are defined by $\mathbf{B}_i \cdot \mathbf{A}_j = 2\pi\delta_{ij}$ as usual. If we define vectors $\boldsymbol{b}_i$ such that $\boldsymbol{b}_i \cdot \boldsymbol{e}_j = 2\pi\delta_{ij}$ for $i, j = 1, 2$, then we can find that $\mathbf{B}_1 = (-2\boldsymbol{b}_1 + \boldsymbol{b}_2)/3$, and $\mathbf{B}_2 = (-2\boldsymbol{b}_2 + \boldsymbol{b}_1)/3$. The resulting Brillouin zone with characteristic

wavevectors labeled is shown in Fig. 6.

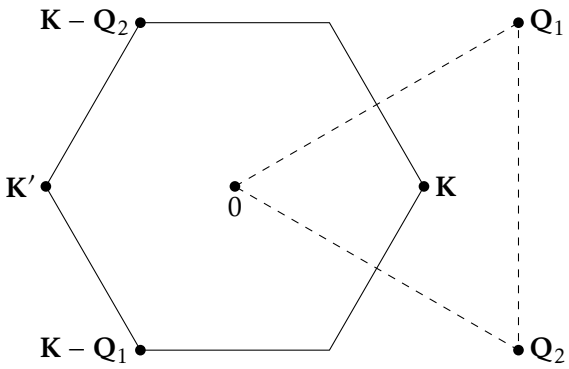


Figure 6: Graphene Brillouin zone and some other useful wavevectors. The wavevectors $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$ *are basis vectors for the reciprocal lattice. The* $\mathbf{K}$ *point is the centroid of the triangle formed by the origin,* $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$*. The two other Brillouin zone corners* $\mathbf{K}-\mathbf{Q}_1$ *and* $\mathbf{K}-\mathbf{Q}_2$ *are equivalent to* $\mathbf{K}$ *as quasimomenta, and are obtained from the latter by* C_3 *rotations.*

With all these definitions, it is straightforward to write down the Bloch Hamiltonian for the nearest-neighbor model. In real space, we define kets $|\mathbf{X}, \mathrm{A}\rangle$ and $|\mathbf{X}, \mathrm{B}\rangle$ labeling states at the A,B sites, respectively, within the unit cell labeled by the Bravais lattice vector $\mathbf{X}$. The tight binding Hamiltonian is

$$\mathrm{H} = -t \sum_{\mathbf{X}\in\mathrm{BL}} \Big[\left(|\mathbf{X}, \mathrm{B}\rangle + |\mathbf{X}-\mathbf{A}_1, \mathrm{B}\rangle + |\mathbf{X}-\mathbf{A}_2, \mathrm{B}\rangle\right) \langle \mathbf{X}, \mathrm{A}| + \left(|\mathbf{X}, \mathrm{A}\rangle + |\mathbf{X}+\mathbf{A}_1, \mathrm{A}\rangle + |\mathbf{X}+\mathbf{A}_2, \mathrm{A}\rangle\right) \langle \mathbf{X}, \mathrm{B}| \Big]. \tag{168}$$

Now we obtain the Bloch Hamiltonian by applying the Bloch ansatz,

$$|\psi_{n\boldsymbol{k}}\rangle = \sum_{\mathbf{X}} e^{i\boldsymbol{k}\cdot\mathbf{X}} \left(u_{\mathrm{A}\boldsymbol{k}}|\mathbf{X}, \mathrm{A}\rangle + u_{\mathrm{B}\boldsymbol{k}}|\mathbf{X}, \mathrm{B}\rangle\right). \tag{169}$$

Then we obtain the Schrödinger equation

$$\mathcal{H}_{\boldsymbol{k}} \begin{pmatrix} u_{\mathrm{A}\boldsymbol{k}} \\ u_{\mathrm{B}\boldsymbol{k}} \end{pmatrix} = \epsilon \begin{pmatrix} u_{\mathrm{A}\boldsymbol{k}} \\ u_{\mathrm{B}\boldsymbol{k}} \end{pmatrix}, \tag{170}$$

with the Bloch Hamiltonian

$$\mathcal{H}_{\boldsymbol{k}} = -t \begin{pmatrix} 0 & 1 + e^{-i\boldsymbol{k}\cdot\mathbf{A}_2} + e^{-i\boldsymbol{k}\cdot\mathbf{A}_1} \\ 1 + e^{i\boldsymbol{k}\cdot\mathbf{A}_2} + e^{i\boldsymbol{k}\cdot\mathbf{A}_1} & 0 \end{pmatrix} = \begin{pmatrix} 0 & f(\boldsymbol{k}) \\ f^*(\boldsymbol{k}) & 0 \end{pmatrix}, \tag{171}$$

with

$$f(\boldsymbol{k}) = -t\left(1 + e^{-i\boldsymbol{k}\cdot\mathbf{A}_2} + e^{-i\boldsymbol{k}\cdot\mathbf{A}_1}\right). \tag{172}$$

We see that $\mathcal{H}_{\boldsymbol{k}}$ is indeed periodic and smooth, as expected. The energy eigenvalues are simply

$$\epsilon_{\pm,\boldsymbol{k}} = \pm|f(\boldsymbol{k})|. \tag{173}$$

Band touching occurs when $f(\boldsymbol{k})$ vanishes identically, i.e. both real and imaginary parts. This occurs at the two inequivalent corners of the Brillouin zone, denoted $\mathbf{K}$ and $\mathbf{K}'$ (see Fig. 6). If we take the length of the nearest-neighbor bond to be unity, then $\mathbf{A}_1 = (\sqrt{3}/2, 3/2)$, $\mathbf{A}_2 = (-\sqrt{3}/2, 3/2)$, and it is easily verified that f vanishes at the points $\mathbf{K} = (4\pi/(3\sqrt{3}), 0)$, $\mathbf{K}' = -\mathbf{K}$. Taylor expanding, we have

$$f(\pm\mathbf{K} + \boldsymbol{k}) \sim \frac{3}{2}t(\pm k_x - ik_y). \tag{174}$$

This implies from Eq. (173) that the bands cross linearly in the vicinity of the touching points $\pm$K,

$$\epsilon_{\pm,\boldsymbol{k}} \sim \pm v|\boldsymbol{k}|, \tag{175}$$

with $v = 3t/2$. This form of dispersion is called a "Dirac cone", due to the similarity with the behavior of a massless relativistic particle described by the Dirac equation. The speed of light is replaced by the "Dirac velocity" v (which in graphene is about 100 times slower).

5.4 *Density of states*

The simplest view of energy bands is to consider only their energies, and ignore the momentum structure. One can define the density of states of a band by summing over all momentum:

$$\mathrm{D}_n(\epsilon) = 2\int \frac{d^d\boldsymbol{k}}{(2\pi)^d}\delta(\epsilon - \epsilon_{n\boldsymbol{k}}). \tag{176}$$

Here we include a factor of 2 for spin degeneracy. In other circumstances, in which there is no spin degeneracy, one would not include this factor. This is defined in such a way that $\mathrm{D}_n(\epsilon)d\epsilon$ gives the number of states *per unit volume in real space* in band n, including spin degeneracy, with energy between ϵ and $\epsilon + d\epsilon$. The full density of states is the sum of the density of states from each band,

$$\mathrm{D}(\epsilon) = \sum_n \mathrm{D}_n(\epsilon). \tag{177}$$

In general, because the energy of each band is bounded on both sides,

$$\epsilon_{n,\min} \leq \epsilon_{n\boldsymbol{k}} \leq \epsilon_{n,\max}, \tag{178}$$

the density of states for a single band has finite support:

$$\mathrm{D}_n(\epsilon) = 0, \qquad \epsilon > \epsilon_{n,\max} \text{ or } \epsilon < \epsilon_{n,\min}. \tag{179}$$

So the support of $\mathrm{D}_n(\epsilon)$ is over some finite band of energy. This means only a small number of bands contribute to $\mathrm{D}(\epsilon)$ at any given energy. The total weight in D_n, i.e. its integral, has a very nice property. Integrating Eq. (441) over energy, one obtains

$$\int d\epsilon\, \mathrm{D}_n(\epsilon) = 2 \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} = \frac{2}{(2\pi)^d} \mathrm{V}_{\mathrm{BZ}}. \tag{180}$$

Using Eq. (42), we find

$$\int d\epsilon\, \mathrm{D}_n(\epsilon) = \frac{2}{\mathrm{V}_{\mathrm{p.u.c.}}}. \tag{181}$$

This is important: **the number of electrons which fill a single band is 2 (or more generally, the degeneracy of the band) per unit cell.**

The importance of this simple counting result becomes clear if one consider the zero temperature limit. In this case, the many body ground state is found by filling the Bloch states from lowest energy on up, until we account for all the electrons. The energy half-way between the highest energy filled state and the lowest energy empty one is called the Fermi energy. There are two possibilities. In the first case, the two aforementioned energies are different, and the Fermi energies lies in a gap, i.e. a region with zero density of states. *This condition defines an insulator in band theory.* We will explore why this is the case below, but intuitively, such a system is resistant to changing its state, because to do so an electron must aquire a non-zero energy to change its state, which can only occur by moving from below to above the gap. This means that a system with the Fermi level in a band gap is unresponsive to weak applied forces, e.g. electric fields, and hence does not conduct.

In the second case, the two energies coincide, and the Fermi energy lies within one or more bands, and/or it lies precisely at the boundary or two or more bands, at least one of which has the Fermi energy as its maximum and another as its minimum. If the Fermi energy lies within at least one band, the system is a metal according to band theory. If it lies precisely at min/max of two or more bands, it is a "zero gap semiconductor" or "semimetal" (both words are used without uniformly accepted definitions so take care).

The counting result in Eq. (181) has strong implications for insulators. For an ideal material, the number of electrons per unit cell is simply equal to the total atomic number of all the atoms in that unit cell. This is necessarily an

integer, but it may be even or odd. Typically it is a large integer, so the first few bands will be completely full. The top-most bands containing electrons may be either full or partly full. There will be infinitely many empty bands. The above result says that every full band contributes an even integer to total number of electrons. In a band insulator, all bands are either full or empty, so the total number of electrons per (primitive) unit cell in a band insulator must be even. Indeed, the same counting applies to the zero gap semi-conductor or semimetal situation above: these too must have an even number of electrons per unit cell. As a result, *any crystal with an odd number of electrons per (primitive) unit cell must be a metal, according to band theory*. This is a remarkable result, and explains the metallicity of many materials, and the commonality of metals. It should be said that the result is not entirely reliable, most importantly because since it assumes band theory, which is itself approximate. It also assumes spin degeneracy, but we will see later that even with spin-orbit coupling, there is often enough degeneracy that the statement still holds. Nevertheless, it is correct enough that it is quite hard to find exceptions. When you think about it, when this criterion predicts a material is a metal, it is doing so entirely due to the spin-1/2 nature of the electron, and due to the Pauli exclusion principle. It is striking that these seemingly exotic properties of electrons (spin arises ultimately from the relativistic Dirac equation, and the Pauli principle from the quantum indistinguishability of identical particles) contribute to something so mundane as a metal.

I would also like to be clear that the above argument does not mean that a material with an even number of electrons per unit cell must be a band insulator or semi-metal. It only means that it can be, within band theory. There are many metals that have an even number of electrons per unit cell. This is totally compatible with band theory. There are also insulators with an odd number of electrons per unit cell. These are usually called Mott insulators, and they are *not* compatible with band theory. This is because they are insulating precisely due to the electron-electron correlations which are neglected in band theory.

6 PHYSICS FROM BANDS

6.1 *Thermodynamics*

6.1.1 Specific heat and Sommerfeld law

Let us return now to band theory and explore some of its implications in more detail. First consider thermodynamics, which are determined entirely by the density of states (in the independent electron approximation). For example, the chemical potential μ is fixed by the electron density,

$$n = \frac{N_e}{V} = \int d\epsilon \, D(\epsilon) n_F(\epsilon - \mu), \tag{182}$$

where $n_F(\epsilon - \mu)$ is the Fermi function

$$n_F(\epsilon - \mu) = \frac{1}{e^{\beta(\epsilon-\mu)} + 1}, \tag{183}$$

with $\beta = 1/(k_B T)$, k_B is Boltzmann's constant. The zero temperature limit of the chemical potential is the Fermi energy, $\epsilon_F = \lim_{T\to 0} \mu(T)$.

We now want to obtain the specific heat, which is the derivative of the internal energy density with respect to temperature. We can in turn obtain the internal energy by differentiating the free energy, and so a nice way to calculate the specific heat is to calculate the free energy and take two derivatives. There is a small subtlety related to the temperature dependence of the chemical potential. A reader not interested in details may skip to Eq. (190).

It is convenient to define the *grand potential* $\Phi_G = -k_B T \ln \mathrm{Tr}[\exp(-\beta(H - \mu N))]$, i.e. the potential obtained from the grand canonical partition function. The corresponding density is $\phi_G = \Phi_G/V$. By expressing the energy (eigenvalues of H) in terms of fermion number, one obtains the expression

$$\phi_G = -k_B T \int d\epsilon\, D(\epsilon) \ln\left(1 + e^{-\beta(\epsilon-\mu)}\right). \tag{184}$$

In principle, we are more interested in the Helmholtz free energy, which is obtained from the canonical ensemble at fixed electron number, $F = -k_B T \ln \mathrm{Tr}_N(\exp(-\beta H))$, and corresponding density $f = F/V$. This is because due to charge neutrality, the density of electrons in the material cannot change. Unfortunately, ϕ_G is much easier to work with than f. However, it turns out that they give the same specific heat at low temperature. To see this, we need to use some thermodynamics. First, we note that F and Φ_G are related by Legendre transforms:

$$F(N, T) = \Phi_G(\mu(N, T), T) + \mu(N, T)N, \tag{185}$$

and importantly that

$$\frac{\partial \Phi_G}{\partial \mu} = -N. \tag{186}$$

It follows that

$$\frac{\partial F}{\partial T} = \frac{\partial \Phi_G}{\partial T} \tag{187}$$

Taking a derivative of this equation and dividing by volume gives

$$\frac{\partial^2 f}{\partial T^2} = \frac{\partial^2 \phi_G}{\partial T^2} + \frac{\partial^2 \phi_G}{\partial T \partial \mu} \left.\frac{\partial \mu}{\partial T}\right|_n = \frac{\partial^2 \phi_G}{\partial T^2} - \left.\frac{\partial n}{\partial T}\right|_\mu \left.\frac{\partial \mu}{\partial T}\right|_n, \tag{188}$$

where $n = N/V$. Now from the definition of the canonical partition function,

the specific heat at constant N, V is

$$c_v = -\mathrm{T}\frac{\partial^2 f}{\partial \mathrm{T}^2} = -\mathrm{T}\frac{\partial^2 \phi_\mathrm{G}}{\partial \mathrm{T}^2} - \mathrm{T}\left.\frac{\partial n}{\partial \mathrm{T}}\right|_\mu \left.\frac{\partial \mu}{\partial \mathrm{T}}\right|_n . \tag{189}$$

One can show that the second term on the right hand side of Eq. (189) is proportional to T^3 in the limit of small temperature. This makes it negligible, as we will see that the first term is linear in temperature. Therefore, to obtain the leading low temperature behavior, we can approximate

$$c_v \sim -\mathrm{T}\frac{\partial^2 \phi_\mathrm{G}}{\partial \mathrm{T}^2} = \frac{1}{4k_\mathrm{B}\mathrm{T}^2}\int d\epsilon\, \mathrm{D}(\epsilon)\frac{(\epsilon-\mu)^2}{\cosh^2(\frac{\epsilon-\mu}{2k_\mathrm{B}\mathrm{T}})}. \tag{190}$$

One can see that at low temperature, the integrand decays exponentially except in the region where $\epsilon \approx \mu$, because the cosh in the denominator grows exponentially at large argument. This leads to an exponentially small electronic specific heat in insulators, for which the Fermi energy is in a gap where $\mathrm{D}(\epsilon) = 0$, and one can see that $c_v \sim e^{-\mathrm{E}_g/(2k_\mathrm{B}\mathrm{T})}$, where E_g is the energy difference between the lowest unoccupied state and the highest occupied one, since $|\epsilon - \mu| = \mathrm{E}_g/2$ at low T. Generally speaking all thermodynamic quantities display such exponential in 1/T behavior, called an activated or Arrhenius law, in band insulators.

In a metal, the density of states is non-zero for $\epsilon = \epsilon_\mathrm{F}$. We can then usefully change variables to to $x = (\epsilon - \mu)/(k_\mathrm{B}\mathrm{T})$ and obtain

$$\begin{aligned} c_v &= \frac{k_\mathrm{B}^2\mathrm{T}}{4}\int dx\, \mathrm{D}(\mu + k_\mathrm{B}\mathrm{T}x)\frac{x^2}{\cosh^2(x/2)} \\ &\sim \frac{k_\mathrm{B}^2\mathrm{T}}{4}\mathrm{D}(\epsilon_\mathrm{F})\int dx\, \frac{x^2}{\cosh^2(x/2)} = \frac{\pi^2}{3}\mathrm{D}(\epsilon_\mathrm{F})k_\mathrm{B}^2\mathrm{T}, \end{aligned} \tag{191}$$

where in the second line we assumed $k_\mathrm{B}\mathrm{T}$ is much smaller than the scale of energy variation of $\mathrm{D}(\epsilon)$, and replaced $\mu(\mathrm{T}) \approx \mu(0) = \epsilon_\mathrm{F}$, and then evaluated the integral. The result is known as the Sommerfeld law of the heat capacity of a metal. The metal has a low temperature specific heat which is linear in temperature, and proportional to the density of states at the Fermi energy. This behavior is often written as

$$c_v \sim \gamma \mathrm{T}, \tag{192}$$

where $\gamma = \frac{\pi^2}{3}k_\mathrm{B}^2\mathrm{D}(\epsilon_\mathrm{F})$ is called the Sommerfeld coefficient.

It is useful to put this in perspective. The specific heat generally is related to the entropy S:

$$S(T) = \int_0^T dT' \frac{C_v(T')}{T'}, \tag{193}$$

using thermodynamics. We see that the Sommerfeld law implies that in a metal, the entropy is linear in temperature at low T ($S/V \sim \gamma T$). The third law of thermodynamics requires that the entropy must vanish as $T \to 0$, and a power-law behavior with temperature generally is indicative of the presence of gapless excitations immediately above the ground state. The smaller the power law, the more low energy excitations there are. In fact, the first power of temperature actually represents one of the smallest exponents for entropy that occur in nature in the zero temperature limit, i.e. the *largest* possible number of low energy excitations in any quantum system. In this sense, the Sommerfeld heat capacity is large.

The linear temperature dependence of the heat capacity has a simple explanation, which elucidates the low energy excitations responsible for it. For a free Fermi system, each Bloch state is occupied with a probability determined by the Fermi function. The Fermi function at low temperature is very close to a step function, so the occupation of states is changed only in a narrow window of energy of width $k_B T$ around the Fermi energy. Within this window, electrons are transferred from below to above the Fermi energy. Each electron transferred in this way increases the energy by of order $k_B T$, and the total number of electrons transferred is the number of states in the narrow energy window proportional to $D(\epsilon_F) k_B T$, so the total energy increase relative to the ground state is $\delta E/V \sim D(\epsilon)(k_B T)^2$. Differentiating this gives the same behavior as Eq. (191).

This argument suggests that in a different sense than the power law, the Sommerfeld heat capacity is small. This is in the sense of the prefactor. Knowing that the density of states at the Fermi energy comes from a small number of bands that overlap that energy, we can estimate that $D(\epsilon_F)$ is of the order of the density of states of a single band. This can be estimated by the counting result of Eq. (181). For a band of width W in energy, the typical magnitude of the density of states should be 1/W times the right hand side of Eq. (181), hence we expect

$$|D_n(\epsilon)| \sim \frac{2}{W V_{\text{p.u.c.}}}. \tag{194}$$

For a *typical* metal, the band-width W of bands near the Fermi energy is of order a few eV. We see that

$$c_v V_{\text{p.u.c.}} \sim \frac{2\pi^2}{3} k_B \frac{k_B T}{W}. \tag{195}$$

The left hand side is the heat capacity per formula unit. It is proportional to the dimensionless ratio of $k_B T$ to W. This can be understood from the

argument above as representing the fraction of the electrons which can be excited at this low temperature. Since W is so large, this ratio is always small, even at room temperature.

How can the electronic specific heat of a metal be both large and small? The answer is that at most temperatures, the electronic specific heat is small because the entropy of the electrons is released very gradually with temperature over a very wide range of energies set by W, which is of order eV. There is a second smallness because even over this temperature range, only the topmost "valence" electrons release their entropy. This is typically just a few electrons per unit cell. The core electrons, i.e. the lower bands, are more strongly bound, and release their entropy only at temperatures sufficient to ionize the core levels of the atoms, i.e. never! There is another source of entropy in solids, which is associated to the positions of the nuclei. This turns out to be much larger in most cases, for most temperatures, because *both* the aforementioned effects are more favorable for them. We will return to this later. The electronic specific heat however becomes dominant at low enough temperature, because it vanishes more slowly as $T \to 0$ than does the lattice contribution. Ultimately this difference is due to Bose statistics of the lattice excitations, called phonons.

6.1.2 Pauli spin susceptibility

Another thermodynamic property of metals is the Pauli spin susceptibility. In general, an applied magnetic field couples to electrons both through the Zeeman and the orbital interactions. The latter is more complicated, and we defer it for now. The former is relatively simple. We add to the Hamiltonian the term

$$H_Z = -g\mu_B B \sum_i S_i^z, \tag{196}$$

taking the field along the z axis. This splits the spin-degenerate bands from $\epsilon_{n\boldsymbol{k}} \to \epsilon_{n\boldsymbol{k}\sigma}$, with

$$\epsilon_{n\boldsymbol{k}\sigma} = \epsilon_{n\boldsymbol{k}} - \frac{1}{2} g\mu_B B\sigma, \tag{197}$$

with $\sigma = \pm 1$. Consequently the density of states must be split into different spin components,

$$D_n(\epsilon) \to \frac{1}{2} \sum_\sigma D_{n\sigma}(\epsilon), \tag{198}$$

where the factor of 1/2 is to remove the spin degeneracy, which is now counted explicitly.

$$D_{n\sigma}(\epsilon) = \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} \delta(\epsilon - \epsilon_{n\boldsymbol{k}\sigma}) = \frac{1}{2} D_n(\epsilon + \frac{1}{2} g\mu_B B\sigma), \tag{199}$$

whence

$$\mathrm{D}_\sigma(\epsilon) = \sum_n \mathrm{D}_{n\sigma}(\epsilon) = \frac{1}{2}\mathrm{D}(\epsilon + \frac{1}{2}g\mu_\mathrm{B}\mathrm{B}\sigma). \tag{200}$$

From this, we obtain the free energy in a Zeeman field

$$\begin{aligned} f(\mathrm{B}) &= -k_\mathrm{B}\mathrm{T}\sum_\sigma \int d\epsilon\, \mathrm{D}_\sigma(\epsilon) \ln\left(1 + e^{-\beta(\epsilon-\mu)}\right) \\ &= -\frac{1}{2}k_\mathrm{B}\mathrm{T}\sum_\sigma \int d\epsilon\, \mathrm{D}(\epsilon + \frac{1}{2}g\mu_\mathrm{B}\mathrm{B}\sigma) \ln\left(1 + e^{-\beta(\epsilon-\mu)}\right) \\ &= -\frac{1}{2}k_\mathrm{B}\mathrm{T}\sum_\sigma \int d\epsilon\, \mathrm{D}(\epsilon) \ln\left(1 + e^{-\beta(\epsilon-\frac{1}{2}g\mu_\mathrm{B}\mathrm{B}\sigma-\mu)}\right). \end{aligned} \tag{201}$$

The spin magnetization is the first derivative $\mathrm{M} = -\partial\mathrm{F}/\partial\mathrm{B}$, and the spin susceptibility (per unit volume) is

$$\chi = \frac{1}{\mathrm{V}}\left.\frac{\partial \mathrm{M}}{\partial \mathrm{B}}\right|_{\mathrm{B}=0} = -\left.\frac{\partial^2 f}{\partial \mathrm{B}^2}\right|_{\mathrm{B}=0} = \frac{(g\mu_\mathrm{B})^2}{16 k_\mathrm{B}\mathrm{T}} \int d\epsilon\, \frac{\mathrm{D}(\epsilon)}{\cosh^2(\frac{\epsilon-\mu}{2k_\mathrm{B}\mathrm{T}})}. \tag{202}$$

Following the same reasoning as above, the low temperature limit of this expression is

$$\chi(\mathrm{T}\to 0) \sim \frac{(g\mu_\mathrm{B})^2}{4}\mathrm{D}(\epsilon_\mathrm{F}) = (g/2)^2\mu_0\mu_\mathrm{B}^2\mathrm{D}(\epsilon_\mathrm{F}), \tag{203}$$

where the latter equality is the transformation to SI units. We see that the spin susceptibility is a constant in the zero temperature limit, and proportional to the density of states. This is simply understood from the fact that the Zeeman interaction shifts opposite spin levels oppositely, which at zero temperature transfers electrons from just below to just above the Fermi energy and vice versa for the two opposite spin polarizations, within a shell of energy of width the Zeeman energy. The number density of electrons transferred in that way is proportional to $\mathrm{D}(\epsilon_\mathrm{F})$ multiplied by the Zeeman energy. This makes the magnetization linear in the field at zero temperature, and results in the above form. This is known as the *Pauli spin susceptibility*, and is another characteristic of metals. In a band insulator, the spin susceptibility will be exponentially small, obeying an Arrhenius law.

6.1.3 Narrow bands and effective mass

Both the Pauli susceptibility and the Sommerfeld coefficient are proportional to the density of states, $\mathrm{D}(\epsilon_\mathrm{F})$, which is typically "small" in the sense of the estimate in Eq. (194). There are, however, exceptions, in which the density of states is anomalously large. This is often phrased in terms of an "effective

mass". The idea is to consider the density of states that would result from a free-electron dispersion with a modified mass,

$$\epsilon(\boldsymbol{k}) = \frac{k^2}{2m^*}. \tag{204}$$

In three dimensions, this leads to the density of states

$$\mathrm{D}(\epsilon) = \frac{m^*\sqrt{2m^*\epsilon}}{\pi^2}. \tag{205}$$

If we evaluate this at the Fermi energy, we can use the standard definition of the Fermi momentum (see Sec. 1.2.3, e.g. Eq. (10) which gives the Fermi energy

$$\epsilon_\mathrm{F} = \frac{k_\mathrm{F}^2}{2m^*}. \tag{206}$$

One can then obtain

$$\mathrm{D}(\epsilon_\mathrm{F}) = \frac{m^* k_\mathrm{F}}{\pi^2}. \tag{207}$$

Since k_F is fixed by the electron density, m^* can be used a proxy for the density of states. Of course, the dispersion in real bands is generally *not* given by Eq. (204), but it is still common to use Eq. (207) to represent the Sommerfeld coefficient in terms of a "thermal effective mass", i.e. a large Sommerfeld coefficient corresponds to a large effective mass. The most dramatic examples occur in the so-called heavy fermion materials, in which m^* may be several hundred times the bare electron mass.

The most common reason for an enhanced density of states is that the bandwidth W is small for some reason. For this reason one often associates narrow bands with "heavy" electrons. In the heavy fermions, narrow bands arise out of nearly localized f-electron states (we will talk about how bands arise from localized orbitals shortly). An enhanced density of states may also arise without narrow bands from special aspects of the band dispersion. Indeed, while Eq. (181) constrains the *typical* value of $\mathrm{D}_n(\epsilon)$ to be of order the inverse bandwidth, larger values may occur for some energies within the band. The density of states may even *diverge* at isolated energies within the band, so long as this divergence is integrable. Such a divergence can arise when the topology of the constant energy surfaces of the band changes at a particular energy, for example if the energy passes through a saddle point of $\epsilon_{n\boldsymbol{k}}$. More generally when such a topology change occurs, the $\mathrm{D}_n(\epsilon)$ is not analytic at this energy. Such a divergence is known as a van Hove singularity. Van Hove singularities occur at band edges and at some special energies within bands. They are of limited relevance unless they occur close to the Fermi energy.

6.2 *Spectroscopy*

There are various experimental techniques that measure energy levels in a solid. The set of energy levels of a quantum system is called its spectrum, and so this is called spectroscopy.

6.2.1 Tunneling

A simple type of spectroscopy involves measuring just the energies of a system, without any reference to the wavefunctions associated to each energy eigenstate. For free fermions this is just the DOS. We have seen that the density of states at the Fermi energy enters thermodynamic quantities. How can one measure the density of states at other energies? There are several techniques that measure this in some ways. One such method, which can be applied to metals, is tunneling spectroscopy. The basic idea is to inject an electron from a known metal into a metal one wants to probe, and by keeping track of the energy of the electron we inject, we learn about the DOS of the probe metal. This works by passing current from one metal to another, across an insulator. The insulator acts as a barrier, almost separating the two metals. Hence it dominates the resistance, and the voltage drops across the barrier. Being separated, we treat each metal as approximately in equilibrium, and we can assign each metal i an electrostatic potential ϕ_i and chemical potential μ_i. Note that in true equilibrium, i.e. without any applied voltage, $\mu_i = \mu$ should be the same in both metals. In general there may be an electrostatic potential difference between the two metals in equilibrium, due to charges in the interface region. The electrostatic potential enters the energies in each metal as a constant, so that, so that $D_i(\epsilon, \phi_i) = D_i(\epsilon + e\phi_i - e\phi_i^{\text{eq}})$, where $D_i(\epsilon)$ is the DOS in metal i in zero applied voltage, when $\phi_i = \phi_i^{\text{eq}}$. The chemical potential only enters the distribution. Then the electron density in metal i is

$$n_i = \int D_i(\epsilon + e\phi_i - e\phi_i^{\text{eq}}) n_{\text{F}}(\epsilon - \mu_i) = \int D_i(\epsilon) n_{\text{F}}(\epsilon - e\phi_i + e\phi_i^{\text{eq}} - \mu_i). \quad (208)$$

The electron density in each metal must remain the same as it is for zero applied voltage, i.e. when $\phi_i = \phi_i^{\text{eq}}$ by definition and both sides have the same chemical potential equal which defines the Fermi energy $\mu_i = \epsilon_{\text{F}}$, so that it equals the ionic density and the metal remains charge neutral (some surface charge can accumulate but not bulk charge). Hence we have

$$e\phi_i + \mu_i - e\phi_i^{\text{eq}} = \epsilon_{\text{F}} \quad (209)$$

This means that the chemical potential shifts in step with the electrostatic potential as the voltage is varied. Consequently

$$D_i(\epsilon, \phi_i) = D_{\text{I}}(\epsilon - \Delta\mu_i), \quad (210)$$

where $\Delta\mu_i = \mu_i - \epsilon_F$. We see that each DOS shifts according to its chemical potential. The voltage is defined by the difference $eV = \mu_1 - \mu_2$. When it is non-zero, the two DOS are shifted relative to one another. The occupation of states on each side is determined by $n_F(\epsilon - \mu_i)$, so the occupation also shifts.

The key observation is that the set of levels with $\min(\mu_1, \mu_2) < \epsilon < \max(\mu_1, \mu_2)$ are empty in one metal and full in the other. The insulating barrier should possess no states in this region. In this case, electrons can move from occupied states on one side of the barrier to empty states on the other by tunneling. This occurs only in one direction at $T = 0$, determined by the sign of the voltage, and hence leads to a net current. As the voltage is increased, the number of states that can tunneling increases, and hence so does the current. We can write the net current as

$$I = -e \int d\epsilon\, T(\epsilon) D_1(\epsilon - \Delta\mu_1) D_2(\epsilon - \Delta\mu_2) \left[n_F(\epsilon - \mu_1)(1 - n_F(\epsilon - \mu_2)) - (1 - n_F(\epsilon - \mu_1)) n_F(\epsilon - \mu_2) \right], \tag{211}$$

where $T(\epsilon)$ is a rate determined from tunneling physics, generally weakly dependent on energy. The terms in the brackets represent the rate to take an electron from side 1 to side 2, from an occupied state (with probability n_F) to an unoccupied one (with probability $1 - n_F$) and vice-versa. This simplifies to

$$\begin{aligned} I &= -e \int d\epsilon\, T(\epsilon) D_1(\epsilon - \Delta\mu_1) D_2(\epsilon - \Delta\mu_2) \left[n_F(\epsilon - \mu_1) - n_F(\epsilon - \mu_2) \right] \\ &= -e \int d\epsilon\, T(\epsilon) D_1(\epsilon - \Delta\mu_1) D_2(\epsilon - \Delta\mu_2) \left[n_F(\epsilon - \epsilon_F - \Delta\mu_1) - n_F(\epsilon - \epsilon_F - \Delta\mu_2) \right] \end{aligned} \tag{212}$$

At this point, let us assume that the rate $T(\epsilon) \approx T$ is constant over the energy window, and we also choose metal 1 to be a "good" metal with approximately constant DOS, so $D_1(\epsilon) = D$ constnat. Then we can shift $\epsilon \to \epsilon + \Delta\mu_2$ to obtain

$$I \approx -eTD_1 \int d\epsilon\, D_2(\epsilon) \left[n_F(\epsilon - \epsilon_F - eV) - n_F(\epsilon - \epsilon_F) \right]. \tag{213}$$

Differentiating with respect to V gives the differential conductance

$$\frac{dI}{dV} \approx e^2 T D_1 \int d\epsilon\, D_2(\epsilon) n_F'(\epsilon - \epsilon_F - eV) \approx e^2 T D_1 D_2(\epsilon_F + eV), \tag{214}$$

where we approximate the derivative of the Fermi function by a delta function at low temperature. We see that the differential conductance is proportional to the DOS of metal 2, and thereby the DOS can be measured.

The treatment above of the rate T was not very rigorous. In general, this depends upon the nature of the tunnel junction. An important case is a point contact, where metal 1 is an scanning tunneling microscope (STM) tip, and the insulator is vacuum. For such a point contact, the tunneling measures the

"local density of states" and T contains information on the Bloch wavefunctions close to the location of the tip. One can also consider planar junctions, and the result depends upon the nature of the interfaces. For very clean interfaces, momentum may be conserved parallel to the interface, leading to highly non-trivial energy dependence of $T(\epsilon)$ and more careful considerations are needed explicitly tracking the momentum.

6.2.2 Angle resolved photoemission

In tunneling, an electron is either added or removed from the metal being probed – both are possible because electrons can be added or removed from the reference metal on the other side of the barrier. In angle resolved photo-emission spectroscopy, an electron is removed from a metal but instead of into another metal it is ejected into free space and then collected at a detector. This is accomplished by directing energetic photons at the sample. The phenomena of electrons being ejected from a solid by incident radiation is the photoelectric effect, the subject of Einstein's nobel prize. Over the past decades the technique has been refined so that it has become a powerful tool for condensed matter physics. A photon is absorbed by the material, causing an electron to transition from a bound state in the material to an unbound one which then propagates to the detector, which registers the energy of the ejected "photo-electron" and its momentum (by identifying the direction of propagation). The result is analyzed by energy and momentum conservation.

Ideally, the surface of the solid is perfectly flat and crystalline. In that case, there is (discrete) translational symmetry in the two directions parallel to the surface. This implies conservation of quasi-momentum in these two directions. One can think of the photoemission as a process in which the initial state consists of a photon and an electron in solid, and the final state is just the photo-electron. Therefore, if we know the momentum of the incident photon $\boldsymbol{k}_{ph}$ and the momentum of the photo-electron $\boldsymbol{k}_f$ (this is called "angle resolved photo-emission" because we must detect the direction of the photo-electron to deduce its momentum), we can deduce the planar components of the quasi-momentum of the initial electron,

$$\boldsymbol{k}_{\|} = \boldsymbol{k}_{f,\|} - \boldsymbol{k}_{ph,\|}. \tag{215}$$

Furthermore, by energy conservation, the energy of the initial electron is the energy of the photo-electron minus the photon energy,

$$\epsilon = \epsilon_f - \epsilon_{ph} = \epsilon_f - \hbar c|k|_{ph}. \tag{216}$$

By collecting photo-electrons and binning the data with respect to momentum and enerby, one arrives at a histogram which approximates the intensity function $I(\boldsymbol{k}_{\|}, \epsilon)$. The technique is known as Angle Resolved Photo-Emission Spectroscopy (ARPES). A peak in this intensity reflects the existence of a state

with energy ϵ and planar momentum $\boldsymbol{k}_{\|}$. Note that the zero of energy must be chosen globally for this to make sense. If we measure the kinetic energy of the electron in the vacuum, then the energy here includes the binding energy of electrons in the solid, the so-called "work function".

In the independent electron picture, the states that must satisfy the momentum and energy conditions of Eq. (215)- (438) are just band states. Hence the photoemission intensity has the form

$$I(\boldsymbol{k}_{\|}, \epsilon) = \sum_n \int \frac{dk_\perp}{2\pi} n_F(\epsilon_{n,\boldsymbol{k},k_z})\, M(\boldsymbol{k}, \epsilon) \frac{\gamma/\pi}{(\omega - \epsilon_{n,\boldsymbol{k},k_z})^2 + \gamma^2}, \tag{217}$$

where M is an amplitude which comes from more detailed calculations, and γ is a phenomenological linewidth or decay rate (more generally this might depend on energy and momentum, and one might also put a width in momentum). A more sophisticated theory relates the ARPES intensity to a somewhat subtle quantity in many-body theory, the one-electron spectral function, which is very useful for diagnosing the effect of correlations/interactions between electrons. However, Eq. (440) is not so far off, and therefore one can often directly visualize the bands of occupied states by determining the peaks in the intensity. The technique works best in quasi-two dimensional materials in which the band dispersions are almost independent of k_z, as this dependence is not really resolved.

It is notable that ARPES measures only the occupied electronic states. To measure the unoccupied bands requires another technique. For example, one may attempt to do "time resolved ARPES", using a "pump" laser to excite electrons into the normally unoccupied bands, and carrying out the photoemission measurement before they relax back.

6.2.3 Friedel oscillations

Friedel oscillations are oscillations in the charge density induced by an applied potential or more often a defect in a metal. They occur because of the sharp jump in the occupation of (quasi-)momentum states inside and outside the Fermi energy. This singular distribution in momentum space leads to oscillations in real space. Characteristically, the scale of the oscillations is of the order of the Fermi wavelength, and the Fourier components of the oscillations are compromised of momenta which can be formed as the difference of two points on the Fermi surface. Friedel oscillations occur as a ground state phenomena (in the presence of defects), so perhaps they should not really be in this spectroscopy section. However, they are often measured via STM, so I grouped it here.

7 TRANSPORT

Transport refers to experiments which study *currents* of conserved quantities like charge, heat, or sometimes spin (which is only approximately conserved at best) induced by applied forces (e.g. electric fields) or other means of forcing a system out of equilibrium (e.g. contact to two reservoirs at different temperatures). This could very well be another topic under the previous section but it is so important and connected to very deep ideas that it merits its own section.

7.1 *Semi-classical dynamics*

As we have repeatedly emphasized, the basic energy and length scales for bands in solids are not too dissimilar from those of atoms. Most forces that we can actually apply to electrons (e.g. external electric fields) are weak in comparison, and are also applied over much longer length scales. We consider the one electron Hamiltonian

$$\mathrm{H} = \frac{|\boldsymbol{p} + e\mathbf{A}(\boldsymbol{x})|^2}{2m} + \mathrm{V}(\boldsymbol{x}) - e\phi(\boldsymbol{x}), \tag{218}$$

where $\mathbf{A}(\boldsymbol{x})$ and $\phi(\boldsymbol{x})$ are vector and scalar potentials associated with *slowly varying* and *weak* electromagnetic field. The electric field is weak if it is small compared to atomic fields, and the magnetic field is weak if the magnetic flux per area of a unit cell, in any direction, is much less than the flux quantum h/e. They are slowly varying in space if they change vary little over the length of a unit cell. They are slowly varying in time if they vary little over the time which is the inverse of the bandwidth. All these conditions are easily satisfied by nearly all electromagnetic fields applied to solids.

With this requirement, one can apply semi-classical dynamics. Why semi-classical? If we assume that all perturbations to the system are slowly-varying in space and time, then over a large local region the problem "looks" like an ideal and almost regular Schrödinger equation in a periodic potential. Even the potentials (not just the fields) can be regarded as almost constant, since the fields themselves are assumed weak. Of course Bloch states are plane waves and so infinite, so do not quite work for even such a slowly-varying problem. However, from Bloch states we can build a *wave packet*, which is localized in space,

$$\psi_{n,\boldsymbol{q},\boldsymbol{y}}(\boldsymbol{x}) = \int_{\mathrm{BZ}} \frac{d^d\boldsymbol{k}}{(2\pi)^d}\, \phi(\boldsymbol{k},\boldsymbol{q}) e^{i\boldsymbol{k}\cdot\boldsymbol{y}} \psi_{n\boldsymbol{k}}(\boldsymbol{x}), \tag{219}$$

where $\phi(\boldsymbol{k},\boldsymbol{q})$ is some amplitude function peaked at $\boldsymbol{k} = \boldsymbol{q}$. The resulting wave-function is peaked near $\boldsymbol{x} = \boldsymbol{y}$. It has indefinite position and momentum, limited by the uncertainty principle. For a situation in which the perturbations to the Bloch problem are slowly varying, then we can increase the "sharpness"

of ϕ until the width in real space becomes comparable to the size of variation of the perturbations. Generally we can make the uncertainty in momentum small compared to the size of the Brillouin zone, while at the same time keeping the uncertainty in position smaller than the scale of variation of the applied potentials.

For such a wave packet, we can study the evolution of the packet's center $\boldsymbol{q}$ in momentum and $\boldsymbol{y}$ in real space as a function of time. The derivation is rather technical and it will not be fully given here (you might still find what is written here technical enough!). Instead, we will quote the results and refer the reader to careful treatment in the literature – see Refs.[5, 4]. The result is

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v}_n = \boldsymbol{\nabla}_{\boldsymbol{k}}\tilde{\epsilon}_{n\boldsymbol{k}} - \frac{d\boldsymbol{k}}{dt} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}, \tag{220}$$

$$\frac{d\boldsymbol{k}}{dt} = \mathbf{F}_n = -e\mathbf{E} - e\frac{d\boldsymbol{x}}{dt} \times \mathbf{B}. \tag{221}$$

Here most of the terms are recognizable from the free electron limit. Indeed, the second equation, for the momentum, is unchanged from that case. The first equation looks a little different. The first term in the first equation, $\boldsymbol{\nabla}_{\boldsymbol{k}}\tilde{\epsilon}_{n\boldsymbol{k}}$, is understandable from the quantum relation $\omega = \epsilon$ (here $\hbar = 1$), and then recalling from elementary physics the formula for the group velocity of a wave packet. So this term is the natural group velocity of a wave in a dispersive medium. There is only one subtlety in that term: the band energy $\epsilon_{n\boldsymbol{k}}$ has been replaced by $\tilde{\epsilon}_{n\boldsymbol{k}}$, which indicates the slight modification:

$$\tilde{\epsilon}_{n\boldsymbol{k}} = \epsilon_n(\boldsymbol{k}) - \mathbf{B} \cdot \boldsymbol{m}_{n\boldsymbol{k}}. \tag{222}$$

Here $\boldsymbol{m}_{n\boldsymbol{k}}$ is a quantity associated with the Bloch states, and can be interpreted as an *orbital magnetic moment*. It is due to rotational motion of charge inside the spatial extent of a wavepacket. It has a slightly complicated form but can be calculated if the Bloch functions are known,

$$m^{\mu}(\boldsymbol{k}) = \frac{-ie}{2}\epsilon_{\mu\nu\lambda}\left\langle \frac{\partial u_{n\boldsymbol{k}}}{\partial k_\nu} | (\mathcal{H}_{\boldsymbol{k}} - \epsilon_{n\boldsymbol{k}}) | \frac{\partial u_{n\boldsymbol{k}}}{\partial k_\lambda} \right\rangle. \tag{223}$$

In this expression, $\mathcal{H}_{\boldsymbol{k}}$ is the Bloch Hamiltonian from Eq. (166), and we used the notation that

$$\langle u_{n'\boldsymbol{k}'} | \mathcal{O} | u_{n\boldsymbol{k}} \rangle \equiv \int_{p.u.c.} u^*_{n'\boldsymbol{k}'}(\boldsymbol{x}) \mathcal{O} u_{n\boldsymbol{k}}(\boldsymbol{x}). \tag{224}$$

Note the integral is just over a single primitive unit cell, and we use the normalization convention

$$\langle u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle = \int_{p.u.c.} |u_{n\boldsymbol{k}}(\boldsymbol{x})|^2 = 1. \tag{225}$$

These definitions are implied any time we write a bra-ket expression with a $u_{n\boldsymbol{k}}$ inside the bras and kets. Eq. (223) looks a little involved but we will at least get a bit more insight into it shortly.

7.1.1 Berry curvature and anomalous velocity

The second term in the first line of Eq. (220) is novel, and is called the *anomalous velocity*. It has a striking similarity to the Lorentz force term in the second line, but it is "dual": momentum and position have exchanged their roles in the anomalous velocity. The quantity which plays the role of the magnetic field is called the Berry curvature,

$$\Omega^{\mu}_{n\boldsymbol{k}} = i\epsilon_{\mu\nu\lambda}\left\langle \frac{\partial u_{n\boldsymbol{k}}}{\partial k_\nu}|\frac{\partial u_{n\boldsymbol{k}}}{\partial k_\lambda}\right\rangle. \tag{226}$$

The Berry curvature is a fundamental quantity not only in band structure but in any quantum mechanics problem which depends upon at least two parameters. We will illustrate how it arises through the specific example of the Bloch hamiltonian $\mathcal{H}_{\boldsymbol{k}}$, which depends upon $\boldsymbol{k}$ as a parameter. The eigenstates are just the periodic parts of the Bloch functions, and we will use bra-ket notations as in the previous two equations,

$$\mathcal{H}_{\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle = \epsilon_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle. \tag{227}$$

As in any quantum mechanics problem, the states are defined only up to a phase. This phase can be chosen independently for every eigenstate, here for every n and $\boldsymbol{k}$. So two phase conventions differ by a "gauge transformation"

$$|u_{n\boldsymbol{k}}\rangle \to e^{i\varphi_{n\boldsymbol{k}}}|u_{n\boldsymbol{k}}\rangle. \tag{228}$$

The choice of phase convention is up to the person solving the Schrödinger equation, Eq. (227). So we should be careful to express physical properties in terms that do not depend upon the phase convention (or to specify it, but this is less elegant). This line of thinking leads to the introduction of the Berry curvature.

To get there, we first ask what information we might like to understand about the wavefunctions $|u_{n\boldsymbol{k}}\rangle$? A natural question is to try to understand how they change as $\boldsymbol{k}$ is varied. To do so, we may try to look at the overlap of two Bloch states with slightly different momenta,

$$\langle u_{n\boldsymbol{k}+\boldsymbol{q}}|u_{n\boldsymbol{k}}\rangle \approx 1 + \boldsymbol{q}\cdot\langle \boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle + \mathrm{O}(k^2). \tag{229}$$

We see that the overlap of nearby states is determined by a single object. This object can be shown to be purely imaginary. To see this, we use the fact that the states are normalized and take a gradient,

$$\langle u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle = 1 \qquad \Rightarrow \qquad \langle \boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle + \langle u_{n\boldsymbol{k}}|\boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}}\rangle = 0. \tag{230}$$

Note that by the definition of hermitian conjugacy,

$$\langle u_{n\boldsymbol{k}}|\boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}}\rangle = (\langle \boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle)^*, \tag{231}$$

which by Eq. (230) implies $\mathrm{Re}\,\langle \boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle = 0$, i.e. this quantity is a pure phase. The math is the same as varying a vector of fixed length, e.g. on a sphere or a circle. As it rotates, the infinitesimal change of the vector is tangent to the sphere, i.e. normal to the instantaneous direction of the vector. Hence we can define a real quantity by multiplying by i,

$$\boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} = i\langle u_{n\boldsymbol{k}}|\boldsymbol{\nabla}_{\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle, \tag{232}$$

which is known as the Berry gauge field or Berry vector potential. The overlap becomes

$$\langle u_{n\boldsymbol{k}+\boldsymbol{q}}|u_{n\boldsymbol{k}}\rangle \approx 1 + i\boldsymbol{q}\cdot\boldsymbol{\mathcal{A}}_{n\boldsymbol{k}}. \tag{233}$$

You can see that the Berry gauge field describes a sort of evolution of the eigenstates of the Bloch Hamiltonian as the parameter $\boldsymbol{k}$ is varied. A famous result from quantum mechanics – the quantum adiabatic theorem – is that if we consider an actual time dependent Hamiltonian with some parameters that vary slowly in time, an initial state which is an eigenstate of the Hamiltonian with the initial parameters remains to a good approximation an instantaneous eigenstate of the Hamiltonian as those parameters change. Here we are applying this result to the situation in which the parameter is the Bloch momentum, which can be a function of time t due to applied forces. Hence $\mathcal{H}_{\boldsymbol{k}(t)}$ defines a time-dependent Hamiltonian. What Michael Berry showed is that the *phase* of the time-dependent state obeying the adiabatic theorem has a geometrical component. Let's make this explicit with some math. The time-dependent Schrödinger equation with a time-dependent (quasi-)momentum is

$$i\partial_t|\psi\rangle = \mathcal{H}_{\boldsymbol{k}(t)}|\psi\rangle. \tag{234}$$

We take the initial state to be an eigenstate,

$$|\psi(t=0)\rangle = |u_{n\boldsymbol{k}(0)})\rangle, \tag{235}$$

with some initial momentum $\boldsymbol{k}(0)$, and in some band n. What Berry showed is that under the adiabatic evolution, the state at time t is

$$|\psi(t)\rangle \approx e^{i\gamma_d} e^{i\gamma_n} |u_{n\boldsymbol{k}(t)}\rangle. \tag{236}$$

Here $\gamma_d = \int_0^t dt' \, \epsilon_{n\boldsymbol{k}(t')}$ a trivial dynamical phase. The second contribution γ_n is *geometrical phase*:

$$\gamma_n(\mathrm{C}) = \int_{\boldsymbol{k}(0)}^{\boldsymbol{k}(t)} d\boldsymbol{k} \cdot \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}}, \tag{237}$$

which depends upon the *path* C taken in momentum space. This is geometrical in that the actual way in which the quasi-momentum varies in time does not enter: only the curve in $\boldsymbol{k}$-space matters, not how fast or slow it is traversed. I will not derive the correspondence between this integral and the adiabatic geometrical phase, but you can find it in many textbooks and reviews, e.g. Ref. [5].

The Berry gauge field is *dependent* upon the phase convention for the Bloch states. Under the gauge transformation in Eq. (228), it transforms according to

$$\boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} \to \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} - \boldsymbol{\nabla}_{\boldsymbol{k}} \varphi_{n\boldsymbol{k}}. \tag{238}$$

This transformation is identical to the transformation of a physical vector potential under a spatial gauge transformation, except it is now in momentum space. One can see that in general the geometric phase is not invariant under such changes, but it becomes invariant if the path is taken to form a closed loop, with $\boldsymbol{k}(t_f) = \boldsymbol{k}(0)$. The phase is then dependent only on the path and independent of the gauge choice.

From the analogy with electromagnetism, it is clear how to define a *local* gauge invariant quantity, which is just the flux associated with this gauge field. This is in fact just the Berry curvature:

$$\boldsymbol{\nabla}_{\boldsymbol{k}} \times \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} = \boldsymbol{\Omega}_{n\boldsymbol{k}}. \tag{239}$$

So this is an exceedingly natural quantity to appear in physical expressions. Why in particular does it enter the wave packet velocity in Eq. (220)?

One way to approach this question is to view the equation of motion for the position as arising from taking the expectation value of the *operator* equation of motion for the position operator (in the Heisenberg representation). Since we are discussing motion within a band, we want to consider the position operator *projected into a single band*.

7.1.2 Less technically-minded readers may want to skip this

We are now going to derive the form of this projected position operator. It is a bit technical and long. Feel free to skip ahead to the result if you do not care.

Projecting the electron operator into a band means understanding how the operator acts on a general electron state within a band. Such a state is an arbitrary superposition of Bloch states, i.e.

$$|\psi\rangle = \int_{\mathrm{BZ}} \frac{d^d\boldsymbol{k}}{(2\pi)^d}\,\psi(\boldsymbol{k})|\psi_{n\boldsymbol{k}}\rangle. \tag{240}$$

Here $\psi(\boldsymbol{k})$ is the amplitude of the Bloch state with quasimomentum $\boldsymbol{k}$. It can be viewed as the electron wavefunction in the quasimomentum representation. Note there is no sum on n here because we deal with one band. In terms of wavefunctions, this is

$$\psi(\boldsymbol{x}) = \int_{\mathrm{BZ}} \frac{d^d\boldsymbol{k}}{(2\pi)^d}\,\psi(\boldsymbol{k})e^{i\boldsymbol{k}\cdot\boldsymbol{x}}u_{n\boldsymbol{k}}(\boldsymbol{x}). \tag{241}$$

We will choose to normalize the periodic parts of the Bloch states such that

$$\langle u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle = \int_{p.u.c.} d^d\boldsymbol{x}\,|u_{n\boldsymbol{k}}(\boldsymbol{x})|^2 = 1. \tag{242}$$

Please take note of the notation: when evaluating a bra-ket product of the periodic parts of Bloch functions, the spatial integral is taken over a single unit cell. When evaluating the bra-ket product of full Bloch states, such as those in Eq. (240), we are to integrate over the full space. Now since the Bloch states are eigenstates of the Hamiltonian with distinct energy and momenta, we expect them to be orthogonal for different quasimomenta. Let us check this. We have

$$\langle\psi_{n\boldsymbol{k}'}|\psi_{n\boldsymbol{k}}\rangle = \int d^d\boldsymbol{x}\,e^{i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{x}}u^*_{n\boldsymbol{k}'}(\boldsymbol{x})u_{n\boldsymbol{k}}(\boldsymbol{x}). \tag{243}$$

Here the integral is over all space. We can divide the integral into a sum over unit cells, and an integration within each cell. Let $\boldsymbol{x} \to \mathbf{X} + \boldsymbol{x}$, where $\mathbf{X}$ are Bravais lattice vectors and now $\boldsymbol{x}$ is integrated over a unit cell,

$$\langle\psi_{n\boldsymbol{k}'}|\psi_{n\boldsymbol{k}}\rangle = \sum_{\mathbf{X}}\int_{p.u.c.} d^d\boldsymbol{x}\,e^{i(\boldsymbol{k}-\boldsymbol{k}')\cdot(\mathbf{X}+\boldsymbol{x})}u^*_{n\boldsymbol{k}'}(\boldsymbol{x})u_{n\boldsymbol{k}}(\boldsymbol{x}). \tag{244}$$

We see that only the exponential depends upon $\mathbf{X}$, so the sum over lattice sites can be carried out explicitly. One has the identity

$$\sum_{\mathbf{X}} e^{i\boldsymbol{q}\cdot\mathbf{X}} = \frac{(2\pi)^d}{V_{p.u.c.}}\sum_{\mathbf{Q}}\delta^{(d)}(\boldsymbol{q}-\mathbf{Q}), \tag{245}$$

where the sum $\mathbf{Q}$ is over reciprocal lattice vectors. Applied to Eq. (244), and assuming we use the reduced zone scheme, then the delta function can be satisfied only when $\boldsymbol{k} = \boldsymbol{k}'$ and we obtain

$$\begin{aligned}\langle \psi_{n\boldsymbol{k}'} | \psi_{n\boldsymbol{k}} \rangle &= \int_{p.u.c.} d^d\boldsymbol{x}\, \frac{(2\pi)^d}{V_{p.u.c.}} \delta^{(d)}(\boldsymbol{k}-\boldsymbol{k}') u^*_{n\boldsymbol{k}'}(\boldsymbol{x}) u_{n\boldsymbol{k}}(\boldsymbol{x}) \\ &= \frac{(2\pi)^d}{V_{p.u.c.}} \delta^{(d)}(\boldsymbol{k}-\boldsymbol{k}'). \end{aligned} \tag{246}$$

So Bloch states with distinct quasimomenta are indeed orthogonal and we have determined the normalization. Note we need to take care with delta functions since we work extended states. Eq. (246) now can be applied to Eq. (240) to invert the relation and obtain $\psi(\boldsymbol{k})$. Take the overlap of both sides with $\langle \psi_{n\boldsymbol{k}}|$, one obtains

$$\psi(\boldsymbol{k}) = V_{p.u.c.} \langle \psi_{n\boldsymbol{k}} | \psi \rangle = V_{p.u.c.} \int d^d\boldsymbol{x}\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} u^*_{n\boldsymbol{k}}(\boldsymbol{x}) \psi(\boldsymbol{x}). \tag{247}$$

This allows us to obtain the quasimomentum representation wavefunction from any real space one. It must be understood as a projection, because the basis of Bloch states of a single band is not complete in the full Hilbert space. That is, if we apply Eq. (247) to an arbitrary real space wavefunction, obtain $\psi(\boldsymbol{k})$ from it, and then go back to use Eq. (240) or Eq. (241) to obtain a new real space wavefunction, this new wavefunction is the projection of the original one into band n. This is what we are after anyway!

Now we are ready to consider the position operator. Acting on an arbitrary real space wavefunction, we have of course that

$$\hat{\boldsymbol{x}}\psi(\boldsymbol{x}) = \psi'(\boldsymbol{x}) = \boldsymbol{x}\psi(\boldsymbol{x}), \tag{248}$$

where here I indicate on the left hand side by a hat that we mean the position operator, and on the far right hand side, the absence of a hat means it is simply the coordinate. We want to obtain the projection of $\psi'(\boldsymbol{x})$ into the Bloch band, i.e. the corresponding momentum wavefunction $\psi'(\boldsymbol{k})$. Since position is a linear operator, we will define the projected position operator $\hat{\boldsymbol{x}}_n$ in the momentum representation by

$$\psi'(\boldsymbol{k}) = \hat{\boldsymbol{x}}_n \psi(\boldsymbol{k}). \tag{249}$$

We need to find $\psi'(\boldsymbol{k})$. According to Eq. (247), it is

$$\begin{aligned}\psi'(\boldsymbol{k}) &= \mathrm{V}_{p.u.c.}\int d^d\boldsymbol{x}\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} u^*_{n\boldsymbol{k}}(\boldsymbol{x})\boldsymbol{x}\psi(\boldsymbol{x}) \\ &= \mathrm{V}_{p.u.c.}\int d^d\boldsymbol{x}\, i\boldsymbol{\nabla}_{\boldsymbol{k}}\left(e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\right) u^*_{n\boldsymbol{k}}(\boldsymbol{x})\psi(\boldsymbol{x}) \\ &= i\boldsymbol{\nabla}_{\boldsymbol{k}}\psi(\boldsymbol{k}) - \mathrm{V}_{p.u.c.}\int d^d\boldsymbol{x}\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\left(i\boldsymbol{\nabla}_{\boldsymbol{k}} u^*_{n\boldsymbol{k}}(\boldsymbol{x})\right)\psi(\boldsymbol{x}). \end{aligned} \tag{250}$$

Now we can use Eq. (241) to express the final integral back in terms of $\psi(\boldsymbol{k})$:

$$\psi'(\boldsymbol{k}) = i\boldsymbol{\nabla}_{\boldsymbol{k}}\psi(\boldsymbol{k}) - \mathrm{V}_{p.u.c.}\int d^d\boldsymbol{x}\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\left(i\boldsymbol{\nabla}_{\boldsymbol{k}} u^*_{n\boldsymbol{k}}(\boldsymbol{x})\right)\int_{\mathrm{BZ}}\frac{d^d\boldsymbol{k}'}{(2\pi)^d}\psi(\boldsymbol{k}')e^{i\boldsymbol{k}'\cdot\boldsymbol{x}}u_{n\boldsymbol{k}'}(\boldsymbol{x}). \tag{251}$$

The double integral can be done by reversing the order of integration and following similar manipulations to those we used to evaluate the overlap of two Bloch states above. Skipping details, one obtains

$$\begin{aligned}\psi'(\boldsymbol{k}) &= i\boldsymbol{\nabla}_{\boldsymbol{k}}\psi(\boldsymbol{k}) - i\langle\boldsymbol{\nabla}_{\boldsymbol{k}}u_{n\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle\psi(\boldsymbol{k}) \\ &= \left(i\boldsymbol{\nabla}_{\boldsymbol{k}} + \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}}\right)\psi(\boldsymbol{k}). \end{aligned} \tag{252}$$

From this we can immediately read off the form of the projected position operator in the momentum representation by comparing to Eq. (249).

7.1.3 The projected position operator and derivation of the anomalous velocity

The result of the previous manipulations is that position operator in the the quasi-momentum representation (i.e. acting on a wavefunction $\psi(\boldsymbol{k})$ which gives the amplitude to find the electron in band b at quasimomentum $\boldsymbol{k}$) is

$$\boldsymbol{x}_n = i\boldsymbol{\nabla}_{\boldsymbol{k}} + \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} = i\left(\boldsymbol{\nabla}_{\boldsymbol{k}} - i\boldsymbol{\mathcal{A}}_{n\boldsymbol{k}}\right). \tag{253}$$

This is a beautiful equation, which should be compared to Eq. (465) for free electrons. We see that in addition to the momentum gradient, the Berry gauge field appears. In the form of the final equality it is clear why this gauge field must appear mathematically: it is just the same minimal coupling form in which the ordinary electromagnetic vector potential appears in combination with a derivative, which is there to ensure gauge invariance. Similarly, this form here ensures invariance under different choices of phase conventions for the Bloch states, i.e. $\boldsymbol{x}_n$ is invariant under

$$\psi(\boldsymbol{k}) \to e^{i\varphi(\boldsymbol{k})}\psi(\boldsymbol{k}), \qquad \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} \to \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} + \boldsymbol{\nabla}\varphi(\boldsymbol{k}). \tag{254}$$

This gives a mathematical reason for the appearance of the Berry gauge field. We will return to the physics shortly. First let us use the result in Eq. (253) to

derive the anomalous velocity. For this purpose, we can consider zero magnetic field and non-zero electric field, for which the semi-classical equations become

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\nabla}_{\boldsymbol{k}}\epsilon_{n\boldsymbol{k}} - \frac{d\boldsymbol{k}}{dt} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}, \tag{255}$$

$$\frac{d\boldsymbol{k}}{dt} = -e\mathbf{E}. \tag{256}$$

We can insert the second equation into the first to obtain

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\nabla}_{\boldsymbol{k}}\epsilon_{n\boldsymbol{k}} + e\mathbf{E} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}. \tag{257}$$

The goal is to derive this. We do so by calculating the operator equation of motion for the position, in the projected band Hamiltonian. We take the Hamiltonian in Eq. (218) for $\mathbf{A} = 0$ and $\phi(\boldsymbol{x}) = -\mathbf{E} \cdot \boldsymbol{x}$, and project it into a single band. This gives

$$\mathrm{H}_n \equiv \mathrm{P}_n \mathrm{H} \mathrm{P}_n = \epsilon_{n\boldsymbol{k}} + e\mathbf{E} \cdot \boldsymbol{x}_n. \tag{258}$$

To compute the equation of motion for the position operator, we need to compute the fundamental position and momentum commutators using Eq. (253). By simple algebra one obtains

$$\left[x_n^\mu, k^\nu\right] = i\delta_{\mu\nu}, \qquad \left[x_n^\mu, x_n^\nu\right] = i\epsilon_{\mu\nu\lambda}\Omega_\lambda. \tag{259}$$

The "canonical" commutation of position and momentum remains as it is in normal quantum mechanics, but the commutator of position with itself is modified and now contains the Berry curvature! From this, we can calculate the equation of motion for position,

$$\begin{aligned}\frac{dx_n^\mu}{dt} &= -i[x_n^\mu, \mathrm{H}] = i\left[x_n^\mu, \epsilon_{n\boldsymbol{k}}\right] + i\left[x_n^\mu, e\mathrm{E}_\nu x_n^\nu\right] \\ &= \frac{\partial \epsilon_{n\boldsymbol{k}}}{\partial k_\mu} + e\mathrm{E}_\nu \epsilon_{\mu\nu\lambda}\Omega_\lambda.\end{aligned} \tag{260}$$

This is an operator equation of motion, and the wave packet dynamics corresponds to taking the expectation value of this in a wave packet initial state (in the Heisenberg picture the states are time independent). This agrees exactly with Eq. (257). Hence we have derived the anomalous velocity (at least for the case $\mathbf{B} = 0$). We have actually used the same method as in the original Karplus and Luttinger paper when this was first derived[1].

7.1.4 Physical meaning of anomalous velocity

Let us now discuss the physical meaning of the anomalous velocity. This is of course encoded in the mathematics. We can see that it arises from the

Berry gauge field $\boldsymbol{\mathcal{A}}_{n\boldsymbol{k}}$. This is nothing but the expectation value of $i\boldsymbol{\nabla}_{\boldsymbol{k}}$ in the periodic part of the Block state. Since this is just a position operator, we can understand this term as representing the "center of mass" of the Bloch state within a unit cell. So roughly, the projected position operator has a contribution due to the envelope of the wave packet, which is the first term in Eq. (253), and a second contribution which is due to the position of the electronic state within a unit cell. I emphasize this is a rough understanding, because actually the first and second terms in Eq. (253) are not really separable: they are not gauge invariant on their own and so the division of each term into a envelope and unit cell contribution is gauge dependent. However, there are indeed two contributions to the spatial motion of the wave packet, and this is why there are two terms.

When we take the time derivative of the position, there are therefore two contributions to the velocity: one arising from the motion of the envelope, and another arising from the motion of the center of mass in the unit cell. It is the latter which should be identified with the anomalous velocity term in the equation of motion in Eq. (220),

$$\boldsymbol{v}_n^{\text{anom}} = -\frac{d\boldsymbol{k}}{dt} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}. \tag{261}$$

This arises because when the quasimomentum depends on time, the mean position of the electron within the unit cell evolves with it. The anomalous velocity is just the time derivative of this position.

7.1.5 Example: uniform electric field

As a simple example of semi-classical dynamics, consider the effect of a uniform electric field, in zero magnetic field. The semi-classical equation for the quasimomentum becomes

$$\frac{d\boldsymbol{k}}{dt} = \mathbf{F}_n = -e\mathbf{E}. \tag{262}$$

This indicates that the quasimomentum simply evolves linearly in time:

$$\boldsymbol{k}(t) = \boldsymbol{k}(0) - e\mathbf{E}t. \tag{263}$$

This is familiar as just acceleration in a constant field in Newtonian mechanics. However, there is an important difference in the Bloch case: the quasimomentum is itself only defined up to a reciprocal lattice vector, so that in this sense, the quasimomentum does not grow without bound. Indeed, if we adopt the reduced zone scheme, the linear growth of $\boldsymbol{k}(t)$ is mapped repeatedly back into the Brillouin zone. The way the trajectory depends somewhat on the orientation of the electric field. If it is oriented parallel to some reciprocal lattice vector $\mathbf{Q}$, then $\boldsymbol{k}(t)$ will be periodic as once $-e\mathbf{E}t = \mathbf{Q}$, the quasimomentum has returned to its initial value. If you remember that the Brillouin

zone can be viewed as a torus, this corresponds to a trajectory which winds around the torus some finite number of times before returning to its original position **[sketch?]**. If the electric field is not parallel to any reciprocal lattice vector, then the trajectory will wind indefinitely and never *exactly* return to its original quasi-momentum.

Once $\boldsymbol{k}(t)$ is known, we can examine the position using the other semi-classical equation:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v}_n = \boldsymbol{\nabla}_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}(0)-e\mathbf{E}t} + e\mathbf{E} \times \boldsymbol{\Omega}_{n\boldsymbol{k}(0)-e\mathbf{E}t}. \tag{264}$$

For short times, the result may look a lot like the acceleration of electrons in free space. For example, near the minimum of a band, $\tilde{\epsilon}$ is quadratic in momentum, and hence its gradient is linear, and inserting this into the first term above will give a velocity which increases linearly in time. However, at fixed electric field this does not reflect the long time behavior.

To see this, recall that all physical quantities are periodic with respect to quasi-momentum. Hence both $\epsilon_{n\boldsymbol{k}}$ and $\boldsymbol{\Omega}_{n\boldsymbol{k}}$ can be expanded in a Fourier series (where the coefficients are associated to Bravais lattice vectors $\mathbf{X}$):

$$\epsilon_{n\boldsymbol{k}} = \sum_{\mathbf{X}} \epsilon_{n\mathbf{X}} e^{i\boldsymbol{k}\cdot\mathbf{X}}, \qquad \boldsymbol{\Omega}_{n\boldsymbol{k}} = \sum_{\mathbf{X}} \boldsymbol{\Omega}_{n\mathbf{X}} e^{i\boldsymbol{k}\cdot\mathbf{X}}. \tag{265}$$

This implies

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v}_n = \sum_{\mathbf{X}} \left(\mathbf{X}\epsilon_{n\mathbf{X}} + e\mathbf{E} \times \boldsymbol{\Omega}_{\mathbf{X}}\right) e^{i(\boldsymbol{k}(0)-e\mathbf{E}t)\cdot\mathbf{X}}. \tag{266}$$

We can see that for long times, any Fourier component $\mathbf{X}$ which is not orthogonal to $\mathbf{E}$ will lead to an oscillatory contribution which averages to zero. For a "generic" direction of electric field, this is all $\mathbf{X}$ except $\mathbf{X} = 0$. For this component the first term above vanishes. For such a generic orientation, one has therefore at long times

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v}_n \sim e\mathbf{E} \times \boldsymbol{\Omega}_{\mathbf{X}=\mathbf{0}}. \tag{267}$$

Only the zeroth Fourier component of the Berry curvature, i.e. its average over the Brillouin zone, contributes – this quantity will arise again soon and we will return tio it! The average velocity remains bounded and proportional to the electric field.

In the standard textbooks where the Berry curvature is neglected, the average velocity therefore simply vanishes. This is fairly astounding from a free electron point of view: instead of accelerating to infinite velocity, the electron does not more on average at all! In fact, it executes in this case oscillatory motion (due to all the $\mathbf{X} \neq 0$ terms). This motion is known as *Bloch oscillations*.

One way to understand Bloch oscillations is as follows. The fundamental

ingredient is that the quasimomentum is periodic: an electron whose quasimomentum exits the Brillouin zone on one side re-enters it on the other. This corresponds to relabeling the quasimomentum by a reciprocal lattice vector. It can be thought of as a Bragg reflection: recall that in the nearly free electron model the Brillouin zone boundaries are Bragg planes, where two bare momentum states become degenerate. So we can think of the electron crossing the zone boundary as Bragg reflected. Correspondingly, the electron reverses its direction in real space. Over long times, this happens repeatedly and the electron does not achieve any net motion.

In reality, Bloch oscillations are difficult to observe because the time needed for an electron's quasimomentum to travel across the Brillouin zone is very long. In this time, the electron's trajectory is typically interrupted by scattering off of impurities or other electrons, which is not included in the semi-classical equations.

7.1.6 Example: uniform magnetic field

Now consider the case of a uniform magnetic field and zero electric field. The semi-classical equations become

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\nabla}_{\boldsymbol{k}}\tilde{\epsilon}_{n\boldsymbol{k}} - \frac{d\boldsymbol{k}}{dt} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}, \tag{268}$$

$$\frac{d\boldsymbol{k}}{dt} = -e\frac{d\boldsymbol{x}}{dt} \times \mathbf{B}. \tag{269}$$

We can use the first equation to eliminate position in the second equation and obtain

$$\begin{aligned}\frac{d\boldsymbol{k}}{dt} &= -e\left(\boldsymbol{\nabla}_{\boldsymbol{k}}\tilde{\epsilon}_{n\boldsymbol{k}} - \frac{d\boldsymbol{k}}{dt} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}\right) \times \mathbf{B} \\ &= -e\tilde{\boldsymbol{v}}_{n\boldsymbol{k}} \times \mathbf{B} - e\boldsymbol{\Omega}_{n\boldsymbol{k}} \cdot \mathbf{B}\,\frac{d\boldsymbol{k}}{dt}.\end{aligned} \tag{270}$$

Here $\tilde{\boldsymbol{v}}_{n\boldsymbol{k}} = \boldsymbol{\nabla}\tilde{\epsilon}_{n\boldsymbol{k}}$. We can solve this equation for $d\boldsymbol{k}/dt$:

$$\frac{d\boldsymbol{k}}{dt} = -\frac{e\tilde{\boldsymbol{v}}_{n\boldsymbol{k}} \times \mathbf{B}}{1 + e\mathbf{B} \cdot \boldsymbol{\Omega}_{n\boldsymbol{k}}}. \tag{271}$$

From this one can notice that the momentum parallel to $\mathbf{B}$ and the energy are constants of the motion:

$$\partial_t\,(\boldsymbol{k} \cdot \mathbf{B}) = \partial_t\tilde{\epsilon}_{n\boldsymbol{k}} = 0. \tag{272}$$

This implies that the quasimomentum evolves on a constant energy contour in a plane perpendicular to $\mathbf{B}$. This is generically a periodic curve – an "orbit" – which either closes within the Brillouin zone or wraps non-trivially around the periodic directions of the zone, which has the topology of a torus since quasimomentum is periodic. This is the generalization of a cyclotron orbit for

free electrons in a magnetic field.

In the usual textbooks, you will not find the denominator in Eq. (271). It has only relatively recently been understood as one of the effects of non-trivial Berry curvature in solids. For the moment, however, let us consider the case in which Berry phase effects vanish, $\Omega_{n\boldsymbol{k}} = \boldsymbol{m}_{n\boldsymbol{k}} = 0$, and take just

$$\frac{d\boldsymbol{k}}{dt} = -e\boldsymbol{v}_{n\boldsymbol{k}} \times \mathbf{B}. \tag{273}$$

This tells us the rate at which the quasi-momentum moves along its curve of constant energy and fixed $\boldsymbol{k} \cdot \mathbf{B}$.

The constant energy curves can be open or closed. A closed curve is one which is periodic in momentum in the extended zone scheme. If drawn in the extended zone scheme as a continuous curve, the curve encloses an area in the plane of motion. An open curve is periodic only in the reduced zone scheme: it consists of a trajectory that wraps around the torus on the Brillouin zone. In the extended zone scheme it is unbounded and drawn on the plane does not close at all but rather forms a "snaking" path extended to infinity. Since bands are typically quadratic near their maxima and minima, in these regions the constant energy curves are closed, and approximate ellipses. Near the middle of a band, the orbits may be open.

From the quasi-momentum solution we can obtain the real space one. Taking the cross product of Eq. (273) with magnetic field we find

$$\hat{\mathbf{B}} \times \frac{d\boldsymbol{k}}{dt} = -e\mathrm{B}\left(\boldsymbol{v}_{n\boldsymbol{k}} - \hat{\mathbf{B}}(\hat{\mathbf{B}} \cdot \boldsymbol{v}_{n\boldsymbol{k}})\right) \equiv -e\mathrm{B}\boldsymbol{v}_{n\boldsymbol{k},\perp}, \tag{274}$$

where $\boldsymbol{v}_{n\boldsymbol{k},\perp}$ is the projection of the velocity to the plane normal to the field. Then the coordinates in real space in the same plane are obtained from

$$\frac{d\boldsymbol{x}_\perp}{dt} = \boldsymbol{v}_{n\boldsymbol{k},\perp} = -\frac{1}{e\mathrm{B}}\hat{\mathbf{B}} \times \frac{d\boldsymbol{k}}{dt}. \tag{275}$$

Integrating this over time we see that

$$\boldsymbol{x}_\perp(t) - \boldsymbol{x}_\perp(0) = -\frac{1}{e\mathrm{B}}\hat{\mathbf{B}} \times (\boldsymbol{k}(t) - \boldsymbol{k}(0)). \tag{276}$$

Now we see that for closed orbits, for which $\boldsymbol{k}(t)$ is periodic, the motion in real space is also periodic. For open orbits, this is not the case, and the real space motion is unbounded. In the case of closed orbits, the motion is fully periodic in phase space because an electron repeatedly revisits the same region in phase space, there is a possibility for quantum interference. This leads to the phenomena of *quantum oscillations*. The basic mechanism is that if the phase accumulated by an electron over one period of the orbit is a multiple of 2π, then there is constructive interference and a discrete state results. The phase accumulated has a dynamical component which is the energy times the period T of the orbit (divided by $\hbar = 1$). So we should expect that in a quantum theory there are discrete energy levels with energy spacing $\Delta\epsilon = \hbar\omega_c$, where

$\omega_c = 2\pi/\mathrm{T}$ is the cyclotron frequency.

To determine this level spacing, and the cyclotron frequency, we need to figure out the period of the orbit. For this we can use Eq. (273), which we can invert to obtain

$$dt = \frac{|d\boldsymbol{k}|}{e|v_{n\boldsymbol{k},\perp}|\mathrm{B}}, \tag{277}$$

The period is the total time for this cyclotron orbit, hence

$$\mathrm{T}_c = \int_0^t dt' = \frac{1}{e\mathrm{B}} \oint \frac{|dk_{\parallel}|}{|v_{n\boldsymbol{k},\perp}|}. \tag{278}$$

Here $dk_{\parallel}$ is the change of $\boldsymbol{k}$ along the orbit, which is of course tangent to it. By definition, the velocity in the denominator is the in-plane component of the gradient of the energy, $v_{n\boldsymbol{k},\perp} = \lim_{\Delta\epsilon\to 0} \Delta\epsilon_n/\Delta k_{\perp}$, so one can write

$$\mathrm{T}_c = \int_0^t dt' = \frac{1}{e\mathrm{B}} \oint \frac{|dk_{\parallel}\Delta k_{\perp}|}{|\Delta\epsilon_n|}. \tag{279}$$

One can simplify this by taking $\Delta\epsilon$ fixed, and pulling it out of the integral. Then the integral gives the area of a ribbon or ring bounded by two the constant energy orbits of energy ϵ and $\epsilon + \Delta\epsilon$. Hence

$$\mathrm{T}_c = \frac{1}{e\mathrm{B}} \lim_{\Delta\epsilon\to 0} \frac{\mathrm{A}_n(\epsilon+\Delta\epsilon) - \mathrm{A}_n(\epsilon)}{\Delta\epsilon} = \frac{1}{e\mathrm{B}} \frac{\partial \mathrm{A}_n(\epsilon)}{\partial\epsilon}. \tag{280}$$

This is a general and compact result for the period of the generalized "cyclotron orbit". It is obvious that this applies only to closed orbits as the open orbits do not enclose an area!

The above analysis applies the semi-classical model purely classically.[1] We can consider quantum interference effects via the Bohr-Sommerfeld type quasi-classical quantization.

The standard Bohr-Sommerfeld condition is:

$$\mathrm{L}_z = \oint (\boldsymbol{k}_{\perp} - e\mathbf{A}) \cdot d\boldsymbol{x}_{\perp} = \frac{-1}{e\mathrm{B}} \oint \boldsymbol{k}_{\perp} \cdot \hat{\mathbf{B}} \times d\boldsymbol{k} - e\Phi = \frac{1}{e\mathrm{B}} \oint (k_x dk_y - k_y dk_x) - e\mathrm{B}\mathrm{A}_{\text{real space}}, \tag{281}$$

where Φ is the flux through the orbit in real space. For a uniform field, this

[1] Sorry for the confusing language. The semi-classical equations are semi-classical in the sense that they use input from the full quantum solution of the Bloch problem of an ideal crystal, e.g. $\epsilon_{n\boldsymbol{k}}$, $\boldsymbol{\Omega}_{n\boldsymbol{k}}$, etc. The equations themselves are however classical equations in that $\boldsymbol{x}$ and $\boldsymbol{k}$ in these equations are just numbers, and there is wavefunction implemented, and no uncertainty principle in these equations. So we can talk about the classical semi-classical equations!—

equals B times the area of the orbit in real space. Here we Eq. (275) to express $d\boldsymbol{x}$ in terms of $d\boldsymbol{k}$. Now the final integral is twice the area in momentum space. Morevoer, from Eq. (276), the area of the orbit in real space is $1/(e\mathrm{B})^2$ times the area in momentum space. Taking both into account, we find that

$$\mathrm{L}_z = \frac{\mathrm{A}(\epsilon, k_z)}{e\mathrm{B}} = h(n+\nu) = 2\pi(n+\nu), \tag{282}$$

in our units with $\hbar = 1$. Here ν is some offset due to geometrical phase contributions beyond the dynamical phase due to just angular momentum.

Strictly speaking, we should be careful in literally interpreting the semi-classical Bohr-Sommerfeld result as true quantization of energy levels; for the latter we should apply a fully quantum approach. The Bohr-Sommerfeld results rather describe the onset of quantum interference effects. One may have in mind that because of scattering, electrons cannot traverse their periodic orbits an infinite number of times before a collision. If the typical time between collisions is τ, then the number of times an orbit can be encircled is $\omega_c\tau$. When $\omega_c\tau$ is finite, the collapse of allowed energies into discrete levels is incomplete and one should instead regard Eq. (282) as defining energies where constructive interference leads to a maximum in the density of states (actually the partial DOS at fixed k_x).

These oscillations in the density of states manifest in many physical quantities. They appear as oscillations of the magnetization in a field, a phenomena known as the de Haas van Alphen effect. They induce oscillations in the resistivity with field, which are known as Shubnikov de Haas oscillations.

In many experiments, the electron density and hence the Fermi level is held fixed, while the magnetic field is varied. In this case, Eq. (282) implies that maxima of the DOS occur at the Fermi energy when $\mathrm{B} = \mathrm{B}_n$ such that this condition holds. This defines a set of discrete magnetic fields. One can see that this condition is linear in n for 1/B, hence inverse magnetic fields are evenly spaced:

$$\Delta\left(\frac{1}{\mathrm{B}}\right) = \frac{2\pi e}{\mathrm{A}(\epsilon_\mathrm{F}, k_z)}. \tag{283}$$

Hence by extracting the maxima in DOS versus 1/B, one can obtain information on the areas of sections of the Fermi surface.

Our discussion does not clarify which values of k_z, i.e. which sections, contribute for a three-dimensional Fermi surface. It turns out that the DOS oscillations are dominated by *extremal orbits*, i.e. the cuts in which the areas are maximal or minimal. The reason is that the full DOS is an integral over the contributions for each k_z, and dominant contributions to this integral arise from the regions in which the oscillations versus 1/B vary least with k_z: these are the extremal orbits. This is a powerful tool because one can rotate the sample or field to choose different k_z, thereby probing different cuts and different extremal areas.

An alternative use of Eq. (282) is to inquire about the oscillations of the DOS versus energy at fixed magnetic field. This is hard to probe in conventional bulk measurements but can be studied using spectroscopic probes and in two dimensional materials. When the magnetic field is small, the energy level spacing must also be small, so we have

$$\Delta\left(\frac{\mathrm{A}(\epsilon, k_z)}{e\mathrm{B}}\right) \approx \frac{\partial \mathrm{A}/\partial\epsilon}{e\mathrm{B}}\Delta\epsilon = 2\pi. \tag{284}$$

Hence the energy level spacing is

$$\Delta\epsilon = \frac{2\pi e\mathrm{B}}{\partial \mathrm{A}/\partial\epsilon}. \tag{285}$$

We see that this spacing corresponds to the cyclotron frequency

$$\Delta\epsilon = \omega_c = \frac{2\pi}{\mathrm{T}_c} = \frac{2\pi e\mathrm{B}}{\partial \mathrm{A}/\partial\epsilon}. \tag{286}$$

Note that unlike Eq. (283), this result is only true for the difference of nearby energy levels, because we Taylor expanded the area (and consequently ω_c depends on energy through the energy dependence of the derivative of the area).

7.2 *Boltzmann equation*

The semi-classical equations, Eq. (220), describe the motion of individual electrons, or more properly single electron wave functions. When we deal with many electrons, we need to account for the simultaneous motion of multiple states. This is often conveniently done using Boltzmann's equation. The Boltzmann equation is derived for a classical ensemble of particles obeying some classical equation of motion. The basic object of the Boltzmann description is the *phase space density*, $f_n(\boldsymbol{x}, \boldsymbol{k}, t)$, which describes the number of electrons with position $\boldsymbol{x}$, momentum $\boldsymbol{k}$, in a given band, and at time t. More precisely,

$$f_n(\boldsymbol{x}, \boldsymbol{k}, t)\, d^d\boldsymbol{x}\, \frac{d^d\boldsymbol{k}}{(2\pi)^d} = \begin{array}{l}\text{occupation of electron wave packet states in}\\ \text{a volume } d^d\boldsymbol{x} \text{ around position } \boldsymbol{x} \text{ in a range}\\ d^d\boldsymbol{k} \text{ around quasi-momentum } \boldsymbol{k} \text{ in band } n \text{ of}\\ \text{a single spin polarization.}\end{array} \tag{287}$$

This clearly requires at least a semiclassical limit to make sense because both quasi-momentum and position are specified, and we should have in mind the wave packet picture. The normalization was chosen (and in particular the factors of 2π) so that, in equilibrium, the distribution function becomes the Fermi-Dirac distribution,

$$f_n(\boldsymbol{x}, \boldsymbol{k})|_{\text{equilibrium}} = n_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}}). \tag{288}$$

From this distribution function, one can obtain various physical quantities, for example the density of electrons as a function of position is

$$n(\boldsymbol{x}, t) = 2 \sum_n \int_{\mathrm{BZ}} d^d \boldsymbol{k} \mathrm{D}_n(\boldsymbol{k}) f_n(\boldsymbol{x}, \boldsymbol{k}, t). \tag{289}$$

Here $\mathrm{D}_n(\boldsymbol{k})$ is the density of levels per unit volume in momentum space and volume in real space. For conventional plane waves, and for Bloch states in zero magnetic field, we have the usual form

$$\mathrm{D}_n(\boldsymbol{k})|_{\mathbf{B}=0} = \frac{1}{(2\pi)^d}. \tag{290}$$

This might seem inviolate, but we will see that in fact Eq. (290) must be modified in the combined presence of both magnetic field and Berry curvature.

The Boltzmann equation describes the time evolution of the distribution function. The basic assumption of the Boltzmann equation is that particles undergo two types of motion. They evolve smoothly according to the semi-classical equations, punctuated by abrupt "collisions" which interrupt the smooth evolution and can be regarded as instantaneous compared to the smooth evolution. Collisions are generally scattering events in which electrons may scatter individually off of impurities, or collectively off of one another, with phonons, etc. They are treated probabilistically. To derive the Boltzmann equation, we consider how the distribution changes in an infinitesimal time dt. For a very short time, the changes in position and momentum of particles due to the smooth evolution are order dt and the number of collisions that occur is also of order dt. Thus for the colliding electrons, we can ignore the smooth evolution for infinitestimal dt, because this would involve order dt^2 changes in the distribution. In other words, we can consider the effect of the two types of evolution independently. Their effects will influence one another as one adds up changes for many infinitesimal time intervals.

7.2.1 Evolution between scattering events

First consider the smooth evolution. Because electrons are not created or destroyed, and move smoothly, a small phase space volume just transforms to a slightly new phase space volume over the time dt. The electrons just move from one volume to another. So we can equate the initial occupation of a single particle wave packet state specified by $\boldsymbol{x}(t), \boldsymbol{k}(t)$ in band n at time t with the final occupation at $\boldsymbol{x}(t+dt), \boldsymbol{k}(t+dt)$ in the same band n at time $t+dt$. Here,

$$\boldsymbol{k}(t+dt) = \boldsymbol{k}(t) + \mathbf{F}_n dt, \qquad \boldsymbol{x}(t+dt) = \boldsymbol{x}(t) + \boldsymbol{v}_n dt, \tag{291}$$

where

$$\mathbf{F}_n = \frac{d\boldsymbol{k}_n}{dt}, \qquad \boldsymbol{v}_n = \frac{d\boldsymbol{x}_n}{dt}. \tag{292}$$

Hence we have

$$f_n(\boldsymbol{x} + \boldsymbol{v}_n dt, \boldsymbol{k} + \mathbf{F}_n dt, t + dt) = f_n(\boldsymbol{x}, \boldsymbol{k}, t). \tag{293}$$

Taylor expand this to order dt and one obtains

$$(\partial_t + \boldsymbol{v}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} + \mathbf{F}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{k}})\, f_n = 0. \tag{294}$$

This holds in the absence of collisions. With collisions, we write

$$(\partial_t + \boldsymbol{v}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} + \mathbf{F}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{k}})\, f_n = \left.\frac{\partial f_n(\boldsymbol{x}, \boldsymbol{k}, t)}{\partial t}\right|_{\text{collisions}}. \tag{295}$$

We will return to the collision term shortly.

In general one should solve the semi-classical equations to obtain $\boldsymbol{v}_n, \mathbf{F}_n$ in terms of $\boldsymbol{x}, \boldsymbol{k}$. To do so, first we take the dot product of the position equation with $\boldsymbol{\Omega}_{n\boldsymbol{k}}$ and of the momentum equation with $\mathbf{B}$ to obtain

$$\frac{d\boldsymbol{x}}{dt} \cdot \boldsymbol{\Omega}_{n\boldsymbol{k}} = \boldsymbol{\Omega}_{n\boldsymbol{k}} \cdot \nabla_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}}, \qquad \frac{d\boldsymbol{k}}{dt} \cdot \mathbf{B} = -e\mathbf{E} \cdot \mathbf{B}. \tag{296}$$

Now we insert the expression for $d\boldsymbol{k}/dt$ into the position equation to obtain a closed equation for the latter,

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \nabla_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}} + e\left(\mathbf{E} + \frac{d\boldsymbol{x}}{dt} \times \mathbf{B}\right) \times \boldsymbol{\Omega}_{n\boldsymbol{k}} \\ &= \nabla_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}} + e\mathbf{E} \times \boldsymbol{\Omega}_{n\boldsymbol{k}} + e\left(\frac{d\boldsymbol{x}}{dt} \cdot \boldsymbol{\Omega}_{n\boldsymbol{k}}\right)\mathbf{B} - e\mathbf{B} \cdot \boldsymbol{\Omega}_{n\boldsymbol{k}} \frac{d\boldsymbol{x}}{dt}. \end{aligned} \tag{297}$$

Now using Eq. (296) and solving for the velocity, we obtain

$$\boldsymbol{v}_n = \frac{d\boldsymbol{x}}{dt} = \frac{\nabla_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}} + e\mathbf{E} \times \boldsymbol{\Omega}_{n\boldsymbol{k}} + e(\boldsymbol{\Omega}_{n\boldsymbol{k}} \cdot \nabla_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}})\mathbf{B}}{1 + e\mathbf{B} \cdot \boldsymbol{\Omega}_{n\boldsymbol{k}}}. \tag{298}$$

Performing parallel manipulation for the rate of change of quasi-momentum, one finds

$$\mathbf{F}_n = \frac{d\boldsymbol{k}}{dt} = \frac{-e\mathbf{E} - e\nabla_{\boldsymbol{k}} \tilde{\epsilon}_{n\boldsymbol{k}} \times \mathbf{B} - e^2 (\mathbf{E} \cdot \mathbf{B})\, \boldsymbol{\Omega}_{n\boldsymbol{k}}}{1 + e\mathbf{B} \cdot \boldsymbol{\Omega}_{n\boldsymbol{k}}}. \tag{299}$$

Eqs. (298,299) give explicit forms for $\mathbf{F}_n$ and $\boldsymbol{v}_n$ to be used in the Boltzmann equation.

If you compare to standard literature and all the textbooks I am aware of, you will not see the denominators in these equations, and most likely some or

all of the Berry curvature terms will be missing. This is because these were not recognized until relatively recently, and a complete theory consistently including all of them is less than a decade old. In fact, when the denominator is not equal to the identity, i.e. when $\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}} \neq 0$, there is a remarkable physical consequence: the phase space density of states, Eq. (290) must be modified to

$$\mathrm{D}_n(\boldsymbol{k}) = \frac{1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}}}{(2\pi)^d}. \tag{300}$$

This is required because the volume element in phase space, $d^d\boldsymbol{x}d^d\boldsymbol{k}$ is not invariant under the semi-classical dynamics when $\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}} \neq 0$, that is, these equations no longer obey Liouville's theorem. To see this explicitly, consider a volume element in momentum space. The volume element $d^d\boldsymbol{k}$ transforms after evolution by a time dt to a volume element $d^d\boldsymbol{k}'$, where $\boldsymbol{k}' = \boldsymbol{k} + \mathbf{F}_n dt$. Using the standard Jacobean for transforming a measure, we have

$$\begin{aligned} d^d\boldsymbol{k}' &= d^d\boldsymbol{k}\left|\det\frac{\partial\boldsymbol{k}'}{\partial\boldsymbol{k}}\right| \\ &= d^d\boldsymbol{k}\,\det\left(\delta_{\mu\nu} + \frac{\partial \mathrm{F}_\mu}{\partial k_\nu}dt\right) = d^d\boldsymbol{k}\left(1 + \frac{\partial \mathrm{F}_\mu}{\partial k_\mu}dt\right). \end{aligned} \tag{301}$$

Now we can evaluate the divergence of the force from Eq. (299). The divergence of the numerator in Eq. (299) is zero, but there is a contribution from the denominator

$$\begin{aligned} \frac{\partial \mathrm{F}_\mu}{\partial k_\mu} &= -\mathrm{F}_\mu\frac{e\mathbf{B}\cdot\partial_\mu\Omega_{n\boldsymbol{k}}}{1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}}} \\ &= -\mathrm{F}_\mu\frac{\partial(1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}})/\partial k_\mu}{1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}}} = -\frac{dk_\mu}{dt}\frac{\partial(1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}})/\partial k_\mu}{1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}}} \\ &= -\frac{d(1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}})/dt}{1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}}}. \end{aligned} \tag{302}$$

This is clearly non-zero so the volume element in momentum space changes under the time evolution. We can find out how it changes by inserting this into Eq. (301). One obtains

$$d^d\boldsymbol{k}' = d^d\boldsymbol{k}\left(1 - \frac{d(1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}})}{1 + e\mathbf{B}\cdot\mathbf{\Omega}_{n\boldsymbol{k}}}\right). \tag{303}$$

Now multiply this equation by $\mathrm{D}_n(\boldsymbol{k}')$ on both sides:

$$\begin{aligned} d^d\boldsymbol{k}'\,\mathrm{D}_n(\boldsymbol{k}') &= d^d\boldsymbol{k}\left(\mathrm{D}_n(\boldsymbol{k}') - \frac{d\mathrm{D}_n(\boldsymbol{k})}{\mathrm{D}_n(\boldsymbol{k})}\mathrm{D}_n(\boldsymbol{k}')\right) \\ &= d^d\boldsymbol{k}\,(\mathrm{D}_n(\boldsymbol{k}') - d\mathrm{D}_n(\boldsymbol{k})) = d^d\boldsymbol{k}\,\mathrm{D}_n(\boldsymbol{k}). \end{aligned} \tag{304}$$

In the second line we used that in the second term of the first line $\mathrm{D}_n(\boldsymbol{k}')$ can be taken equal to $\mathrm{D}_n(\boldsymbol{k})$ to first order in the infinitesimal since it multiplies an infinitesimal. So we see that the measure modified by $\mathrm{D}_n(\boldsymbol{k})$ is invariant under the evolution.

Consequently, the "stretching" of volumes in phase space under time evolution means that the when we count the number of electrons in a volume, we must take into account of the change in the local density of states. One can check that using the proper phase space density of states in Eq. (300) ensures proper conservation laws, e.g. that the particle density in Eq. (289) obeys a continuity equation (Try multiplying Eq. (466) by $\mathrm{D}_n(\boldsymbol{k})$ and integrating over position and space, and show that the spatial integral of the density, i.e. the electron number, is time independent). The modified density of states must be used when integrating over quasi-momentum to obtain any other physical quantity from its value in a single wave packet, for example the electric current is

$$\boldsymbol{j}_e(\boldsymbol{x},t) = 2\sum_n \int d^d\boldsymbol{k}\, \mathrm{D}_n(\boldsymbol{k})(-e\boldsymbol{v}_n(\boldsymbol{k},\boldsymbol{x}))f_n(\boldsymbol{x},\boldsymbol{k},t). \tag{305}$$

7.2.2 Collisions

Now we turn to the collisions. In general, we treat the effects of scattering statistically in terms of rates. The collision term gives the rate at which electrons are scattered *into* the quasi-classical state with the given set of phase space quantum numbers (band index, quasi-momentum, position), *minus* the rate at which electrons are scattered out from this quasi-classical state into all other states. In general both the "incoming" and "outgoing" rates depend upon the occupation of one or more states.

EXAMPLE: ELASTIC SINGLE-PARTICLE SCATTERING It is helpful to consider an example. The simplest and canonical case is elastic scattering due to impurities/defects in the solid. An electron in band n and quasi-momentum $\boldsymbol{k}$ is scattered at a rate $\Gamma_{n'\boldsymbol{k}'n\boldsymbol{k}}$ to band n' and quasi-momentum $\boldsymbol{k}'$ (we will ignore spatial dependence, i.e. assume the scattering is local, and suppress the explicit position label). In this case the collision term is

$$\left.\frac{\partial f_n(\boldsymbol{x},\boldsymbol{k},t)}{\partial t}\right|_{\text{collisions}} = \sum_{n'}\int \frac{d^d\boldsymbol{k}'}{(2\pi)^d}\left[\Gamma_{n\boldsymbol{k}n'\boldsymbol{k}'}(1-f_n(\boldsymbol{k}))f_{n'}(\boldsymbol{k}') - \Gamma_{n'\boldsymbol{k}'n\boldsymbol{k}}(1-f_{n'}(\boldsymbol{k}'))f_n(\boldsymbol{k})\right]. \tag{306}$$

Here the first term represents scattering *into* the state $n, \boldsymbol{k}$, and is hence positive, and carries a factor of $f_{n'}(\boldsymbol{k}')$ which counts the number of electrons in the initial state, and a factor $1 - f_n(\boldsymbol{k})$ which ensures that the process is not Pauli blocked. The second term represents scattering *out* of the state $n, \boldsymbol{k}$: it is negative, proportional to the occupation $f_n(\boldsymbol{k})$ of the initial state, and has the $1 - f_{n'}(\boldsymbol{k}')$ to account for Pauli blocking.

It is often but not necessarily the case that the rate is symmetric, i.e. $\Gamma_{n\boldsymbol{k}n'\boldsymbol{k}'} = \Gamma_{n'\boldsymbol{k}'n\boldsymbol{k}}$[2]. This is the case for example for elastic scattering in the first Born approximation, within which

$$\Gamma_{n'\boldsymbol{k}'n\boldsymbol{k}}|_{\text{1st Born, elastic}} = 2\pi \left|\langle \psi_{n'\boldsymbol{k}'}|\hat{\mathrm{H}}_{\text{impurity}}|\psi_{n\boldsymbol{k}}\rangle\right|^2 \delta(\epsilon_{n\boldsymbol{k}} - \epsilon_{n\boldsymbol{k}}). \tag{307}$$

When the rate is symmetric, the terms quadratic in the occupation cancel and one has simply

$$\left.\frac{\partial f_n(\boldsymbol{k}, t)}{\partial t}\right|_{\text{collisions}} = \sum_{n'} \int \frac{d^d \boldsymbol{k}'}{(2\pi)^d} \Gamma_{n\boldsymbol{k}n'\boldsymbol{k}'} \left[f_{n'}(\boldsymbol{k}') - f_n(\boldsymbol{k})\right]. \tag{308}$$

Inspecting Eq. (306) or the simplification Eq. (308), you can observe that the collision term takes the form of an integral transform of the distribution function. For more complex types of scattering, e.g. electron-electron interactions, it takes the form of a non-linear integral operator acting on the distribution. The collision term is therefore often call the "collision integral". The full equation, putting the collision term back into Eq. (295), is a non-linear integro-differential equation.

Please note that in writing down Eq. (306) we have treated the transitions between states entirely classically, and added the probabilities (i.e. rates) for transitions from different states. At one level, this is a fundamental assumption of the Boltzmann equation, and we cannot depart too far from the classical treatment and still use the semi-classical dynamics. However, there can be some deviations from classical behavior, e.g. during the scattering events, or as small corrections in multiple scattering. Consequently one may in some circumstances need to amend or modify the description.

7.2.3 Relaxation time approximation

The collision integral can contain many complexities which are not always central. To uncover some general features and to enable simpler calculations, it is useful to introduce a phenomenological form of the collision term known as the relaxation time approximation. The idea is by design to make the collision term local, i.e. a multiplication operator instead of an integral one. The most general such form is

[2] This is definitely the case if one has time reversal and inversion symmetry, and the scattering is elastic. I am not sure if one needs all these conditions.

$$\left.\frac{\partial f_n(\boldsymbol{k},t)}{\partial t}\right|_{\text{collisions}} = -\frac{1}{\tau_{n\boldsymbol{k}}}\left[f_n(\boldsymbol{k}) - f_n^{(0)}(\boldsymbol{k})\right], \tag{309}$$

where $\tau_{n\boldsymbol{k}}$ is called a relaxation time. The constant distribution function $f_n^{(0)}(\boldsymbol{k})$ is determined by the requirement that in the absence of applied forces, $\mathbf{F}_n = 0$ the distribution relaxes to the (spatially uniform) equilibrium distribution. Therefore

$$f_n^{(0)}(\boldsymbol{k}) = n_{\text{F}}(\epsilon_{n\boldsymbol{k}}). \tag{310}$$

(I waffle about whether to include μ in the definition of n_{F} or not...)

The relaxation time approximation is a crude approximation, and fails to account for numerous features, but it does allow a rough understanding of the interplay of scattering and forces on electrons. In the relaxation time approximation, the full Boltzmann equation is

$$(\partial_t + \boldsymbol{v}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} + \mathbf{F}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{k}})\, f_n = -\frac{1}{\tau_{n\boldsymbol{k}}}\left[f_n(\boldsymbol{k}) - f_n^{(0)}(\boldsymbol{k})\right]. \tag{311}$$

7.3 *Zero field conductivity in the relaxation time approximation*

Let's take a break from developing formalism to actually calculate a conductivity. The simplest case we can take is a uniform applied electric field $\mathbf{E}$ which is constant in space and time, and assume zero magnetic field $\mathbf{B} = 0$, and we use the relaxation time approximation. Then we can assume the distribution function is independent of $\boldsymbol{x}$, and we seek a steady-state solution $\partial_f f_n = 0$. Then we have $\mathbf{F}_n = -e\mathbf{E}$, and so

$$-e\mathbf{E} \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} f_n(\boldsymbol{k}) = -\frac{1}{\tau_{n\boldsymbol{k}}}\left[f_n(\boldsymbol{k}) - f_n^{(0)}(\boldsymbol{k})\right]. \tag{312}$$

This equation is simple enough to be solvable exactly. But it is not really so useful to do so. Let us instead assume a weak electric field, and compute the linear response of the system to that field. Then we can expand the distribution function in a perturbation series in the field strength

$$f_n(\boldsymbol{k}) = f_n^{(0)}(\boldsymbol{k}) + \delta f_n(\boldsymbol{k}) + \cdots, \tag{313}$$

where δf_n is O(E). Inserting this into Eq. (312) and collecting terms of O(E), one obtains

$$-e\mathbf{E} \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} f_n^{(0)}(\boldsymbol{k}) = -\frac{1}{\tau_{n\boldsymbol{k}}}\delta f_n(\boldsymbol{k}), \tag{314}$$

so using $f_n^{(0)}(\boldsymbol{k}) = n_{\text{F}}(\epsilon_{n\boldsymbol{k}})$ we directly obtain the solution

$$\delta f_n(\boldsymbol{k}) = e\tau_{n\boldsymbol{k}} n'_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}}) \boldsymbol{v}_{n\boldsymbol{k}} \cdot \mathbf{E}, \tag{315}$$

where $\boldsymbol{v}_{n\boldsymbol{k}} = \nabla_{\boldsymbol{k}} \epsilon_{n\boldsymbol{k}}$ is the band group velocity. Note that factor of the derivative of the Fermi function means that the modifications to the distribution function occur only very near the Fermi energy. Now we can compute the current using Eq. (305), which in this case simplifies to

$$\boldsymbol{j}_e = -2e \sum_n \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} \left(\boldsymbol{v}_{n\boldsymbol{k}} + e\mathbf{E} \times \boldsymbol{\Omega}_{n\boldsymbol{k}}\right) \left(n_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}}) + e\tau_{n\boldsymbol{k}} n'_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}}) \boldsymbol{v}_{n\boldsymbol{k}} \cdot \mathbf{E}\right). \tag{316}$$

Here we can consistently keep terms only to linear order in the electric field. Multiplying out the product above, the zeroth order term in the electric field vanishes under integration (it is a total derivative), which ensures that the current density is zero in equilibrium. We can drop the $\mathrm{O}(\mathrm{E}^2)$ term. We are left with two contributions:

$$\begin{aligned} \boldsymbol{j}_e =& 2e^2 \sum_n \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} \boldsymbol{v}_{n\boldsymbol{k}} \left(\boldsymbol{v}_{n\boldsymbol{k}} \cdot \mathbf{E}\right) \tau_{n\boldsymbol{k}} [-n'_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}})] \\ &+ 2e^2 \sum_n \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} n_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}}) \boldsymbol{\Omega}_{n\boldsymbol{k}} \times \mathbf{E}. \end{aligned} \tag{317}$$

This can be written as a conductivity tensor

$$\boldsymbol{j}_e^{\mu} = \sigma_{\mu\nu} \mathrm{E}_{\nu}, \tag{318}$$

where $\sigma_{\mu\nu} = \sigma^s_{\mu\nu} + \sigma^a_{\mu\nu}$ and

$$\sigma^s_{\mu\nu} = 2e^2 \sum_n \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} v^{\mu}_{n\boldsymbol{k}} v^{\nu}_{n\boldsymbol{k}} \tau_{n\boldsymbol{k}} [-n'_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}})], \tag{319}$$

and

$$\sigma^a_{\mu\nu} = -2e^2 \left[\sum_n \int \frac{d^d \boldsymbol{k}}{(2\pi)^d} n_{\mathrm{F}}(\epsilon_{n\boldsymbol{k}}) \Omega^{\lambda}_{n\boldsymbol{k}}\right] \epsilon_{\lambda\mu\nu}. \tag{320}$$

We will discuss these two contributions one by one.

7.3.1 Symmetric/dissipative conductivity:

The contribution $\sigma^s_{\mu\nu}$ first term is symmetric and corresponds to the "longitudinal" conductivity. In many textbooks this is the only contribution you will find. For an isotropic system, or for high enough crystalline (e.g. cubic) symmetry, it is diagonal and proportional to the identity matrix. More generally, this term is dissipative, in that it contributes to ohmic heating, since the rate of power dissipation, from elementary electromagnetism, is $\boldsymbol{j} \cdot \mathbf{E} = \sigma_{\mu\nu} \mathrm{E}_{\mu} \mathrm{E}_{\nu}$. One might worry about this: if energy is being constantly pumped into the

system, it should heat up. That is indeed true, and in reality there needs to be some heat sink to balance the joule heating: this role is typically played by coupling to the lattice, which we have not included. However, because the joule heating is quadratic in the electric field, this physics does not modify the linear response result.

Note that σ^s contains the derivative of the Fermi function. Therefore it is dominated by the region near the Fermi energy, i.e. the Fermi surface. Indeed, at low temperature, it can be approximated as

$$\sigma^s_{\mu\nu} = e^2 \sum_n D_n(\epsilon_F) \left\langle v^\mu_{n\boldsymbol{k}} v^\nu_{n\boldsymbol{k}} \tau_{n\boldsymbol{k}} \right\rangle_{FS,n}, \tag{321}$$

where the angular brackets here denote an average over the Fermi surface in band n.

Drude conductivity: For a simple understanding, consider the case of a free electron band in three dimensions, with $\epsilon = k^2/(2m^*)$, and momentum-independent relaxation time τ. Then we can use Eq. (207) to write $D(\epsilon_F) = m^* k_F/\pi^2$, and

$$\begin{aligned}\sigma^s_{\mu\nu} &= e^2 \frac{m^* k_F}{\pi^2} \langle v^\mu v^\nu \rangle_{FS} \tau = e^2 \frac{m^* k_F}{\pi^2} \frac{v_F^2}{3} \delta_{\mu\nu} \tau \\ &= e^2 \frac{m^* k_F}{\pi^2} \frac{k_F^2}{3(m^*)^2} \tau \delta_{\mu\nu} = \frac{k_F^3}{3\pi^2} e^2 \frac{\tau}{m^*} \delta_{\mu\nu} = \frac{n e^2 \tau}{m^*} \delta_{\mu\nu}.\end{aligned} \tag{322}$$

The final result is the well-known form from the simple Drude-Sommerfeld theory. It can be obtained by writing the equation of motion for the average velocity of the electrons, $\boldsymbol{v}_d = \frac{1}{N} \sum_i \boldsymbol{v}_i$, known as the drift velocity,

$$m^* \left(\frac{d\boldsymbol{v}_d}{dt} + \frac{\boldsymbol{v}_d}{\tau} \right) = -e\mathbf{E}. \tag{323}$$

One can obtain this from the Boltzmann equation in the relaxation time approximation by assuming the quadratic dispersion, and defining the drift velocity as the average of $\boldsymbol{k}/m^*$ over the momentum distribution. Solving this equation in the steady state gives $\boldsymbol{v}_d = -(e\tau/m^*)\mathbf{E}$ and one can readily obtain the conductivity from $\boldsymbol{j}_e = -ne\boldsymbol{v}_d$.

Temperature dependence of resistivity in metals: From both Eq. (321) and Eq. (322) we see that the resistivity (e.g. $\rho = m/(ne^2\tau)$ in the Drude approximation) becomes temperature-independent at low temperature, and is limited by the zero temperature scattering rate $1/\tau(T = 0)$. In general, this zero temperature scattering is due to elastic scattering by impurities, so the zero temperature resistivity, known as the *residual resistivity* is entirely dependent upon the purity of the material. A typical value for a "good" bulk

three dimensional metal at low temperature is in the μΩ-cm range.

The temperature dependence of resistivity in metals arises mainly from other sources of scattering.[3] In addition to elastic impurity scattering, electrons may scatter with one another, with phonon excitations of the lattice, or with any other excitations that might be present. The relaxation time should include the total rate of scattering, which generally one may write as

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{imp}}} + \frac{1}{\tau_{\text{inelastic}}}. \tag{324}$$

The additivity of scattering rates is known as Matthiessen's rule. An important property of the inelastic scattering rate, which includes both electron-electron and electron-lattice scattering, is that it vanishes as $T \to 0$, usually as a power law. To see this, we need to clarify what is meany by "inelastic" scattering. This is a scattering process which is due to interaction terms in the Hamiltonian. As such, it of course still conserves the total energy in the system. It is inelastic only in the sense that it does not conserve the energy of individual electrons. For example, an inelastic electron-electron scattering process would involve two electrons with energy ϵ_1 and ϵ_2 scattering to states with energy ϵ_3 and ϵ_4, with the sum conserved but none of the four energies equal to one another. For such a process to be possible, we need that there are electrons present in states 1 and 2, and not present in states 3 and 4. At zero temperature, this is not possible, because if states 1 and 2 are occupied they are below the Fermi energy, and thereby at least one of state 3 and 4 must be below the Fermi energy by energy conservation. The process becomes possible for $T > 0$ because the occupation of the states below the Fermi energy is not unity. Hence generally at low temperature

$$\frac{1}{\tau_{\text{inelastic}}} \sim \sum_i c_i T^{a_i}, \tag{325}$$

a sum of terms corresponding to different types of scattering processes, where the coefficients $c_i > 0$ and a_i are some exponents determined by phase space arguments. Standard arguments give $a = 2$ for electron-electron interactions, $a = 5$ for low temperature scattering off of acoustic phonons. At higher temperature, e.g. comparable to characteristic phonon energies, there is less justification for power-law behavior, but $a = 1$ is sometimes seen and can be argued for. Fractional powers can also arise, e.g. $a = 3/2$ in ferromagnets.

Putting this together, the typical resistivity in a metal starts at the residual resistivity at low temperature, and rises continuously up to high temperature, roughly as

[3] One might also think of temperature dependence arising from the Fermi function in Eq. (319). In bulk metals, this temperature dependence is very weak, because temperature is much smaller than the Fermi energy even at room temperature. The temperature dependence of scattering rates is much stronger and dominates the temperature dependence of the resistivity. This may no longer be true in low density electron systems.

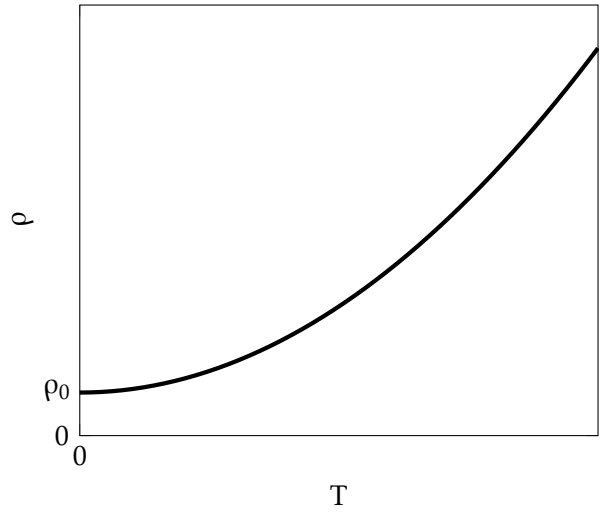


Figure 7: Schematic form of resistivity in a metal

$$\rho(T) = \rho_0 + \sum_i A_i T^{a_i}, \tag{326}$$

where ρ_0 is the residual resistivity. A typical measure of quality is to compare the room temperature resistivity to the low temperature one, which is known as the *Residual Resistivity Ratio*, or RRR,

$$RRR = \frac{\rho(T_{room})}{\rho(T=0)} \simeq \frac{\rho(273K)}{\rho(4.2K)}. \tag{327}$$

The idea here is that the room temperature resistivity is nominally intrinsic, controlled by the inelastic scattering rates which do not depend on impurities, while the residual resistivity is determined by impurities. A larger RRR means a better quality material.

Einstein relation: Eq. (321) connects directly to a very general identity which connects the conductivity to the density of states and diffusion constant, which is known as the Einstein relation. The Einstein relation says the conductivity is given by

$$\sigma = e^2 \frac{\partial n}{\partial \mu} D, \tag{328}$$

where $\partial n/\partial \mu$ is the electronic compressibility, and D is the diffusion constant. This is typically derived for an isotropic medium, but one can generalize to the anisotropic case by

$$\sigma^s_{\mu\nu} = e^2 \frac{\partial n}{\partial \mu} D_{\mu\nu}. \tag{329}$$

In general the diffusion constant is a symmetric tensor. To obtain the Einstein relation, we first must define the diffusion constant. It is defined by considering the particle current driven by a particle density gradient. In reality a gradient of electron density inevitably induces an electric field, so there will always be some current driven by electric field as well in this situation. However, we presume that the net current in general is the sum of a contribution driven by a density gradient and one driven by an electric field. We focus first on the former, called a diffusion current,

$$j^{\mu}_{n,\text{diff}} = -\mathrm{D}_{\mu\nu}\partial_{\nu} n. \tag{330}$$

This relation is called "Fick's law". The subscript n indicates that this is a particle current density. Now by the continuity equation,

$$\partial_t n = -\boldsymbol{\nabla}\cdot\boldsymbol{j}_{n,\text{diff}} = \mathrm{D}_{\mu\nu}\partial_{\mu}\partial_{\nu} n. \tag{331}$$

This is the (anisotropic) diffusion equation, and this defines the diffusion constant (tensor). Taking $\mathrm{D}_{\mu\nu}$ from this equation only defines a symmetric tensor, since $\partial_{\mu}\partial_{\nu}$ is symmetric. So the diffusion tensor is symmetric and the Einstein relation defines only the symmetric part of the conductivity (any anti-symmetric term in the relation of current to density gradient will vanish when the divergence is taken in the continuity equation).

If you are not familiar with the diffusion equation, Eq. (331) describes the spread of a initial non-uniform density profile in the absence of any applied forces. It applies for example to molecules in a gas, or to heat in a solid, and in many other situations. Microscopically diffusion arises out of random walks. In the simplest case, one may consider an ensemble of particles, each of which propagates ballistically with some velocity v in a random direction, and scatters every time τ_{rw}, emerging from the scattering in a new random direction. Because of the randomness in direction, the typical distance traveled after a time t much longer than τ_{rw} is much less than vt. Specifically, the final coordinate after a time $t = n\tau_{rw}$ is

$$\boldsymbol{x}(t) - \boldsymbol{x}(0) = \sum_{i=1}^{n} v\hat{\boldsymbol{v}}_i\tau_{rw}, \tag{332}$$

where $\hat{v}_i$ is the random direction for scattering i. The variance of the displacement defines a matrix

$$\begin{aligned}\overline{\big(x_{\mu}(t) - x_{\mu}(0)\big)\big((x_{\nu}(t) - x_{\nu}(0)\big)} &= \overline{v^2\tau_{rw}^2\sum_{i,j=1}^{n}\hat{v}_i^{\mu}\hat{v}_j^{\nu}} = n\ell_{rw}^2\,\overline{v_i^{\mu}v_i^{\nu}} \\ &= n\frac{\ell_{rw}^2}{d}\delta_{\mu\nu} = \frac{\ell_{rw}^2}{d\tau_{rw}}t\,\delta_{\mu\nu}.\end{aligned} \tag{333}$$

in d dimensions for an isotropic medium, since each component of v_μ^2 has the same average over the sphere, and they must sum to unity. We introduced the mean free path $\ell_{rw} = v\tau_{rw}$, which is the distance traveled between scattering events, i.e. the length of the "steps" in the random walk. The square of the displacement grows linearly in time. This is a famous result for random walks.

Now we can compare this with the diffusion equation. Consider an initial density in Eq. (331) $n(\boldsymbol{x}, t = 0) = \delta^{(d)}(\boldsymbol{x})$, corresponding to placing a particle at the origin. The solution for scalar D is

$$n(\boldsymbol{x}, t) = \frac{1}{(2\pi \mathrm{D}t)^{d/2}} e^{-\frac{|x|^2}{4\mathrm{D}t}}. \tag{334}$$

The density can be viewed as a probability distribution for the particle. So we can calculate the variance at time t as

$$\overline{x_\mu(t)x_\nu(t)} = \int d^d\boldsymbol{x}\, n(\boldsymbol{x}, t)\, x_\mu x_\nu = 2\mathrm{D}t\delta_{\mu\nu}. \tag{335}$$

We see that this indeed gives the same linear growth of the variance of the position. Comparing to Eq. (333) one finds $\mathrm{D} = \ell_{rw}^2/(2d\tau_{rw}) = v_\mathrm{F}^2\tau_{rw}/(2d)$. Clearly τ_{rw} and τ in the Boltzmann equation play similar roles, but they are not quite identical: in the relaxation time approximation, we treat collisions as occuring at random times, with a probability per unit time of $1/\tau$. In the random walk, we took collisions to occur regularly at time intervals of τ_{rw}. Both models reproduce diffusion, and it turns out that they give the same diffusion constant if we take $\tau_{rw} = 2\tau$. Hence we should expect, in the isotropic case $\mathrm{D} = \ell^2/(d\tau) = v_\mathrm{F}^2\tau/d$, and the mean free path $\ell = v_\mathrm{F}\tau$.

OK now back to the Einstein relation. To obtain it, consider a closed system with an applied electrostatic potential at two ends, i.e. an "open circuit", in the steady state. Upon initially applying the voltage, charges will flow until they pile up near the contacts and in the bulk their resulting electric field will cancel the initial applied one any the flow will stop. In this situation there can be no net electron current anywhere (or at least the divergence must be zero everywhere). However, there must be electric fields somewhere within the sample. In a metal these occur within some thin layer near the two contacts. At a generic point in the sample, there will be some combination of density gradient and electric field. Each of these drives some current, but they must cancel:

$$\mathbf{J}_e = -e\mathbf{J}_{n,\mathrm{diff}} + \sigma\mathbf{E} = 0. \tag{336}$$

We have

$$e\mathrm{D}\nabla n + \sigma\mathbf{E} = 0. \tag{337}$$

Now to obtain the Einstein relation we need to relate the density gradient to the electric field. We write

$$\nabla n = \nabla\left(\frac{\partial n}{\partial \mu}\mu_{\rm ec}\right) = \frac{\partial n}{\partial \mu}\nabla\mu_{\rm ec}. \tag{338}$$

Here we introduced the "electrochemical potential". The idea is that in general we treat the electrons as in local equilibrium at some chemical potential μ which might depend on position, and with an energy that is shifted by the local electrostatic potential ϕ according to $\Delta\epsilon = -e\phi$. Then the density is obtained by the Fermi-Dirac distribution

$$n_{\rm F}(\epsilon, \mu, \phi) = \frac{1}{\exp\left[-\beta(\epsilon - e\phi - \mu)\right] + 1}. \tag{339}$$

We see that the density is determined by the combined effect of the electrostatic potential and the chemical potential, whence we define

$$\mu_{\rm ec} = \mu + e\phi. \tag{340}$$

Using Eq. (339) in Eq. (337) we obtain

$$e{\rm D}\frac{\partial n}{\partial \mu}\nabla\mu_{\rm ec} + \sigma\mathbf{E} = 0. \tag{341}$$

In the open circuit configuration, after it reaches a steady state, the entire system is in thermal equilibrium, and the true chemical potential is constant. Hence the gradient of the electrochemical potential is just due to the electrostatic potential gradient, which is just (minus) the electric field. This gives

$$-e^2{\rm D}\frac{\partial n}{\partial \mu}\mathbf{E} + \sigma\mathbf{E} = 0. \tag{342}$$

The Einstein relation, Eq. (328), follows.

Now we can compare Eq. (328) to the Boltzmann equation result in Eq. (321). One can see that it matches perfectly. The Einstein relation contains the thermodynamic compressibility, which in the independent electron theory is just the DOS. The angular average gives the diffusion constant ${\rm D}_{\mu\nu} = \langle v_{\rm F}^{\mu} v_{\rm F}^{\nu} \tau\rangle$.

It is worth thinking about how the combination of diffusion and electric field driven currents works out in our Boltzmann calculation. There we solved the Boltzmann equation for a spatially uniform distribution, and found a non-zero current in the presence of an electric field. Therefore in this calculation there is zero density gradient, and the diffusion current is zero. This means that the electrochemical potential must be constant. Since the electric field is non-zero, there is an electrostatic potential gradient $\nabla\phi = -\mathbf{E}$ and consequently there must be a chemical potential gradient, $\nabla\mu = e\mathbf{E}$. This is consistent in several ways. First, the system is truly out of equilibrium when

the current is flowing, so it is natural that the chemical potential is not constant. Furthermore, by maintaining a constant electro-chemical potential, the electron density remains uniform. This is the only way to maintain charge neutrality in the thermodynamic limit.

Warning: the terminology of chemical potential and electro-chemical potential is not uniform. You will sometimes (often?) find the terms used with exactly the opposite meaning. I like the meaning used here because in this way the chemical potential μ has purely statistical meaning: it appears only in the distribution function, and it is constant when the system is in equilibrium, as it should be in the grand canonical ensemble where it is defined as a single number for a closed system.

7.3.2 Anti-symmetric Hall conductivity:

The second term in Eq. (467) is anti-symmetric and is non-dissipative. It corresponds to a "Hall conductivity": the current generated by an electric field is normal to the applied field.

The appearance of a Hall effect is quite odd and striking here, since we assumed zero magnetic field. Consequently, the Hall effect appearing here is not the conventional Hall effect (which I hope you heard about and is due to bending of electron trajectories by the orbital effect of a magnetic field), but what is called an *anomalous Hall effect*. The anomalous Hall effect is a common one in ferromagnetic metals, and is identified experimentally as a distinct contribution to the Hall conductivity unrelated to the applied magnetic field.

The expression we found for the anomalous Hall conductivity was discovered by Karplus and Luttinger in 1954[1], and contains the seeds of much of topological band theory. For decades it was however believed by much of the community that these results were irrelevant, and that the anomalous Hall effect in experiment was caused by other effects related not to topology but to scattering. We now know that this belief was largely unfounded, and Karplus+Luttinger's theory is now the basis of most theory of anomalous Hall effects.

At this point, we will hold off on discussing the topological implications of Eq. (467) a bit longer, and just comment on the symmetry aspects of the anomalous Hall conductivity. In particular, the important properties are the transformations under time-reversal and inversion symmetries. They are determined from the Berry curvature. We can calculate the transformations of the Berry curvature from its definition, but a quick and dirty method is to deduce them from the relation to the position operator, Eq. (253). We see that the Berry vector potential $\mathcal{A}$ must have the same transformation properties as the position operator.

First consider time-reversal symmetry. The position operator is invariant under this, but momentum changes sign. Hence we deduce that under time-reversal, $\boldsymbol{\mathcal{A}}_{n\boldsymbol{k}} \to \boldsymbol{\mathcal{A}}_{n-\boldsymbol{k}}$. Since the Berry curvature is $\boldsymbol{\Omega}_{n\boldsymbol{k}} = \boldsymbol{\nabla}_k \times \boldsymbol{\mathcal{A}}_{n\boldsymbol{k}}$, the extra derivative implies that

$$\text{time-reversal T}: \qquad \boldsymbol{\Omega}_{n\boldsymbol{k}} \to -\boldsymbol{\Omega}_{n-\boldsymbol{k}}. \tag{343}$$

Now consider inversion. Under inversion, which we also call parity (P), the position and momentum operators both change sign. So there is an additional minus sign in the transformation of Berry gauge field, and hence the Berry curvature:

$$\text{parity P}: \qquad \boldsymbol{\Omega}_{n\boldsymbol{k}} \to \boldsymbol{\Omega}_{n-\boldsymbol{k}}. \tag{344}$$

If either symmetry is present, the arrow in the corresponding equation, Eq. (343) and/or Eq. (344), becomes an equality. Moreover, the energy, being time-reversal invariant, is guaranteed to be an even function of momentum when either symmetry is present.

This allows us to draw some simple conclusions. First, we see that *when time-reversal is present, the Berry curvature must be an odd function of quasi-momentum*. Since the energy is then an even function of momentum, the integral of the product of the Fermi function and the Berry curvature vanishes. Thus the anomalous Hall conductivity vanishes in the presence of time-reversal symmetry. This should not be surprising because even the classical Hall effect is odd under time-reversal. The presence of an anomalous Hall in ferromagnetic metals means that ferromagnetism imprints time-reversal symmetry breaking on the electrons in some fashion that results in a non-zero net Berry curvature. How this happens in detail is an interesting subject.

Inversion symmetry does not force a vanishing Hall conductivity. However, it does require that the Berry curvature be an even function of momentum. The interesting consequence is that *if both inversion and time-reversal symmetries are present, the Berry curvature must be identically zero*, $\boldsymbol{\Omega}_{n\boldsymbol{k}} = 0$.

7.4 *Filled bands and holes*

7.4.1 Filled bands are inert

One of the basic results of standard band theory is that filled bands are "inert". This is true in the following sense: if we ignore collisions, it is generally true that if $f_n(\boldsymbol{k}, \boldsymbol{x}) = 1$ for all $\boldsymbol{k}, \boldsymbol{x}$, then

$$(\partial_t + \mathbf{F}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} + \boldsymbol{v}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{x}})\, f_n = 0. \tag{345}$$

This implies that, provided the filled band is also preserved by the collision term, it is maintained by the full Boltzmann equation. This is generally expected to be true at T = 0 for bands which are entirely below the Fermi energy, since the collision term must vanish for the equilibrium distribution.

In textbooks, you will also find the statement that filled bands are also inert in the sense that they do not carry any current. This is true in conventional band theory but not strictly true when the effects of Berry curvature are taken into account. Take the contribution from a single band from Eq. (305),

$$\boldsymbol{j}_{e,n} = 2\int d^d\boldsymbol{k}\, \mathrm{D}_n(\boldsymbol{k})(-e\boldsymbol{v}_n(\boldsymbol{k},\boldsymbol{x}))f_n(\boldsymbol{k}). \tag{346}$$

For a filled band, we can take $f_n(\boldsymbol{k}) = 1$ and use the expressions for $\mathrm{D}_n(\boldsymbol{k})$ in Eq. (300) and the velocity $\boldsymbol{v}_n$ in Eq. (298). We get

$$\boldsymbol{j}_{e,n}^{\text{filled}} = -2e\int_{\text{BZ}} \frac{d^d\boldsymbol{k}}{(2\pi)^d}\, \left(\tilde{\boldsymbol{v}}_{n\boldsymbol{k}} + e\mathbf{E}\times\boldsymbol{\Omega}_{n\boldsymbol{k}} + e\left(\Omega_{n\boldsymbol{k}}\cdot\tilde{\boldsymbol{v}}_{n\boldsymbol{k}}\right)\mathbf{B}\right). \tag{347}$$

Here we wrote the group velocity $\tilde{\boldsymbol{v}}_{n\boldsymbol{k}} \equiv \boldsymbol{\nabla}_{\boldsymbol{k}}\tilde{\epsilon}_{n\boldsymbol{k}}$. Because the Brillouin zone is periodic and the group velocity is a total derivative, the integral of the first term is zero. The integral of the last term is also zero, because since the Berry curvature is a curl, it is divergenceless, which means that $\left(\Omega_{n\boldsymbol{k}}\cdot\tilde{\boldsymbol{v}}_{n\boldsymbol{k}}\right)\mathbf{B} = \frac{\partial}{\partial k_\nu}\left(\Omega^\nu_{n\boldsymbol{k}}\tilde{\epsilon}_{n\boldsymbol{k}}\mathbf{B}\right)$. We find that

$$\boldsymbol{j}_{e,n}^{\text{filled}} = 2e\mathbf{E}\times\int_{\text{BZ}} \frac{d^d\boldsymbol{k}}{(2\pi)^d}\boldsymbol{\Omega}_{n\boldsymbol{k}}. \tag{348}$$

In the presence of an electric field, this current is not necessarily zero. It is, however, always a dissipationless Hall current since it is normal to the electric field. So the correct statement is that *filled bands carry no longitudinal (dissipative) current, but may carry dissipationless transverse (Hall) currents*. This is a puzzling fact, and we see shortly that this is related to the quantum Hall effect and band topology. We continue to postpone this a bit, but we can say from what we already know also that *in the presence of time-reversal symmetry, filled bands carry no current*, since the Berry curvature is odd in quasi-momentum under those conditions.

7.4.2 Almost filled bands and holes

In the previous (subsub-)section, we showed that when the Fermi energy lies near the bottom of a band, and the dispersion can be approximated as quadratic around the minimum, the conductivity of that band reproduces the simple Drude result in the relaxation time approximation. The electrons near the band bottom behave almost like electrons in free space, albeit with an effective mass rather than the bare electron mass. The conductivity from these electrons is proportional to their density, and hence vanishes when the band empties. It is of course natural that an empty band carries no current, and hence that the almost empty band can be described in terms of dilute electrons.

We just learned that a full band carries no (dissipative) current, or no current at all in the presence of time-reversal symmetry. So what about a *nearly* full band? The natural description of such a nearly full band is in terms of the states that are *not* full, since these are dilute. Such a missing electron is

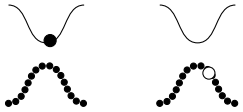


Figure 8: Electron (panel a) and hole (panel b) excitations, respectively.

called a "hole", and naturally enough acts like a positively charged particle. To prove this mathematically, it is convenient to examine the Boltzmann equation in this limit. By definition, $f_n(\boldsymbol{k})$ gives the occupation of an electron state with quasi-momentum $\boldsymbol{k}$, i.e. $f_n(\boldsymbol{k}) = 1$ if there is an electron in this state. An missing electron with quasi-momentum $\boldsymbol{k}$ is described by $f_n(\boldsymbol{k}) = 0$, or equivalently $1 - f_n(\boldsymbol{k}) = 1$. When an electron with quasi-momentum $\boldsymbol{k}$ is removed from the system, the net quasi-momentum left behind is $-\boldsymbol{k}$. So it is natural to define the occupation factor for holes as

$$\breve{f}_n(\boldsymbol{k}, \boldsymbol{x}, t) = 1 - f_n(-\boldsymbol{k}, \boldsymbol{x}, t), \qquad \Rightarrow \qquad f_n(\boldsymbol{k}, \boldsymbol{x}, t) = 1 - \breve{f}_n(-\boldsymbol{k}, \boldsymbol{x}, t). \tag{349}$$

Inserting this into the Boltzmann equation (let us take the relaxation time approximation) gives

$$\begin{aligned} &(\partial_t + \mathbf{F}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} + \boldsymbol{v}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{x}})\,(1 - \breve{f}_n(-\boldsymbol{k}, \boldsymbol{x}, t)) = -\frac{1}{\tau_{n\boldsymbol{k}}}\left(1 - \breve{f}_n(-\boldsymbol{k}) - f_n^{(0)}(\boldsymbol{k})\right) \\ \Rightarrow &-(\partial_t + \mathbf{F}_n(\boldsymbol{k}, \boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} + \boldsymbol{v}_n(\boldsymbol{k}, \boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\boldsymbol{x}})\,\breve{f}_n(-\boldsymbol{k}) = \frac{1}{\tau_{n\boldsymbol{k}}}\left(\breve{f}_n(-\boldsymbol{k}) - \breve{f}_n^{(0)}(-\boldsymbol{k})\right), \end{aligned} \tag{350}$$

where we defined $\breve{f}_n^{(0)}(\boldsymbol{k}) = 1 - f_n^{(0)}(-\boldsymbol{k})$. Multiply this equation by an overall minus sign, and change variables from $\boldsymbol{k} \to -\boldsymbol{k}$. One obtains

$$(\partial_t - \mathbf{F}_n(-\boldsymbol{k}, \boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} + \boldsymbol{v}_n(-\boldsymbol{k}, \boldsymbol{x}) \cdot \boldsymbol{\nabla}_{\boldsymbol{x}})\,\breve{f}_n(\boldsymbol{k}) = -\frac{1}{\tau_{n\boldsymbol{k}}}\left(\breve{f}_n(\boldsymbol{k}) - \breve{f}_n^{(0)}(\boldsymbol{k})\right). \tag{351}$$

We can rewrite this exactly like the original Boltzmann equation

$$\left(\partial_t + \breve{\mathbf{F}}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{k}} + \breve{\boldsymbol{v}}_n \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}\right)\breve{f}_n = -\frac{1}{\tau_{n\boldsymbol{k}}}\left(\breve{f}_n - \breve{f}_n^{(0)}\right), \tag{352}$$

where

$$\breve{\mathbf{F}}_n(\boldsymbol{k}, \boldsymbol{x}) = -\mathbf{F}_n(-\boldsymbol{k}, \boldsymbol{x}), \qquad \breve{\boldsymbol{v}}_n(\boldsymbol{k}, \boldsymbol{x}) = \boldsymbol{v}_n(-\boldsymbol{k}, \boldsymbol{x}). \tag{353}$$

Eq. (352) and Eqs. (353) define a general "particle-hole transformation" for a single band. Going any further requires specifying some details of the band structure. Let us now focus on states near the top of the band, so that the

energy can be expanded as $\epsilon_{n\boldsymbol{k}} = \epsilon_{n,\max} - k^2/(2m^*)$ (for simplicity we take the band maximum at $\boldsymbol{k} = 0$ but this does not matter much). We also assume that the Berry curvature vanishes or is negligible here (this is common). Then we have

$$\mathbf{F}_n(\boldsymbol{k}, \boldsymbol{x}) = -e\mathbf{E} + e\frac{\boldsymbol{k}}{m^*} \times \mathbf{B}, \qquad \boldsymbol{v}_n(\boldsymbol{k}, \boldsymbol{x}) = -\frac{\boldsymbol{k}}{m^*}. \tag{354}$$

Hence using Eqs. (353),

$$\check{\mathbf{F}}_n(\boldsymbol{k}, \boldsymbol{x}) = e\mathbf{E} + e\frac{\boldsymbol{k}}{m^*} \times \mathbf{B}, \qquad \check{\boldsymbol{v}}_n(\boldsymbol{k}, \boldsymbol{x}) = \frac{\boldsymbol{k}}{m^*}. \tag{355}$$

We see that these are just the forces and velocity expected for a free particle with mass m^* and *positive* charge $+e$.

We can also consider the current in the band, generalizing Eq. (346) to a partially filled band and using Eq. (349),

$$\begin{aligned} \boldsymbol{j}_{e,n} &= -2e \int d^d\boldsymbol{k}\, \mathrm{D}_n(\boldsymbol{k})\boldsymbol{v}_n(\boldsymbol{k})(1 - \check{f}_n(-\boldsymbol{k})) \\ &= 2e\mathbf{E} \times \int_{\mathrm{BZ}} \frac{d^d\boldsymbol{k}}{(2\pi)^d}\boldsymbol{\Omega}_{n\boldsymbol{k}} + 2e \int d^d\boldsymbol{k}\, \check{\mathrm{D}}_n(\boldsymbol{k})\check{\boldsymbol{v}}_n(\boldsymbol{k})\check{f}_n(\boldsymbol{k}), \end{aligned} \tag{356}$$

with $\check{\mathrm{D}}_n(\boldsymbol{k}) = \mathrm{D}_n(-\boldsymbol{k})$. We see that the current in a nearly filled band can be written as the anomalous current in the band due to Berry curvature, plus a contribution which appears identical to the original electron one but due to positively charged holes. If we make the same assumptions for the top of the band as led to Eq. (355), we find

$$\boldsymbol{j}_{e,n} = 2e \int \frac{d^d\boldsymbol{k}}{(2\pi)^d} \frac{\boldsymbol{k}}{m^*} \check{f}_n(\boldsymbol{k}), \tag{357}$$

showing that in this case the current is precisely that expected due to positively charged "free" holes.

7.5 *What lies beyond*

We will conclude this section with a discussion of a few aspects of conduction beyond the steady state relaxation time approximation.

7.5.1 Equilibrium and detailed balance

First, let us comment on the treatment of collisions beyond the relaxation time approximation. More generally, scattering should induce relaxation to the equilibrium distribution in a manner consistent with conservation laws. Whatever the form of the collision integral, it should therefore satisfy several constraints:

- Number conservation: (assuming) scattering does not remove electrons

from the system, the number $\mathrm{N} = 2\sum_n \int d^d\boldsymbol{x} d^d\boldsymbol{k}\, \mathrm{D}_n(\boldsymbol{k}) f_n(\boldsymbol{x}, \boldsymbol{k})$ should be conserved. This requires that the integral of the collision term multiplied by $\mathrm{D}_n(\boldsymbol{k})$ should vanish.

- Equilibrium condition: the collision term should vanish when the distribution function is the equilibrium one. Since it is in general the integral of some function, one can achieve this by a more strict condition which is often satisfied, called *detailed balance*. The detailed balance condition is that *for every scattering process between a pair of states* i, j (here i, j can refer for example to band and quasi-momentum quantum numbers of individual electrons, or more general processes involving more electrons), the ratio of the rate of scattering from state i to state j to the rate of the inverse process is equal to the ratio of occupation in equibrium of state j to state i, so that the two processes exactly balance in equilibrium. Specifically for one-electron scattering, this implies that

$$\frac{\Gamma_{n\boldsymbol{k}n'\boldsymbol{k}'}}{\Gamma_{n'\boldsymbol{k}'n\boldsymbol{k}}} = e^{-\beta(\epsilon_{n\boldsymbol{k}} - \epsilon_{n'\boldsymbol{k}'})}. \tag{358}$$

  In the first Born approximation this holds because the scattering is elastic, and the ratio is equal to unity. It is true however with considerably more generality.

7.5.2 Angle dependent scattering and the transport scattering rate

An important example of collision physics beyond the relaxation time occurs when impurities scatter very differently at small and large angles. This can happen for example if the impurity potential is very slowly varying, such as occurs in "modulation doped" two dimensional electron gases in semiconductors, where the defects are separated relatively far away in the third dimension from the plane of the electrons. Any time the impurity potential varies slowly compared to the Fermi wavelength, its Fourier transform is confined to momenta small compared to the Fermi wavevector, which means that scattering occurs mainly between states nearby on the Fermi surface, and rarely across it. The former is small angle scattering, i.e. the angle between the initial and final momenta is small. Small angle scattering is relatively poor at relaxing the current, since it only slightly changes the electron velocity, which varies continuously along the Fermi surface. This leads to a reduction of the scattering rate that enters the conductivity, called the *transport* scattering time, as compared to the total rate of scattering events.

This is captured by treating the collision integral more correctly. Let us see how it works out for a simple case in which we assume spherical symmetry and a single band, i.e. we take $\epsilon_{n\boldsymbol{k}} = \epsilon(k)$. Moreover, we assume elastic scattering in the first Born approximation, in which case we can write the scattering rate as

$$\Gamma_{\boldsymbol{k}'\boldsymbol{k}} = \Gamma(\epsilon(k), \boldsymbol{k}\cdot\boldsymbol{k}')\frac{\delta(\epsilon(k)-\epsilon(k'))}{\mathrm{D}(\epsilon)}., \tag{359}$$

which expresses the isotropy and energy conservation. We included a factor of the DOS in the denominator so that $\Gamma(\epsilon, x)$ has dimensions of a relaxation time. Now let us examine the Boltzmann equation, Eq. (308) with Eq. (295), for the case of zero magnetic field and spatially uniform electric field, in the steady state. This gives instead of Eq. (312),

$$-e\mathbf{E}\cdot\boldsymbol{\nabla}_{\boldsymbol{k}} f(\boldsymbol{k}) = \int \frac{d^d\boldsymbol{k}'}{(2\pi)^d}\,\Gamma(\epsilon(k), \boldsymbol{k}\cdot\boldsymbol{k}')\frac{\delta(\epsilon(k)-\epsilon(k'))}{\mathrm{D}(\epsilon)}\left[f(\boldsymbol{k}')-f(\boldsymbol{k})\right]. \tag{360}$$

We again linearize this equation around the equilibrium solution, and note that right hand side vanishes at zeroth order because of energy conservation. Therefore we have

$$-e\boldsymbol{v}_{\boldsymbol{k}}\cdot\mathbf{E}\,n_{\mathrm{F}}'(\epsilon(k)) = \int \frac{d^d\boldsymbol{k}'}{(2\pi)^d}\,\Gamma(\epsilon(k), \boldsymbol{k}\cdot\boldsymbol{k}')\frac{\delta(\epsilon(k)-\epsilon(k'))}{\mathrm{D}(\epsilon)}\left[\delta f(\boldsymbol{k}')-\delta f(\boldsymbol{k})\right]. \tag{361}$$

We can express δf as a function of energy and direction of momentum. We have in these variables $\boldsymbol{v}_{\boldsymbol{k}} = v_{\mathrm{F}}\hat{\boldsymbol{k}}$ and

$$-ev_{\mathrm{F}}\hat{\boldsymbol{k}}\cdot\mathbf{E}\,n_{\mathrm{F}}'(\epsilon) = \int \frac{dk'(k')^{d-1}d\hat{\boldsymbol{k}}'}{(2\pi)^d}\Gamma(\epsilon, \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\frac{\delta(\epsilon-\epsilon(k'))}{\mathrm{D}(\epsilon)}\left[\delta f(\epsilon,\hat{\boldsymbol{k}}')-\delta f(\epsilon,\hat{\boldsymbol{k}})\right]. \tag{362}$$

Collapsing the energy delta function gives

$$-ev_{\mathrm{F}}\hat{\boldsymbol{k}}\cdot\mathbf{E}\,n_{\mathrm{F}}'(\epsilon) = \frac{k_{\mathrm{F}}^{d-1}}{(2\pi)^d v_{\mathrm{F}}\mathrm{D}(\epsilon)}\int d\hat{\boldsymbol{k}}'\,\Gamma(\epsilon, \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\left[\delta f(\epsilon,\hat{\boldsymbol{k}}')-\delta f(\epsilon,\hat{\boldsymbol{k}})\right]. \tag{363}$$

We can again use spherical symmetry to proceed. The left hand side transforms as a vector under rotations of $\hat{\boldsymbol{k}}$, so clearly so must the right hand side. The only consistent way to achieve this is to take

$$\delta f(\epsilon, \hat{\boldsymbol{k}}) = \hat{\boldsymbol{k}}\cdot\mathbf{E}\delta f(\epsilon). \tag{364}$$

Inserting this into Eq. (363) gives

$$-ev_{\mathrm{F}}\hat{\boldsymbol{k}}\cdot\mathbf{E}\,n_{\mathrm{F}}'(\epsilon) = \frac{k_{\mathrm{F}}^{d-1}}{(2\pi)^d v_{\mathrm{F}}\mathrm{D}(\epsilon)}\int d\hat{\boldsymbol{k}}'\,\Gamma(\epsilon, \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\left(\mathbf{E}\cdot\hat{\boldsymbol{k}}'-\mathbf{E}\cdot\hat{\boldsymbol{k}}\right)\delta f(\epsilon). \tag{365}$$

In the integral of the $\mathbf{E}\cdot\hat{\boldsymbol{k}}'$ term, only the component of $\hat{\boldsymbol{k}}'$ parallel to $\hat{\boldsymbol{k}}$ can contribute (since the rest of the integrand is indepdent of rotations of $\hat{\boldsymbol{k}}'$ around the $\hat{\boldsymbol{k}}$ axis and reflections through planes containing this axis). Hence

we can replace in the integrand $\mathbf{E}\cdot\hat{\boldsymbol{k}}' \to \mathbf{E}\cdot\hat{\boldsymbol{k}}(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')$. We therefore obtain

$$-ev_{\mathrm{F}}\hat{\boldsymbol{k}}\cdot\mathbf{E}\,n'_{\mathrm{F}}(\epsilon) = \hat{\boldsymbol{k}}\cdot\mathbf{E}\delta f(\epsilon)\frac{k_{\mathrm{F}}^{d-1}\mathrm{S}_d}{(2\pi)^d v_{\mathrm{F}}\mathrm{D}(\epsilon)}\int\frac{d\hat{\boldsymbol{k}}'}{\mathrm{S}_d}\,\Gamma(\epsilon,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\left(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}'-1\right). \quad (366)$$

Here we also multiplied and divided by the area of the sphere S_d, so that the integral becomes an angular average. We also recognize that

$$\mathrm{D}(\epsilon_{\mathrm{F}}) = \frac{2k_{\mathrm{F}}^{d-1}\mathrm{S}_d}{(2\pi)^d}. \quad (367)$$

So we can now solve the equation to obtain

$$\delta f(\epsilon) = \frac{ev_{\mathrm{F}}n'_{\mathrm{F}}(\epsilon)}{\bar{\Gamma}}, \quad (368)$$

where

$$\bar{\Gamma} = \left\langle\Gamma(\epsilon,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\frac{1-\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}'}{2}\right\rangle_{\hat{\boldsymbol{k}}'}. \quad (369)$$

Here the angular brackets indicate the angular average. Note that $\bar{\Gamma}$ is independent of $\hat{\boldsymbol{k}}$ and is just a number. Now we can evaluate the current

$$\begin{aligned} j_e^{\mu} &= -2e\int\frac{d^d\boldsymbol{k}}{(2\pi)^d}v_{\boldsymbol{k}}^{\mu}\delta f(\epsilon,\hat{\boldsymbol{k}}) \\ &= -e\int d\epsilon\,\mathrm{D}(\epsilon)\left\langle v_{\mathrm{F}}\hat{k}^{\mu}\hat{k}^{\nu}\mathrm{E}_{\nu}\frac{ev_{\mathrm{F}}n'_{\mathrm{F}}(\epsilon)}{\bar{\Gamma}}\right\rangle_{\hat{\boldsymbol{k}}}. \end{aligned} \quad (370)$$

Only the $\hat{k}^{\mu}\hat{k}^{\nu}$ factor depends on $\hat{k}$, and gives the angular average $\left\langle\hat{k}^{\mu}\hat{k}^{\nu}\right\rangle = \delta_{\mu\nu}/d$. Putting it all together, taking againg $\mathrm{T}\ll\epsilon_{\mathrm{F}}$, we obtain the conductivity

$$\sigma_{\mu\nu} = e^2\mathrm{D}(\epsilon_{\mathrm{F}})\frac{v_{\mathrm{F}}^2\tau_{\mathrm{tr}}}{d}, \quad (371)$$

where we defined the transport relaxation time

$$\frac{1}{\tau_{\mathrm{tr}}} = \bar{\Gamma} = \left\langle\Gamma(\epsilon,\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\frac{1-\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}'}{2}\right\rangle_{\hat{\boldsymbol{k}}'}. \quad (372)$$

This result can be compared to the earlier ones in the relaxation time and Drude approximations. The only distinction (apart from the specialization to spherical symmetry) is the weighting in the angular average defining the transport scattering time/rate. We note that the factor $\frac{1-\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}'}{2}$ is equal to unity when $\hat{\boldsymbol{k}}' = -\hat{\boldsymbol{k}}$, which corresponds to 180 degree scattering which reverses

the direction of propagation of the electron. This backscattering is the most effective in relaxing the current, and receives the largest weight in the angular average. By contrast, the angular factor vanishes if $\hat{\boldsymbol{k}}' = \hat{\boldsymbol{k}}$, reflecting the fact that small-angle scattering does not slow down the electron. So in general the transport scattering rate $1/\tau_{\rm tr}$ is smaller than the total scattering rate $1/\tau$, i.e.

$$\frac{\tau}{\tau_{\rm tr}} = \frac{\int d^d\hat{\boldsymbol{k}}'\,\Gamma(\epsilon, \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')\frac{1-\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}'}{2}}{\int d^d\hat{\boldsymbol{k}}'\,\Gamma(\epsilon, \hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}}')} < 1. \tag{373}$$

7.5.3 Optical conductivity

Another interesting problem to consider is the response of metals to ac electromagnetic fields. Here the semi-classical approach is a bit more suspect, and in particular the assumption of no inter-band transitions is definitely wrong if the frequency of the light becomes comparable to the energy difference between bands. However, we will for the moment just examine the intra-band contribution to the ac conductivity, which limits the description to low frequencies, e.g. typically less than a few 100 meVs in metals. Working in the relaxation time approximation and assuming a uniform but time-dependent electric field and zero magnetic field, we have

$$\left(\partial_t - e\mathrm{Re}\left[\mathbf{E}e^{-i\omega t}\right]\cdot\boldsymbol{\nabla}_{\boldsymbol{k}}\right) f_n = -\frac{1}{\tau_{n\boldsymbol{k}}}\left[f_n(\boldsymbol{k}) - f_n^{(0)}(\boldsymbol{k})\right]. \tag{374}$$

Note that in reality the electric field will not be uniform at non-zero frequency, nor will the magnetic field be zero: an electromagnetic wave has a definite dispersion relation and a magnetic field transverse to the electric one. However, because of the large speed of light, the wavelength is very long compared to atomic scales so we can treat it as approximately infinite, and the magnetic field is smaller by a factor of one over the speed of light, so it has little effect. Now we linearize the equation by writing

$$f_n(\boldsymbol{k}) = n_{\rm F}(\epsilon_{n\boldsymbol{k}}) + \mathrm{Re}\left[e^{-i\omega t}\delta f_n(\boldsymbol{k})\right], \tag{375}$$

which gives

$$-i\omega\delta f_n(\boldsymbol{k}) - e\boldsymbol{v}_{n\boldsymbol{k}}\cdot\mathbf{E}n'_{\rm F}(\epsilon_{n\boldsymbol{k}}) = -\frac{\delta f_n(k)}{\tau_{n\boldsymbol{k}}}, \tag{376}$$

whence

$$\delta f_n(\boldsymbol{k}) = \frac{e\boldsymbol{v}_{n\boldsymbol{k}}\cdot\mathbf{E}}{1/\tau_{n\boldsymbol{k}} - i\omega}n'_{\rm F}(\epsilon_{n\boldsymbol{k}}). \tag{377}$$

This leads to the ac conductivity

$$\sigma_{\mu\nu}(\omega) = e^2 \sum_n \mathrm{D}_n(\epsilon_\mathrm{F}) \left\langle \frac{v^\mu_{n\boldsymbol{k}} v^\nu_{n\boldsymbol{k}}}{1/\tau_{n\boldsymbol{k}} - i\omega} \right\rangle_{\mathrm{FS},n}. \tag{378}$$

This can be compared to Eq. (321). If we assume a constant relaxation time (independent of n and $\boldsymbol{k}$) then we can write this simply as

$$\sigma(\omega) = \frac{\sigma_\mathrm{dc}}{1 - i\omega\tau}, \tag{379}$$

where σ_dc is the DC conductivity. This last is the Drude result. If you examine the real part of the conductivity, you see that it has a Lorentzian structure peaked at zero frequency:

$$\mathrm{Re}\left[\sigma(\omega)\right] = \frac{\sigma_\mathrm{dc}}{1 + \omega^2\tau^2}. \tag{380}$$

This is the *Drude peak*. Its width is of the order of $1/\tau$, which is very narrow if we have a metal in the semi-classical approximation, which requires $1/\tau \ll \epsilon_\mathrm{F}$ for consistency. Note that the integral of the conductivity is

$$\int_0^\infty d\omega \mathrm{Re}\left[\sigma(\omega)\right] = \frac{\pi\sigma_\mathrm{dc}}{2\tau} = \frac{\pi e^2}{2} \sum_n \mathrm{D}_n(\epsilon_\mathrm{F}) \left\langle v^\mu_{n\boldsymbol{k}} v^\nu_{n\boldsymbol{k}} \right\rangle_{\mathrm{FS},n} \to_\mathrm{free} \frac{\pi n e^2}{2m}. \tag{381}$$

This weight is independent of the relaxation time if the relaxation time is constant, and is basically a measure of the electron's kinetic energy. This can be understood from general "sum rules" which we will not discuss here.

The ac conductivity can be used to understand the propagation of electromagnetic waves inside the metal. The easiest way to do this is to view the currents described by the conductivity as "bound" currents, described by a time-dependent polarization. By definition, the bound currents and charges are

$$\mathbf{J}_b = \boldsymbol{\nabla} \times \mathbf{M} + \frac{\partial \mathbf{P}}{\partial t}, \qquad \rho_b = -\boldsymbol{\nabla} \cdot \mathbf{P}. \tag{382}$$

So if we equation $\mathbf{J}_b = \sigma\mathbf{E}$, we obtain

$$\mathbf{P} = \frac{i\sigma}{\omega}\mathbf{E}. \tag{383}$$

Then by definition the displacement field is

$$\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} = \left(\epsilon_0 + \frac{i\sigma}{\omega}\right)\mathbf{E}. \tag{384}$$

Then by definition $\mathbf{D} = \epsilon\mathbf{E}$ so we obtain the dielectric constant

$$\epsilon(\omega) = \epsilon_0 + \frac{i\sigma(\omega)}{\omega}. \tag{385}$$

The dielectric constant here is in general complex. This implies that electromagnetic waves are not fully propagating but in general decay in the medium. This sounds a little strange but just means that electric fields in the material induce currents, which in turn generate back fields and screen the fields of the incoming wave. Let us use the Boltzmann form for the conductivity in Eq. (385). This gives

$$\epsilon(\omega) = \epsilon_0 + \frac{i\sigma_{\mathrm{dc}}}{\omega(1 - i\omega\tau)}. \tag{386}$$

This becomes simple if $\omega\tau \gg 1$, i.e. outside the Drude peak. Then we can approximate

$$\epsilon(\omega) = \epsilon_0 - \frac{\sigma_{\mathrm{dc}}/\tau}{\omega^2}\left(1 - \frac{i}{\omega\tau} + \mathrm{O}\left(\frac{1}{(\omega\tau)^2}\right)\right). \tag{387}$$

Noting that $\sigma_{\mathrm{dc}}/\tau$ is independent of the relaxation time, we define

$$\omega_p^2 = \frac{\sigma_{\mathrm{dc}}}{\epsilon_0\tau}, \tag{388}$$

The quantity ω_p is called the plasma frequency. This gives

$$\epsilon(\omega) = \epsilon_0\left(1 - \frac{\omega_p^2}{\omega^2}\left(1 - \frac{i}{\omega\tau} + \cdots\right)\right), \qquad \omega\tau \gg 1. \tag{389}$$

The dielectric constant is predominantly real outside the Drude peak, but is negative for $\omega < \omega_p$ and positive for $\omega > \omega_p$. A negative dielectric constant implies non-propagating waves in the solid, so incoming light is reflected for $\omega < \omega_p$. We know it is reflected rather than absorbed because the dielectric constant is still real. This explains the shine of metals. Above the plasma frequency light can propagate. In either case, we see that there is a small imaginary part so there is still some absorption but relatively little. The small absorption makes sense outside the Drude peak because the real part of the conductivity is very small there, and the power dissipated is due to the real part of the conductivity.

The plasma frequency is an intrinsic quantity independent of relaxation time. In our Boltzmann calculation, it is given by

$$\omega_p^2 = \frac{e^2}{d\,\epsilon_0}\sum_n \mathrm{D}_n(\epsilon_{\mathrm{F}})\langle|\boldsymbol{v}_{n\boldsymbol{k}}|^2\rangle_{\mathrm{FS},n}, \tag{390}$$

where we assumed isotropy. In the Drude approximation it is simple

$$\omega_p^2 = \frac{ne^2}{\epsilon_0 m}. \tag{391}$$

One should take this analysis, at least on the scale of the plasma frequency, with a grain of salt, because if you actually put in numbers, for typical metals the plasma frequency is huge - multiple eV, e.g. about 15eV in aluminum. Clearly inter-band transitions are also possible in this energy range. The present treatment completely ignores inter-band transitions. We can imagine that the full response to an electromagnetic wave is a combination of the intra-band response, which can be approximated in this Drude-like fashion, and additional inter-band transitions.

7.5.4 Quantum interference effects

The Boltzmann treatment completely neglects quantum interference effects. We already discussed one such effect, the quantum oscillations in the density of states induced by magnetic fields. This is really an equilibrium effect, which can be measured without transport, for example by the de Haas van Alphen effect. There are, however, also effects of quantum interference on transport, and these are present in zero magnetic field as well.

The classical limit applies when the quantum uncertainty in the particle position is small compared to the distances between scattering events. The wave packet size is limited by the electron wavelength λ, so this means that the mean free path $\ell \gg \lambda$. In momentum space, $\lambda \sim 1/k_F$, where k_F should be interpreted as describing the linear size of the Fermi surface in momentum space. For a metal it is typically of the order of the Brillouin zone since, or the inverse lattice spacing. In low density electron systems, k_F may be smaller. The criteria for classical behavior is them $k_F \ell \gg 1$. We should expect that quantum corrections arise as powers of the small parameter $\frac{1}{k_F \ell}$. Note that increasing mean free path means both smaller quantum corrections and smaller resistivity. So stronger quantum effects are generally expected to be associated with higher resistivity.

Let us see how this looks slightly more quantitatively. Using the Einstein relation, we can estimate the conductivity in general dimensions by the diffusion constant and density of states. Up to order one factors, the density of states can be estimated as

$$D(\epsilon_F) \sim \frac{A_{FS}}{v_F} \sim \frac{k_F^{d-1}}{v_F}, \tag{392}$$

which follows by collapsing volume integral over momentum to a surface one using the energy delta function. We further estimate the diffusion constant by $D \sim v_F^2 \tau = v_F \ell$ (we will neglect the order one $1/d$). Putting this together, we obtain

$$\sigma \sim e^2 v_F \ell \frac{k_F^{d-1}}{v_F} \sim \frac{e^2}{h} k_F^{d-1} \ell. \tag{393}$$

Here in the last line we restored the factor of Planck's constant required for

dimensions. In terms of resistivity,

$$\rho \sim \frac{h}{e^2} \frac{1}{k_F^{d-1} \ell}. \tag{394}$$

The prefactor e^2/h in Eq. (394) is a simple combination of fundamental constants, and has the dimensions of resistance. Its value is

$$\frac{h}{e^2} \approx 25k\Omega. \tag{395}$$

This is known as the resistance quantum. From Eq. (394), we see that in two dimensions, the resistivity is just this resistance quantum multiplied by the dimensionless small parameter for quantum effects $1/(k_F \ell)$. Hence in two dimensions, there is a very simple criteria for the semi-classical limit: quantum effects are small when the resistivity is small compared to h/e^2.

In three dimensions, the resistivity no longer has the dimensions of resistance, and so the criterion that $k_F \ell \gg 1$ no longer corresponds to a definite resistivity, rather one should have that $\rho \ll \frac{h}{k_F e^2}$. This sets an upper limit on resistivity, which however depends upon the Fermi momentum, which, for a typical solid, would be of order the inverse lattice spacing. This upper limit is called the Ioffe-Regel limit. Metals whose resistivity approaches the Ioffe-Regel limit are sometimes called "bad metals". It is not always clear theoretically how such metals should behave, since manifestly the semi-classical model does not apply.

It is worthwhile to consider the simplest situation of non-interacting electrons with static impurities, and ask in this context what occurs as $k_F \ell$ is decreased from large values. When $0 < \frac{1}{k_F \ell} \ll 1$, one can ask about perturbative quantum effects on the resistivity. As we saw, the classical effect of the impurities on the motion of electrons is to turn their ballistic propagation into diffusion. We are then interested in quantum interference corrections to this classical diffusion.

Unfortunately due to an extreme lack of time, we cannot really go through a careful exposition of how this worksm but here is the idea. The basic phenomena is known as *weak localization*: this means that quantum interference effects tend to *reduce* the conductivity and make the electrons less mobile. A very crude idea of how this works is as follows. The classical random walk can be understood as a "diagonal" approximation. Consider the probability $P(\boldsymbol{x}, t)$ for an electron to propagate from the origin to position $\boldsymbol{x}$ in time t. In quantum mechanics, this probability is the square of the amplitude $\Psi(\boldsymbol{x}, t)$ for this process. The amplitude is the sum of amplitudes for all the different paths with these boundary conditions,

$$\Psi(\boldsymbol{x}, t) = \sum_{\text{paths } \Gamma} \psi_\Gamma(\boldsymbol{x}, t). \tag{396}$$

The probability is the square of this amplitude. Each of these individual amplitudes ψ_Γ has a phase which is complicated and *almost* effectively random when the time is long and it contains many scattering events. Consequently, one may approximate the square of the total amplitude as the sum of the squares of each individual amplitude, assuming the cross-terms average to zero. This diagonal approximation reproduces the classical sum of probabilities.

However, in the presence of time-reversal symmetry, there is a special case, in particular when $\boldsymbol{x}$ is very close to the origin, for which the phases are not entirely random. In particular, if the system is time-reversal invariant (as generically true in zero magnetic fields and without magnetism), when a path is traversed in reverse, i.e. backward, the phase is exactly the same. This means that when $\boldsymbol{x}$ is very close to zero, the probability is enhanced because instead of summing the squares for two contributions (forward and backward for each path), we get twice the amplitude squared for each such path, which gives a factor of 4 instead of 2. Thus there is an enhancement of the return probability to the origin. The electron is thus less mobile, which leads to a reduction of the conductivity.

Going beyond this rather hand-waving argument will take us too far afield. Let me instead just summarize the main conclusions. The aforementioned quantum interference effect gives a reduction of the *conductance* G (not conductivity!) of the order of e^2/h, the conductance quantum,

$$\mathrm{G} = \mathrm{G}_{\text{classical}} - c\frac{e^2}{h}. \tag{397}$$

This might be expected by dimensional analysis, since G has dimensions of e^2/h. Since in d dimensions, the conductivity (for an isotropic sample) is classically equal to $\sigma = \mathrm{G}\mathrm{L}^{2-d}$, where L is the sample dimensions. In general, the length L should be replaced by a dephasing length L_φ which sets the distance beyond which quantum interference is destroyed, and behavior becomes classical. Generally L_φ grows as temperature is reduced, diverging as some power law as $\mathrm{T} \to 0$. This gives

$$\begin{aligned}\sigma &= \sigma_{\text{classical}} - c\frac{e^2}{h}\mathrm{L}_\varphi^{2-d} \\ &= \frac{e^2}{h}\left(k_\mathrm{F}^{d-1}\ell - c\mathrm{L}_\varphi^{2-d}\right).\end{aligned} \tag{398}$$

Notably the corrections become small when L_φ is large in 3 dimensions, so quantum corrections are typically not very important in three dimensions, provided at least the classical conductivity is not too small. In two dimensions, the quantum correction is roughly independent of the dephasing length, and the order one reduction should be compared to $k_\mathrm{F}\ell$. So when $k_\mathrm{F}\ell$ is large, the correction is small, and we can talk of "weak" localization. When $k_\mathrm{F}\ell$ becomes smaller, then the interference effects become stronger. In one dimension, we see that the quantum correction grows with increasing dephasing length, and

becomes large as soon as the dephasing length becomes comparable to the mean free path. In the low temperature limit, it is clear that this is always the case.

It turns out that this means that quantum corrections always induce *localization* in one dimension. The classical picture is really qualitatively incorrect, and electrons do not in the end diffuse at low temperature. They instead occupy localized bound-state-like single particle states that have a finite size. In two dimensions, the situation is the most subtle, and it is believed that ultimately localization still occurs, due to effects beyond the simplistic Eq. (398). This localization is however extremely weak, and only occurs when L_φ exceeds a length which is exponentially long in $k_F\ell$, i.e. the localization length $\xi_{\text{loc}} \sim \ell e^{k_F\ell}$. When $k_F\ell \gg 1$, it is difficult to actually see localization, but a logarithmic increase of resistivity with decreasing temperature is observed. Because this weak localization is due to the constructive addition of time-reversal paths, the introduction of a magnetic field destroys this constructive interference and increases the conductivity. A rapid drop in resistivity with very small magnetic fields is a characteristic feature of weak localization.

I would also like to comment on *strong* localization. If $k_F\ell \lesssim 1$, then we cannot trust the semi-classical approximation at all. This corresponds to very strong disorder. In that case, localization occurs in all cases, even in three dimensions.

I apologize for the very brief treatment here but we need to move on!!

8 TOPOLOGICAL INSULATORS

8.1 *Basic ideas of topology*

Until the mid-2000s, band theory had been regarded as an "old and crusty" subject by most researchers in condensed matter physics: a necessary but tedious part of the fundamentals of the physics of solids. "Real theorists" would not touch the stuff, and regarded the band structure problem as a (conceptually) trivial one: just the solution of a one-particle Schrödinger equation.

The viewpoint has changed drastically since then, as work of Charlie Kane, Gene Mele, Shoucheng Zhang, and others uncovered surprising connections between band structure and arguably the most interesting single electron problem in solid state physics: Landau levels and the corresponding integer quantum Hall effect. Now we understand that there are a broad range of topological aspects of bands, which lead to a variety of physical phenonena.

First let us introduce some simple ideas from topology. Topology is a subfield of mathematics. This is not a math course, and I am not a mathematician, so you will get a rough and non-rigorous summary. The basic problem of topology is to answer the question: given a set of objects, which can be "smoothly" deformed into another, and which cannot? In any given problem, we will need to define what we mean by the set of objects, and what we mean by smooth-

ness. Once we have done that, we can group the objects into subsets, such that every object within one subset can be deformed into another in the same subset. These discrete subsets can be called *topological classes*. There might be a finite number of these, or an infinite number. We can assign a discrete label to these classes, for example an integer. In principle, any given object belongs to just one class, and hence can be asssigned the specific discrete label of that class. Thus the discrete label is a function of the object, such that for every object in the same class, the label is the same, i.e. the function does not change under smooth deformations. This function is called a *topological invariant* or a *topological index*. Often we can at least partly determine the topological classes by first discovering a topological invariant, i.e. an explicit function of the objects, which gives a discrete output, which is unchanged by smooth deformations.

Let's just go through a very simple toy example, before trying to use these ideas in band theory. We'll take our objects to be continuous closed oriented curves in a plane which avoid the origin. Mathematically, we describe such an oriented curve by a continuous vector function $\boldsymbol{x}(s) = (x(s), y(s))$, where s parametrizes the coordinate along the curve, and we can take $0 \leq s < 1$. Since the curve is closed, we have the periodic boundary condition $\boldsymbol{x}(0) = \boldsymbol{x}(1)$. Since we specify that the curves avoid the origin, we have $\boldsymbol{x}(s) \neq (0, 0)$. In this case, a smooth deformation is any continuous deformation of the function $\boldsymbol{x}(s)$ that preserves the periodic boundary conditions, and avoids the origin.

Consider the following integral:

$$\mathrm{W} = \frac{1}{2\pi} \int_0^1 ds \, \frac{x \frac{dy}{ds} - y \frac{dx}{ds}}{x^2 + y^2}. \tag{399}$$

This is well-defined because of the requirement that the origin is excluded. Now write this in radial coordinates, $x = r\cos\phi$, $y = r\sin\phi$. One gets

$$\mathrm{W} = \frac{1}{2\pi} \int_0^1 ds \, \frac{r^2(\cos^2\phi + \sin^2\phi)\frac{d\phi}{ds}}{r^2} = \frac{1}{2\pi} \int_0^1 ds \, \frac{d\phi}{ds} = \frac{\phi(1) - \phi(0)}{2\pi}. \tag{400}$$

Because of periodic boundary conditions, $\phi(1) - \phi(0) \in 2\pi\mathbb{Z}$ must be an integer multiple of 2π, hence we find that W must be an integer. Because it is discrete, i.e. quantized, W cannot vary when the curve is smoothly deformed (just because an integer cannot change smoothly). Hence it is a topological invariant. It is called the *winding number* of the curve. Physically, it just counts how many times the curve encircles the origin in the counter-clockwise direction, as we traverse the curve in the increasing s sense. A curve cannot "smoothly" be deformed from one winding number to another because we imposed the condition that the origin is avoided in our definition of smoothness.

This is just one simple example of a topological index and topological

classes, for this simple set of objects. Here the objects were particular types of functions. Band structure is rich with similar objects: the Bloch wave-functions of all the bands at all the quasi-momenta in the Brillouin zone. This allows a lot of different uses of the idea of topology.

8.2 *How to apply topology to insulators*

We define insulators in band theory as electronic systems in which all bands are either empty or full, and the Fermi level lies in a gap separating the highest occupied states (a filled "valence" band) from the lowest unoccupied states (an empty "conduction" band). In conventional band theory, this is a boring situation, since the presence of a band gap makes the system unresponsive to weak fields, and any thermodynamic contributions of electrons are exponentially weak at low temperatures.

Since these properties are really independent of any details of the filled and empty bands, it is natural to ask the question: are all band insulators are alike? Surprisingly, the answer is no. It turns out that there are distinct classes of insulators, which are fundamentally different from one another. The distinctions between them are topological.

To apply the ideas of topology, we need to decide what sort of objects we are comparing. Since in an insulator, all bands are either empty or full, it seems most natural to compare entire bands. We will take our objects then to be the *set* of Bloch wavefunctions comprising a band. We should also specify the dimensionality, so we know what sort of functions these are. Next, we need to decide what it means to smoothly deform a band. Since bands are the solution of the Schrödinger equation for periodic function $u_{n\boldsymbol{k}}(\boldsymbol{x})$, we can think of deformations corresponding to continuous changes of the Bloch Hamiltonian, for example changes in the periodic potential. By construction when we speak of bands, we assume periodicity, so we do assume that periodicity remains fixed when we deform the bands, i.e. the deformations preserve the translational symmetry of the solid. We should then also specify if the deformations should preserve any other symmetries.

8.3 *Chern insulators*

8.3.1 Quantization of the Chern number:

Let us consider the assume for the moment there are no other symmetries imposed. Then in two dimensions there is a topological invariant which is rather closely analogous to the winding number, and moreover, we have already encountered it! It is the so-called "Chern number"

$$\mathrm{C}_n = \frac{1}{2\pi} \int_{\mathrm{BZ}} d^2\boldsymbol{k}\, \Omega_n(\boldsymbol{k}). \tag{401}$$

This is an integral of the Berry curvature over the Brillouin zone. We saw in Sec. 7.3.2 that the Berry curvature is an odd function of momentum unless time-reversal symmetry is broken, so consequently $C_n = 0$ if time-reversal symmetry is preserved. Hence we will have non-trivial physics only with broken time-reversal. In two dimensions, there is just a single component of the Berry curvature, which we thereby wrote as a scalar,

$$\Omega_n(\boldsymbol{k}) = \frac{\partial \mathcal{A}_y}{\partial k_x} - \frac{\partial \mathcal{A}_x}{\partial k_y}. \tag{402}$$

To see that this is a topological invariant, we can show that it is quantized. The quantization is not too hard to guess, when the Berry curvature is written in the form of Eq. (402). Since it is a curl, we can apply Stokes' theorem to write the area integral in Eq. (401) as a line integral of the Berry gauge field around the boundary. If you are not careful, you conclude from this argument that not only is the Chern number quantized, it is just zero! (zero is quantized!). That is because the Brillouin zone is periodic and does not really have a boundary. Indeed if $\boldsymbol{\mathcal{A}}$ is well-defined and fully periodic, then this argument is true and the Chern number must vanish.

However, there is no guarantee that $\boldsymbol{\mathcal{A}}$ is well-defined and periodic. Recall that in fact the Berry gauge field is not itself gauge-invariant: it depends upon our choice of convention for the phases of the Bloch functions. The formula in Eq. (401) is well-defined because it involves only the curl of the gauge field, which is gauge invariant. Employing Stokes' theorem is suspect. We need to be a bit more careful.

Let us think about what is safe to assume about the gauge field. Ultimately, it is obtained from a solution of the Schrödinger equation defined by the Bloch Hamiltonian $\mathcal{H}_{\boldsymbol{k}}$, which is a function of quasi-momentum. We can solve this equation for a given momentum $\boldsymbol{k}$, and choose some arbitrary phase for the Bloch function. Now by applying perturbation theory, we can obtain the solution for a nearby quasi-momentum, $\boldsymbol{k} + d\boldsymbol{k}$, where $d\boldsymbol{k}$ is an infinitesimal displacement in some particular direction in momentum space. We assume that the band n is not degenerate with any other bands – this is part of the smoothness requirement: perturbation theory in quasi-momentum is well-defined under these conditions, which guarantees that the bands are "smooth". Through perturbation theory we will naturally choose a phase such that the Bloch function at $\boldsymbol{k} + d\boldsymbol{k}$ is smoothly connected to that at $\boldsymbol{k}$. By continuing in this direction, we can find a one-dimensional set of solutions for the Bloch functions along some line in momentum space.

For convenience, let's establish coordinates k_1, k_2 along reciprocal lattice vectors, $\boldsymbol{k} = k_1\boldsymbol{b}_1 + k_2\boldsymbol{b}_2$, where $\boldsymbol{b}_{1,2}$ are basis vectors for the reciprocal lattice (notice that the Chern number is dimensionless, and is in fact invariant under any linear coordinate transformation). Then the BZ can be considered the space with $0 \le k_1 < 1$, $0 \le k_2 < 1$. Now let us choose $d\boldsymbol{k}$ along the k_1 direction. By perturbation theory, we can find a one-dimensional set of solutions for

fixed k_2, starting from $(0, k_2)$ and ending at $(1, k_2)$. There is no guarantee that when we perturbatively work our way from $k_1 = 0$ to $k_1 = 1$, that we return to the same Bloch wavefunction, but if, as assumed, the bands are non-degenerate, then we must return to the same state up to a phase (it is not the periodic part but the full state $|\psi_{n\boldsymbol{k}}\rangle$ which must be periodic up to a phase). That is, we will find $|\psi_n(1, k_2)\rangle = e^{i\phi(k_2)}|\psi_n(0, k_2)\rangle$. We can easily now make these states fully periodic in k_1 by letting $|\psi_n(k_1, k_2)\rangle \to e^{-i\phi(k_2)k_1}|\psi_n(k_1, k_2)\rangle$, which is just another gauge transformation. In this way, we are guaranteed to be able to form states which are smooth and periodic in the k_1 direction. Having made such a periodic loop, we can define a Berry phase for that loop:

$$\theta(k_2) = \int_0^1 dk_1 \mathcal{A}_1(k_1, k_2). \tag{403}$$

This is a pure phase and is gauge invariant and hence physical, in the sense of a phase. It is called the "Zak phase". It has the usual phase ambiguity in that we could multiply our Bloch states by a phase that winds by an integer multiple of 2π in the k_1 direction, which would cause θ to change by an integer multiple of 2π.

Now we can use the above procedure to obtain $\theta(k_2)$ at $k_2 = 0$, and then use perturbation theory to smoothly extend this to $k_2 = dk_2$, with $dk_2 \ll 1$. We get a new set of Bloch functions which are again periodic in k_1, at now $k_2 = dk_2$. We can repeat this process to slowly obtain a smooth function $\theta(k_2)$. When we reach $k_2 = 1$, we have obtained a new set of Bloch functions, which must be same as those for $k_2 = 0$, up to a phase. Consequently, we must find $\theta(1) - \theta(0) \in 2\pi\mathbb{Z}$. That is, the Zak phase must wind by an integer multiple of 2π on going from $k_2 = 0$ to $k_2 = 1$.

The winding number of the Zak phase is in fact just the Chern number. To see this, we can write Eq. (401) as a series of small integrals over rectangles that extend over the full k_1 direction and have width $dk_2 = \epsilon$:

$$\begin{aligned} C_n &= \frac{1}{2\pi} \lim_{\epsilon\to 0} \sum_{k_2=0,\epsilon,\cdots}^{1-\epsilon} \int_0^1 dk_1 \ (\partial_1 \mathcal{A}_2(k_1, k_2) - \partial_2 \mathcal{A}_1(k_1, k_2))\, \epsilon \\ &= -\frac{1}{2\pi} \lim_{\epsilon\to 0} \sum_{k_2=0,\epsilon,\cdots}^{1-\epsilon} \int_0^1 dk_1 \ (\mathcal{A}_1(k_1, k_2 + \epsilon) - \mathcal{A}_1(k_1, k_2)) \\ &= -\frac{1}{2\pi} \lim_{\epsilon\to 0} \sum_{k_2=0,\epsilon,\cdots}^{1-\epsilon} (\theta(k_2 + \epsilon) - \theta(k_2)) = -\frac{1}{2\pi} (\theta(1) - \theta(0)). \end{aligned} \tag{404}$$

In going from the first line to the second, we dropped the first term in the parenthesis because it vanishes due to periodic boundary conditions in the k_1 direction. We see that the Chern number is indeed the winding number of the

Zak phase, and hence must be an integer.

The conclusion is that the Chern number is an integer for any smooth band. The set of all bands will divide into different topological classes characterized by the Chern number as a topological invariant.

8.3.2 Physical meaning of the Zak phase in one dimension

To better understand the previous argument, it is very helpful to develop a physical understanding of the Zak phase. It turns out that it is proportional to what is called the "Wannier center", which is something like the center of mass of the Bloch state within the unit cell.

To get there, we need to understand a little better how to relate Bloch states, which are delocalized like plane waves, to localized atomic-like orbitals. The latter are best described by *Wannier states*. A Wannier state is a superposition of Bloch states that best approximates an atomic state. It is a kind of inverse Fourier transform. We consider the one dimensional case, and focus on a single band whose Bloch states are

$$|k_x\rangle = \psi_{k_x}(x) = e^{ik_x x} u_{k_x}(x) \equiv e^{ik_x x}|u_{k_x}\rangle. \tag{405}$$

These are normalized so that

$$\langle u_{k_x}|u_{k_x}\rangle = \int_0^a dx\, |u_{k_x}(x)|^2 = 1, \tag{406}$$

where a is the lattice constant. Now we construct Wannier states as

$$|\mathrm{X} = na\rangle = \phi_{\mathrm{X}}(x) = \frac{a}{2\pi}\int_0^{2\pi/a} dk_x e^{ik_x(x-\mathrm{X})} u_{k_x}(x). \tag{407}$$

The Wannier state can be regarded as a Fourier transform of $\psi_{k_x}(x)$, regarded as a function of k_x. Note that from the definition in Eq. (407), using the periodicity according to Bloch theorem, we can write $u_{k_x}(x) = u_{k_x}(x - \mathrm{X})$. Therefore, we see from inspection that $\phi_{\mathrm{X}}(x)$ depends upon X *only* through the combination $x - \mathrm{X}$. That is, $\phi_{\mathrm{X}}(x) = \phi_x(x - \mathrm{X}) \equiv \phi(x - \mathrm{X})$, where

$$\phi(x) = \frac{a}{2\pi}\int_0^{2\pi/a} dk_x e^{ik_x x} u_{k_x}(x). \tag{408}$$

Generally, such Fourier integrals exhibit exponential decay in the Fourier transform variable (here $x - \mathrm{X}$), provided the function being Fourier transformed is smooth and analytic. The arguments of the previous section establish that for a one-dimensional band, it is always possible to choose the Bloch states

to be periodic in k_x, and they are therefore analytic and smooth provided the band remains non-degenerate for all k_x. So we expect that the wavefunction $\phi_X(x)$ is exponentially localized.

We can compute the overlap

$$\begin{aligned}
\langle X'|X\rangle &= \int_{-\infty}^{\infty} dx\, \phi^*_{X'}(x)\phi_X(x) = \int_{-\infty}^{\infty} dx \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk'_x dk_x\, e^{i(k_x-k'_x)x} e^{ik'_x X' - ik_x X} u^*_{k'_x}(x) u_{k_x}(x) \\
&= \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk'_x dk_x \sum_n \int_0^a dx\, e^{i(k_x-k'_x)(x+na)} e^{ik'_x X' - ik_x X} u^*_{k'_x}(x) u_{k_x}(x) \\
&= \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk'_x dk_x \frac{2\pi}{a} \sum_m \delta(k_x - k'_x - \frac{2\pi m}{a}) \int_0^a dx\, e^{i(k_x-k'_x)x} e^{ik'_x X' - ik_x X} u^*_{k'_x}(x) u_{k_x}(x) \\
&= \frac{a}{2\pi} \int_0^{2\pi/a} dk_x\, e^{-ik_x(X-X')} \int_0^a dx\, u^*_{k_x}(x) u_{k_x}(x) \\
&= \frac{a}{2\pi} \int_0^{2\pi/a} dk_x\, e^{-ik_x(X-X')} = \delta_{X,X'}. \qquad (409)
\end{aligned}$$

In the second line we interchanged the order of spatial and momentum integration, and decomposed the infinite integral over x into a discrete sum of finite intervals of size a. In the third line we used $\sum_n e^{iqna} = \frac{2\pi}{a} \sum_m \delta(q - \frac{2\pi m}{a})$. In the fourth and final line we collapsed the delta function and then used the normalization condition on the periodic part of the Bloch states, and finally recognized the Fourier representation of the Kronecker delta (using the fact that X, X' are integer multiples of a).

This establishes that the Wanner states form an orthonormal basis. They can be viewed as an alternative basis to the Bloch states, but which, like the Bloch states, span the band. Indeed, in a finite system with periodic boundary conditions containing N unit cells, there are N Wannier states, just as there are N distinct values of k_x labeling Bloch states. In a situation, i.e. a metal, in which a band is partially occupied, the Bloch states are a preferred basis, because they are energy eigenstates, and occupation is decided by this energy. In the case of an entirely filled band, however, the two become equivalent': the *many body* state with all Bloch states occupied is equivalent to the many body state of all Wanner states occupied.

In this way, Wannier states are a sort of natural description for band insulators, in which all electrons reside in filled bands. They give an intuitive understanding of why insulators are insulating: in the Wannier basis, all the occupied states are localized!

Now let us consider the "Wannier center", i.e. the center of mass of a Wanner state. Since a Wannier state decays exponentially with distance, it is well-defined to write

$$\bar{x} \equiv \langle \mathrm{X}|(\hat{x} - \mathrm{X})|\mathrm{X}\rangle = \int_{-\infty}^{\infty} dx\, |\phi_{\mathrm{X}}(x)|^2 (x - \mathrm{X}). \tag{410}$$

To evaluate this, we insert the explicit form of the Wannier state and then use a small trick:

$$\begin{aligned}
\bar{x} &= \int_{-\infty}^{\infty} dx \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk_x' dk_x\, e^{i(k_x - k_x')x} e^{-i(k_x - k_x')\mathrm{X}} u^*_{k_x'}(x) u_{k_x}(x)(x - \mathrm{X}) \\
&= \int_{-\infty}^{\infty} dx \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk_x' dk_x \left(-i\frac{\partial}{\partial k_x} e^{i(k_x - k_x')(x-\mathrm{X})}\right) u^*_{k_x'}(x) u_{k_x}(x) \\
&= \int_{-\infty}^{\infty} dx \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk_x' dk_x\, e^{i(k_x - k_x')(x-\mathrm{X})} u^*_{k_x'}(x) i\frac{\partial}{\partial k_x} u_{k_x}(x).
\end{aligned} \tag{411}$$

The derivative in the second line generates the $x - \mathrm{X}$ factor needed for the Wannier center. To go from the second line to the third, we integrate by parts in k_x. It is crucial in this step that the Bloch function is chosen periodic in k_x (otherwise there will be boundary terms). From here, we can proceed very similarly to how we did with the overlap of Wannier states:

$$\begin{aligned}
\bar{x} &= \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk_x' dk_x \sum_n \int_0^a dx\, e^{i(k_x - k_x')(x - \mathrm{X} + na)} u^*_{k_x'}(x) i\frac{\partial}{\partial k_x} u_{k_x}(x) \\
&= \left(\frac{a}{2\pi}\right)^2 \int_0^{2\pi/a} dk_x' dk_x \frac{2\pi}{a} \sum_m \delta(k_x - k_x' - \frac{2\pi m}{a}) \int_0^a dx\, e^{i(k_x - k_x')(x-\mathrm{X})} u^*_{k_x'}(x) i\frac{\partial}{\partial k_x} u_{k_x}(x) \\
&= \frac{a}{2\pi} \int_0^{2\pi/a} dk_x \int_0^a dx\, u^*_{k_x}(x) i\frac{\partial}{\partial k_x} u_{k_x}(x) = \frac{a}{2\pi} \int_0^{2\pi/a} dk_x\, \mathcal{A}_x(k_x).
\end{aligned} \tag{412}$$

We see that the result is just proportional to the Zak phase, i.e.

$$\bar{x} = \frac{a}{2\pi}\theta. \tag{413}$$

This is a beautiful result. The Zak phase just gives the location of the Wannier center. We can now understand that the 2π phase ambiguity of the Zak phase just corresponds to a relabeling of Wannier states, $\mathrm{X} \leftrightarrow \mathrm{X} + a$. What is really

physical, i.e. gauge invariant, is the full set of Wannier centers.

8.3.3 Chern number in terms of hybrid Wannier functions

In one dimension, we can always find a smooth periodic gauge for the Bloch states of a band, and hence can always find exponentially localized Wannier states. The existence of bands with non-zero Chern number in two dimensions means this is not always possible in $d \geq 2$ (there is general recipe to form Wannier functions in any dimension, but they are not guaranteed to be exponentially localized). Instead, we understood the Chern number in Sec. 8.3.1 as a winding number of the Zak phase. We can now reinterpret this in terms of Wannier centers.

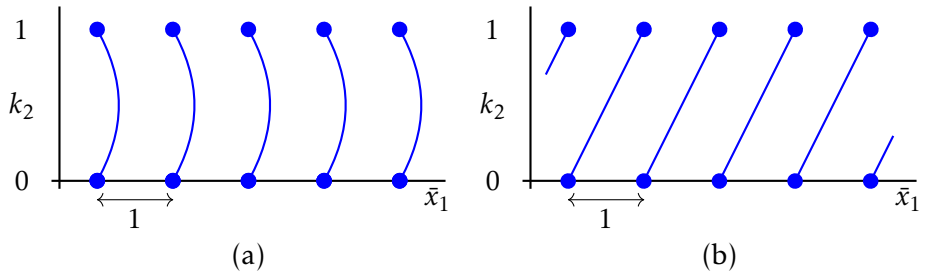


Figure 9: Motion of Wannier centers $\bar{x}_1$ (in dimensionless units with lattice spacing 1) as a function of the orthogonal momentum k_2. In case (a), the Wannier centers return to their original locations upon varying k_2 from 0 to 1, i.e. across the k_2 direction of the Brillouin zone. This corresponds to the case of zero Chern number. In case (b), each Wannier center moves to the position of the next center upon the same variation of k_2. This corresponds to Chern number $\mathrm{C} = +1$.

To do so, we should construct what are called hybrid Wannier states, which are localized in one dimension and delocalized in the other. They are just Fourier transforms in one coordinate but not the other, e.g. k_x but not k_y. To do this in some generality, we can adopt lattice coordinates, as in Sec. 8.3.1:

$$\boldsymbol{k} = k_1 \boldsymbol{b}_1 + k_2 \boldsymbol{b}_2, \qquad \boldsymbol{x} = x_1 \boldsymbol{a}_1 + x_2 \boldsymbol{a}_2. \tag{414}$$

Here x_i are the fractional coordinates; in these units the lattice spacing is unity. Now we will Fourier transform in k_1 but not k_2, to define the hybrid Wannier state:

$$|\mathrm{X}_1, k_2\rangle = \psi_{\mathrm{X}_1,k_2}(x_1, x_2) = \int_0^1 dk_1 \, e^{2\pi i k_1 (x_1 - \mathrm{X}_1) + 2\pi i k_2 x_2} u_{k_1,k_2}(x_1, x_2), \tag{415}$$

where X_1 is an integer. For a fixed k_2, we can think of these as a set of Wannier states for an effective one-dimensional system. In this way, we can define the 1d Wannier center of this 1d band *as a function of* k_2

$$\bar{x}_1(k_2) = \langle X_1, k_2|(\hat{x}_1 - X_1)|X_1, k_2\rangle = \frac{\theta_1(k_2)}{2\pi}. \tag{416}$$

Then following through the manipulations we just illustrated in one dimension gives

$$\theta_1(k_2) = \int_0^1 dk_1 \, \mathcal{A}_1(k_1, k_2). \tag{417}$$

The definition coincides with $\theta(k_2)$ in Sec. 8.3.1, and we can therefore see that the Chern number corresponds to the winding number of $\theta_1(k_2)$, and hence to a "winding" of the Wannier center $\bar{x}_1(k_2)$. The behavior of the Wannier centers on varying k_2 for the cases $C = 0$ and $C = 1$ is shown in Fig. 9.

8.3.4 Quantum Hall effect:

Now that we recognize that the integral of the Berry curvature of a band over the entire Brillouin zone is a topological invariant, we can connect this to our discussion of transport in the previous section. In particular, we showed that the linear response conductivity generally contains the anti-symmetric Hall term in Eq. (467). Here we rewrite this for the case of a band insulator at $T = 0$, where the Fermi function is 1 for every state in an occupied band, and 0 for every state in an empty band. Furthermore, instead of including a factor of 2 for spin degeneracy, we allow for spin states to be non-degenerate, and more generally for spin not even to be a good quantum number. Then there is no degeneracy factor and we have, in two dimensions, the Hall conductivity

$$\sigma_H(T = 0) = \frac{\sigma_{yx} - \sigma_{xy}}{2} = e^2 \sum_{n \text{ occ.}} \int \frac{d^2\boldsymbol{k}}{(2\pi)^2} \Omega_n(\boldsymbol{k}) = \frac{e^2}{2\pi} \sum_{n \text{ occ.}} C_n = \frac{e^2}{h} \sum_{n \text{ occ.}} C_n. \tag{418}$$

In the final equality we restored the factor of $\hbar$ to get physical units. This relationship between the Hall conductivity and the Chern number is known as the Thouless-Kohmoto-Nightingale-de Nijs (TKNN) formula. Note that the prior formula for spin degenerate states is restored by simply counting each degenerate band as two non-degenerate ones with equal Chern numbers. The above formula is more general, and indeed appropriate since the Chern number is non-zero only when time-reversal symmetry is broken, and generally this also implies breaking of spin degeneracy. We see that the Hall conductivity is equal to the quantum of conductance, e^2/h times the sum of the Chern numbers of the occupied bands. This implies quantization of the Hall conductivity! This is the celebrated Integer Quantum Hall Effect (IQHE). The IQHE above can be considered an "anomalous" IQHE because no where do we invoke an orbital magnetic field. An insulator for which the total Chern number of the occupied bands is non-zero is known as a *Chern insulator* or IQHE state. When the total Chern number is zero, is may be called a trivial or

non-topological insulator.

Please note the connection to the discussion in Sec. 7.4.1, where we saw that the current carried by a filled band is not necessarily zero. If you check Eq. (348), you will see that the current of a filled band is exactly proportional to its Chern number. The exception to the standard rule that filled bands carry no current is nothing but the topological contribution to the current. (Please note also from this section where we allowed non-zero magnetic field that quantization remains in the presence of an orbital magnetic field).

One can directly understand the quantum Hall effect from the motion of the Wannier centers discussed in Sec. 8.3.3 and illustrated in Figure 9. Consider an applied electric field $\mathbf{E} = \mathrm{E}\hat{\boldsymbol{b}}_2$ ($\hat{\boldsymbol{b}}_2 = \boldsymbol{b}_2/|\boldsymbol{b}_2|$) , which is perpendicular to $\boldsymbol{a}_1$. Take the dot product of the Hamiltonian equation for the momentum with $\boldsymbol{a}_2$, which gives

$$2\pi\frac{dk_2}{dt} = -e\mathrm{E}\boldsymbol{a}_2\cdot\hat{\boldsymbol{b}}_2 = -e\mathrm{E}\frac{|\boldsymbol{a}_1\times\boldsymbol{a}_2|}{|\boldsymbol{a}_1|}. \tag{419}$$

The second equality follows by a little algebra. Averaged over all the different values of k_2, the winding of the Wannier centers implies that

$$\frac{d\bar{x}_1}{dt} = \mathrm{C}\frac{dk_2}{dt} = -e\mathrm{EC}\frac{|\boldsymbol{a}_1\times\boldsymbol{a}_2|}{2\pi|\boldsymbol{a}_1|}. \tag{420}$$

Now the current density perpendiculat to the applied field is

$$j = -ne\frac{d\boldsymbol{x}}{dt}\cdot\hat{\boldsymbol{a}}_1 = -ne|\boldsymbol{a}_1|\frac{d\bar{x}_1}{dt}. \tag{421}$$

Using Eq. (420) we see that the Hall conductivity is

$$\sigma_{\mathrm{H}} = ne^2|\boldsymbol{a}_1\times\boldsymbol{a}_2|\frac{\mathrm{C}}{2\pi} = \frac{e^2}{h}\mathrm{C}, \tag{422}$$

using the fact that for a filled band in two dimensions, the density times the volume of the unit cell is unity, and $|\boldsymbol{a}_1\times\boldsymbol{a}_2|$ is that volume, and inserting the needed factor of $\hbar$ to restore units.

8.3.5 Quantum Hall effect due to Landau levels

We have discussed the integer quantum Hall effect as due to Chern number of a Bloch band. But the integer quantum Hall effect was actually discovered in the much simpler context of an electron gas in a magnetic field,

$$\mathrm{H} = -\frac{1}{2m}\left(\boldsymbol{\nabla} - ie\mathbf{A}\right)^2. \tag{423}$$

In that, more common, situation, the IQHE arises due to the formation of Landau levels. It is well-known that the a state with one filled Landau level has a unit quantized Hall conductance, $\sigma_H = e^2/h$. Where does this come from, and does this mean that a Landau level has Chern number C = 1?

It is not obvious that we can define a Chern number of a Landau level in the same way as we did for a Bloch band. This is because, in the Landau level problem, any vector potential satisfying $\nabla \times \mathbf{A} = B \neq 0$ necessarily breaks translational symmetry, and so the Hamiltonian is not periodic. We will see that this issue can be overcome, and that we can eventually form Bloch-like states regardless.

First, though, what we can do easily is to make the Hamiltonian periodic in one direction, by choosing the Landau gauge $A_y = Bx$. Then the Hamiltonian Eq. (423) is periodic in y, and the standard solution for the Landau levels is

$$\psi_{n,k_y}(x, y) = e^{ik_y y}\phi_n(x - k_y\ell^2), \tag{424}$$

where n is the Landau level index, $\ell = \sqrt{\hbar/eB}$ is the magnetic length, and $\phi_n(x)$ is a simple harmonic oscillator wavefunction, i.e. Gaussian multiplied by a Hermite polynomial. The function ψ has a striking similarity to a hybrid Wannier function! It is periodic in y but localized in x. The difference from the Bloch case is that k_y is a true momentum in this case, and a quasi-momentum for the hybrid Wannier state.

An important feature the Landau gauge state Eq. (424) shared with the hybrid Wannier state is that its *center*, $\bar{x}(k_y) = k_y\ell^2$ shifts with the momentum k_y. Thus we can repeat the arguments beginning with Eq. (419) to relate this connection of x position and y momentum to the Hall conductivity. We have

$$\frac{d\bar{x}}{dt} = \ell^2\frac{dk_y}{dt} = -e\ell^2 E_y. \tag{425}$$

Note a direct comparison with Eq. (419) is a little misleading because there we used dimensionless units while here we retain standard units of length and momentum. Writing the current

$$j_x = -ne\frac{d\bar{x}}{dt} = e^2 n\ell^2 E_y. \tag{426}$$

To obtain quantization, we need to insist that the Landau level is full. This corresponds to the density $n = 1/(2\pi\ell^2)$, which then gives the desired IQHE.

Note that in this argument we only use the slope of $\bar{x}$ versus k_y, which is a linear relation for a Landau level. In a general Chern band, the Wannier center does not vary linearly with the quasimomentum, but it is the *average* slope that governs the Hall conductivity (since we sum over all the occupied states).

We can try to sharpen the connection to the Chern band by constructing linear combinations of the states in Eq. (424). These are still eigenstates

since the Landau level is degenerate (but from the topological perspective we do not really care if these are eigenstates, only that we find a new basis which still spans the full space of the Landau level). Specifically, consider the superposition

$$\tilde{\psi}_{k_x,k_y}(x,y) = \sum_m e^{i2\pi k_x \ell m} \psi_{n,k_y+\frac{2\pi m}{\ell}}(x,y) = e^{ik_y y}\left(\sum_m e^{i2\pi k_x \ell m} e^{i\frac{2\pi m y}{\ell}} \phi_n(x - k_y\ell^2 - 2\pi m\ell)\right). \tag{427}$$

Because we superimposed states with different k_y, this is no longer a momentum eigenstate in the y direction. However, we carefully chose the factors so that it retains the notion of quasi-momentum. Indeed, the function $\tilde{\psi}$ has the standard Bloch properties that

$$\tilde{\psi}_{k_x,k_y}(x+a,y) = e^{ik_x a}\tilde{\psi}_{k_x,k_y}(x,y), \qquad \tilde{\psi}_{k_x,k_y}(x,y+b) = e^{ik_y b}\tilde{\psi}_{k_x,k_y}(x,y), \tag{428}$$

if we take the "lattice constants" equal to

$$a = 2\pi\ell, \qquad b = \ell. \tag{429}$$

Consequently in this new form we can regard the Landau levels as forming a band with this unit cell. Actually we could freely define other unit cells with different dimensions so long as they satisfied $ab = 2\pi\ell^2$, which is the area which encloses a single flux quantum.

We can also calculate the Berry curvature with these Bloch functions we defined. Let us extract the plane wave part to define the periodic part of the Bloch function,

$$\tilde{\psi}_{k_x,k_y}(x,y) = e^{ik_x x + ik_y y} u_{n\boldsymbol{k}}(x,y), \tag{430}$$

with

$$u_{n\boldsymbol{k}}(x,y) = \sum_m e^{-ik_x(x-2\pi\ell m)} e^{i\frac{2\pi m y}{\ell}} \phi_n(x - k_y\ell^2 - 2\pi m\ell). \tag{431}$$

Standard normalization for the periodic part of a Bloch function is that

$$\begin{aligned}
&\int_0^a dx \int_0^b dy\, |u_{n\boldsymbol{k}}(x,y)|^2 = 1, \\
&= \sum_m \int_0^{2\pi\ell} dx \int_0^{\ell} dy\, |\phi_n(x - k_y\ell^2 - 2\pi m\ell)|^2 = 1, \\
&= \ell \int_{-\infty}^{\infty} dx\, |\phi_n(x - k_y\ell^2)|^2 = \ell \int_{-\infty}^{\infty} dx\, |\phi_n(x)|^2 = 1.
\end{aligned} \tag{432}$$

For the $n = 0$ lowest Landau level, the explicit function is

$$\phi_0(x) = \frac{1}{(2\pi)^{1/4}\ell} e^{-\frac{x^2}{4\ell^2}}. \tag{433}$$

We need some derivatives to evaluate the Berry curvature:

$$\partial_{k_y} u_{n\boldsymbol{k}} = -\ell^2 \sum_m e^{-ik_x(x-2\pi\ell m)} e^{i\frac{2\pi m y}{\ell}} \phi_n'(x - k_y\ell^2 - 2\pi m\ell),$$

$$\partial_{k_x} u_{n\boldsymbol{k}} = -i \sum_m e^{-ik_x(x-2\pi\ell m)} e^{i\frac{2\pi m y}{\ell}} (x - 2\pi\ell m)\phi_n(x - k_y\ell^2 - 2\pi m\ell). \tag{434}$$

Then the Berry curvature is

$$\begin{aligned}\Omega_{n\boldsymbol{k}} &= 2\mathrm{Im}\langle \partial_{k_x} u | \partial_{k_y} u\rangle \\ &= 2\int_0^a dx \int_0^b dy\, \mathrm{Im}\left[\partial_{k_x} u^*_{n\boldsymbol{k}} \partial_{k_y} u_{n\boldsymbol{k}}\right] \\ &= -2\ell^2 \sum_m \int_0^{2\pi\ell} dx \int_0^{\ell} dy\, (x - 2\pi\ell m)\phi_n^*(x - k_y\ell^2 - 2\pi m\ell)\phi_n'(x - k_y\ell^2 - 2\pi m\ell).\end{aligned} \tag{435}$$

Here we used the orthonormality of the plane waves in the y direction to collapse a double sum from the two Bloch functions. Now we can carry out the y integral and also change variables to $x' = x - 2\pi m\ell$. The integral then is over $-2\pi m\ell < x' < 2\pi\ell(1-m)$. Summing over m converts this to an infinite integral. So we have

$$\begin{aligned}\Omega_{n\boldsymbol{k}} &= -2\ell^3 \int_{-\infty}^{\infty} dx\, \phi_n^*(x - k_y\ell^2) x \phi_n'(x - k_y\ell^2) = -2\ell^3 \int_{-\infty}^{\infty} dx\, \phi_n^*(x) x \phi_n'(x) \\ &= \ell^2\end{aligned} \tag{436}$$

The last equality on the first line is obtained by shifting $x \to x + k_y\ell^2$, and shows that the Berry curvature is constant (i.e. independent of $\boldsymbol{k}$). On the second line we evaluated it in the lowest Landau level using Eq. (433) (presumably the answer is the same for all Landau levels).

We can now integrate this over the Brillouin zone corresponding to the real space unit cell in Eq. (429), i.e. $0 < k_x < 1/\ell$ and $0 < k_y < 2\pi/\ell$. We see that

$$\mathrm{C}_n = \frac{1}{2\pi}\int d^2\boldsymbol{k}\, \Omega_{n\boldsymbol{k}} = \frac{1}{2\pi} \times \frac{2\pi}{\ell^2} \times \ell^2 = 1. \tag{437}$$

So following the same definitions as for Bloch bands, we indeed recover unit

Chern number for a Landau level!

8.3.6 Laughlin argument

The quantization of the IQHE is more general and robust than the above derivation may suggest. For example, while the above discussion assumed that there are no crossings between bands, it is actually allowed for bands below the Fermi energy to cross. In this case, the individual Chern numbers C_n may not be well-defined or quantized, but their sum remains quantized. Moreover, while the derivation of Eq. (467) was based on the semi-classical model, the relation in Eq. (418) can be shown fully quantum mechanically based on the linear response formalism, and is correct so long as there is a gap at the Fermi level. Furthermore, quantization of the Hall conductivity can be shown to be true in insulators even beyond the non-interacting electron model, and remains also in the presence of disorder (and in fact the quantization is enhanced in many cases by disorder).

Some of the robustness can be understood by thinking in detail about the spectrum and states of a large finite system, and in particular about what occurs at the boundary of the material with the vacuum (or more generally with an interface to a material with a different Chern number). Indeed, there is a very general argument that the boundary of an insulator with a non-zero (quantized) Hall conductivity and another insulator with a different (quantized, and possibly zero) Hall conductivity must have gapless states localized to it. These are called "edge states".

The argument is due to Robert Laughlin, and not only requires edge states but explains the quantization itself. Laughlin's argument neglects interactions between electrons, but not disorder. It goes something like this. We consider a two dimensional system at zero temperature, and assume that it has a well-defined local conductivity tensor $\sigma_{\mu\nu}$. We further assume that there are no extended states at the Fermi energy. In a clean system, this implies that the system is a band insulator, but we can also allow for disorder, which may induce states at the Fermi energy, provided that in the bulk – i.e. away from any boundaries – those states are localized. In condensed matter physics we say a state is localized if its wavefunction decays exponentially in space away from some region specific to that state. One can think roughly of localized states as states bound to some impurities. We may talk about localization in more detail later. The assumption is basically that the system has no mobile states at the Fermi energy with which to dissipate energy.

This implies that the symmetric parts of the conductivity tensor vanish, because the power dissipated in an electric field E_μ is $\sigma_{\mu\nu}E_\mu E_\nu$ which vanishes by assumption. This implies the conductivity tensor in 2d has just two elements $\sigma_H = \sigma_{xy} = -\sigma_{yx}$, the Hall conductivity. Now Laughlin's argument further constrains the magnitude of the Hall conductivity under these assumptions. We can use the conductivity tensor to compute linear response in any geometry, and so choose what is sometimes called the Corbino geometry, which consists

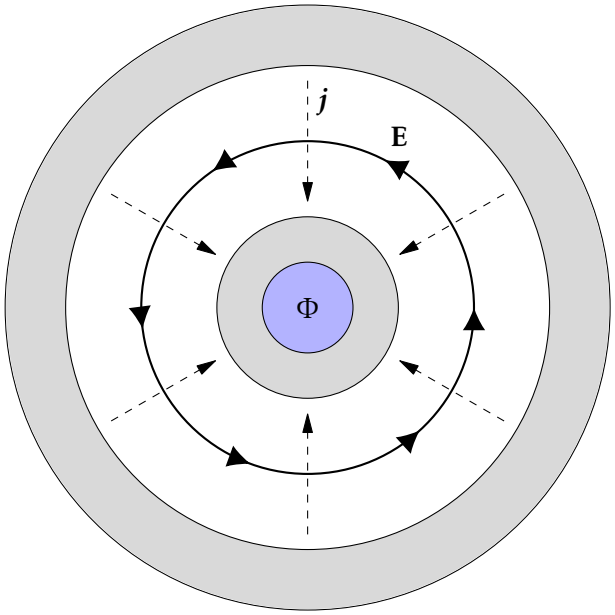


Figure 10: Corbino geometry: the sample is an annulus, with a flux Φ inserted inside the inner hole. No magnetic field penetrates the sample. The time-dependence of the flux during the insertion creates a circumferential electric field **E**. *Due to the Hall conductivity a radial current* $\boldsymbol{j}$ *is produced.*

of a ring-shaped sample or annulus (one can also formulate the argument using a cylinder). Imagine slowly turning on a magnetic field inside the inner hole of the ring, so that no field at all penetrates the sample itself, but a net flux $\Phi(t)$ passes through the hole. The flux is increased from zero to the flux quantum $\Phi = h/e$ very slowly, so that the response of the system is adiabatic.

Let us first analyze the effect of the flux using linear transport. A time dependent flux is accompanied according to Faraday's law by an electric field, in the azimuthal direction. The electric field is $E_\phi = -\partial_t \Phi/(2\pi r)$, at a radius r from the center of the hole. Accordingly, this creates a radial current $I = 2\pi r \sigma_H E_\phi = -\sigma_H \partial_t \Phi$. Integrating the current over time, we obtain a total transfer of charge from the outer to inner radius of

$$Q = -\sigma_H(\Phi(t_f) - \Phi(0)) = -\frac{h}{e}\sigma_H. \tag{438}$$

Note that the above argument is just linear in the flux: if we double the flux, we double the charge transferred.

Now we use special properties of the situation in which the flux is equal to the flux quantum. We will see that the edge must be gapless, and moreover that the transferred charge must be an integer multiple of the elementary charge e. To see this we use quantum mechanics. The flux is included in quantum mechanics by a vector potential $A_\phi(r) = \Phi(t)/(2\pi r)$ along the tangential direction at radius r, which is included via minimal coupling as usual. After

the flux is increased to h/e, the Hamiltonian reaches a form which is equivalent up to a gauge transformation, $\psi \to e^{i\phi}\psi$, where ϕ is the azimuthal angle in the plane, to the one with zero flux. At the single-particle level, we may write that

$$\mathcal{H}(\Phi = \frac{h}{e}) = e^{-i\phi}\mathcal{H}(\Phi = 0)e^{i\phi}, \tag{439}$$

where $\mathcal{H}$ is the single-particle Hamiltonian, and ϕ is the operator representing the azimuthal angle. This is a unitary transformation, which implies that the energy levels and single particle states (up to phases) are the same before and after the flux insertion. However, in the middle of the insertion process, the energies and states can have evolved. Since the process is assumed adiabatic, we can follow these individual levels through the flux evolution, and they must evolve in such a way that each eigenstate at zero flux evolves into another eigenstate at one flux quantum, i.e. the levels may permute. Note that this argument works for the full finite system, edges included. Now the assumption

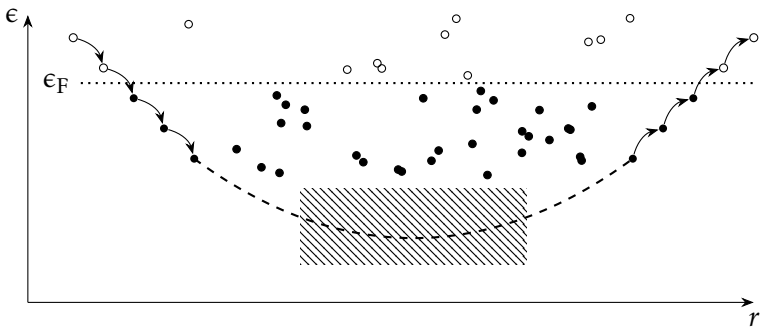


Figure 11: Sketch of spectral flow of single particle levels in the flux insertion process. Initially empty and full states are shown as open and filled circles, respectively. The horizontal axis is the radial distance, with circles showing the centroid of the corresponding levels. States which are extended around the annulus undergo spectral flow near the Fermi energy at the boundaries of the sample. Levels that are localized do not undergo spectral flow, and are indicated as circles without arrows. Some extended states must persist in the bulk, indicated by the dashed line. The spectral flow of levels across the Fermi level determines the number of electrons transferred, n.

that any states at the Fermi level in the bulk are localized comes into play. The Aharonov-Bohm effect is only operative for states which are extended fully around the circumference of the disk, so that an electron in this state is able to sense the phase. For any localized state, the energy and wavefunction must be, up to a phase factor, completely independent of the flux. This implies there is no spectral flow for the levels at the Fermi energy in the bulk.

Knowing this, we therefore understand that any spectral flow at the Fermi level comes entirely from levels at the two edges. Moreover, this spectral flow requires states arbitrarily close to the Fermi level. The conclusion is there

must be gapless states localized at the edges of the sample. These are the edge states. In the next section, we study edge states explicitly.

Laughlin's argument goes further and dictates the quantization of the Hall conductivity, by relating it to the charge transfer. The net result of spectral flow can only be a change of population of electrons by some integer n at either edge, and by charge conservation the change must be equal and opposite, so that n electrons are transferred from one edge to another. Equating the charge Q transferred in Eq. (438) to $-ne$ we obtain finally the Hall quantization condition

$$\sigma_{\mathrm{H}} = n\frac{e^2}{h}. \tag{440}$$

The Laughlin argument is powerful because it includes the effects of disorder. It helps to understand the global structure of the extended states, and makes a connection between the IQHE and pumping.

8.4 *Graphene and Haldane model*

It is instructive to see some of the above features in action. This is most easily done in a simple model, for which we will take the modified model of graphene due to Haldane. Haldane's model starts with the graphene model used in Sec. 5.3 but include spin-orbit coupling.

Recall the features of Sec. 5.3. The Bloch Hamiltonian appears as a 2×2 matrix whose index we denote a, b etc. and which lies in the sublattice space. To include spin, we would also require a spin-1/2 index $\sigma = \uparrow, \downarrow$.

We will adopt a low energy description, which linearizes the dispersion around the Dirac points, there is an additional "valley" index A = 1, 2, which labels the two Brillouin zone corners. Then continuum fields are defined simply by separating the momentum components which are in a small neighborhood of the Dirac points:

$$u_{a\alpha,\mathbf{K}+\boldsymbol{k}} = \psi_{1a\alpha,\boldsymbol{k}}, \qquad \text{for } |\boldsymbol{k}| \ll 1, \tag{441}$$

$$u_{a\alpha,\mathbf{K}'+\boldsymbol{k}} = \psi_{2a\alpha,\boldsymbol{k}}, \qquad \text{for } |\boldsymbol{k}| \ll 1. \tag{442}$$

Putting this all together, we can write a low-energy continuum Hamiltonian which acts on the 8-component spinor $\psi_{\mathrm{A}a\alpha}$,

$$\tilde{\mathcal{H}}_{\boldsymbol{k}}\psi = \epsilon\psi, \tag{443}$$

The fermions are described by a spinor $\psi = \psi_{\mathrm{A}a\alpha}$, where $\boldsymbol{\tau}$ Pauli matrices act on the sublattice a space, $\boldsymbol{\mu}$ act on the valley space (you can see Eq. (444) is diagonal in the valley space because it only involves μ^z), and $\boldsymbol{\sigma}$ act on the spin space (these matrices aare not present in the Hamiltonian because spin-orbit coupling is weak and can be neglected). We keep the indices implicit for

compactness as much as possible. In this notation, the continuum Hamiltonian is

$$\tilde{\mathcal{H}} = v(\mu^z \tau^x k_x + \tau^y k_y). \tag{444}$$

The Hamiltonian in Eq. (444) is easy to diagonalize. It is already diagonal in the valley and spin subspaces, so we can treat μ^z in the Hamiltonian as a constant $= \pm 1$ (its eigenvalues). We are left with a 2×2 matrix in the $\boldsymbol{\tau}$ space. The eigenvalues of this matrix are easy to find by for example rotating it by an SU(2) rotation to the τ^z direction. One has

$$\epsilon_\pm = \pm v\sqrt{k_x^2 + k_y^2} = \pm v|k|. \tag{445}$$

We see the dispersion is just a relativistic "light-cone" of conduction and valence bands intersecting at $\boldsymbol{k} = 0$.

8.4.1 Stability of the Dirac point

A key question is whether this behavior is generic. That is, we may have made a small mistake in our Hamiltonian by neglecting some term, and would correcting this lead to the removal of this intersection point and fundamental modification of the spectrum? This certainly appears possible, since the 2d Dirac equation allows a mass term. Even if we "freeze" the valley and spin degrees of freedom, i.e. just consider the two component Dirac equation for fixed spin and valley quantum numbers, one can add a term of the form $m\tau^z$ to the Hamiltonian above which is known as a Dirac mass, and will indeed remove the intersection point. Maybe some principle (symmetry?) prohibits adding this term? Perhaps there are other allowed perturbations?

Perturbative argument Let us start with the simple-minded answer, which is just based on the Dirac Hamiltonian, Eq. (444) and symmetries. We ask what might prevent adding a term proportional to $M\tau^z$ to the Hamiltonian? Here M could be a matrix in the spin and valley spaces. So long as this anti-commutes with the two matrices inside Eq. (444), it will serve as a mass. First, it is natural to assume spin rotation symmetry, $SU(2)_\sigma$, because spin-orbit coupling is weak in graphene so this is a good approximation microscopically. This requires M to not contain any of the spin Pauli matrices. Second, we impose translational symmetry, which means that quasi-momentum is conserved up to a reciprocal lattice vector, and so there is no scattering between valleys. This requires M to not contain μ^x or μ^y. These two constraints allow matrices of the form $M \sim \mathbb{1}, \mu^z$, i.e. mass terms τ^z and $\mu^z\tau^z$. Now consider time-reversal symmetry. This changes the sign of momentum, and hence interchanges the two Dirac points, which means $\mu^z \to -\mu^z$. The matrix τ^z simply labels the sublattices, so it is time-reversal invariant. Thus of the two remaining options, only the

$M \sim 1$ or pure τ^z mass is time-reversal invariant. If we assume time-reversal symmetry, we still need one more symmetry to prevent the addition of τ^z. For this, we need to require some symmetry which interchanges the two sublattices, for example inversion around the center of a bond. Under this operation, $\tau^z \to -\tau^z$, and so inversion symmetry removes the finally remaining mass term. We conclude that the Dirac points remain intact if we maintain four conditions: 1. spin rotation symmetry $SU(2)_\sigma$, 2. inversion symmetry, 3. the translation symmetry of the honeycomb lattice, and 4. time-reversal symmetry.

TOPOLOGICAL ARGUMENT It turns out that the stability of the Dirac point can be understood more deeply in topological terms. We need to introduce a different topological argument, since the system is not gapped, and there is also no Chern number since with time-reversal and inversion symmetry there is zero Berry curvature.

However, we can consider a loop encircling a Dirac point,

$$\Theta = \oint_{\mathrm{K}} d\boldsymbol{k} \cdot \boldsymbol{\mathcal{A}}_n(\boldsymbol{k}), \tag{446}$$

where the subscript K indicates the line integral is taken around the Dirac point at (say) the K point of the Brillouin zone. This is a Berry phase: it gives the phase evolved under adiabatic evolution of a wavefunction through this loop. The quantity is invariant under single-valued gauge transformations, because it is the integral of a gradient, but changes under large but smooth gauge transformations, e.g. $\chi(\boldsymbol{k}) = p\theta(\boldsymbol{k})$, where $\theta(\boldsymbol{k})$ is the angle of the $\boldsymbol{k}$ point measured from the location of the Dirac point, and p is an integer so that the gauge transformation is single valued. This means that Θ is defined modulo 2π. It turns out that for a Dirac electron, the value of theta is actually

$$\Theta = \pi (\mathrm{mod}\ 2\pi). \tag{447}$$

This is easily worked out from the eigenfunction of the Dirac Hamiltonian, or microscopically from the Bloch Hamiltonian of the graphene model.

The non-trivial value of π for the Berry phase implies the stability of the band touching, as can be argued as a proof by contradition. Suppose the band touching were to be removed by a small perturbation. Then we would require Eq. (120) to hold everywhere, including at the former Dirac point. Then we could use Stokes' theorem to express Θ in Eq. (121) as the area integral of $\mathcal{B}_n$ inside the loop, which would immediately have to vanish. However, a small perturbation can make only small changes in the Bloch states far from a degeneracy point, and so the loop integral cannot change discontinuously. We conclude that the band touching cannot be lifted by any small perturbation preserving time-reversal and inversion symmetry. What can in fact happen,

if symmetry allows it, is for the Dirac point to move in $\boldsymbol{k}$ space under the effect of perturbations (this is allowed if we break the 3-fold lattice rotation symmetry). Then the two Dirac points can drift and annihilate.

By similar arguments, we can obtain the quantization of Θ, which must be a multiple of π (including 0) around *any* loop, modulo 2π, and also the "fermion doubling" result that there must be an even number of such Dirac points. I leave these as exercises to the reader.

8.4.2 Two Dirac masses

Now we are ready to return to build a model of a topological insulator. We take the graphene Hamiltonian and add some perturbations that turn the Dirac semimetal into an insulator. For the moment, we will consider spinless Dirac electrons, governed by Eq. (444). We saw that, for spin-independent interactions, there are two possible "mass" terms which could be added that maintain the translational symmetry of the lattice. Consider the Hamiltonian with both these terms added:

$$\tilde{\mathcal{H}}_{\boldsymbol{k}} = v(\mu^z \tau^x k_x + \tau^y k_y) + m_1 \tau^z + m_2 \mu^z \tau^z \tag{448}$$

Here the mass m_1 is time-reversal invariant, and could be realized by adding a staggered potential of opposite sign on the A and B sublattices. Such a potential would induce a corresponding modulated charge density, so this can be called a "charge density wave" state. The mass m_2 is odd under time-reversal, and can be realized by adding a second neighbor hopping (dashed lines in Fig. 5) which is pure imaginary and has a positive (negative) sign for second neighbors reached by "turning" right (left) when walking two steps on the lattice. The honeycomb model with imaginary second neighbor hopping is known as the *Haldane model*, after Haldane introduced it for reasons to become clear below. Since μ^z commutes with the one-particle terms in Eq. (448), it is a constant of the motion and can be treated as equal to ± 1. Then the energy dispersion is easily calculated as a function of μ^z to be

$$\epsilon_{\pm,\boldsymbol{k}}(m_1, m_2) = \pm\sqrt{v^2 k^2 + (m_1 + \mu^z m_2)^2}. \tag{449}$$

We see that either m_1 or m_2 alone introduces a gap; however, the gap vanishes if $|m_1| = |m_2|$, by taking $\mu^z = -\text{sign}(m_1/m_2)$. If one plots a "phase diagram" in the $m_1 - m_2$ plane, there are four gapped regions separated by "phase boundaries". At least in this model, it is not possible to pass from the "charge density wave" insulator with $m_1 \neq 0, m_2 = 0$ to the "time-reversal broken" insulator with $m_1 = 0, m_2 \neq 0$, without passing through a model in which the gap vanishes. When the gap does vanish, on the $|m_1| = |m_2|$ lines, it does so for just one of the two Dirac points.

What is the physical meaning of this? It turns out that the two separated gapped insulators are indeed physically distinct phases. The "charge density

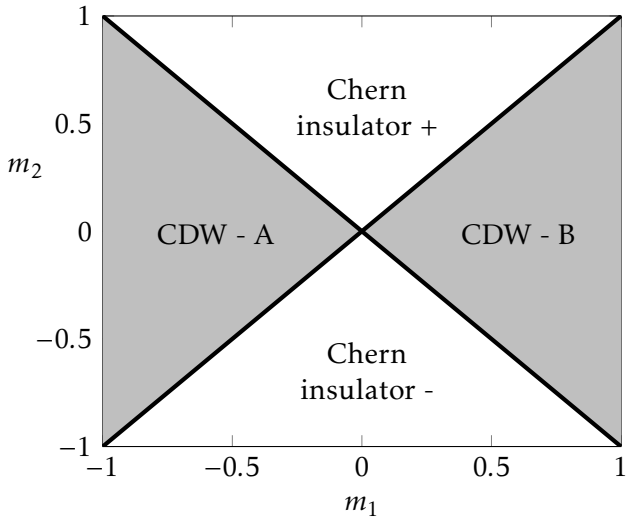


Figure 12: "Phase diagram" showing effects of masses on 2d Dirac fermions. The labels "CDW-A" and "CDW-B" indicate charge density wave regions in which the electrons are localized preferentially on the A or B sublattice sites, respectively.

wave" insulator is a simple band insulator, while the "time-reversal broken" insulator is the Chern insulator or quantum Hall state. The simplicity of the charge-density wave insulator can be seen by simply going back to the lattice model and increasing the staggered potential until it is very large. This process is smooth and no phase transitions occur: the gap increases monotonically as the potential is increased. When the potential is very strong, the insulator itself becomes atomic in nature: one electron resides each site of one of the sublattices (the one with much lower energy), while the other sublattice is empty. There is virtually no motion of the electrons.

8.5 *Edge state*

The Chern insulator, by contrast, does not have a simple atomic limit. This leads to interesting phenomena at an interface between the two. Let us consider modeling such an interface by the Dirac Hamiltonian but with masss $m_1(y), m_2(y)$ that are functions of y, with the interface located at $y = 0$. For $y \to \infty$, we have the charge density wave, and $m_1 > 0, m_2 = 0$, while for $y \to \infty$, we have the Chern insulator, and $m_1 = 0, m_2 > 0$. We assume the masses vary smoothly between the two regions, and write the Dirac equation in the position representation in the y direction:

$$\tilde{\mathcal{H}}_{k_x} = vk_x\mu^z\tau^x - iv\tau^y\partial_y + (m_1(y) + m_2(y)\mu^z)\tau^z. \tag{450}$$

The single-particle eigenfunctions that diagonalize the Hamiltonian obey

$$\left[vk_x\mu\tau^x - iv\tau^y\partial_y + (m_1(y) + m_2(y)\mu)\tau^z\right]\phi_{k_x}(y) = \epsilon_{k_x}\phi_{k_x}(y). \tag{451}$$

Here we replaced $\mu^z \to \mu = \pm 1$ to indicate that we can treat the two eigenvalues of μ^z independently, as constants. Let us seek a solution in which ϕ is an eigenstate of τ^x, i.e. $\tau^x \phi = \tau\phi$, with $\tau = \pm 1$. We can rewrite $i\tau^y = \tau^z\tau^x = \tau^z\tau$ when acting on ϕ. Hence we have

$$\left[\tau\mu v k_x + \tau^z\left(-\tau v\partial_y + (m_1(y) + \mu m_2(y))\right)\right]\phi_{k_x}(y) = \epsilon_{k_x}\phi_{k_x}(y). \tag{452}$$

This is consistent under the conditions:

$$\epsilon_{k_x} = \tau\mu v k_x, \tag{453}$$

$$\left(-\tau v\partial_y + (m_1(y) + \mu m_2(y))\right)\phi_{k_x}(y) = 0. \tag{454}$$

The second equation has a formal solution:

$$\phi_{k_x}(y) = \mathrm{A}e^{\int_0^y dy' \frac{m_1(y') + \mu m_2(y')}{\tau v}}. \tag{455}$$

This solution is formal because this function is only normalizable if the exponential becomes large and negative at *both* $y \to +\infty$ and $y \to -\infty$. This requires $(m_1 + \mu m_2)/\tau < 0$ as $y \to +\infty$ and $(m_1 + \mu m_2)/\tau > 0$ as $y \to -\infty$. In turn this implies that the sign of $m_1 + \mu m_2$ is opposite at $y = \pm\infty$. The interface between the Chern and band insulator satisfies this condition. Specifically, in the band insulator at $y = +\infty$, $m_1 > 0$ and $m_2 = 0$, so the sign $m_1 + \mu m_2$ is positive, while in the Chern insulator, $m_1 = 0$ and $m_2 > 0$, so the sign of $m_1 + \mu m_2$ is the sign of μ. Hence for $\mu = -1$, the sign is different, and we obtain convergence for $\tau = -1$. So we obtain a *single branch* of modes (those with $\mu = \tau = -1$ which have such a special eigenstate, for which the dispersion relation is

$$\epsilon_{k_x} = v k_x. \tag{456}$$

This is a one-dimensional branch of states, whose wavefunction is exponentially localized at the interface between the Chern and band insulator. The mode resides *inside* the band gap, which is required for exponential localization, and which means that its low energy states reside at the Fermi energy even when the bulk of both insulators exhibit a gap. Importantly, the mode is *chiral*, in that the group velocity in the x direction, parallel to the interface, is positive. This is called a chiral edge state, and is characteristic of the integer quantum Hall effect.

The chirality of the edge state gives it a great deal of robustness. Perturbations at the edge, including disorder, cannot make a low energy electron turn around, because all the available states propagate in the same direction. We say that there is no backscattering possible. This means that the edge state cannot become localized by disorder, and in fact has an *infinite conductivity* (but not infinite conductance - we will come back to this).

8.6 *Chern number*

We are now in a position to bring the discussion full circle. We argued that in general, a Chern insulator has a quantized Hall effect, and through Laughlin's argument must have gapless excitations at its edge. We have seen that there is such a gapless edge state at the boundary of one of the phases of the gapped Dirac model. Can we actually see that in this phase there is a non-zero Chern number?

Recall that each massless Dirac point has a π Berry phase associated with loops encircling it. This is almost like having large Berry curvature within the loop: if it were permissible to use Stokes' theorem, then we could conclude that the Berry curvature within integrates to π. This is not correct, but it becomes correct when a small mass term is added. With a mass, it becomes possible to define the Berry curvature everywhere, and we can take its integral in a large area containing the Dirac point. By Stokes' theorem, this must be equal to Θ. The latter cannot change by some large amount for a loop far from the Dirac point, when a very small mass is introduced. Therefore a small mass makes the integrated Berry curvature "jump" to $\pm\pi$.

One may wonder what resolves the sign ambiguity? It is straightforward to just add a mass to the Dirac Hamiltonian and calculate the Berry curvature. For $\tilde{\mathcal{H}} = v\mu^z\tau^x k_x + v\tau^y k_y + m\tau^z$, one obtains for the valence band

$$\Omega = \mu_z \frac{mv^2}{2(v^2k^2 + m^2)^{3/2}}, \tag{457}$$

The sign is opposite for the conduction band. For a given band, the sign of the Berry curvature is determined by the sign of μ_z and the sign of the mass:

$$\int d^2k\, \Omega(\boldsymbol{k}) = \pi\mu_z \operatorname{sgn}(m). \tag{458}$$

Note the appearance of μ_z. This appears because the valley determines the sense of winding of the Dirac point, or chirality. For a given sign of mass, opposite chirality gives opposite Berry curvature. The integrand is strongly peaked in a region of width m/v in momentum space around the Dirac point. So when the Fermi level lies in the gap formed by the mass, we can say, using the general formula of Eq. (401)), that each Dirac point contributes plus or minus *half* an integer to the Chern number. This must be added for every distinct Dirac point, i.e. for each spin and valley. Therefore there is a general formula for the Chern number for a set of massive Dirac points with the Fermi level in the gap:

$$\mathrm{C} = \sum_i \frac{1}{2} \operatorname{sgn}(m_i \mu_i^z). \tag{459}$$

Here the sum is over all Dirac points, i.e. for our model of graphene it includes

four such points, for spin and valley. One might be worried about Eq. (459), because it looks like it can give a half-integer quantum Hall effect. However, for any physical two dimensional system, there is a famous theorem that there must always be an even number of Dirac points. This guarantees an integer result for an insulator.

Let's apply the formula to the two states we discussed. For the CDW, we had the mass m_1, which is the same for both valleys.

$$C^{CDW} = 2_{spin} \times \left(\frac{1}{2}\text{sgn}(m_1) - \frac{1}{2}\text{sgn}(m_1)\right) = 0. \tag{460}$$

The CDW has zero Hall conductivity because the contributions from the two valleys have opposite sign, as expected since it is time-reversal invariant.

Next consider the QAHE phase. Now we have mass m_2 which is opposite for the two valleys. We obtain

$$C^{QAHE} = 2_{spin} \times \left(\frac{1}{2}\text{sgn}(m_2) + \frac{1}{2}\text{sgn}(m_2)\right) = 2\text{sgn}(m_2). \tag{461}$$

Indeed the QAHE state has a non-zero Chern number, which is consistent with the edge state as expected.

8.7 *Chern insulators: summary and bulk-boundary correspondence*

In the prior parts of this section we have come to understand that there are classes of insulators in two dimensions, called Chern insulators, that are non-trivial and cannot be deformed into trivial ones. We described them in a number of ways:

- An example of a Chern insulator occurs in the Haldane model of a honeycomb lattice, which realizes time-reversal breaking opposite mass terms for the two Dirac fermions at the K and K′ points.
- The Chern insulator in this example has a chiral edge state.
- The Hall conductivity is quantized and equal to e^2/h times an integer C known as the Chern number. The Chern number may be expressed in various ways, for non-interacting and interacting electrons.
- Laughlin's argument shows that this quantization can be understood as a consequence of spectral periodicity under insertion of a pure flux equal to the flux quantum, and that C describes a spectral flow at the edges of the sample. Equivalently, C gives the number of electrons pumped across the sample when a flux quantum is inserted. A non-zero C requires that there are extended states at the boundaries of the sample.
- The quantization of the Hall conductivity is robust to both disorder and interactions, but occurs only in the limit of zero temperature.

One point we did not comment on explicitly is the *bulk-boundary correspondence*, which is implied in some way by Laughlin's argument. We saw that the Chern insulator in graphene has a single chiral fermion edge state at the boundary to a trivial insulator. It is not too hard to show that it also has a unit Chern number $\mathrm{C} = \pm 1$. In fact, this correspondence is general, and there is an identity relating the Chern number to the number of left and right moving modes, $\mathrm{N_R}$ and $\mathrm{N_L}$, respectively, at a boundary:

$$\mathrm{C} = \mathrm{N_R} - \mathrm{N_L}. \tag{462}$$

(This equation requires some definition of what "right" and "left" mean but let us not belabor it).

One way to argue for Eq. (462) is to use the fact that the Hall conductivity is given by $e^2/h \times \mathrm{C}$, and then to calculate the Hall voltage directly from the low energy model of edge states, and compare the two results. The latter calculation is quite simple. Consider a Hall bar which is infinite in the x direction and boundary by $y = 0, \mathrm{L}$ in the vertical direction. Suppose a voltage V_y applied between the top and bottom edges. This induces a shift in the chemical potential for the top modes from equilibrium of μ_L and those at the bottom of μ_0, with $\mu_\mathrm{L} - \mu_0 = e\mathrm{V}_y$. Now for each mode, the shifted chemical potential induces a change in the density of electrons. This occurs because when the chemical potential is shited by μ, the states between $k = 0$ and $k = -\mu/\hbar|v|$ change their occupation (v is the velocity of the mode). The change in the electron density for mode a is

$$n_a = -\frac{\mu_a}{2\pi\hbar|v_a|}. \tag{463}$$

Note that electrons are always added with negative μ_a, irrespective of the direction of the velocity of the mode. This is why there is an absolute value here. Now the current induced in this mode is given by $\mathrm{I}_a = -n_a e v_a$, which implies

$$\mathrm{I}_a = \frac{e}{h}\frac{v_a}{|v_a|}\mu_a \tag{464}$$

Then the total current on a single edge is

$$\mathrm{I}_{\mathrm{L}/0} = \frac{e}{h}\mu_{\mathrm{L}/0}\sum_a \frac{v_a}{|v_a|}. \tag{465}$$

Now we can get the full current in the x direction by taking $\mathrm{I}_x = \mathrm{I}_\mathrm{L} - \mathrm{I}_0$,

$$\mathrm{I}_x = \frac{e}{h}(\mu_\mathrm{L} - \mu_0)\sum_a \frac{v_a}{|v_a|} = \frac{e^2}{h}\mathrm{V}_y \sum_a \frac{v_a}{|v_a|}. \tag{466}$$

The final sum is exactly the difference in the number of right and left moving modes, so we see that $I_x = G_{xy} V_y$ with

$$G_{xy} = \frac{e^2}{h}(N_R - N_L). \tag{467}$$

The quantity G_{xy} is the Hall *conductance* rather than the Hall *conductivity*, i.e. it is the ratio of the current to the voltage, rather than the ratio of the current density to electric field. However, one can easily show that these are equal in two dimensions. Thus $\sigma_{xy} = G_{xy}$ in this case, and by comparing to the formula Eq. (233) of the Hall conductivity in terms of Chern number, we prove the bulk-boundary correspondence, Eq. (462).

8.8 *Time-Reversal Symmetric $\mathbb{Z}_2$ TI*

Having understood the Chern insulator from many angles, we will now discuss the case of time-reversal symmetry. As shown in Sec. 7.3.2 that the Berry curvature is an odd function of momentum when time-reversal is present, this is enough to force zero Chern number. So any topological physics is something different. Following seminar work of Kane and others, we know since around 2005 that there is a $\mathbb{Z}_2$ topological distinct amongst time-reversal symmetric insulators in two and three dimensions.

A complication is that, with time-reversal symmetry, we actually generally need to consider a *pair* of bands, because Kramer's theorem implies that "spin reversed" bands cannot be separated from one another. Consider a Bloch state $|\psi_{n\boldsymbol{k}}$. Kramer's theorem generates a new state $|\psi'_{n,-\boldsymbol{k}}\rangle$ which is guaranteed orthogonal to the original one and with the same energy (this is the content of Kramer's theorem). For certain special quasi-momenta, the reversed momentum $-\boldsymbol{k}$ is equivalent to the original one by a reciprocal lattice vector, i.e. $\boldsymbol{k} = \mathbf{Q} - \boldsymbol{k}$, where $\mathbf{Q}$ is a reciprocal lattice vector. In dimensionless coordinates, Eq. (414), these are when $k_i = 0, \frac{1}{2}$ for all i. At these "time reversal invariant momenta" (TRIM), there must be two degenerate orthogonal states, so two bands must pass through these momenta at the same energy. One either has a pair of entirely degenerate bands (which occurs if in addition to time-reversal one also has inversion symmetry), or one has a pair of bands which cross at these TRIM points (which occurs otherwise). In the following, we will successively make two assumptions: (1) two dimensionality, and (2) no inversion symmetry. The former is important mainly for simplicity, and we will comment on the extension to three dimensions later. The latter condition is actually unnecessary, but relaxing it requires discussion which is overly technical.

Let us now specialize to the case of two dimensions. While the full Brillouin zone is spanned by $0 < k_i < 1$, it is enough to consider half the zone because time-reversal symmetry fully determines the states at $-\boldsymbol{k}$ from those at $\boldsymbol{k}$. For convenience, we will therefore take $0 \leq k_2 \leq 1/2$. For fixed $k_2 \neq 0, 1/2$, time-

reversal symmetry does not constrain the bands as a function of k_1. Now we assume no inversion symmetry. In this case, for $k_2 \neq 0, 1/2$, there are two 1d non-degenerate bands (regarded as functions of k_1) whose states and energies are unique (up to the usual phase ambiguity) at each k_1. This allows us to form the Wannier centers of Sec. 8.3.3, which are localized in x_1. In this way, we obtain a pair of Wannier centers, $\bar{x}_{1,\pm}(k_2)$, where $\pm$ indexes the two centers arising from the two bands. We note in passing that this may also be achieved in the case where the bands are degenerate, but requires constructing either maximally localized Wannier functions, or calculating a non-abelian Berry phase (see e.g. Ref.[3]).

Now as $k_2 \to 0$ or $k_2 \to 1/2$, we should recover time-reversal symmetry. Technically, we group the states in such a way that for $k_2 = 0, 1/2$, the two bands (as functions of k_1) are time-reversed copies of one another, e.g. $\epsilon_+(k_1, k_2) = \epsilon_-(-k_1, k_2)$. Then the Wannier centers from each band must coincide:

$$\bar{x}_{1+}(0) = \bar{x}_{1-}(0), \qquad \bar{x}_{1+}(\tfrac{1}{2}) = \bar{x}_{1-}(\tfrac{1}{2}). \tag{468}$$

As k_2 evolves from 0 to 1/2, the paired Wannier centers split apart, move in some fashion, and then recombine into identical pairs. It turns out that there are two inequivalents ways in which this evolution can occur.

8.9 *From Chern to Time-Reversal Symmetric Topological Insulators*

We can use the bulk-boundary correspondence in different ways. One way is to regard the boundary property, i.e. the difference $N_R - N_L$, as the definition of the topological invariant. If we can argue independently of the bulk that this quantity is itself indeed topologically invariant, i.e. it is unchanged by smooth deformations of the Hamiltonian which do not cause a bulk phase transition, then we may not need the bulk definition. This turns out to be possible, and when we generalize beyond the Chern insulators, may be much easier than the bulk approach. This will lead us to the Z_2 topological insulator with time-reversal symmetry in two dimensions.

8.9.1 Chern insulator and chirality of the edge

To do so, we first think through how we can argue for the edge invariant in the case of the Chern insulator, in a non-interacting picture. This is the "chirality" $N_R - N_L$. Consider a semi-infinite sample in the upper half-plane $y > 0$, with axes chosen so that translational symmetry is maintained along the x direction. Then we can still label states by quasimomentum k_x. The spectrum at a fixed k_x will consist of bulk states, which are extended (scattering) wavefunctions that are not bound to the wall, and bound states. The bulk states can have variable energy even at fixed k_x because the momentum transverse to the wall can change, so these appear as continuous regions in the k_x–ϵ plane (ϵ is the single-particle energy). Since we consider an insulator, the bulk states are

separated at all k_x by a non-zero gap. The bound states appear as discrete states at fixed k_x, which then form dispersing curves $\epsilon_n(k_x)$. They must lie within the gap, or they would mix with the continuum states and lose their identity. Since we are interested in topological features, we can imagine deforming

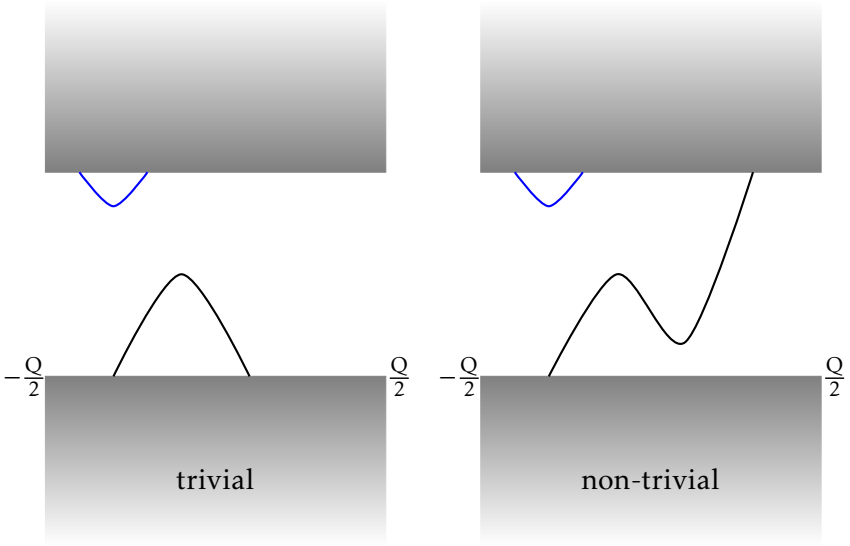


Figure 13: Schematic spectra of semi-infinite 2d insulators without time-reversal symmetry, where the horizontal axis is the momentum parallel to the edge, and the vertical axis is energy. For simplicity the conduction and valence band edges have been flattened. The trivial insulator is characterized by $C = N_R - N_L = 0$ *for all energies within the gap. The non-trivial insulator shown has* $C = N_R - N_L = 1$. *The blue bound state dispersion near the conduction edge can be deformed away, and does not change* C.

the Hamiltonian so that the conduction and valence bands become horizontal at their edges – this does not close the gap so it is allowed. Then a given set of edge modes consists of a set of curves, the bound state dispersions, lying within the gap. These curves must obey certain rules:

1. A curve cannot end except by passing into the continuum.

2. The total number of discrete bound states at fixed k_x changes only when the end of a curve tied to the continuum passes through this k_x. This just means that discrete states do not appear or disappear at energies away from the continuum.

3. Curves must be smooth except when they intersect (this is just the smoothness of non-degenerate eigenvalues we have already discussed several times).

4. Subject to these rules, the curves may be smoothly deformed, and new curves may be added by smoothly drawing them out of a continuum (the latter corresponds to formation of a bound state).

We may start by considering a "trivial" insulator, by which we mean one which can be deformed to the case with no bound states at all. From this we may pull edge state branches out of the conduction or valence bands. Now we study the numbers N_R and N_L of edge states crossing a particular energy insider the gap. As a new branch is created and pulled across this energy, N_R and N_L change but always do so together, so that $N_R - N_L$ is unchanged. One can convince one's self that this remains true for all deformations allowed by the above rules.

In a similar way, we can consider starting with a situation with one right-moving edge state crossing the gap, so that $N_R - N_L = 1$. Once again, deforming this mode or adding new modes may give rise to additional pairs of right and left moving states at some energies, but the chirality $N_R - N_L$ remains fixed. At least at the level of pictures, we can convince ourselves that $N_R - N_L$ is a topological invariant.

8.9.2 Time-reversal invariant TIs and Z_2 invariant

For a time-reversal invariant system, the Chern number must be zero. One can readily see that time-reversal symmetry (TRS) implies $\mathcal{B}(\boldsymbol{k}) = -\mathcal{B}(-\boldsymbol{k})$, which forces $C = 0$. Similarly, under TRS, a right-moving edge mode becomes a left-moving edge mode and so $N_R - N_L = 0$ (so the bulk-boundary correspondence in Eq. (462) is still valid but trivial). However, it turns out that there is a still a topological invariant that survives in the presence of TRS. This is easiest to understand in terms of edge modes. Consider again the semi-infinite sample with translational symmetry and momentum k_x a good quantum number. The presence of the boundary does not spoil TRS, which takes $k_x \to -k_x$. Thus edge modes must come in degenerate pairs at k_x and $-k_x$. In general there are two values of k_x which are time-reversal invariant: $k_x = 0$ and $k_x = Q/2$ where Q is the smallest reciprocal lattice vector of the boundary Brillouin zone. At these time-reversal invariant wavevectors, a two-fold Kramer's degeneracy is required. Apart from these conditions we require the same ones as for the prior case without TRS.

Since the spectrum at k_x is identical to that at $-k_x$, it is sufficient to plot the spectrum for $0 \le k_x \le Q/2$. At both ends of this interval, any bound state modes must occur in pairs. Out of the energies within the gap at $k_x = 0$ and $k_x = Q/2$, two edge modes must emanate. Additional modes may emerge from the conductance and/or valence bands. Consider a fixed energy within the band gap and count the number of modes crossing the horizontal line at that energy. Now imagine varying that energy, which sweeps that line up or down. The number of modes crossing may change as that line crosses the local maxima or minima of edge modes, or by crossing the endpoints at $k_x = 0, Q/2$. However, when it does so, the number always changes by a multiple of 2. Thus the parity of the number of modes crossing the line is independent of the energy within the gap. Similarly, we may vary the edge modes rather than the energy at which we count, and the parity conservation holds. Thus we have

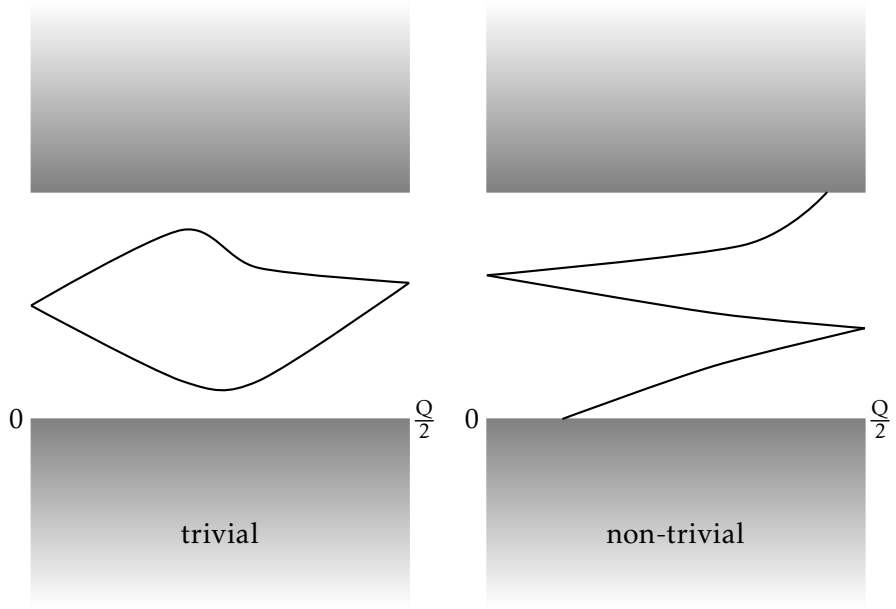


Figure 14: Schematic spectra of semi-infinite 2d time-reversal symmetric insulators, where the horizontal axis is the momentum parallel to the edge, and the vertical axis is energy. Only half the edge Brillouin zone is shown, between two time-reversal invariant momenta. For simplicity the conduction and valence band edges have been flattened. The trivial insulator is characterized by an even number of crossings of bound states at a fixed energy within the gap. The non-trivial insulator has an odd number of such crossings.

identified a $\mathbb{Z}_2$ topological invariant, which is just the parity of the number of modes crossing a constant energy line over half the surface Brillouin zone.

This argument leaves many things unresolved. It is not immediately obvious that the invariant defined this way is independent of the choice of surface (it is – though the generalization to three dimensions is not). What is the bulk definition of the invariant, and a bulk-boundary correspondence? Most importantly, what are the physical consequences of a non-trivial $\mathbb{Z}_2$ topological invariant?

8.10 Z_2 *Topological insulator in graphene*

We now go back to our graphene model and show that we can realize the $\mathbb{Z}_2$ topological insulator there. We consider the possible mass terms that could result if we drop our assumption of spin-rotational symmetry. Choosing a spin quantization axis along z, there are two obvious terms:

$$\mathcal{H}' = m_3 \sigma^z \tau^z + m_4 \sigma^z \mu^z \tau^z. \tag{469}$$

Here the Pauli matrix σ^z acts on electron spin. These masses are obviously not spin-rotationally invariant. How about under time-reversal and inver-

sion? Under inversion or C_2, we know that $\tau^z \to -\tau^z$ and $\mu^z \to -\mu^z$, as this exchanges sublattices and valleys. Inversion does nothing to spin, since it is a pseudovector. Equivalently, a rotation around the z axis does not change σ^z. This means that m_3 is odd under I/C_2 while m_4 is invariant under I/C_2.

Next consider time-reversal. Sublattices are not interchanged but valleys are, so τ^z is invariant under TR while μ^z changes sign under it. Spin is of course odd under TR, so σ^z change sign. We see that m_3 is odd under TR while m_4 is invariant under it. All this may be a little confusing so you may want to check Table 1 to see how each Pauli matrix transforms. Also shown in the last three columns of the table are the transformations of the three mass terms m_2, m_3, m_4.

Symmetry	τ^z	μ^z	σ^z	$\mu^z\tau^z$ (m_2)	$\sigma^z\tau^z$ (m_3)	$\mu^z\sigma^z\tau^z$ (m_4)
I/C_2	-1	-1	1	1	-1	1
TR	1	-1	-1	-1	-1	1
SF	1	1	-1	1	-1	-1

Table 1: Transformation of the different diagonal Pauli matrices under inversion/two-fold rotation (I/C_2), time-reversal (TR), and spin-flip (SF). Here τ^z denotes sublattice, μ^z denotes valley, and σ^z denotes spin. A 1 indicates the Pauli matrix is invariant under this operation, while a -1 indicates that it is odd under it.

Based on this table, we can discuss the nature of the new states created by m_3 and m_4. We see that m_3 breaks time-reversal symmetry, spin symmetry, and inversion symmetry. These are the properties of an *antiferromagnet*. Indeed we could have guessed this because it is just the same as making a CDW (mass m_1) for up spins and a CDW of opposite sign for down spins. Thus it corresponds to spins up on the A sublattice and down on the B sublattice, or vice-versa, depending upon the sign of m_3.

From the table, we see that m_4 breaks only the spin-flip symmetry. Actually spin-flip symmetry is not a symmetry of nature. It is just a consequence of neglecting spin-orbit coupling. In fact, the mass m_4 is completely allowed for graphene. This fact was recognized by Kane and Mele who included it and saw that it generated what they called the quantum spin Hall effect. This actually realizes precisely the situation of the previous subsection – a time-reversal invariant topological insulator. We can see that this is the case because the m_4 mass has just the same form as in the Haldane model, if we look only at one spin polarization. Together, it corresponds to two copies of a spin-less Haldane model with opposite masses. So each spin corresponds to a Chern number ± 1, and this leads to one edge state for each spin, but with opposite chirality for up and down spins. This is exactly the picture we envisioned above.

This was the first (theoretical) discovery of a time-reversal invariant topological insulator. For graphene, however, it turns out that spin-orbit coupling is extremely weak (it is estimated that m_4 is of order micro-eV), so for all

practical purposes the quantum spin Hall effect does not occur there. Still, we now know it occurs in many other materials. It is detectable by the formation of helical edge states, but we will not get into that here.

REFERENCES

[1] Robert Karplus and JM Luttinger. Hall effect in ferromagnetics. *Physical Review*, 95(5):1154, 1954.

[2] Elliott H. Lieb. The stability of matter. *Rev. Mod. Phys.*, 48:553–569, Oct 1976.

[3] Alexey A. Soluyanov and David Vanderbilt. Wannier representation of F_2 topological insulators. *Phys. Rev. B*, 83:035108, Jan 2011.

[4] Ganesh Sundaram and Qian Niu. Wave-packet dynamics in slowly perturbed crystals: Gradient corrections and berry-phase effects. *Phys. Rev. B*, 59:14915–14925, Jun 1999.

[5] Di Xiao, Ming-Che Chang, and Qian Niu. Berry phase effects on electronic properties. *Rev. Mod. Phys.*, 82:1959–2007, Jul 2010.