Lecture Notes on

Solid State Physics

Kevin Zhou

kzhou7@gmail.com

These notes comprise an undergraduate-level introduction to solid state physics. Results from undergraduate quantum mechanics are used freely, but the language of second quantization is not. Nothing in these notes is original; they have been compiled from a variety of sources. The primary sources were:

- Kittel, *Introduction to Solid State Physics*. The standard undergraduate-level introduction to solid state physics, with consistently good quality.
- Ashcroft and Mermin, *Solid State Physics*. The standard graduate-level introduction to solid state physics. Relatively dry and difficult to read. Covers essentially the same conceptual material as Kittel, with more detail on specific properties of solids and experimental techniques.
- Simon, *The Oxford Solid State Basics*. A new book that covers most of the material in Kittel. More modern than any other solid state book at this level, with clear, clean explanations throughout. Also has the huge advantage of using Dirac notation over wavefunctions, which shortens and clarifies almost every equation. Has well-written, helpful problems.
- David Tong's lecture notes on Applications of Quantum Mechanics. Covers much of the material of Simon's book, also in clear modern notation, with a conversational tone. Also contains nice asides on topics like graphene and the Peierls instability.
- Annett, *Superconductivity, Superfluids, and Condensates*. A very clear first book on these phenomena, both describing their basic phenomenology and connecting it to microscopic theory. Uses many body theory, but with a very light touch.
- The Feynman Lectures on Physics, volume 2. About ten chapters cover properties of materials. Entertaining bedtime reading, though a number of statements are out of date.

The most recent version is here; please report any errors found to kzhou7@gmail.com.

Contents

1 Free Electron Models

1.1 Introduction

In these notes, we investigate properties of solids. We would like to ask:

- What is the global ground state of the atoms? Is it a periodic crystal, and if so, what is the crystal structure? Does it agree with the results of X-ray diffraction?
- Given the crystal structure, what are the properties of the solid? For example, can we calculate the thermal and electrical conductivity, color, hardness, magnetic susceptibility, resistivity, etc.?
- Resistivity is an especially interesting quantity, since it can range over 30 orders of magnitude between metals and insulators. Can we explain this dramatic difference in behavior?

In principle, we have a "theory of everything" for solids, given by the Hamiltonian

$$H = -\sum_j \frac{\hbar^2}{2m}\nabla_j^2 - \sum_\alpha \frac{\hbar^2}{2M_\alpha}\nabla_\alpha^2 - \sum_{j,\alpha} \frac{Z_\alpha e^2}{|\mathbf{r}_i - \mathbf{R}_\alpha|} + \sum_{j<k} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_k|} + \sum_{\alpha<\beta} \frac{Z_\alpha Z_\beta e^2}{|\mathbf{R}_\alpha - \mathbf{R}_\beta|}$$

where capital letters/Greek indices denote lattice ions and lowercase letters/Latin indices denote electrons. However, in a solid, with $O(10^{23})$ nuclei and electrons, solving this Hamiltonian exactly is infeasible. In fact, the situation is more like QCD than perturbative QFT. Couplings are generally strong, and perturbation theory can fail. Instead, we must use better approximations.

- First, we can use adiabatic approximations. Since the electrons are much less massive than the ions, the ions move very slowly, and we may treat their effect on the electrons adiabatically. This yields the Born-Oppenheimer approximation.
- Second, we may use independent particle approximations. By neglecting electron-electron interactions, we may approximate the behavior of the full N-body interacting system by considering the behavior of a single electron.
- Third, we may use field theory methods. Suppose we can write the Hamiltonian in the form

  $$H' = E_0 + \sum_k \epsilon'_k \alpha_k^\dagger \alpha_k + f(\ldots, \alpha_k, \ldots, \alpha_{k'}^\dagger, \ldots).$$

  Then the low-lying degrees of freedom behave like independent harmonic oscillators. If f is small, we can treat it perturbatively. This approach is useful because many properties of solids only depend on the low-lying excitations.
- There will generally be two kinds of excitations: quasiparticles, or 'dressed' particles, that resemble a single free particle, and 'collective excitations' which are due to many particles.

First, we'll consider very basic 'free electron' models, which completely neglect interactions of the electrons with the lattice ions; such an approach can only be a good approximation for metals. Next, we reintroduce the lattice ions, leading to a discussion of band structure. Much later, we'll reintroduce interactions between electrons, leading to many body/field theory methods. We'll see that the structure of the Fermi sea tends to make interactions unimportant in some contexts, retroactively justifying the neglect of interactions.

1.2 Drude Theory

The Drude theory of metals, introduced in 1900, models a metal as a classical gas of electrons, assumed to be the valence electrons of the atoms used to form the metal.

- We assume the electrons don't interact with each other at all, the 'independent electron approximation'. However, we will allow collisions with the lattice ions. We take the 'free electron approximation', assuming that in between collisions, the electrons are completely free, with the exception that the ions act as a 'wall' preventing the electrons from leaving the metal.
- We assume that collisions instantaneously randomize the velocity of an electron, so that its mean final velocity is zero, and that they occur in a time dt with probability dt/τ, where τ is the relaxation time.
- If we wanted to treat the collisions more carefully, we could choose the speed distribution after a collision so that the electrons thermalize appropriately over time.
- It is difficult to calculate the collision rate τ^{-1}. There are many contributions, including scattering off impurities, phonons, and other electrons. One might estimate $\tau = 1/n\sigma v$ where σ is the cross section, but the cross section is infinite for the Coulomb fields of electron-electron scattering; the collisions simply aren't sharp as assumed by Drude theory. Instead, we take τ as a phenomenological parameter.

Next, we consider the effects of static fields.

- Suppose the electrons experience an external force $\mathbf{F}$ and have average momentum $\langle \mathbf{p} \rangle$. Then

$$\frac{d\langle \mathbf{p} \rangle}{dt} = -\frac{\mathbf{p}}{\tau} + \mathbf{F}$$

  where the first term accounts for collisions. For simplicity we drop the brackets below.

- The average current density is given by

$$\mathbf{j} = -\frac{ne\mathbf{p}}{m}.$$

  For a static electric field, we thus have

$$\mathbf{p} = -e\mathbf{E}\tau, \quad \mathbf{j} = \sigma\mathbf{E}, \quad \sigma_{\mathrm{DC}} = \frac{ne^2\tau}{m}.$$

- Next, we consider the Hall effect. Suppose we apply an electric field E_x and a transverse magnetic field B_z. Then the Lorentz force should deflect electrons in the y direction, causing a buildup of charge on the side of the metal and creating a transverse field E_y.
- More formally, let $\mathbf{E} = \rho\mathbf{j}$ where ρ is the resistivity tensor. Again working in the steady state, and restricting to a two-dimensional sample in the xy plane,

$$\mathbf{E} = \left(\frac{1}{ne}\mathbf{j} \times \mathbf{B} + \frac{m}{ne^2\tau}\mathbf{j} \right)$$

from which we read off ρ,

$$\rho = \begin{pmatrix} m/ne^2\tau & B/ne \\ -B/ne & m/ne^2\tau \end{pmatrix} = \frac{1}{\sigma_{\text{DC}}} \begin{pmatrix} 1 & \omega_B\tau \\ -\omega_B\tau & 1 \end{pmatrix}$$

where $\omega_B = eB/m$ is the cyclotron frequency. Note that ρ had to be antisymmetric by rotational invariance.

- We conventionally report the Hall coefficient

$$R_H = \frac{\rho_{xy}}{B} = \frac{1}{ne}.$$

This is especially useful because τ, which depends on messy details, cancels out. Another useful fact is that in practice, we measure transverse resistances, but

$$R_{xy} = \frac{V_y}{I_x} = -\frac{LE_y}{LJ_x} = -\frac{E_y}{J_x} = \rho_{xy}$$

so this is equivalent to a measurement of R_H. Note this is particular to two-dimensional samples.

- Finally, it is sometimes useful to know the conductivity,

$$\sigma = \frac{\sigma_{\text{DC}}}{1 + \omega_B^2\tau^2} \begin{pmatrix} 1 & -\omega_B T \\ \omega_B\tau & 1 \end{pmatrix}.$$

- Strikingly, the Hall coefficient depends on the sign of the charge carriers, so it can be used to show that charge carriers have negative charge. But even more strikingly, this fails! The Hall coefficient is measured to have the opposite of the expected sign in some common materials, such as Be and Mg. It is also measured to be anisotropic. These results will be explained by crystal and band structure below.

- In practice, the Hall effect can be used in reverse to detect magnetic fields, using a Hall sensor. To make the measured voltages large, Hall sensors use materials with a low density of electrons, such as semiconductors. Drude theory works very well for semiconductors.

Next, we consider AC fields.

- We consider an electric field with frequency ω, so

$$\frac{d\mathbf{p}}{dt} = -\frac{\mathbf{p}}{\tau} - e\mathbf{E}_0 e^{-i\omega t}.$$

The momentum is also sinusoidal, and we find the frequency-dependent conductivity

$$\sigma(\omega) = \frac{\sigma_{\text{DC}}}{1 - i\omega\tau}, \quad \mathbf{j}(\omega) = \sigma(\omega)\mathbf{E}(\omega).$$

Note that this is only sensible if $\lambda \gg \ell$ where ℓ is the mean free path, since we've neglected the spatial variation of the field.

- It is useful to consider the limit $\tau \to \infty$. Naively we would have

$$\lim_{\tau\to\infty} \sigma(\omega) = -\frac{ne^2}{i\omega m}$$

but the real part requires a bit more care. We have

$$\operatorname{Re}\sigma(\omega) = \frac{ne^2}{m}\frac{\tau}{1+\omega^2\tau^2}$$

which is a Lorentzian with width $1/\tau$ and total area $\pi ne^2/m$. Hence we have

$$\lim_{\tau\to\infty} \sigma(\omega) = \frac{\pi n_s e^2}{m_e}\delta(\omega) - \frac{ne^2}{i\omega m}.$$

This indicates that $\mathbf{j}$ and $\mathbf{E}$ are 90° out of phase except for DC fields, where they are in phase.

- Now suppose we pass an electromagnetic wave through the material. It can be shown that the dielectric constant is related to the conductivity by

$$\epsilon(\omega) = 1 + \frac{4\pi i}{\omega}\sigma(\omega).$$

At low frequencies, the imaginary part of ϵ yields an exponential damping, corresponding to absorption of light. For high frequencies, we have

$$\epsilon(\omega) \approx 1 - \frac{\omega_p^2}{\omega^2}, \quad \omega_p^2 = \frac{4\pi ne^2}{m}$$

where ω_p is called the plasma frequency, and is notably independent of τ. In this regime, the metal is transparent. For intermediate frequencies, ϵ has a negative real part, which means there are only evanescent waves, so the metal is reflective.

- In the more realistic Lorentz oscillator model, we account for the lattice ions by putting every charge carrier on a spring, and replace the collisions with a damping term. At low frequencies, light is transmitted, because the spring suppresses the charges' response. At the resonant frequency ω_0 of the spring, we get strong absorption, while for higher frequencies we find the same features as above, since the spring won't matter.

- We can also consider an AC electric field in the presence of a transverse DC magnetic field. One finds that the conductivity sharply peaks when $\omega = eB/m$, diverging in the limit $\tau \to 0$. This is called the cyclotron resonance; a large current is built up by resonance with the cyclotron frequency. The cyclotron resonance may be used to measure m, and it is generally found that $m \neq m_e$, another consequence of band structure.

Finally, we turn to the thermal conductivity.

- The thermal conductivity is defined as

$$\mathbf{j}_q = \kappa\nabla T$$

where $\mathbf{j}_q$ is the heat current. We may calculate it crudely for gases using kinetic theory.

- In one dimension, suppose that particles have a typical velocity v and move a distance τ before being scattered. If the energy per particle is $E(T(x))$, then

$$j^q \sim \frac{nv}{2}\left(E(T(x-v\tau)) - E(T(x+v\tau))\right) = -nv^2\tau\frac{dE}{dT}\frac{dT}{dx}.$$

  Here $dE/dT = c_v$ is the heat capacity per particle, so

$$\kappa = \frac{\langle v^2\rangle}{3}nc_v\tau$$

  where we divided by three to account for the three spatial dimensions.

- Next, using the standard results

$$\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_BT, \quad c_v = \frac{3}{2}k_B$$

  we conclude that

$$\kappa = \frac{3}{2}\frac{n\tau k_B^2T}{m}.$$

  This still depends on the unknown parameter τ, but it cancels out in the ratio

$$L = \frac{\kappa}{\sigma T} = \frac{3}{2}\frac{k_B^2}{e^2}$$

  called the Lorenz number.

- The Wiedemann-Franz law states that L is approximately constant and temperature-independent across many metals. In the Drude model, the entire analysis above carries over for metals without any change, so it explains the law.

- The derivation above is off by constant factors. More seriously, it doesn't account for the fact that electrons are fermions, which makes c_v much smaller than expected and v much higher than expected. Luckily, both of these errors mostly cancel.

- To see this more sharply, we can consider the thermoelectric effect, which is the fact that an electrical current also transports heat,

$$\mathbf{j}^q = \Pi\mathbf{j}.$$

  To understand this classically, note that in regions of lower temperature, the typical velocities are lower; in the Drude model this appears in the speed distribution after scattering. Hence electrons pile up in regions of lower temperature, producing a gradient of electric potential alongside a gradient of temperature. On the quantum level, the change in temperature instead changes the Fermi level, providing the imbalance of electrons.

- Various aspects of the thermoelectric effect are called the Peltier, Seebeck, and Thomson effects. They may be used for refrigeration devices.

- Since both $\mathbf{j}^q$ and $\mathbf{j}$ are proportional to $\mathbf{v}$, it cancels to yield

$$\Pi = -\frac{c_vT}{3e} = -\frac{k_BT}{2e}.$$

  However, this prediction is off by a factor of 100, and has the wrong sign in some metals.

1.3 Sommerfeld Theory

Sommerfeld theory fixes some of the problems of Drude theory by replacing its Maxwell-Boltzmann distribution with a Fermi-Dirac distribution.

- Considering a box of volume V with N electrons at zero temperature, we have

$$N = 2V \int^{k_F} đ\mathbf{k}$$

  where k_F is the Fermi wavevector and the 2 accounts for the two spin states of the electron. For convenience we define the Fermi energy, temperature, momentum, and velocity by

$$E_F = \frac{\hbar^2 k_F^2}{2m}, \quad T_F = \frac{E_F}{k_B}, \quad p_F = \hbar k_F, \quad v_F = \frac{\hbar k_F}{m}.$$

- Performing the integral, we find

$$k_F = (3\pi^2 n)^{1/3}.$$

  Numerically, for a typical metal, this implies

$$E_F \sim 10\,\text{eV}, \quad T_F \sim 10^5\,\text{K}, \quad v_F \sim \frac{c}{100}.$$

  We see $T \ll T_F$ is always true for metals, since they melt at a far lower temperature.

- It is also useful to define the density of states per unit volume,

$$g(E) = \frac{m^{3/2}}{\pi^2 \hbar^3} \sqrt{2E} = \frac{3n}{2} \sqrt{\frac{E}{E_F^3}}$$

  so that at finite temperatures,

$$\frac{E}{V} = \int dE \frac{E g(E)}{1 + e^{\beta(E-\mu)}}, \quad \frac{N}{V} = \int dE \frac{g(E)}{1 + e^{\beta(E-\mu)}}$$

- Sommerfeld theory correctly predicts the electronic heat capacity, as

$$E(T) - E(T=0) \sim V g(E_F)(k_B T)^2, \quad C \sim (N k_B) \frac{T}{T_F}.$$

  Hence the heat capacity is linear in the temperature, and smaller than the Drude result by a factor of T/T_F. This may be measured at low temperatures, where the T^3 phonon contribution, which is typically much larger, is negligible.

- Using this result gives a reasonable result for Π. As for the Wiedemann-Franz law, we note that $\langle v^2 \rangle$ is larger than in the Drude model by a factor of T/T_F, which is why the Drude result is approximately right.

- Sommerfeld theory also explains Pauli paramagnetism. Upon turning on a magnetic field, the Fermi surfaces for spin up and spin down electrons shift in/out by energy $\mu_B B$, which gives

$$M = g(E_F) \mu_B^2 B$$

  which is on the right order of magnitude, though again the sign is occasionally wrong.

Next, we reflect on the successes and failures of these models.

- Some quantities calculated in the Drude model, such as R_H, are on the right order of magnitude without Sommerfeld theory. At the simplest level, this is simply because they don't depend directly on $\mathbf{v}$. At a deeper level, it's because they only require the equation of motion

$$\frac{d\mathbf{p}}{dt} = -\frac{\mathbf{p}}{\tau} + \mathbf{F}$$

  and in the context of Sommerfeld theory, we can take $\mathbf{p}$ to be the mean momentum of the entire Fermi sea!

- In this context, collisions are quite violent, since they require going from one end of the Fermi sea to the other. In practice, such a process may be composed of many small scattering events which move an electron around the Fermi sea.

- Neither Drude nor Sommerfeld theory cannot account for the number of conduction electrons, and hence cannot describe nonmetals. Similarly, they cannot account for the occasionally opposite sign or modified mass of the charge carriers. Both are resolved by band structure.

- Unlike the predictions of free electron theory, the DC conductivity can be temperature-dependent and anisotropic. Moreover, the frequency dependence of the AC conductivity and hence the optical spectrum is much more complicated than predicted.

- Thermodynamic properties such as the specific heat and equation of state are far off; the lattice ions and electron-electron interactions are important here.

- Free electron theories cannot account for permanent magnetism, which crucially depends on interactions. Moreover, they can't explain why interactions can be ignored, since the Coulomb interaction energies are very large, on the order of the Fermi energy. The resolution is provided by Landau Fermi liquid theory.

2 Crystal Structure

2.1 Bloch's Theorem

We now refine our theory of solids by accounting for the ionic lattice, which produces bands of allowed and forbidden energy. For simplicity, we first show the physical effects in one dimension.

- Consider an electron in a periodic potential U of period a. The translation operator $T(a)$ thus commutes with the Hamiltonian, so they may be simultaneously diagonalized. The energy eigenfunctions hence take the form

  $$\psi(x) = u_k(x)e^{ikx}, \quad u_k(x+a) = u_k(x).$$

  This result is called Bloch's theorem, the eigenfunctions are called Bloch waves, and k is called the crystal momentum; note that it is only defined up to factors of $2\pi/a$.

- Conventionally, we take the crystal momenta to lie in the first Brillouin zone $[-\pi/a, \pi/a]$. The crystal momenta are ambiguous by $k \to k+G$, where the reciprocal lattice G contains elements of the form $2\pi n/a$.

- For a crystal of finite size, say with N lattice ions, the values of k are also quantized. There are N allowed values of k within the Brillouin zone.

- For each value of k, the function $u_k(x)$ obeys a Schrodinger equation with periodic boundary conditions. Hence there are really multiple solutions, with quantized energy. In the limit $N \to \infty$, the energy levels fill out "bands" of allowed and forbidden energy.

- Since the potential is periodic, its Fourier transform has support on the reciprocal lattice

  $$U(x) = \sum_G U_G e^{iGx}$$

  and since $U(x)$ is real, $U_G^* = U_{-G}$. Plugging in the Fourier transform

  $$\psi(x) = \sum_k C(k)e^{ikx}, \quad k = \frac{2\pi n}{L}$$

  the Schrodinger equation becomes

  $$\sum_k \frac{k^2}{2m} C(k)e^{ikx} + \sum_{G,k'} U_G e^{iGx} C(k')e^{ik'x} = E\sum_k C(k)e^{ikx}.$$

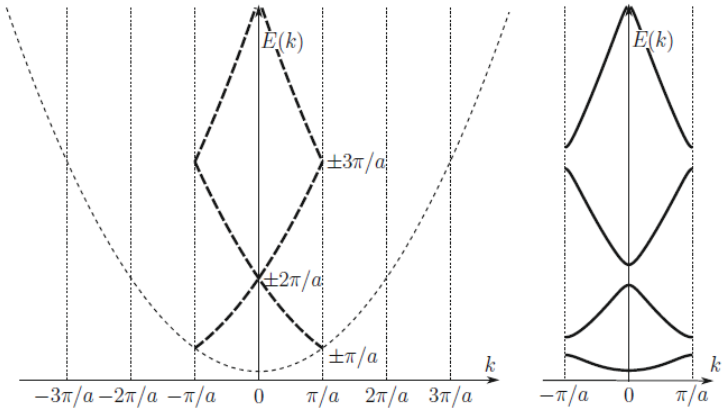
  Setting $k' = k - G$ and setting the e^{ikx} Fourier coefficient to zero,

  $$\left(\frac{k^2}{2m} - E\right) C(k) + \sum_G U_G C(k-G) = 0.$$

- This result gives us a better understanding of Bloch's theorem. We see that scattering off the lattice can only exchange momenta in the reciprocal lattice. Therefore, every stationary state can be built out of a linear combination of plane waves whose momenta differ by reciprocal lattice vectors.

- The result above is not unfamiliar. An electron in an external field $e^{i\omega t}$ may absorb energy from the field in multiples of $\hbar\omega$, so its energy is only conserved up to multiples of $\hbar\omega$. The lattice ions simply do the same with spatial variation.
- The crystal momentum is not the momentum of the electron. If an electron has crystal momentum k, it has Fourier components at all $k + G$. However, crystal momentum behaves somewhat like momentum. If a photon carries crystal momentum q, then in an electron-phonon interaction $k + q$ is conserved, again up to multiples of G. These multiples of G correspond to momentum being transferred to the lattice ions.
- Conceptually, momentum is the conserved quantity associated with spatial translations, while crystal momentum is the conserved quantity (up to G) associated with translation of excitations within a medium. In other words, the difference is between translating the crystal and everything in it, and translating the phonon within the crystal.
- A particle that interacts weakly with the crystal, such as an X-ray photon, has equal momentum and crystal momentum. However, a phonon with definite crystal momentum has precisely zero momentum, because the only Fourier component with any momentum is the $k = 0$ component. Hence when an X-ray photon is scattered off a phonon, we can conclude momentum is transferred, but it is not given to the phonon but instead to the lattice as a whole.
- This remains true even if we account for momentum uncertainties. Suppose the phonon and scattered photon are both reasonably well-localized. Then the lattice configuration after the scattering is the superposition of wavepackets peaked near wavenumbers 0 and some $k \neq 0$. However, since phonons generically have nontrivial dispersion, these wavepackets separate quickly, so it is not reasonable to attribute the momentum to the phonon.
- In the theory of liquids, one often works in an approximation where the dispersion is linear. In that case the wavepackets don't separate, so one may attribute momentum to phonons.

Note. To understand the appearance of band gaps, we can use the "nearly free electron model". Starting from free electrons with a continuous spectrum $\omega = k^2/2m$, turning on the potential splits up the spectrum. Small band gaps appear because of the general phenomenon of level repulsion.



More explicitly, suppose the potential is $U(x) = U_0 \cos(2\pi x/a)$, and consider the initially degenerate states $e^{\pm ik\pi/a}$. The new eigenstates are cosines and sines, and their probability accumulates either close to, or far away from the atoms. These represent bonding and antibonding states.

Note. To understand the appearance of bands, we can use the "tight binding model". We begin with electrons tightly bound to each lattice ion. Then we have discrete atomic energy levels. Allowing the electrons to 'hop' between the atoms broadens these levels into small bands.

Example. The Kronig-Penney model is a simple model of a solid, where we choose the potential to be periodic and stepwise constant. More specifically, the 'unit cell' for the potential is

$$V(x) = \begin{cases} 0 & -w \le x \le 0 \\ V_0 & 0 \le x \le b \end{cases}$$

We assume that $N \to \infty$, so we don't have to worry about the periodic boundary conditions. The period is $a = w + b$. The solutions in these two regions, assuming $E < V_0$, are

$$\psi_{\mathbf{I}}(x) = Ae^{iqx} + Be^{-iqx}, \quad q(E) = \sqrt{2mE}$$

and

$$\psi_{\mathbf{II}}(x) = Ce^{\beta x} + De^{-\beta x}, \quad \beta(E) = \sqrt{2m(V_0 - E)}.$$

The first two conditions are continuity of the wavefunction and its derivative,

$$\psi_{\mathbf{I}}(0) = \psi_{\mathbf{II}}(0), \quad \psi'_{\mathbf{I}}(0) = \psi'_{\mathbf{II}}(0).$$

We also have the same conditions at $x = -w$ and $x = b$, up to the Bloch phase factor e^{ika},

$$\psi_{\mathbf{II}}(b) = e^{ika}\psi_{\mathbf{I}}(-w), \quad \psi'_{\mathbf{II}}(b) = e^{ika}\psi'_{\mathbf{I}}(-w).$$

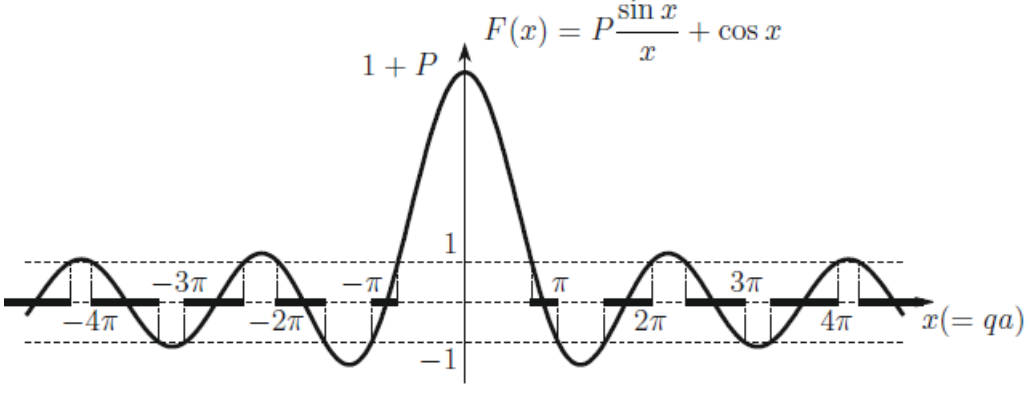
We find four linear equations for the coefficients A, B, C, and D. Since zero is a solution, the only way to get a nontrivial solution is if the determinant of the matrix of coefficients is zero. Evaluating this condition, we find

$$\frac{\beta^2 - q^2}{2q\beta} \sinh \beta b \sin qw + \cosh \beta b \cos qw = \cos ka.$$

As a further simplification, we may narrow the potentials to delta functions, taking $b \to 0$ while $V_0 b$ is held constant; this gives

$$P\frac{\sin qa}{qa} + \cos qa = \cos ka, \quad P = mV_0 ba.$$

As a function of qa, the left-hand side interpolates between sinc-like and cos-like. We only have a solution for k if the left-hand side is between -1 and 1. Therefore we find bands separated by gaps in q (and hence in E), as expected. For high q, the separation between the bands decreases, as the effect of the potential becomes relatively weaker. This result is shown below.



2.2 Bravais Lattices

We now investigate 3D crystal structure in detail.

- A three-dimensional Bravais lattice is the set of all points of the form

$$\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$$

  where the $\mathbf{a}_i$ are linearly independent and the n_i are integers. Alternatively, a Bravais lattice is a set of discrete points that looks exactly the same from every point, or a set of vectors closed under addition.

- We say the $\mathbf{a}_i$ are primitive lattice vectors and they generate/span the lattice. Note that primitive vectors are not unique; different equivalent sets of primitive vectors are related by matrices A with integer entries and unit determinant.

- The coordination number of a Bravais lattice is the number of nearest neighbors of each point.

- A unit cell is a pattern that, when translated by a sublattice of a Bravais lattice, tiles the space perfectly. A primitive unit cell does the same when translated by any vector in the lattice. Primitive unit cells are far from unique, but all have the same volume, since they contain one lattice point each.

- One example of a primitive unit cell is the set

$$\{\mathbf{r} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 \,|\, 0 \leq x_i \leq 1\}.$$

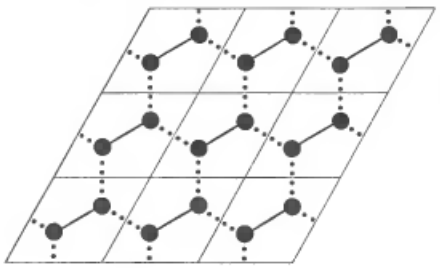
  However, note that this will often be less rotationally symmetric than the lattice as a whole. A non-primitive 'conventional' unit cell is often picked to have the full symmetry of the lattice.

- The Wigner-Seitz cell about a lattice point is the set of points that are closer to that point than to any other point; it is a special case of a Voronoi cell. The cells are all identical by the definition of a Bravais lattices, so the Wigner-Seitz cell is indeed a primitive cell.

- Geometrically, the Wigner-Seitz cell may be constructed by drawing the planes of equidistant points between our chosen point and each other point in the lattice. It can be shown that the Wigner-Seitz cell has the same symmetries as the lattice.

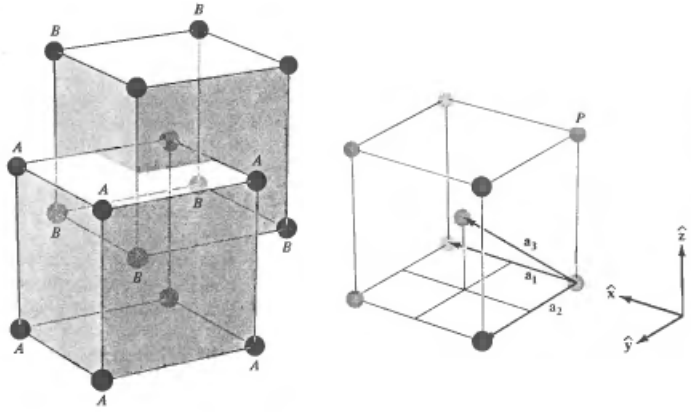
- For a real crystal, the Bravais lattice only reflects the underlying translational symmetry, but we must also specify the positions of the atoms with respect to a unit cell; such a description with respect is called a basis.

Example. The honeycomb lattice of graphene is not a Bravais lattice, since the lattice does not look the same from any two neighboring points, but it is a Bravais lattice with a basis.



The underlying Bravais lattice is called a hexagonal, or triangular lattice. If the primitive vectors are $\mathbf{a}_1$ and $\mathbf{a}_2$, the basis is $\{(1/3)(\mathbf{a}_1+\mathbf{a}_2),(2/3)(\mathbf{a}_1+\mathbf{a}_2)\}$.

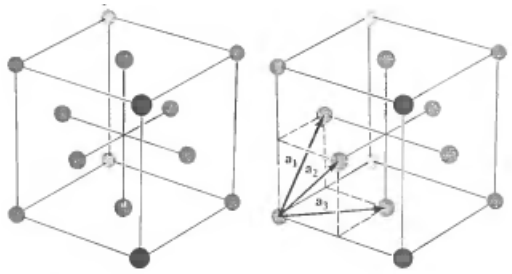
Example. A simple cubic lattice has primitive vectors $\hat{x}$, $\hat{y}$, and $\hat{z}$. A body-centered cubic (bcc) lattice is the same, but adds an extra point in the center of each of the cubes.



This is a Bravais lattice, since every point is surrounded by 8 neighbors in a cubic arrangement. One set of primitive vectors is

$$\mathbf{a}_1=\hat{x},\quad \mathbf{a}_2=\hat{y},\quad \mathbf{a}_3=\frac{1}{2}(\hat{x}+\hat{y}+\hat{z}).$$

Next, the face-centered cubic lattice (fcc) is obtained by adding a point to the center of each face of the simple cubic lattice. Its conventional unit cell is shown below.

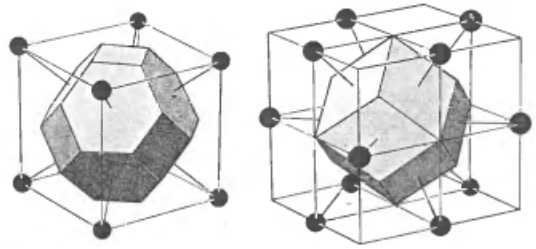


To see this is a Bravais lattice, note that from the perspective of each of the new points, the old points are the ones in the centers of the faces. A set of primitive vectors is

$$\mathbf{a}_1=\frac{1}{2}(\hat{y}+\hat{z}),\quad \mathbf{a}_2=\frac{1}{2}(\hat{z}+\hat{x}),\quad \mathbf{a}_3=\frac{1}{2}(\hat{x}+\hat{y}).$$

Both the bcc and fcc lattices are very common in nature, much more so than the simple cubic lattice, because they pack the atoms closer together.

Example. The Wigner-Seitz cells for the bcc and fcc lattices are shown below.



Note that the surrounding cell for the fcc lattice is not the conventional unit cell for the fcc lattice, which we have shown previously; instead, it is translated to center the Wigner-Seitz cell. In this picture the atoms are at the midpoints of the sides of the surrounding cube.

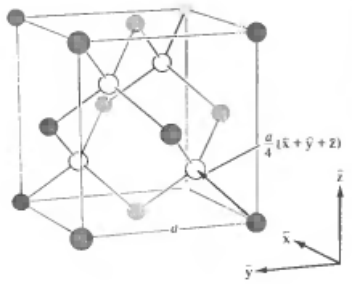
Example. Every Bravais lattice is trivially a lattice with a one-point basis. In addition, the bcc and fcc lattices may be described by a cubic conventional cell with the bases

$$\text{bcc: } \left\{\mathbf{0}, \frac{1}{2}(\hat{x}+\hat{y}+\hat{z})\right\}, \quad \text{fcc: } \left\{\mathbf{0}, \frac{1}{2}(\hat{x}+\hat{y}), \frac{1}{2}(\hat{y}+\hat{z}), \frac{1}{2}(\hat{z}+\hat{x})\right\}.$$

From now on, we simply use the word 'lattice' to refer to a Bravais lattice or a lattice with a basis.

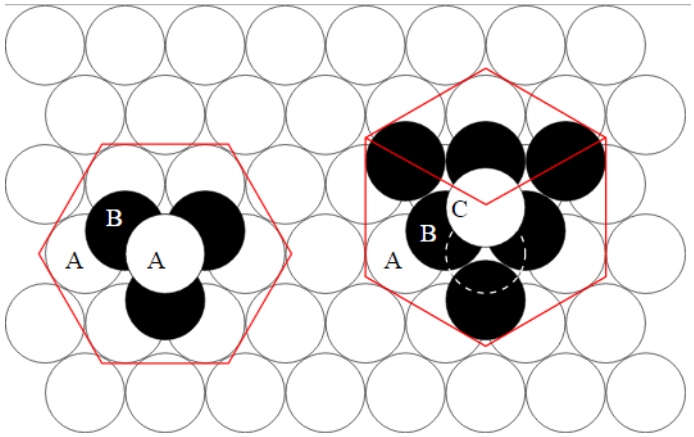
Example. Salt (sodium chloride) has a simple cubic structure with sodium and chloride atoms in a checkerboard pattern. Then the lattice is fcc with a two-point basis, $\mathbf{0}$ and $(\hat{x}+\hat{y}+\hat{z})/2$.

Example. The diamond lattice also has the form of two interpenetrating fcc lattices.



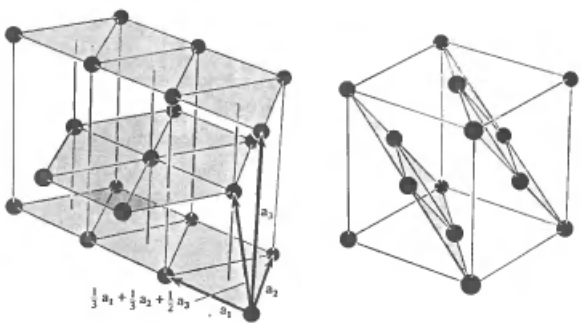
Its two-point basis is $\mathbf{0}$ and $(\hat{x}+\hat{y}+\hat{z})/4$. Silicon and germanium also crystallize in this form.

Example. Sphere packing. In three dimensions, there are several ways to get the densest possible packing of identical spheres. To start, note that the densest packing for a single layer has six spheres around each sphere, as shown below.



Write the first layer of spheres as A. There are two different ways to stack a second layer of spheres on top, written as B and C in the figure. Then we can form a close packed structure by taking any string of A's, B's, and C's where no two adjacent letters are identical.

For example, the pattern AB is formed by two interspersed simple hexagonal Bravais lattices, shown at left below; each of these lattices has horizontal side length a and height c, where $c = \sqrt{8/3}a$. This is called the hexagonal close packed (hcp) structure. The hcp structure is not a Bravais lattice, but it is very common in real materials.



The next simplest pattern is ABC, shown at right above, which turns out to simply be the fcc lattice viewed at an angle. Some rare earth metals also close-pack in the structure $ABAC$.

We now consider symmetries of Bravais lattices beyond translations.

- The space group of a Bravais lattice is the group of isometries that takes the lattice to itself. These include translations, reflections, and inversions.
- Consider some isometry S that takes 0 to $\mathbf{R}$. Now, translation by $-\mathbf{R}$ must also be a symmetry, since the lattice is Bravais, so $T_{-\mathbf{R}}S$ is a symmetry/isometry that fixes the point at the origin. Similarly, all isometries of a Bravais lattice may be factored in this way.

  Note that this argument fails if the lattice if not a Bravais lattice; for instance, we may have glide-reflections, which cannot be translated away.
- The set of isometries that fix a given point form the point group. We will show that there are seven distinct point groups, yielding seven 'crystal systems'. Considering the full space group, we will find that there are fourteen distinct possible space groups/Bravais lattices.

  When we say space groups are the same, we mean they are obtainable from each other by continuous lattice deformations without losing or gaining any symmetries.
- More formally, we say a space group is symmorphic if it is the direct product of a point group and a translation group. All space groups of Bravais lattices are symmorphic, but only a small fraction of crystal structures (lattices with bases) are, due to glide-rotations (screw axes) and glide-reflections (glide planes).

We now enumerate the seven crystal systems and fourteen Bravais lattices.

1. Cubic. The point group is the symmetry group of the cube. There are simple, body-centered, and face-centered cubic lattices.
2. Tetragonal. We stretch the cube so that the c-axis is a different length than the rest. We find a simple and a centered tetragonal lattice. There is no distinction between the body-centered and face-centered tetragonal lattices, because they can be converted into each other by redefining the unit cell.
3. Orthorhombic. We further stretch the cube so all sides are different lengths. There are simple, body-centered, face-centered, and base-centered orthorhombic lattices.
4. Monoclinic. We distort angles so that square faces become parallelograms.

5. Triclinic. We further distort angles so there are no right angles.
6. Trigonal. This is obtained by stretching a cube along a space diagonal. It is generated by three primitive vectors of equal length, that make equal angles with one another.
7. Hexagonal. This is unrelated to the cubic lattices. The lattice is made up of hexagonal prisms.

Next, we can consider general crystal structures. In this case, there are 32 point groups, called the crystallographic point groups, and 230 space groups.

- Our previous Bravais lattices were effectively lattices with bases, where the bases had complete spherical symmetry. In a general crystal structure, the basis may have lesser symmetry, so we obtain the 32 point groups by removing parts of the original point groups.
- For example, the cubic crystal system has full octahedral symmetry, giving a symmetry group with 48 elements. However, there are four other cubic point groups, which have 12, 24, 24, and 24 elements. (We call these point groups cubic, because the full cubic symmetry is the smallest group they fit into.)
- The Schoenflies notation labels crystallographic point groups by the group generators. These can include n-fold rotations (with $n = 2, 3, 4, 6$ by the crystallographic restriction), reflections, inversions, and rotation-reflections and rotation-inversions. 73 of the 230 space groups are symmorphic, while the rest are not.
- There is a subtlety here: a crystal without cubic symmetry but with a cubic lattice does not count as cubic. This is because there is no reason for the lattice to be cubic; it is an "accidental" symmetry and closer measurement would show that the lattice was not perfectly cubic.

2.3 The Reciprocal Lattice

We now define the reciprocal lattice, which will be useful for analyzing diffraction.

- Given a Bravais lattice, a point $\mathbf{G}$ is in the reciprocal lattice if

  $$e^{i\mathbf{G}\cdot\mathbf{R}} = 1$$

  for all points $\mathbf{R}$ in the direct lattice. The reciprocal lattice is a Bravais lattice, and by the symmetry of the definition, taking the reciprocal lattice twice recovers the original/direct lattice.
- Equivalently, the reciprocal lattice has lattice vectors $\mathbf{b}_i$ where

  $$\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}.$$

  To show the $\mathbf{b}_i$ span the reciprocal lattice, note that for a point $\sum_i m_i\mathbf{b}_i$ in the reciprocal lattice, we require $\sum_i m_i n_i$ to be integral for all integers n_i. This holds precisely when the m_i are integers.
- Explicitly, we have

  $$\mathbf{b}_1 = \frac{2\pi\mathbf{a}_2 \times \mathbf{a}_3}{V}, \quad V = \mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)$$

  where V is the volume of a primitive cell of the direct lattice, and the rest are defined cyclically. The volume of a primitive cell in the reciprocal lattice is hence $(2\pi)^3/V$. Note that the vectors in the reciprocal lattice are really dual vectors, so they have dimensions of inverse length.

- The reciprocal lattice has the same rotational symmetries as the direct lattice. To see this, note that if $\mathbf{b}_j$ satisfies the relation $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$, then $\mathbf{b}'_j = R_g^{-1}\mathbf{b}_j$ satisfies the same relation with the rotated vectors $\mathbf{a}'_i = R_g\mathbf{a}_i$. If R_g is a symmetry of the direct lattice, then $\mathbf{a}_i$ and $\mathbf{a}'_i$ generate the same direct lattice, so $\mathbf{b}_i$ and $\mathbf{b}'_i$ generate the same reciprocal lattice, and R_g^{-1} is a symmetry of the reciprocal lattice.

- Another interpretation of the reciprocal lattice is that it is the Fourier transform of the direct lattice. More specifically, we have the Fourier pairs

$$\rho(\mathbf{r}) = \sum_{\mathbf{R}} \delta(\mathbf{r} - \mathbf{R}), \quad \widetilde{\rho}(\mathbf{k}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} = \frac{1}{V}\sum_{\mathbf{G}} \delta\!\!\!^{-}(\mathbf{k} - \mathbf{G})$$

  where the $\mathbf{R}$ are in the direct lattice and the $\mathbf{G}$ are in the reciprocal lattice.

- For a general periodic structure, we consider the Fourier transform of a periodic function $\rho(\mathbf{r})$ with the translational symmetry of the direct lattice, $\rho(\mathbf{r}) = \rho(\mathbf{r}+\mathbf{R})$. Then by similar reasoning,

$$\widetilde{\rho}(\mathbf{k}) = \frac{1}{V}\sum_{\mathbf{G}} \delta\!\!\!^{-}(\mathbf{k} - \mathbf{G})S(\mathbf{k}), \quad S(\mathbf{k}) = \int_V d\mathbf{x}\, e^{i\mathbf{k}\cdot\mathbf{x}}\rho(\mathbf{x})$$

  where $S(\mathbf{k})$ is integrated over a unit cell and is called the structure factor.

- In practice, we often let the $\mathbf{a}_i$ span a conventional unit cell, which means the $\mathbf{b}_i$ will generate a lattice that contains the reciprocal lattice as a sublattice. For example, for a face-centered cubic lattice we typically use the conventional unit cell because the $\mathbf{a}_i$ and $\mathbf{b}_i$ are then orthogonal.

Example. The reciprocal of a simple cubic lattice is another simple cubic lattice, and the reciprocal of a fcc lattice is a bcc lattice.

Next, we construct some quantities derived from the reciprocal lattice.

- A Brillouin zone is any primitive cell of the reciprocal lattice. Like in the one-dimensional case, a Brillouin zone contains all physically distinct values of crystal momentum.

- The Wigner-Seitz primitive cell of the reciprocal lattice is called the first Brillouin zone. The n^{th} Brillouin zone is the set of points for which the origin is the n^{th} closest lattice point.

- To construct the n^{th} Brillouin zone, draw the set of all perpendicular bisectors between reciprocal lattice points. Then the n^{th} Brillouin zone is the set of points that require at least $n-1$ crossings but not n crossings to reach from the origin.

- The physical interpretation of the Brillouin zone remains the same: for a crystal with N sites, the (first) Brillouin zone contains N allowed momenta. Moreover, each higher Brillouin zone also contains N allowed momenta. Hence all Brillouin zones have the same volume.

- A lattice plane is any plane containing at least three noncollinear Bravais lattice points. A family of lattice planes is a set of parallel, equally spaced lattice planes, which contain all points in the Bravais lattice.

- Every family of lattice planes corresponds to a vector in the reciprocal lattice perpendicular to each plane, with length $2\pi/d$ where d is the plane separation. This follows directly from the definition of the reciprocal lattice. For example, suppose we have a family of lattice planes with unit normal $\hat{n}$ and separation d. Then the vector $2\pi\hat{n}/d$ is in the reciprocal lattice by definition, since it takes the same value within each plane, and also the same value between each plane.

- We may this to specify lattice planes using reciprocal lattice vectors. We say that a lattice plane has Miller indices (hkl) if the resulting reciprocal lattice vector is

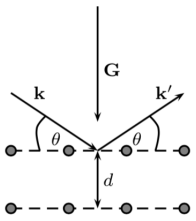
$$h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3.$$

  By construction, h, k, and l have no common factors. Note that the Miller indices depend on the choice of reciprocal lattice basis.

- Bravais' law states that crystals cleave most readily along lattice planes that maximize the distance to adjacent lattice planes. These correspond to families of lattice planes with small Miller indices.

2.4 X-ray Diffraction

We can probe the structure of the reciprocal lattice using diffraction of photons; the photons with wavelengths on the order of the atomic spacing are in the X-ray range. Experimentally, we find sharp diffraction peaks at locations corresponding to reciprocal lattice vectors.



- We assume that scattering is elastic; this is true for the majority of the radiation. The basic equation is Fermi's golden rule,

$$\Gamma(\mathbf{k}', \mathbf{k}) = \frac{2\pi}{\hbar}|\langle \mathbf{k}'|V|\mathbf{k}\rangle|^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}})$$

  which gives the rate of scattering from momentum $\mathbf{k}$ to $\mathbf{k}'$, and V is the potential the photons experience. Since V has the periodicity of the lattice,

$$\mathbf{k} - \mathbf{k}' = \mathbf{G}$$

  for $\mathbf{G}$ in the reciprocal lattice; this is the Laue condition. The momentum the photon loses is gained by the crystal; this does not cost any energy in the limit of infinite crystal mass.

- Alternatively, we can consider how the scattering happens in detail. Consider a family of lattice planes with separation d, and let θ be the angle of incidence. Then assuming specular reflection off of every plane, demanding that the reflected waves interfere constructively gives

$$n\lambda = 2d\sin\theta.$$

  This is the Bragg condition.

- These two conditions are equivalent. Starting from the Laue condition,

$$\frac{2\pi}{\lambda}(\hat{\mathbf{k}} - \hat{\mathbf{k}}') = \mathbf{G}$$

and dotting both sides with $\hat{\mathbf{G}}$ gives

$$\frac{2\pi}{\lambda}(2\sin\theta) = |\mathbf{G}|$$

where θ is the angle of $\mathbf{k}$ to a lattice plane in a family of lattice planes. The separation d between the planes obeys $|\mathbf{G}| = 2\pi n/d$, giving the Bragg condition. The two conditions both say scattering must be in phase; the Bragg condition simply sums over planes first.

- As we've seen, $\tilde{V}(\mathbf{k})$ has delta peaks at the reciprocal lattice vectors times a structure factor

$$S(\mathbf{k}) = \int_V d\mathbf{x}\, e^{i\mathbf{k}\cdot\mathbf{x}} V(\mathbf{x})$$

where the integral is over a unit cell. The most important factor here typically comes from the basis, if there is one, though for larger $\mathbf{k}$ the structure of the individual atoms is also important.

Next we make some qualitative remarks on scattering.

- For X-ray photons, the scattering potential is effectively the electron density; this may be proven classically by treating the scattering as Thomson scattering. **(how to prove?)** Then we pick up structure factors from both the unit cell structure and the shape of the orbitals, called atomic form factors.
- The total amount of scattering from an atom is roughly proportional to the number of electrons. Hence X-ray diffraction patterns can look quite complicated, and have trouble distinguishing light atoms.
- Scattering can also be done with neutrons. Here the potential is sourced by the atomic nuclei. The nuclear form factors are negligible, since the nuclei are very small, and the strength of scattering varies erratically with the atomic number, so that even light nuclei can be easily identified and distinguished.
- A more recent approach is electron diffraction crystallography, which has been used to determine the structure of some very complicated biological structures.
- There are many corrections we haven't discussed above. One of the most important is the Lorentz polarization correction, which gives an extra angular dependence to the scattering intensity, computed by accounting for the polarization of the X-rays. There are also Debye-Waller factors that reduce the scattering intensity due to the thermal motion of the crystal.
- Diffraction techniques can also be used to investigate amorphous solids or liquids; this works because they have short-range order, so the diffraction peaks still exist, though they are broadened.
- We may also consider inelastic scattering, where the incident wave excites an internal degree of freedom. In such processes, the energy and crystal momentum are conserved, allowing one to determine the dispersion relation of the internal excitation.

- For example, neutron scattering is often used to determine phonon dispersion relations. This is easier than using X-ray scattering, because the speed of light is very large, and hence the wavenumbers k of the X-rays can only change by a small amount for a given energy transfer.

Example. Consider a bcc lattice, as a simple cubic lattice with basis. Then the structure factor is

$$S_{\mathbf{G}} = 1 + \exp(-i\mathbf{G} \cdot (a/2)(\hat{x} + \hat{y} + \hat{z})).$$

Plugging in $\mathbf{G} = \sum v_i \mathbf{b}_i$, where the $\mathbf{b}_i$ are the reciprocal lattice vectors of the simple cubic lattice,

$$S_{\mathbf{G}} = 1 + (-1)^{\sum v_i}.$$

Therefore, the form factor is nonzero if $\sum v_i$ is even; this phenomenon is called a "systematic absence", and would hold even if there were a nontrivial unit cell. It is simply the statement that the reciprocal lattice is fcc.

Example. Let's repeat the exercise with the fcc lattice. Then the structure factor is

$$S_{\mathbf{G}} = 1 + \exp(-i\mathbf{G} \cdot (a/2)(\hat{x} + \hat{y})) + \exp(-i\mathbf{G} \cdot (a/2)(\hat{y} + \hat{z})) + \exp(-i\mathbf{G} \cdot (a/2)(\hat{z} + \hat{x})).$$

which yields

$$S_{\mathbf{G}} = 1 + (-1)^{v_x + v_y} + (-1)^{v_y + v_z} + (-1)^{v_z + v_x}.$$

This is nonzero if the v_i are all even or all odd, giving a bcc lattice as expected.

Now we consider experimental techniques for observing X-ray diffraction.

- The main issue is that, since the incoming momentum must lie on a Bragg plane, and these planes do not fill space, we generically get no strong scattering at all if we send in a monochromatic plane wave.
- To see this more generally, we can use the Ewald construction. For an incoming plane wave with wavevector $\mathbf{k}$, draw a sphere of radius k centered at its tip, called the Ewald sphere. Then we observe a diffraction peak if a reciprocal lattice vector lies on this sphere.
- In the Laue method, we send in a range of incoming wavelengths, giving the Ewald sphere nonzero thickness.
- In the rotating crystal method, we continuously rotate the crystal. Reciprocal lattice points rotate with it, intersecting the Ewald sphere at some points.
- Both of these methods rely on the ability to make a large crystal of a material. In the powder, or Debye-Scherrer method, we diffract the beam through a powder whose grains are randomly oriented, giving a similar result to the rotating crystal method. Alternatively, we may use a 'bad sample' with crystal grains pointing in all directions.
- A typical lab produces X-rays by accelerating electrons to hit a metal target. This produces a continuous spectrum of X-ray radiation from Bremsstrahlung, as well as a discrete spectrum from X-ray fluorescence, when the electrons eject electrons from low-lying orbitals, and electrons from higher-lying orbitals fall in to replace them. Usually, the X-rays are detected using semiconductor devices.

- Higher intensity X-ray sources are provided by large facilities called synchrotron light sources, where electrons are accelerated in circles at GeV energies. They emit highly collimated synchrotron radiation in the X-ray range.
- Neutron diffraction is only practical at large facilities. In the past, byproduct neutrons from nuclear reactors were used. Modern facilities use the technique of spallation, where protons are accelerated into a target and neutrons are emitted. There are many methods to detect the outgoing neutrons, though they all involve interactions with nuclei.
- Also note that a continuous spectrum can be converted into a discrete spectrum by diffraction from a known crystal. In the case of neutrons, one can also use time of flight.

3 Band Structure

3.1 Bloch Electrons

In three dimensions, Bloch's theorem states

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

where n indexes the band, $\mathbf{k}$ is the crystal momentum, and $u_{n,\mathbf{k}}$ is periodic,

$$u_{n,\mathbf{k}}(\mathbf{r}+\mathbf{a}) = u_{n,\mathbf{k}}(\mathbf{r})$$

for $\mathbf{a}$ in the Bravais lattice. As before, the volume of k-space per allowed value of $\mathbf{k}$ is $(2\pi)^3/V$, where V is the total crystal volume, equal to N times the volume of a primitive cell. We now consider the dynamics of Bloch electrons in one dimension.

- Let $\hat{p}$ denote momentum (i.e. the operator $-i\hbar\nabla$) and let k be the crystal momentum. Consider a wavepacket built out of $\exp(i(kx-\omega(k)t))$ waves. Then

  $$v_p = \frac{\omega}{k}, \quad v_g = \frac{d\omega}{dk}.$$

  Now, for a free particle we define the velocity operator $\hat{v} = \hat{p}/m$. Using the dispersion relation, we find $\langle\hat{v}\rangle = v_g$. This makes physical sense, because the quantum velocity operator should measure probability flux, and v_g does the same.

- More generally, in the spirit of canonical quantization, the velocity operator must be defined by the equation $v = \partial H/\partial p$ in Hamiltonian mechanics. That is,

  $$\hat{v} = \left.\frac{\partial H}{\partial p}\right|_{x\to\hat{x},p\to\hat{p}}.$$

  Hence, $\langle\hat{v}\rangle$ is always equal to the group velocity.

- In the special case of a Bloch electron, we simply have $\hat{v} = \hat{p}/m$ as usual. But by the reasoning above, this implies

  $$\langle\hat{v}\rangle = \frac{dE}{dk}.$$

  Physically, the electron can move smoothly, without abruptly colliding with the ions.

- More formally, we can show this using perturbation theory. The periodic function $u(k)$ has an effective Hamiltonian obtained by substituting $p \to p + \hbar k$ in the usual Hamiltonian, so

  $$\langle u(k)|\frac{(p+\hbar k)^2}{2m} + V(x)|u(k)\rangle = E(k).$$

  By first-order perturbation theory, this implies

  $$\frac{dE(k)}{dk} = \langle u(k)|\frac{\hbar}{m}(p+\hbar k)|u(k)\rangle = \frac{\hbar}{m}\langle\psi(k)|p|\psi(k)\rangle = \hbar\langle v\rangle$$

  as desired.

- Next, we consider the effect of an external force F on the Bloch electron, e.g. by a linear potential. Semiclassically, assuming the electron has a well-defined position and momentum,

$$dE = Fvdt = F\frac{dE}{dk}dt, \quad F = \frac{dk}{dt}.$$

Hence an external force acts simply on the crystal momentum.

- More formally, take the Hamiltonian

$$H = \frac{p^2}{2m} + V(x) - Fx$$

and let the state at time $t = 0$ be a Bloch state $\psi(k_0, x)$. Then

$$\psi(x,t) = \exp\left(-\frac{i}{\hbar}\left[\frac{p^2}{2m} + V(x) - Fx\right]t\right)\psi(k_0, x)$$

Shifting $x \to x + a$, we have

$$\psi(x+a,t) = \exp\left(-\frac{i}{\hbar}\left[\frac{p^2}{2m} + V(x) - Fx\right]t\right)e^{iFta/\hbar}\psi(k_0, x+a)$$

which implies that

$$\psi(x+a,t) = e^{i(k_0+Ft/\hbar)a}\psi(x,t).$$

That is, the state remains a Bloch state, but the crystal momentum changes at rate $F/\hbar$, as desired. For sufficiently strong forces, there will also be interband transitions, which invalidate the semiclassical approximation.

- If there are no interband transitions, the semiclassical approximation yields

$$\frac{d\langle v\rangle}{dt} = \frac{d}{dt}\frac{dE}{dk} = \frac{d^2E}{dk^2}\frac{dk}{dt} = F\frac{d^2E}{dk^2}.$$

To get Newton's second law, $F = ma$, we need to use an effective mass

$$\frac{1}{m_*} = \frac{d^2E}{dk^2}.$$

The electron's mass has been 'renormalized' by the crystal lattice. For nearly free electrons, $m_* \approx m$ near the bottom of the band, but at the top of a band, m_* is negative, and in the middle, it can be infinite!

- In three dimensions, we have a mass tensor, with elements

$$(M_*^{-1})_{ij} = \frac{1}{\hbar^2}\frac{\partial^2 E}{\partial k_i \partial k_j}.$$

- To get another characterization of the effective mass, note that

$$\frac{dk}{dt} = m_*\frac{d\langle v\rangle}{dt}.$$

Integrating both sides and setting $\langle v\rangle = 0$ when $k = 0$,

$$k = m_*\langle v\rangle.$$

This is to be contrasted with $\langle p\rangle = m\langle v\rangle$. We see that effective mass is to crystal momentum as bare mass is to real momentum.

- In three dimensions, we have a mass tensor, with elements

$$(M_*^{-1})_{ij} = \frac{1}{\hbar^2}\frac{\partial^2 E}{\partial k_i \partial k_j}.$$

- To summarize, the crystal momentum and effective mass are "renormalized" quantities that include the effect of the crystal lattice. The "bare" momentum satisfies $F_{\text{tot}} = d\langle\hat{p}\rangle/dt$ where F_{tot} includes the external force and the effects of the lattice, while the crystal momentum accounts for the latter automatically.

3.2 Tight Binding

We now consider the tight binding model.

- Suppose the atomic spacing in a crystal was artificially increased, so that atomic orbitals barely overlapped. Then the electron states would look nothing like linear combinations of a few plane waves, as in the previous section. Instead, they would be approximately linear combinations of the atomic orbitals (LCAO), which themselves are approximately orthogonal.

- At any distance, this is not exactly true; what we are really doing is using a special trial wavefunction to approximate the energy variationally. That is, suppose we take a variational ansatz $|\psi\rangle = \psi_i|i\rangle \in \mathcal{H}'$ for the ground state. The variational ground state is the one that minimizes the energy $\langle\psi|H|\psi\rangle$, which means it obeys $H'|\psi\rangle = E|\psi\rangle$ where E is the lowest eigenvalue of H', the Hamiltonian H restricted to $\mathcal{H}'$.

- Continuing, our variational guess for the first excited state will be the state satisfying $H'|\psi\rangle = E|\psi\rangle$ where E is the second lowest eigenvalue of H', and so on. Hence to get predictions from a variational approach, we simply solve a Schrodinger equation as usual.

- In general, the states $|i\rangle$ will not be orthogonal, though this doesn't modify the validity of the variational approximation; it just makes the computations a bit more complicated. The real issue is that in general H will contain off-diagonal terms between states inside and outside $\mathcal{H}'$. The variational approach works when these terms are small. It may be systematically improved by adding more atomic orbitals to $\mathcal{H}'$.

- In a real solid, this approximation is good for the $1s$, $2s$, and $2p$ atomic orbitals. It also works for partially filled d orbitals in transition metals.

- Tight binding works well when interactions between atomic orbitals are weak; this implies the energies of the tight binding states are close to the atomic orbital energy. Hence to get a good result, we only need to include sets of orbitals that are nearly degenerate with each other, and can ignore others.

- For example, if we consider p orbitals, we have to consider the three degenerate p orbitals; after guessing a Bloch wavefunction we must diagonalize a 3×3 matrix for each value of $\mathbf{k}$. Similarly, for d orbitals we have a 5×5 matrix, though in practice we must also include the outer s orbital to get an accurate result. The resulting states are "hybridized".

Example. A toy example of tight binding in one dimension. Suppose we use one atomic orbital per atom, assumed orthogonal for simplicity, and we label the states as $|n\rangle$. Furthermore, suppose

$$\langle n|H|m\rangle = \epsilon\delta_{n,m} - t(\delta_{n+1,m} + \delta_{n-1,m}).$$

This is called the tight binding chain. By translational symmetry, the eigenfunctions are

$$|\psi\rangle = \psi_n|n\rangle, \quad \psi_n = \frac{e^{-ikna}}{\sqrt{N}}, \quad E(k) = \epsilon - 2t\cos(ka).$$

This gives a single, sinusoidal band whose width is determined by t. Even in this simple model, we find nontrivial physics. For instance, at the bottom of a band the electrons have the effective mass

$$m_* = \frac{1}{2ta^2}$$

In the case above, if the valence of the metal is one, then the band is half filled. Since the bottom half of the band lies below the energy ϵ, the overall energy of the electrons is lowered by the lattice; this is the principle behind metallic bonding.

Now suppose the metal has valence two. Then this single band is completely filled, and we naively have an insulator, because a filled band carries no current. However, we should also account for other bands due to other atomic orbitals. When t grows large enough, the bands will overlap and a Fermi surface will appear, allowing conduction.

Example. In a three-dimensional crystal using a single s-orbital, labeling the states by their positions in the lattice, splitting $H = H_0 + U$ where H_0 is the Hamiltonian for a single atom, and again assuming orthogonality for simplicity, we have

$$E(\mathbf{k}) = E_0 + \sum_{\mathbf{R}} \langle \mathbf{0}|U|\mathbf{R}\rangle\, e^{i\mathbf{k}\cdot\mathbf{R}}.$$

where the sum is over the Bravais lattice. Counting only nearest neighbor overlaps,

$$E(\mathbf{k}) = E + \sum_i \gamma_i \cos(\mathbf{k}_i \cdot \mathbf{a}_i)$$

where the $\mathbf{a}_i$ generate the lattice.

Example. A 2D square lattice with sp^3 hybridization. We consider the s, p_x, p_y, and p_z orbitals. We again assume orthogonality and count only nearest neighbor interactions for simplicity. The matrix elements of the variational Hamiltonian are hence given by the overlaps $\langle \mathbf{0}|U|\mathbf{R}\rangle$ as above.

By the same reasoning as before, we have

$$H_{ss} = E_s + V_{ss}(e^{ik_x a} + e^{-ik_x a} + e^{ik_y a} + e^{-ik_y a}) = E_s + 2V_{ss}(\cos k_x a + \cos k_y a).$$

Taking the z direction to be out of the lattice, by similar reasoning,

$$H_{zz} = E_p + 2V_{pp\pi}(\cos k_x a + \cos k_y a).$$

Here, the π in $V_{pp\pi}$ comes from the fact that the overlapping p_z orbitals form π bonds. Now, the p_x orbitals form σ bonds along the x axis and π bonds along the y axis, yielding

$$H_{xx} = E_p + 2V_{pp\sigma}\cos k_x a + 2V_{pp\pi}\cos k_y a.$$

The situation for H_{yy} is similar. Next, by symmetry, we have $H_{xy} = H_{xz} = H_{yz} = 0$, since positive and negative overlapping regions cancel out. Finally, we have

$$H_{sx} = V_{sp}(e^{ik_x a} - e^{-ik_x a}) = 2i\sin(k_x a)V_{sp}$$

where the sign flip is because the overlapping lobes on each side have opposite sign. The final matrix element H_{sy} is similar. We find the states by diagonalizing the Hamiltonian. In certain situations, the resulting solutions look like the standard sp^3 hybridized orbitals from chemistry, repeated throughout the lattice with phase factor $e^{i\mathbf{k}\cdot\mathbf{R}}$.

Next, we introduce Wannier functions.

- Tight binding assumes a trial wavefunction of the form

  $$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}\phi(\mathbf{r}-\mathbf{R})$$

  where ϕ is a linear combination of atomic orbitals. But in general, we have

  $$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}w(\mathbf{r}-\mathbf{R})$$

  exactly, where w is called a Wannier function. More generally, we would have a band index n, which we are suppressing for simplicity.

- To see this, note that since $\psi_{\mathbf{k}} = \psi_{\mathbf{k}+\mathbf{G}}$ for $\mathbf{G}$ in the reciprocal lattice, the inverse Fourier transform with respect to $\mathbf{k}$ has support only on the direct lattice $\mathbf{R}$, and we may define

  $$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} w_0(\mathbf{r},\mathbf{R})e^{i\mathbf{k}\cdot\mathbf{R}}.$$

  Now we have

  $$w_0(\mathbf{r},\mathbf{R}) = \frac{1}{N}\sum_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{R}}\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{N}\sum_{\mathbf{k}}\psi_{\mathbf{k}}(\mathbf{r}-\mathbf{R})$$

  where we applied Bloch's theorem. However, since the right-hand side is only a function of $\mathbf{r}-\mathbf{R}$, we can simply define the Wannier functions as $w(\mathbf{r}-\mathbf{R}) = w_0(\mathbf{r},\mathbf{R})$.

- The Wannier functions are orthogonal, by the orthogonality of the original states,

  $$\begin{aligned}\int d\mathbf{r}\, w_n^*(\mathbf{r}-\mathbf{R}_i)w_m(\mathbf{r}-\mathbf{R}_j) &\propto \int d\mathbf{r}d\mathbf{k}_1 d\mathbf{k}_2\, e^{-i\mathbf{k}_1\cdot\mathbf{R}_1+i\mathbf{k}_2\cdot\mathbf{R}_2}\psi^*_{n,\mathbf{k}_1}(\mathbf{r})\psi_{m,\mathbf{k}_2}(\mathbf{r}) \\ &\propto \int d\mathbf{k}_1 d\mathbf{k}_2\, e^{-i\mathbf{k}_1\cdot\mathbf{R}_1+i\mathbf{k}_2\cdot\mathbf{R}_2}\delta_{n,m}\delta_{\mathbf{k}_1,\mathbf{k}_2} \\ &\propto \delta_{n,m}\int d\mathbf{k}\, e^{-i\mathbf{k}\cdot(\mathbf{R}_i-\mathbf{R}_j)} \\ &\propto \delta_{n,m}\delta_{\mathbf{R}_i,\mathbf{R}_j}.\end{aligned}$$

- The Wannier functions are not unique, since we may redefine the Bloch states as

  $$\psi_{\mathbf{k}}(\mathbf{r}) \to e^{i\varphi(\mathbf{k})}\psi_{\mathbf{k}}(\mathbf{r}).$$

  This may be used to maximize their locality. In analogy with tight binding, we would like $w(\mathbf{r})$ to be sharply peaked about $\mathbf{r} = 0$.

- By completeness, we may always write Wannier functions in terms of a linear combination of the eigenfunctions of an atomic Hamiltonian, including all its atomic orbitals, plus a continuum of ionized states. The tight binding method simply states that we can get a good approximation by only using a few of the atomic orbitals.

3.3 Nearly Free Electrons

Next, we consider the opposite limit to tight binding.

- In some metals, the potential experienced by conduction electrons due to the lattice ions is weak. Physically, this is because the lattice charges are mostly shielded by bound electrons, and the conduction electrons cannot penetrate the shield by Pauli exclusion.
- Without the potential, the eigenstate are plane waves $|\mathbf{k}\rangle$. The matrix elements of the potential in this basis are simply the Fourier components of the potential,

  $$\langle \mathbf{k}'|V|\mathbf{k}\rangle = \frac{1}{L^3}\int d\mathbf{r}\, e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} V(\mathbf{r}) = V_{\mathbf{k}'-\mathbf{k}}.$$

  Then the matrix elements are zero unless $\mathbf{k}$ and $\mathbf{k}'$ differ by a reciprocal lattice vector; it is only these combinations of levels that are mixed by perturbation theory.
- At first order, the energy shift is

  $$\epsilon(\mathbf{k}) = \epsilon_0(\mathbf{k}) + \langle \mathbf{k}|V|\mathbf{k}\rangle = \epsilon_0(\mathbf{k}) + V_0$$

  which is an uninteresting constant shift that we ignore. At second order, we have

  $$\epsilon(\mathbf{k}) = \epsilon_0(\mathbf{k}) + \sum_{\mathbf{k}'=\mathbf{k}+\mathbf{G}} \frac{|\langle \mathbf{k}'|V|\mathbf{k}\rangle|^2}{\epsilon_0(\mathbf{k}) - \epsilon_0(\mathbf{k}')}.$$

  where the sum is over all nonzero reciprocal lattice vectors $\mathbf{G}$. In general, we see the energy shift is quadratic in V, but it blows up when levels are degenerate, in which case we must use degenerate perturbation theory. When this is done, it turns out that the energy shift is linear in V, so we will only consider the degenerate case, as it is the leading effect.
- Consider degenerate levels that differ by a reciprocal lattice vector $\mathbf{G}$. Since energy is quadratic,

  $$|\mathbf{k}| = |\mathbf{k}+\mathbf{G}|.$$

  This is the condition for Bragg scattering, so the momenta that are most affected are the ones that can undergo Bragg scattering, which makes physical sense.
- Geometrically, this implies $\mathbf{k}$ must lie on a Bragg plane, the set of points equidistant from $\mathbf{0}$ and $\mathbf{G}$. But Bragg planes separate Brillouin zones, so a weak periodic potential separates the original parabolic levels into bands, where the n^{th} band lies in the n^{th} Brillouin zone.

We now carry out the degenerate perturbation theory.

- Consider two nearly-degenerate levels, $|\mathbf{k}\rangle$ and $|\mathbf{k}'\rangle = |\mathbf{k}+\mathbf{G}\rangle$. In the subspace they span,

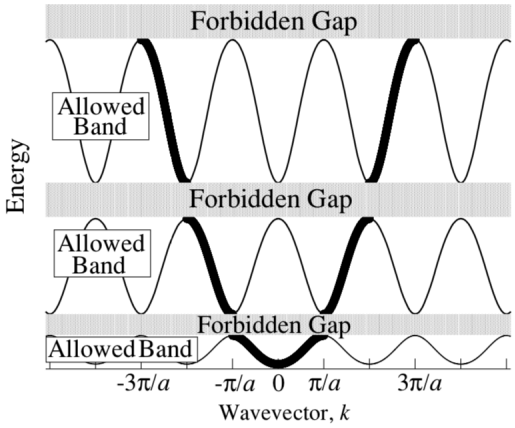
  $$H = \begin{pmatrix} \epsilon_0(\mathbf{k}) & V_{\mathbf{G}}^* \\ V_{\mathbf{G}} & \epsilon_0(\mathbf{k}+\mathbf{G}) \end{pmatrix}$$

  where we used the property $V_{-\mathbf{G}} = V_{\mathbf{G}}^*$, since V is real.

- The energies satisfy
$$(\epsilon_0(\mathbf{k}) - E)(\epsilon_0(\mathbf{k}+\mathbf{G}) - E) - |V_\mathbf{G}|^2 = 0.$$
If $\mathbf{k}$ lies precisely on a Bragg plane, we have
$$E_\pm = \epsilon_0(\mathbf{k}) \pm |V_\mathbf{G}|$$
so the shift is linear, as expected earlier.

- More generally, if we are near a Bragg plane, the energies are approximately $E_0 \pm \delta$, which gives
$$E_\pm = E_0 \pm \sqrt{|V_\mathbf{G}|^2 + \delta^2} \approx E_0 \pm |V_\mathbf{G}| \pm \frac{\delta^2}{2|V_\mathbf{G}|}$$
where $\delta \approx (k/m)\Delta k$. Then the dispersion relation breaks into two parabolas, opening upward and downward, separated by $2|V_\mathbf{G}|$. Moreover, at the bottom of the n^{th} band we have
$$m_* \approx \frac{m^2|V_\mathbf{G}|}{(n\pi/a)^2}$$
for $n \neq 0$.

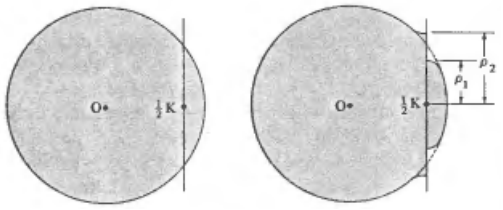
Example. Nearly free electrons in one dimension. The bands are created by breaking apart a parabola into segments, and rounding off the segments with more parabolas, as shown.



There are several ways to display the results. In the extended-zone scheme, we draw each band lying in its appropriate Brillouin zone. In the reduced-zone scheme, we translate everything to the first Brillouin zone. In the repeated-zone scheme used above, we use the periodicity to show each band for all k, though this is somewhat redundant.

Note that since the splittings depend on U_{nK}, it is possible for the splitting to be zero. For example, in the example of a cosine potential, which has only one nonzero Fourier component, only one band splitting occurs; the rest of the dispersion relation is continuous.

Example. The situation is more complicated in higher dimensions, since we can have more than two-fold degeneracy. For example, in a square lattice of side a, the four points $(\pm\pi/a, \pm\pi/a)$ are all degenerate; they are the 'corner' of the boundary between the first and second Brillouin zones. Generally, we may have high-symmetry points where n Bragg planes meet, giving $(n+1)$–fold degeneracy.

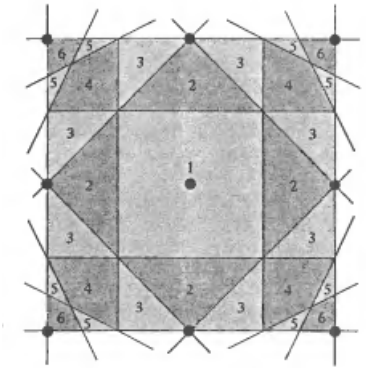


The characteristic splitting at a single Bragg plane is shown above. It distorts the Fermi surface by lowering the energy on one side and raising the energy on the other. Note that the Fermi surface is drawn perpendicular to the plane; this must occur if we have inversion symmetry about a plane parallel to that plane, as in the 1D case. This often occurs, but not always.

Note. The algorithm to construct the Fermi surface is as follows:

1. Draw the free electron Fermi sphere. Remember that the volume of this sphere is half of what one would naively think, because of the two spin states of the electron.
2. Deform it in the vicinity of every Bragg plane, as shown above.
3. Take the portion of the resulting surface lying in the n^{th} Brillouin zone and translate it back to the first Brillouin zone.

Example. The Brillouin zones for a two-dimensional square lattice are shown below.



For a solid of valence 2 with nearly free electrons, there will be branches of the Fermi surface for the first and second Brillouin zones.

Note. The reasoning above is not restricted to electrons. For example, it explains why band gaps appear in a photonic crystal. When a photon with a frequency in the gap strikes the material, it is perfectly reflected; alternatively the transmitted light experiences perfect destructive interference.

3.4 Phonons

In this section, we consider phonons, the quantum vibrations of lattice ions in a solid. For simplicity, we start in one dimension.

- Consider a set of ions with equilibrium positions $x_n^{\text{eq}} = na$. For small displacements

$$x_n = x_n^{\text{eq}} + u_n$$

we have the equation of motion

$$\frac{d^2u_n}{dt^2} = \frac{\kappa}{m}(u_{n+1} + u_{n-1} - 2u_n)$$

where κ is an empirical parameter, which requires a detailed understanding of atomic physics to calculate.

- By guessing a sinusoid $u_n = e^{i\omega t - ikna}$, we find

$$\omega = 2\sqrt{\frac{\kappa}{m}}\left|\sin\frac{ka}{2}\right|.$$

For small k, we find a linear dispersion relation $\omega = vk$ with speed of sound

$$v = a\sqrt{\frac{\kappa}{m}}.$$

One can also drive with higher ω, yielding a complex k indicating evanescent waves.

- For larger k, the solutions are periodic with period $k = 2\pi/a$, simply because two solutions with k related by a multiple of $2\pi/a$ are exactly the same. Hence we restrict to the Brillouin zone $k \in [-\pi/a, \pi/a]$. The group velocity vanishes at the edge of the Brillouin zone.

- Taking periodic boundary conditions for simplicity, the allowed values of k are

$$k = \frac{2\pi n}{L}$$

where $L = Na$ is the total length, so there are N modes, as expected.

- In the classical limit, these N modes each contribute $k_B T$ to the heat capacity. As a quantum system, the normal modes correspond to decoupled quantum harmonic oscillator whose excitations are photons. We may also take the continuum limit $N \to \infty$ and represent the lattice excitations with a quantum field, with

$$\mathcal{L} = \frac{\rho}{2}\dot{u}^2 - \frac{\lambda}{2}u'^2.$$

However, using this approach we get only the linear part of the dispersion relation.

- Given the dispersion relation, we can directly compute the heat capacity. The Debye model essentially assumes a linear dispersion relation, and hence is correct at low temperatures, while the Einstein model assumes a constant dispersion relation.

Example. Phonons in two dimensions. Consider a triangular lattice with separation a, with neighboring ions connected by springs of spring constant κ. As before, letting

$$\mathbf{u}(\mathbf{x}) = \mathbf{u}_0 e^{i\omega t - i\mathbf{k}\cdot\mathbf{x}}$$

we have the relation

$$\omega^2\mathbf{u}_0 = \frac{\kappa}{m}\sum_{i=1}^{3}\mathbf{a}_i(\mathbf{a}_i\cdot\mathbf{u}_0)(e^{i\mathbf{k}\cdot\mathbf{a}_i} + e^{-i\mathbf{k}\cdot\mathbf{a}_i} - 2).$$

Hence, to find the frequencies, we must diagonalize the matrix

$$D_{\alpha\beta}(\mathbf{k}) = \frac{\kappa}{m}\sum_{i=1}^{3}a_i^\alpha a_j^\beta \sin^2\frac{\mathbf{k}\cdot\mathbf{a}_i}{2}, \quad \omega^2\mathbf{u}_0 = D_{\alpha\beta}\mathbf{u}_0.$$

For each value of $\mathbf{k}$, there are two eigenvectors, corresponding to the transverse mode ($\mathbf{u}_0 \parallel \mathbf{k}$) and longitudinal modes ($\mathbf{u}_0 \perp \mathbf{k}$). The explicit frequencies are quite complicated.

In the continuum limit, the phonons are described by a field $u_i(\mathbf{x}, t)$ and the most general possible quadratic Lagrangian is

$$\mathcal{L} = \frac{1}{2}\left(\rho \dot{u}_i \dot{u}_i - 2\mu u_{ij} u_{ij} - \lambda u_{ii} u_{jj}\right), \quad u_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i)$$

where μ and λ are called Lame coefficients. In particular, u_{ii} cannot serve as a term since it is a total derivative, and the quadratic combinations of the antisymmetric part of $\partial_i u_j$ either vanish or are redundant. The transverse and longitudinal phonons both have linear dispersion relations with

$$v^2 = \begin{cases} (2\mu + \lambda)/\rho & \text{longitudinal,} \\ \mu/\rho & \text{transverse.} \end{cases}$$

The excitations of the field are described by a wavevector and polarization vector.

Example. A one-dimensional diatomic chain. We consider a unit cell of length a containing two masses m connected with springs κ_1 and κ_2. Let the displacements of these masses from equilibrium be x_n and y_n. Then

$$m\ddot{x_n} = \kappa_2(y_n - x_n) + \kappa_1(y_{n-1} - x_n), \quad m\ddot{y_n} = \kappa_1(x_{n+1} - y_n) + \kappa_2(x_n - y_n).$$

As usual, we guess an exponential,

$$x_n = A_x e^{i\omega t - ikna}, \quad y_n = A_y e^{i\omega t - ikna}$$

which yields the eigenvalue equation

$$m\omega^2 \begin{pmatrix} A_x \\ A_y \end{pmatrix} = \begin{pmatrix} \kappa_1 + \kappa_2 & -\kappa_2 - \kappa_1 e^{ika} \\ -\kappa_2 - \kappa_1 e^{-ika} & \kappa_1 + \kappa_2 \end{pmatrix} \begin{pmatrix} A_x \\ A_y \end{pmatrix}.$$

The solutions for ω in terms of k are

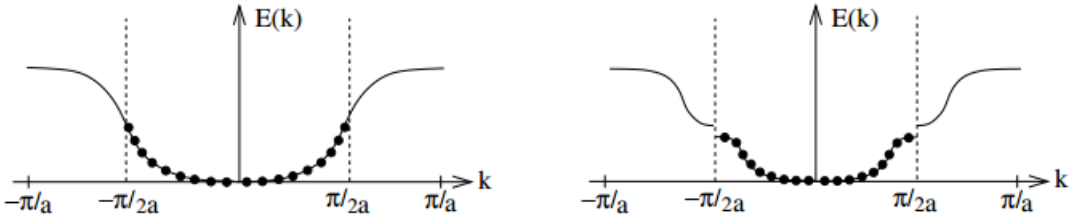
$$\omega_\pm = \sqrt{\frac{\kappa_1 + \kappa_2}{m} \pm \frac{1}{m}\sqrt{(\kappa_1 + \kappa_2)^2 - 4\kappa_1\kappa_2 \sin^2(ka/2)}}.$$

The dispersion relation now contains two branches. The lower-frequency branch contains modes where the two ions oscillate in phase; it is called the acoustic branch since it has a linear dispersion relation for small k. The higher-frequency branch contains modes where the ions oscillate against each other; it is called the optical branch because it may absorb light. (More correctly, photons can scatter inelastically by emitting acoustic branch phonons.) This is because light has the dispersion relation $\omega = ck$, which does not intersect the acoustic branch because $v_s \ll c$, but does intersect the optical branch because it has a gap. Another way of saying this is that the light can interact with the oscillating dipole moment formed within each unit cell.

In the limit $\kappa_1 \gg \kappa_2$, we can view the acoustic branch as a kind of molecular vibration (between the ions in each cell) whose spectrum is broadened by interactions between cells. In the limit $\kappa_1 \approx \kappa_2$, we can view the two branches as the result of a single branch (with unit cell size $a/2$) which is split by a perturbation of wavenumber $2\pi/a$.

Example. The Peierls instability. Consider a monatomic 1D lattice with one electron per lattice site, yielding a half-filled band. Now consider a distortion of the lattice, where atoms 'pair up' with each other by moving an amount δx. The total energy cost of this motion is proportional to $(\delta x)^2$.

Now consider the band structure. Since the period of the potential has been doubled from a to $2a$, the lowest band splits into two, lowering the energies of the states near the new band maximum.



To compute how much, we focus on the energy shifts of states near the band edge. Without the distortion, we have the linear dispersion relation

$$E_0(k) \approx \mu + \nu q, \quad q = k - \frac{\pi}{2a}$$

Now suppose the distortion creates a band gap of Δ. Then near the band edge we have

$$(E_0 - E)^2|_{k=\pi/2a+q} = \frac{\Delta^2}{4}$$

which breaks the linear dispersion relation into two branches. The occupied branch has

$$E(q) = \mu - \sqrt{\nu^2 q^2 + \frac{\Delta^2}{4}}.$$

The total shift in the energy of the electrons is

$$\Delta U = -2\frac{Na}{2\pi}\int dk\,(E_0(k) - E(k)) \propto \int_0^\Lambda dq\,\left(\sqrt{\nu^2 q^2 + \frac{\Delta^2}{4}} - \nu q\right)$$

where Λ is a cutoff with $\Delta \ll \nu\Lambda$. Integrating and then applying the approximation, we have

$$\Delta U \sim \Delta^2 \log(1/\Delta).$$

Supposing that $\delta x \propto \Delta$, the energy cost of the distortion is proportional to Δ^2. Then for sufficiently small Δ, the energy savings wins out, so the lattice is unstable! This experimental effect has been shown in one-dimensional polymer chains such as polyacetylene, which switch from conducting to insulating as the temperature drops below Δ, by the Peierls instability forming a band gap. Explicitly, the chains may be modeled using tight binding with alternating bond lengths; this is known as the Su–Shrieffer–Heeger (SSH) model.

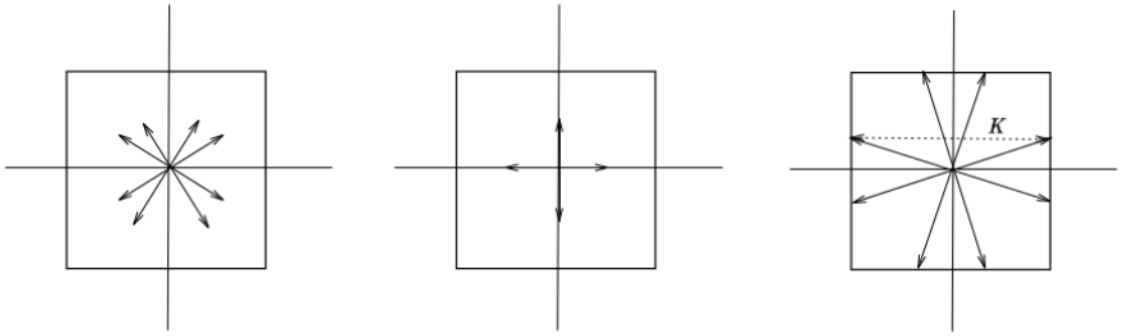
3.5 Band Degeneracy

In this section, we apply representation theory to find where Bloch bands touch.

- A crystal lattice has a space group of symmetries, realized on the Hilbert space in the usual way. For example, a rotation R is realized by $\hat{R}|\mathbf{k}\rangle = |R\mathbf{k}\rangle$.
- Consider a specific crystal momentum $\mathbf{k}$. The little group $G_\mathbf{k}$ is the set of symmetry operations that fixes $\mathbf{k}$, or sends $\mathbf{k}$ to $\mathbf{k} + \mathbf{G}$ for $\mathbf{G}$ in the reciprocal lattice.

- As a result, the states $|n, \mathbf{k}\rangle$ for all n form a representation of $G_{\mathbf{k}}$. If $G_{\mathbf{k}}$ is nonabelian, we can find degenerate Bloch bands at this point.
- In the case of a symmorphic space group, we can focus on the point group P. Define the star of $\mathbf{k}$ as the set of vectors $R_g\mathbf{k}$ for $g \in P$ (up to reciprocal lattice translation). Then the size of the little group is equal to $|P|$ divided by the size of the star.

Example. The square lattice. The point group is the set of symmetries of the square, i.e. the dihedral group D_4. The operations include the identity, 90 degree rotations, horizontal/vertical flips, and diagonal flips. Some of the stars are shown below.



arbitrary $\vec{k}$ to BZ interior: 8 $\vec{k}$ in star; (E) **symmetrical $\vec{k}$ to BZ interior: 4 $\vec{k}$ in star; (E, C_2)** **arbitrary $\vec{k}$ to BZ boundary: 4 $\vec{k}$ in star; (E, C_2)**

Generic $\mathbf{k}$ have trivial little group. A point with $k_x = 0$ has a four-element star and a two-element little group. A generic point on the Brillouin zone boundary has a four-element star (since pairs of the points shown are related by reciprocal lattice translations) and a two-element little group. As an extreme example, the corner of the Brillouin zone (and the center) both have one-element stars and little group D_4.

The characters of the irreps of D_4 are displayed below, where the multiplicities of the conjugacy classes are shown in parentheses.

	$1(1)$	$R_{\pi/2}(2)$	$R_\pi(1)$	$S_+(2)$	$S_\times(2)$
Γ_1	1	1	1	1	1
Γ_2	1	-1	1	-1	1
Γ_1'	1	1	1	-1	-1
Γ_2'	1	-1	1	1	-1
Γ_3	2	0	-2	0	0

Now, we apply our symmetry analysis to the point $\mathbf{k} = 0$. In the absence of a potential, we know that the four states $e^{\pm i(2\pi/a)x}$ and $e^{\pm i(2\pi/a)y}$ are all degenerate and have crystal momentum $\mathbf{k} = 0$, and hence they form a representation of D_4. Explicitly calculating, we find

	$1(1)$	$R_{\pi/2}(2)$	$R_\pi(1)$	$S_+(2)$	$S_\times(2)$
D	4	0	0	2	0

Taking the inner product, we find $n_1 = 1$, $n_{2'} = 1$, and $n_3 = 1$, so we have a two-dimensional irrep. Then the four-fold degeneracy turns into a two-fold degeneracy when the potential is turned on.

Example. Diamond. Looking at the diamond lattice, we see the point group is the symmetry group of a cube, O_h, with 48 elements. (The space group is not symmorphic, and we will return to

this point later.) The group factors as $O_h = O \times \mathbb{Z}_2$, where O only includes proper rotations, and the $\mathbb{Z}_2$ is generated by an inversion. Therefore, it suffices to look at the representations of O and $\mathbb{Z}_2$ individually.

The group O has five conjugacy classes:

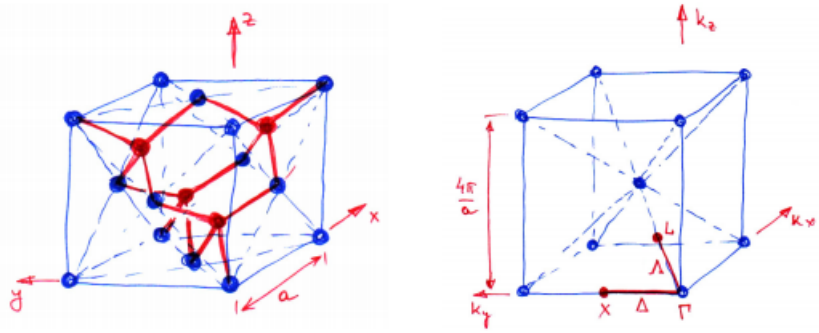
- 1, the identity element.
- 6_{90°, the six rotations by 90 degrees.
- 3_{180°, the three rotations by 180 degrees about a face of the cube.
- 6_{180°, the six rotations by 180 degrees about an edge center.
- 8_{120°, the eight rotations by 120 degrees about a cube diagonal.

We thus have to find five irreps. We know we have the trivial representation. We can guess another 1D representation, P, by coloring the vertices alternately white and black; then we assign -1 to the operations that swap the colors. We know there is a 3D representation from physical rotations of the cube, and it must be an irrep since no proper subspace is fixed. We also have the representation $3P$ by multiplying by P above. This leaves a single two-dimensional irrep which we find by orthogonality, giving the following character table.

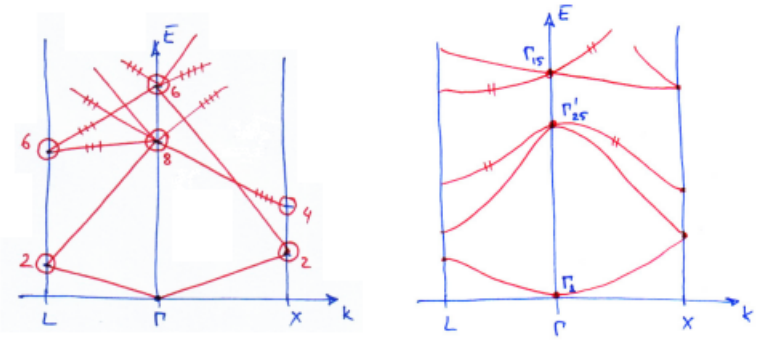
	1	6_{90°	3_{180°	6_{180°	8_{120°
1	1	1	1	1	1
P	1	-1	1	-1	1
2	2	0	2	0	-1
3	3	1	-1	-1	0
$3P$	3	-1	-1	1	0

We also know the character table of $\mathbb{Z}_2$, so the full character table of O_h is found by taking tensor products of pairs of irreps, giving ten irreps in total.

We now find degeneracies along points of high symmetry. The direct lattice and reciprocal lattice are shown below.



We are particularly interested in degeneracies along the segment $X - \Gamma - L$, where Γ is the origin. The free-electron and actual (numerically calculated) bandstructure are shown below. The free electron energy levels are highly degenerate; much of the degeneracy is broken by the potential, but a fair amount remains.



We first consider the origin Γ. Here, the little group is the entire point group O_h, and our table above tells us we can have two-fold or three-fold degeneracy. Numerically we find the lowest-lying irreps are 1D, 3D, and 3D, respectively. The glide-reflection doesn't matter here, because $\mathbf{k} = 0$, and hence there is no additional phase picked up from the translation. For other points, we must restrict to the little group, and account for this phase explicitly.

4 Applications of Band Structure

4.1 Electrical Conduction

Now we use band theory to understand conduction in real materials.

- We always assume that electrons in a periodic potential are in superpositions of Bloch states in a narrow window of crystal momentum $\mathbf{k}$. The semiclassical approximation holds when we may neglect the extent of this wavepacket.
- For this to hold in (crystal) momentum space, we need the total spatial extent of the wavepacket to be much greater than the lattice spacing. For this to hold in position space, the width of the wavepacket must be much less than the wavelength of any applied fields. If these conditions hold, we may speak of "the" position and crystal momentum of a Bloch electron, and hence its velocity (i.e. the group velocity), acceleration, and so on.
- The semiclassical approximation also neglects interband transitions, so Bloch electrons have a band index n. Such transitions are negligible if

  $$eEa \ll \frac{E_{\text{gap}}^2}{E_F}, \quad \hbar\omega_c \ll \frac{E_{\text{gap}}^2}{E_F}$$

  where ω_c is the cyclotron frequency. When they are violated, we get electric and magnetic breakdown, respectively. Additionally, we require $\hbar\omega \ll E_{\text{gap}}$ where ω is the frequency of the applied field. Otherwise, a single photon can cause an interband transition.
- Above, we have found that individual Bloch states have a nonzero velocity, yet they are stationary states. This implies that the conductivity of a perfect crystal is infinite.
- Using our previous results for Bloch electrons, we have the equations of motion

  $$\mathbf{v} = \frac{1}{\hbar}\frac{\partial E_n(\mathbf{k})}{\partial \mathbf{k}}, \quad \hbar\dot{\mathbf{k}} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}).$$

- In thermal equilibrium, the Bloch electron states have the usual Fermi distribution,

  $$f(E_n(\mathbf{k})) = \frac{1}{e^{(E_n(\mathbf{k})-\mu)/k_B T} + 1}.$$

  We ignore the spin, which couples to the magnetic field. Thermal effects are small, so we set $T = 0$ below. We drop the band index, since the occupancy of each band is fixed.

We now show filled bands do not contribute to conduction, leading to the hole picture.

- In such a band, the $(\mathbf{r}, \mathbf{k})$ phase space is filled uniformly with density $2/(2\pi)^3$. By Liouville's theorem, phase space density is preserved under time evolution. Therefore, external fields cannot change the distribution at all.
- The net electric current is

  $$\mathbf{j} = -2e \int đ\mathbf{k}\, \mathbf{v}.$$

  However, since the integral over a primitive cell of the gradient of a periodic function is zero, we have $\mathbf{j} = 0$, so a filled band carries no current. This confirms our earlier statement that only partially filled bands are important for conduction.

- Intuitively, the above results seem unreasonable, since we can shift all the momenta $\mathbf{k}$ with a uniform field. The resolution is that the velocity $\mathbf{v}(\mathbf{k})$ is periodic in $\mathbf{k}$, and it is the velocity distribution that remains the same.

- In one dimension, the above logic shows that $\int v(k)\,dk = 0$. If a single Bloch electron is pushed by a uniform DC electric field, then the net displacement after one cycle through the Brillouin zone is proportional to this integral, so the electron must oscillate back and forth! This is called a Bloch oscillation. Since real metals contain a substantial amount of impurities, such oscillations are hard to observe in practice, but have been seen in optical lattices.

- Note that this argument doesn't apply in more than one dimension, since the displacement is an integral of $\mathbf{v}(\mathbf{k})$ over a path in $\mathbf{k}$–space, not over the entire $\mathbf{k}$–space. Indeed, orbits of Bloch electrons in this context can be 'open', with nonzero net displacement, as we'll see.

- Also note that a filled band carries no heat current, where

$$\mathbf{j}_E = 2\int đ\mathbf{k}\, E\mathbf{v} \propto \int đ\mathbf{k}\frac{\partial(E^2)}{\partial \mathbf{k}}$$

  by a similar argument.

- Since a fully occupied band produces no current, the current may be calculated in two ways:

$$\mathbf{j} = (-2e)\int_{\text{occupied}} đ\mathbf{k}\,\mathbf{v}(\mathbf{k}) = 2e\int_{\text{unoccupied}} đ\mathbf{k}\,\mathbf{v}(\mathbf{k}).$$

  That is, we may regard the unoccupied states as the degrees of freedom; they are 'holes' with positive charge. In this picture, we regard the occupied states as simply unoccupied by holes!

- To correctly compute the current due to a band, we can use the electron picture or the hole picture, but not both, which would double count.

- Under an external field, unoccupied states behave exactly the same way as occupied states do. This is because a field simply induces a flow in the $(\mathbf{r}, \mathbf{k})$ phase space, which doesn't depend on the occupancy of the states.

- Therefore hole states evolve as if they have mass m_* and negative charge $-e$. But since the right-hand side of the equation $a = F/m$ only depends on the charge to mass ratio, holes can also be regarded as evolving with mass $-m_*$ and positive charge $+e$. This is more convenient, because in a nearly filled band, where the only unoccupied states are near the energy maximum, it ensures the holes have positive mass.

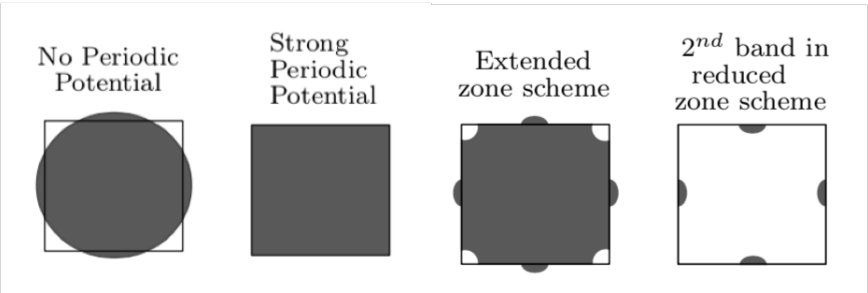
- Looking at the total energy, holes have negative energy $-E(\mathbf{k})$. Similarly, they have negative crystal momentum. Since $\mathbf{v} \propto \partial E/\partial \mathbf{k}$, holes have the same velocity as electrons.

- The hole picture resolves the mystery of the sign of the Hall coefficient mentioned in the beginning of the course: depending on the band structure, the relevant degrees of freedom can have positive charge! We have also resolved the issue of the number of valence electrons: electrons in filled bands can't conduct.

- Intuitively, we expect that electrons that are more tightly bound conduct less. This is true, because smaller hopping elements give a flatter band, and $\mathbf{v} \propto dE/d\mathbf{k}$. Typically the greatest contribution to conduction comes from nearly free electrons.

Now we turn to real examples.

- A solid is a conductor if it has partially filled bands, which means that it has Fermi surfaces. Upon application of an electric field, the Fermi surface deforms, producing a net current. In a band insulator, there is a gap between completely filled (valence) bands and unfilled (conduction) bands, called the band gap.
- We say a band gap is direct if the valence band maximum has the same crystal momentum as the conduction band minimum, and indirect otherwise.
- A semiconductor is a band insulator with a band gap less than about 2 eV, so a small amount of electrons can conduct due to thermal excitations. Notably, the conductivity of a semiconductor increases with temperature while the conductivity of a metal decreases.
- Consider an element such as sodium with one valence electron. Then there are enough electrons to fill precisely half a band, so we expect the solid to be a conductor.
- In more detail, we expect the $3s$ valence electron in sodium to be nearly free, because it is much larger than the lower orbitals, but those orbitals control the lattice spacing. Hence $3s$ electrons can't be treated by tight binding. However, in the nearly free electron picture, bands are always repelled apart by the potential, so we can be certain that only the lowest band plays a role, and it is precisely half occupied.
- There are some materials with valence one that are insulators, such as nickel oxide NiO. This occurs when there is a high energy cost for two electrons to occupy the same orbital; instead, every valence orbital has exactly one electron, and all electrons are locked into place. Such a system is called a Mott insulator. Seeing this effect requires accounting for electron electron interactions and the electron spin.
- We expect that electrons that fit the nearly free electron model most conduct the best. Indeed, the most conductive metals are silver, copper, and gold, each of which have a single valence s electron that is very nearly free. They perform better than sodium because the s orbital is even larger relative to the others, since the nuclear charge is very well shielded. In general, almost all of the most conductive metals have just one or two valence s electrons.
- In the case of first row transition metals, the $n = 3$ orbitals determine metallic bonding, and the $3d$ electrons contribute to conduction. In this case, the relevant orbitals are on the same order of the atomic spacing, and only nearest neighbor overlaps are significant; hence tight binding is a good model.
- Now we consider the case of valence two. If these electrons are approximately free, we get a picture like the one below, in two dimensions.



The electrons occupy two separate bands, forming a conductor. As the periodic potential's strength increases, these bands move apart, so that in the limit of a strong potential the electrons fully occupy the bottom band, forming an insulator. Usually we occupy an intermediate case shown at right, where the solid remains a conductor. Note that the Fermi surface of the second band consists of disconnected pieces, and we describe conduction in the first band with holes.

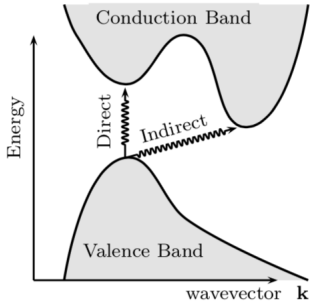
Note. Chemistry versus solid state. We describe sodium as having "one valence electron" which is free to populate bands, with the rest of the electrons "locked up in chemical bonds". But in reality, all the electrons fundamentally behave the same way, obeying Bloch's theorem. In sodium, with $2+8+1$ electrons, the $1s$, $2s$, and $2p$ electrons may be treated by the tight binding model; they fill extremely narrow bands far from the Fermi level and don't contribute to conduction. That is, being "locked up in bonds" is just the chemist's way of saying the electrons fill up a low energy band.

One might think there still is a difference, because we imagine the $1s$ electrons as localized to atoms rather than delocalized like Bloch waves. However, this is a meaningless statement, because the state of the $1s$ electrons must be described by a totally antisymmetric wavefunction; no individual electron has a well-defined state. This wavefunction is a Slater determinant of any combination of the original $1s$ states, including either the localized $1s$ atomic orbitals or the Bloch waves. For a nearly free electron, we would strongly prefer to use the Bloch waves because the orbitals are not nearly energy eigenstates. But for a tightly bound $1s$ electron there is no disadvantage to using the orbitals, and we do so because it's easier to picture.

Finally, we link band structure to optical properties.

- The color of a semiconductor depends on the band gap, as only photons with greater energy can be absorbed. Photons of visible light have energies ranging from 1.65 eV to 3.1 eV. Hence diamond, with a large band gap, is transparent, while GaAs, Si, and Ge, with small band gaps, are transparent. CdS has a band gap of approximately 2.6 eV, so violet and blue light are absorbed, making the material appear red.

- For the purposes of the discussion above, it matters whether the band gap is direct or indirect.



It is more difficult to perform an indirect transition because photons have very little momentum relative to their energy. Such processes require, e.g. a phonon to carry away some crystal momentum, and are suppressed by a few orders of magnitude.

- Metals are more complicated; here the photons excite the electrons, which can then quickly reemit them, leading to reflection. This makes metals look generically shiny. The "noble metals" gold, silver, and platinum look more shiny simply because their surfaces do not form insulating oxides when exposed to air.

Note. Diamond and silicon are very similar; both have valence four and are semiconductors, but diamond has a much larger band gap. In the tight binding picture, this is because the valence electrons for diamond are $2p$, while those for silicon are $3p$, and higher atomic energy levels are generally closer together. Alternatively, since the lattice ions for diamond are so much closer together, the periodic potential experiences by the valence electrons is much stronger.

4.2 Magnetic Fields

We now consider the motion of Bloch electrons in a uniform magnetic field.

- The equations of motion, from our previous results, are

$$\hbar\dot{\mathbf{k}} = -e\mathbf{v} \times \mathbf{B}, \quad \mathbf{v} = \frac{1}{\hbar}\frac{\partial E}{\partial \mathbf{k}}.$$

  Combining these equations, we immediately find that

$$\mathbf{k} \cdot \mathbf{B} = \text{const}, \quad E = \text{const}.$$

  Since Fermi surfaces are surfaces of constant energy, we see that the orbits are intersections of Fermi surfaces and planes perpendicular to the magnetic field.

- Next, note that

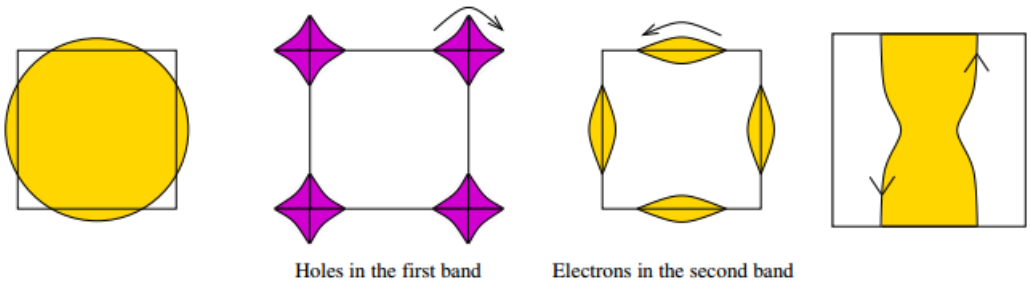
$$\mathbf{B} \times \hbar\dot{\mathbf{k}} = -e\mathbf{B} \times (\dot{\mathbf{r}} \times \mathbf{B}) = -eB^2\dot{\mathbf{r}}_\perp$$

  where $\mathbf{r}_\perp$ is the component perpendicular to the magnetic field. Integrating, we find

$$\mathbf{r}_\perp(t) = \mathbf{r}_\perp(0) - \frac{\hbar}{eB}\hat{\mathbf{B}} \times (\mathbf{k}(t) - \mathbf{k}(0)).$$

  Therefore, the shape of the orbit in real space looks identical to the shape of the orbit in momentum space (when both are projected perpendicular to the field), but rotated and scaled by $l_B^2 = \hbar/eB$.

- Depending on the dynamics, the orbits can be 'closed' or 'open', as shown below.



  At left, we see closed orbits in the nearly free electron model; at right, we see open orbits. Electrons with very low crystal momentum move approximately in circles, as we expect.

We now investigate the period of the motion.

- Next, note that the time required to go from $\mathbf{k}_1 = \mathbf{k}(t_1)$ to $\mathbf{k}_2 = \mathbf{k}(t_2)$ is

$$\Delta t = \int_{\mathbf{k}_1}^{\mathbf{k}_2} \frac{|d\mathbf{k}|}{|\dot{\mathbf{k}}|}$$

where

$$|\dot{\mathbf{k}}| = |\dot{\mathbf{k}_\perp}| = \frac{eB}{\hbar}|\dot{\mathbf{r}}_\perp| = \frac{eB}{\hbar^2}\left|\left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp\right|.$$

- Now consider a displaced orbit at energy $E + \Delta E$. The new path satisfies

$$\mathbf{k}' = \mathbf{k} + \left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp \Delta(\mathbf{k}), \quad \Delta E = \left(\frac{\partial E}{\partial \mathbf{k}}\right)_\perp \Delta(\mathbf{k}).$$

Substituting into the integral, we have

$$\Delta t = \frac{\hbar^2}{eB}\frac{1}{\Delta E}\int_{\mathbf{k}_1}^{\mathbf{k}_2} \Delta(\mathbf{k})\,|d\mathbf{k}|.$$

However, this is simply the area of the strips that separates the orbits in k–space!

- Therefore in general the period of an orbit is

$$T = \frac{\hbar^2}{eB}\frac{\partial A(E)}{\partial E}$$

where $A(E)$ is the area enclosed by the orbit. Here, we are imagining E changing, while $\mathbf{k}\cdot\mathbf{B}$ remains fixed. Geometrically, the Fermi surface grows while the plane perpendicular to the field remains invariant.

- A little calculation shows that A is proportional to a classical action variable, so finding the period here is just the same as the usual action-angle procedure. Experimentally, the above result is significant since it gives us a way to directly measure the Fermi surface.

Quantizing this system is already challenging with no lattice, which gives Landau levels. For our purposes, it suffices to use a semiclassical approximation.

- Bohr–Sommerfeld quantization states that quantum states satisfy

$$\frac{1}{2\pi}\int \mathbf{p}\cdot d\mathbf{r} = \hbar(n+\gamma), \quad n \in \mathbb{Z}$$

where γ is an arbitrary constant. This is exact in the semiclassical limit $n \to \infty$.

- Onsager quantization extends this to the case with a magnetic field, where

$$\frac{1}{2\pi}\int \mathbf{p}\cdot d\mathbf{r} = \frac{\hbar}{2\pi}\int \mathbf{k}\cdot d\mathbf{r} = \frac{\hbar^2}{2\pi eB}\int \mathbf{k}\cdot(d\mathbf{k}\times\hat{\mathbf{B}}).$$

But the integral is just the area of the orbits in phase space, so we have found that the orbit area is quantized as

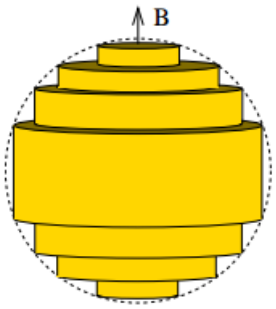
$$A_n = \frac{2\pi eB}{\hbar}(n+\gamma).$$

- In the case of no lattice, this reduces to our Landau level result. To see this, note that

$$\omega_c = \frac{2\pi eB}{\hbar^2}\frac{E_{n+1} - E_n}{A_{n+1} - A_n} = \frac{E_{n+1} - E_n}{\hbar}$$

  which shows that Landau levels are evenly spaced, by the cyclotron frequency.

- Moreover, if we perform the integral in real space, we instead find that the magnetic flux is quantized in units of $2\pi\hbar/e$, the quantum of flux. Hence Onsager quantization implies the states in momentum space form 'Landau tubes', shown below.



  For nearly free electrons, the Fermi surface is the intersection of a sphere with these tubes.

- Now note that when the radius of the Fermi surface is equal to the radius of one of the tubes, the density of states becomes very large. This enhancement occurs whenever

$$A_n = \frac{2\pi eB}{\hbar}(n + \gamma) = A_{\text{ext}}$$

  where A_{ext} is the area of the extremal Fermi surface. Therefore the oscillations happen periodically in $1/B$ with period

$$\Delta\left(\frac{1}{B}\right) = \frac{\hbar}{2\pi e}\frac{1}{A_{\text{ext}}}.$$

  This variation in the density of states affects nearly all observables, so we can measure the extremal area of the Fermi surface by varying the magnetic field.

- For example, the de Haas–van Alphen effect is a periodic variation in the magnetization, while the Shubnikov–de Haas effect is a periodic variation in the conductivity.

4.3 Semiconductor Devices

Next, we apply band structure to the design of semiconductor devices.

- A semiconductor can be doped by adding impurities with extra electrons, called n-dopants, or fewer electrons, called p-dopants. The result is an extrinsic semiconductor. Non-extrinsic semiconductors are called intrinsic.

- Naively, doping simply adds extra electrons or holes that can contribute to conduction. However, it's worth taking a closer look. Suppose we add a single P atom to an Si lattice, yielding one extra electron. There is now a new state where this electron is bound to the P atom.

- In hydrogen, the energy and length scales are the Rydberg and Bohr radius

$$\mathrm{Ry} = \frac{me^2}{8\epsilon_0^2 h^2} = 13.6\,\mathrm{eV}, \quad a_0 = \frac{4\pi\epsilon_0\hbar^2}{me^2} = 0.5 \times 10^{-10}\,\mathrm{m}.$$

  The same logic applies here, but we must replace m with the effective mass m_*, which is often slightly smaller, and ϵ_0 with $\epsilon_0\epsilon_r$. In a typical semiconductor, $\epsilon_r \approx 10$, so the binding energies are much smaller than the band gap. For many impurity atoms, we find an impurity band lying just below the bottom of the conduction band.

- Similarly, for p-doping, we add holes, and find an impurity band lying just above the top of the valence band. We will assume that the temperature is high enough to ignore the effects of these bands, restoring the naive picture; for low temperatures the impurities instead "freeze out". Without freeze out, the effect of the doping is simply to change the chemical potential.

- Now consider a conduction band starting at energy E_c. Then

$$g(E) = \frac{(2m_e^*)^{3/2}}{2\pi^2\hbar^3}\sqrt{E - E_c}$$

  and the number of electrons in the conduction band is

$$n(T) = \int_{E_c}^{\infty} \frac{g(E)\,dE}{e^{\beta(E-\mu)} + 1}.$$

  In the limit $\beta(E - \mu) \ll 1$ where occupancies are low, the Fermi-Dirac distribution reduces to the Maxwell distribution, so

$$n(T) \approx \int_{E_c}^{\infty} dE\, g(E) e^{-\beta(E-\mu)} = \frac{1}{4}\left(\frac{2m_e^* k_B T}{\pi\hbar^2}\right)^{3/2} e^{-\beta(E_c-\mu)}.$$

  The most important factor here is simply the intuitive exponential dependence.

- Similarly, suppose the valence band has maximum energy E_v. Then

$$p(T) \approx \frac{1}{4}\left(\frac{2m_h^* k_B T}{\pi\hbar^2}\right)^{3/2} e^{-\beta(\mu-E_v)}.$$

  In particular, multiplying these equations gives the law of mass action

$$n(T)p(T) = \frac{1}{2}\left(\frac{k_B T}{\pi\hbar^2}\right)^3 (m_e^* m_h^*)^{3/2} e^{-\beta E_{\mathrm{gap}}}$$

  where E_{gap} is the band gap. The name is from chemistry, because this equation can also be derived by considering equilibrium for the chemical reaction $e + \text{hole} \leftrightarrow \text{nothing}$.

- For an intrinsic semiconductor, $n(T) = p(T)$, which yields the equation

$$1 = \left(\frac{m_e^*}{m_h^*}\right)^{3/2} e^{-\beta(E_v + E_c - 2\mu)}$$

  from which we may solve for μ. The law of mass action gives

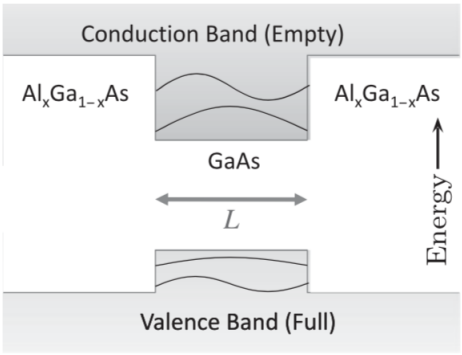
$$n(T) = p(T) = \frac{1}{\sqrt{2}}\left(\frac{k_B T}{\pi\hbar^2}\right)^{3/2} (m_e^* m_h^*)^{3/4} e^{-\beta E_{\mathrm{gap}}/2}.$$

- For an extrinsic semiconductor, the situation is more complicated, but for strong n/p-doping $n(T)/p(T)$ is set by the doping, while the other is determined by the law of mass action.

Next, we turn to band structure engineering.

- Two common semiconductor materials are GaAs and AlAs, with direct band gaps at $\mathbf{k} = 0$ of 1.4 eV and 2.7 eV respectively. One can also use a mixture of Ga and Al in $Al_xGa_{1-x}As$, whose band gap interpolates between the two, allowing the band gap to be set, e.g. to produce a laser with a given color.

- Next, we can consider a semiconductor heterostructure, where the composition varies from place to place. Then the band energies also vary. From the point of view of a single conduction electron, this is equivalent to a potential equal to the bottom of the local conduction band.



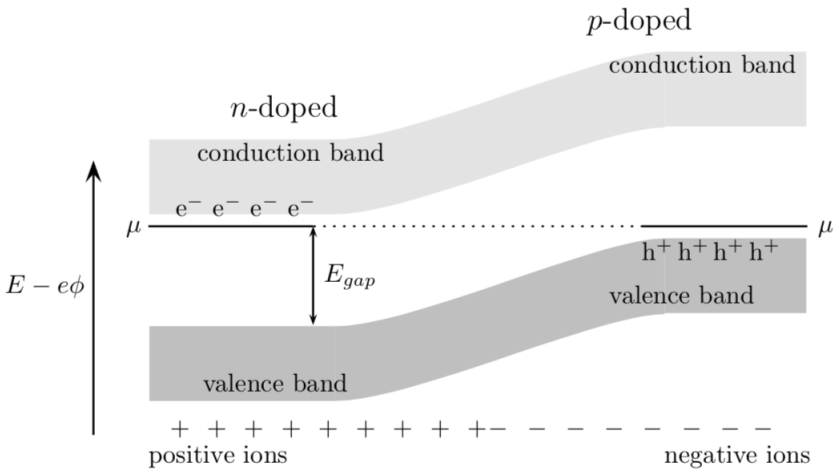
  For example, a thin layer of GaAs within AlGaAs creates a "quantum well", within which electrons can move freely in two dimensions. These materials are commonly used to build heterostructures because their lattice constants are very close.

- A useful trick is "modulation doping", where the n-dopants are placed outside of the quantum well. When the extra electrons then drop into the quantum well, they see no impurities, so the semiconductor device has low dissipation.

Now we consider the p-n junction, where a p-doped and n-doped semiconductor are placed together.

- For simplicity, suppose the semiconductors are made from the same material. The chemical potential for an n-doped semiconductor is at the bottom of the conduction band, while it is at the top of the valence band for a p-doped semiconductor. Hence electrons flow from the n-doped semiconductor to the p-doped semiconductor to annihilate with holes.

- As a result, the charges of the lattice ions on both sides will not be balanced; this creates an electric field much like a parallel plate capacitor.



In equilibrium, the potential energy $-e\phi$ balances the difference in chemical potential, yielding a finite-sized depletion region.

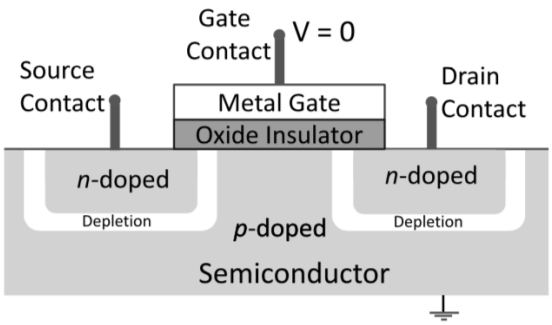
- If one applies light to a semiconductor, electron-hole pairs can be created, but they typically quickly reannihilate. However, if a pair if created in the depletion region, then the electric field separates them, producing a net current flow. This is the principle behind solar cells.
- A p-n junction functions as a diode, which allows current to flow preferentially from p to n. To understand this, we consider the current in equilibrium. There is a rightward current due to thermal excitation of electrons in the p-doped region and holes in the n-doped region. They require no energy to cross the depletion region, so the current is proportional to e^{-E_{gap}/k_BT}.
- This current is balanced by a leftward current from the existing electrons in the n-doped region and holes in the p-doped region. They must cross a potential barrier of height E_{gap} created by the electric field in the depletion region, so the current is proportional to e^{-E_{gap}/k_BT}.
- Hence in equilibrium, the two currents balance. The point is that the leftward current can be dramatically increased by an applied potential, which converts it to $e^{-(E_{\text{gap}}+eV)/k_BT}$, while the rightward current can't. Hence we have

$$I = J_s(T)(e^{-eV/k_BT} - 1), \quad J_s \propto e^{-E_{\text{gap}}/k_BT}$$

where the current rises exponentially in one direction and approaches a constant in the other.

Possibly the most important application is the transistor.

- The most common type of transistor is the Metal-Oxide Semiconductor Field Effect Transistor, or MOSFET. An n-MOSFET consists of two back-to-back p-n junctions, with contacts on both n-doped regions and a contact on the intermediate p-doped region separated by an oxide insulator.
- By default, little current can flow through the transistor. However, when a positive voltage is applied to the p-doped region, it induces a negative charge on the other end of the oxide layer.



This converts the p-doped region between the source and gate into an n-doped region, allowing current to flow. Hence a transistor allows a large current to be controlled by a small input, allowing amplification and digital logic.

- In terms of band structure, the applied voltage "bends" the band structure diagram, just as we saw for a diode. Hence the current flow is exponentially sensitive to the applied voltage.
- One can also build p-MOSFETs which allow current to flow with a sufficiently negative voltage. Typically circuit designs use a balance of the two; this is called complementary MOS (CMOS).

5 Magnetism

5.1 Paramagnetism and Diamagnetism

We begin with a qualitative overview of paramagnetism and diamagnetism.

- Paramagnetism and diamagnetism can be largely understood ignoring interactions between electrons. Generally only electrons matter for magnetism in solids, because spins have magnetic moment $\mu = e\hbar/2m$, and the nuclei are much more massive.

- Thinking classically, paramagnetism arises from magnetic dipoles begin rotated to align with the external field, while diamagnetism comes from Lenz's law, with dipole moments changing to oppose external fields. However, these classical arguments are simply incorrect; neither paramagnetism nor diamagnetism exist outside of quantum mechanics!

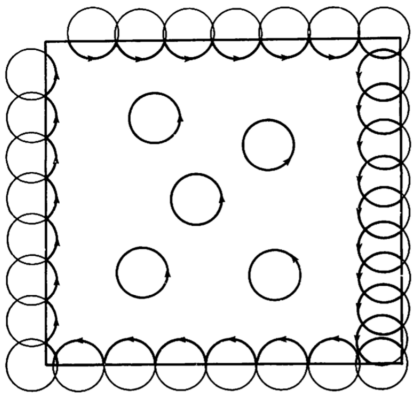
- This result is called the Bohr-van Leeuwen theorem. To prove it, note that in statistical mechanics we have $\boldsymbol{\mu} = -\partial F/\partial \mathbf{B}$ where F is defined in terms of

$$Z = \int d\mathbf{r} d\mathbf{p}\, e^{-\beta H(\mathbf{r},\mathbf{p})}, \quad H = \frac{(\mathbf{p} - e\mathbf{A})^2}{2m} + V(\mathbf{r}).$$

  The integral is independent of $\mathbf{A}$ by shifting the integration variable, and hence F is independent of $\mathbf{B}$. Hence in thermal equilibrium we must have $\boldsymbol{\mu} = 0$.

- Now one might wonder where the classical argument went wrong. In the case of paramagnetism, the issue is that there are no classical permanent magnetic dipoles; work must be done to keep the currents flowing. Accounting for this work, the energy of interaction between a magnetic dipole and external field is $U = 0$ rather than $U = -\boldsymbol{\mu} \cdot \mathbf{B}$, as magnetic fields do no work.

- In the case of classical diamagnetism, there are indeed induced current loops.

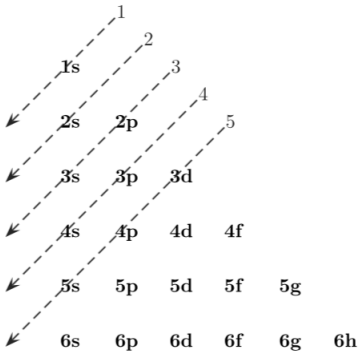


  However, if we take any subset of a larger conductor, the effects of the current loops inside the bulk are canceled by the partial current loops which straddle the edge because their area is much larger, giving $\boldsymbol{\mu} = 0$. This occurs no matter what the boundary conditions are. For instance, for hard wall boundary conditions, some electrons would continually bounce off the wall, traversing a large clockwise current loop which cancels the electrons in the bulk.

- In both cases, quantum mechanics allows magnetism to exist because it provides permanent magnetic dipoles, associated with electrons bound to atoms.

Next, we turn to some basic atomic physics.

- The aufbau principle states that orbitals are filled starting with the lowest available energy state. An entire shell is filled before another shell is started. The partially filled shell is called the valence shell, and typically determines bonding properties.
- Madelung's rule states that the shells are ordered in energy as shown below.



  This is not intuitive, given the results for hydrogen, because it depends on how the electrons all interact with each other; Madelung's rule should be regarded as an empirical result. There are many exceptions; for example, copper fills its $3d$ shell by taking an electron from the $4s$ shell.

- Hund's rules deal with the spins of the electrons. Roughly speaking, they are:

  1. Electrons try to align their spins.
  2. Given this, electrons try to maximize their total orbital angular momentum.
  3. Given both of these, the orbital and spin angular momentum align when the shell is more than half filled, and antialign when the shell is less than half filled, so that $J = |L \pm S|$.

  As an example, for filling $2p$ orbitals, the order is

$$|\ell = 1, \downarrow\rangle, \quad |\ell = 0, \downarrow\rangle, \quad |\ell = -1, \downarrow\rangle, \quad |\ell = -1, \uparrow\rangle, \quad |\ell = 0, \uparrow\rangle, \quad |\ell = 1, \uparrow\rangle.$$

- Hund's third rule results from the spin orbit coupling $\mathbf{L} \cdot \mathbf{S}$, one of the components of atomic fine structure. One might think there are separate behaviors for shells below and above half filling, but the rule is that it is always energetically favorable for $\mathbf{L}$ and $\mathbf{S}$ to antialign. In the above example, once the shell is half-filled, we have $L = 0$, and the next electron must be spin up by Pauli exclusion. It takes negative ℓ to antialign with its own spin, but this means it is aligned with the net spin. That doesn't cost energy, because the half-shell has $L = 0$.
- Now we justify Hund's first rule. Write the wavefunction as $\Psi = \psi_{\text{pos}}\psi_{\text{spin}}$. A symmetric spin wavefunction yields an antisymmetric spatial wave function, so the electrons are further away from each other. More quantitatively, we have

$$E_{\text{anti}} = (\langle AB| - \langle BA|)V(|AB\rangle - |BA\rangle), \quad E_{\text{sym}} = (\langle AB| + \langle BA|)V(|AB\rangle + |BA\rangle)$$

  and the difference in energies is $-4\,\text{Re}\langle AB|V|BA\rangle < 0$ where V is the Coulomb potential between the electrons; this is called an exchange interaction because it comes from the amplitude for exchange. It yields an effective spin-spin interaction $\mathbf{S}_1 \cdot \mathbf{S}_2$.

- One must be careful applying Hund's rules for molecules. For example, spins are typically antialigned in a covalent bond, because the energy of a bonding orbital is much less that of an antibonding orbital. The interaction effects above win out for atoms, because orbitals of the same n are almost exactly degenerate when interactions are ignored. Hund's rules are also unreliable for, e.g. 3d electrons in a crystal, due to crystal field effects.

- Next we consider electrons in atoms in an external field. We have

$$H = \frac{(\mathbf{p} + e\mathbf{A})^2}{2m} + g\mu_B \mathbf{B} \cdot \boldsymbol{\sigma} + V(\mathbf{r})$$

  where $\boldsymbol{\sigma}$ is the electron spin, and $e > 0$. In a uniform magnetic field, we may take $\mathbf{A} = (\mathbf{B} \times \mathbf{r})/2$, and expanding gives

$$H = H_0 + \frac{e}{2m} \epsilon_{ijk} B_i r_j p_k + \frac{e^2}{2m} \frac{1}{4} |\mathbf{B} \times \mathbf{r}|^2 + g\mu_B \mathbf{B} \cdot \boldsymbol{\sigma}$$

  where H_0 is the Hamiltonian for no external field, and we used $\epsilon_{ijk} r_j p_k = \epsilon_{ijk} p_k r_j$.

- The second term can be rewritten as $(e/2m)\mathbf{B} \cdot (\mathbf{r} \times \mathbf{p})$, so that

$$H = H_0 + \mu_B \mathbf{B} \cdot (\mathbf{l} + g\boldsymbol{\sigma}) + \frac{e^2}{2m} \frac{1}{4} (|\mathbf{B} \times \mathbf{r}|)^2.$$

  The second term here is called the paramagnetic term, because it makes the magnetic moments align with the external field, and lowers the energy, causing an attraction towards the field. The sign is flipped from what would be intuitive, because the magnetic moments are aligned opposite to $\mathbf{l}$ and $\boldsymbol{\sigma}$, due to the choice $e > 0$.

- The final term is called the diamagnetic term, because it causes an increase in the energy of the atom when a magnetic field is applied, causing a repulsion away from the field.

Now we consider the basic mechanisms of paramagnetism and diamagnetism.

- Generally the paramagnetic term dominates, leading to Curie/Langevin paramagnetism. In the special case of a spin 1/2 particle with no orbital angular momentum,

$$Z = e^{-\beta\mu_B B} + e^{\beta\mu_B B}, \quad \mu = -\frac{\partial F}{\partial B} = \mu_B \tanh(\beta\mu_B B).$$

  Then the susceptibility is

$$\chi = \left.\frac{\partial M}{\partial H}\right|_{H=0} = \frac{n\mu_0\mu_B^2}{k_B T}$$

  for a density n of spins. The fact that $\chi \propto 1/T$ is called Curie's law.

- Now for a general atom, summing over the electrons gives

$$H = \mu_B \mathbf{B} \cdot (\mathbf{L} + g\mathbf{S}).$$

  It turns out, after some calculation, that

$$H = \tilde{g}\mu_B \mathbf{B} \cdot \mathbf{J}, \quad \tilde{g} = \frac{1}{2}(g+1) + \frac{1}{2}(g-1)\frac{S(S+1) - L(L+1)}{J(J+1)}$$

  where $\tilde{g}$ is called the Lande g-factor. This result is the weak-field limit of the Zeeman effect. The reason that H had to be directly proportional to $\mathbf{J}$ is because we are dealing with a single $\mathfrak{su}(2)$ irrep, so all vector operators must be proportional to $\mathbf{J}$.

- Hence by the same idea as before, we find

$$Z = \sum_{J_z=-J}^{J} e^{-\beta\tilde{g}\mu_B B J_z}, \quad \chi = \frac{n\mu_0(\tilde{g}\mu_B)^2}{3}\frac{J(J+1)}{k_B T}.$$

  In general, the Curie paramagnetism dominates all other effects, unless it vanishes.

- The dependence of μ on B for general spin is rather complicated, but it is simple in the limit $J \to \infty$, where we recover a "classical" vector-valued spin,

$$\langle\mu\rangle = \frac{\int_0^\pi \mu_B \cos\theta\, e^{\mu B\cos\theta/k_B T}\sin\theta\, d\theta}{\int_0^\pi e^{\mu B\cos\theta/k_B T}\sin\theta\, d\theta} = \mu_B \frac{\int_{-1}^1 x e^{yx}\, dx}{\int_{-1}^1 e^{yx}\, dx}$$

  where we substituted $x = \cos\theta$ and defined $y = \mu_B B/k_B T$. This gives

$$\langle\mu\rangle = \mu_B L(y), \quad L(y) = \coth y - \frac{1}{y}$$

  which is called the Langevin function. For general J, we instead get the Brillouin function,

$$\langle\mu\rangle = \mu_B B_J(y), \quad B_J(y) = \frac{2J+1}{2J}\coth\left(\frac{2J+1}{2J}y\right) - \frac{1}{2J}\coth\frac{y}{2J}, \quad y = \tilde{g}J\frac{\mu_B B}{k_B T}.$$

- Larmor diamagnetism can be measured when $J = 0$, and the Curie paramagnetism vanishes; this occurs when the valence shell is filled or one electron away from half-filled. Then we have

$$\delta H = \frac{e^2 B^2}{8m}\langle x^2 + y^2\rangle = \frac{e^2 B^2}{12m}\langle r^2\rangle$$

  where we used spherical symmetry, since $J = 0$. Hence we have

$$\mu = -\frac{dE}{dB} = -\left(\frac{e^2}{6m}\langle r^2\rangle\right)B, \quad \chi = -\frac{ne^2\mu_0\langle r^2\rangle}{6m}.$$

- There is also van Vleck paramagnetism, which results from the paramagnetic term in second-order perturbation theory; it is only visible when $J = 0$.

- Relatively large and anisotropic diamagnetic susceptibilities are also observed in molecules with delocalized π electrons, such as graphite, as these electrons cover a large circular area.

So far we've only considered isolated atoms; now we turn to solids.

- We have already seen Pauli paramagnetism in Sommerfeld theory, where

$$\chi = \mu_0\mu_B^2 g(E_F).$$

  As we saw earlier, this is smaller than the result for Curie paramagnetism by a factor of $k_B T/E_F$. As a result, it does not diverges when $T \to 0$ but instead approaches a constant. The susceptibility is limited by the fact that almost all the electrons are buried in the Fermi sea.

- Larmor diamagnetism also applies to core electrons, giving

$$\chi = -\frac{Zne^2\mu_0\langle r^2\rangle}{6m}.$$

  However, for conduction electrons, we must account for the Landau level structure, leading to the Landau diamagnetism

$$\chi = -\frac{1}{3}\mu_0\mu_B^2 g(E_F)$$

  In solids with many core electrons, Larmor diamagnetism can win out against Pauli paramagnetism, which only affects the conduction electrons.

- Curie paramagnetism can also occur in solids for "free" spins. This occurs when we have a Mott insulator, where Coulomb repulsion effects prevent valence electrons from hopping between lattice sites. Then these electrons behave as if they are on isolated atoms, in the sense that the electrons on each atom are free to flip their spins independently, yielding Curie paramagnetism.

- This picture is corrected by the effects of other atoms in the lattice, leading to a "crystal field" splitting, which breaks the degeneracy of levels with the same n but different ℓ. In particular, Hund's rules may break down. These effects are most important for transition metals, where the $3d$ electrons are only shielded by the $4s$ electrons; in contrast, for the rare earths the $4f$ shell is shielded by $6s$ and $5p$.

5.2 Ferromagnetism

Now we consider a basic model for ferromagnetism.

- Ferromagnetism is even more difficult to understand classically, because in a general situation $\mathbf{M}$ is not parallel to $\mathbf{B}$, so we cannot think of spins being aligned by fields from distant spins. Nor can they be aligned by adjacent spins; the classical dipole-dipole interaction is far too weak and doesn't favor aligning the moments for all relative orientations.

- Instead, the spins are aligned by the exchange interaction, which, as we have seen, arises from Pauli exclusion and electrostatic repulsion. This is responsible for their strength, since generically magnetic forces are smaller than electric ones by $(v/c)^2$, and the electrons in many solids move nonrelativistically.

- We consider a toy model that considers only the spin degrees of freedom in an insulator and the exchange interaction,

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} J_{ij}\mathbf{S}_i\cdot\mathbf{S}_j + \sum_i g\mu_B\mathbf{B}\cdot\mathbf{S}_i$$

  where the sum is over neighboring i and j, and every pair is counted twice. The J_{ij} are also defined with an extra factor of 2 in other sources. While the simple atomic exchange interaction has $J_{ij} > 0$, it may also be negative, due to the other effects we've seen.

- For simplicity we consider zero external field and all J_{ij} equal,

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} J\mathbf{S}_i\cdot\mathbf{S}_j$$

yielding the Heisenberg model. The dynamics of the model can be found quantum mechanically using the Heisenberg equation of motion.

- The $\mathbf{S}_i$ above are properly quantum spin operators, but we can also take the limit of infinite spin; in this semiclassical limit they may be treated like classical vectors.
- In a ferromagnet, $J > 0$ and neighboring spins align. In an antiferromagnet, $J < 0$ and there is zero net magnetization, but there is still magnetic order; such a state is called a Neel state. Classically, a Neel state simply has spins pointing in alternating directions, but in quantum mechanics the situation is more subtle. For example, for spin 1/2, the state $|\uparrow\downarrow\uparrow\downarrow \ldots\rangle$ is not an eigenstate.
- Antiferromagnetism can be established by measuring the susceptibility as a function of temperature. It can also be detected directly using neutron diffraction; fixing the neutron spin, the scattering is spin-dependent, so the neutrons see a unit cell of doubled size, producing new diffraction peaks.
- A frustrated antiferromagnet is an antiferromagnet on a lattice where the ground state cannot "satisfy" the interaction for all pairs of spins; one example is a triangular lattice. Classically, for a single triangle the ground state has the spins pointing 120° apart, and a full triangular lattice has this pattern repeated.
- The elementary excitations in a ferromagnet are spin waves, which in the classical limit are of the form $\delta\mathbf{S} = \mathbf{A}e^{i\omega t - i\mathbf{k}\cdot\mathbf{r}}$ and have a quadratic dispersion relation. They are typically detected by inelastic neutron scattering, and become magnons upon quantization.
- A ferrimagnet has a unit cell with more than one variety of atom, and antiferromagnetic ordering. Since the moments have different magnitudes, the material has a net magnetization. Most common ferromagnets, such as magnetite (Fe_3O_4) are ferrimagnetic.
- The Heisenberg model has rotational symmetry, but it generically is broken by the lattice. For example, lattice effects may yield

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} J\mathbf{S}_i \cdot \mathbf{S}_j - \kappa \sum_i (S_i^z)^2.$$

  Classically, this yields magnetization along the $\pm\hat{\mathbf{z}}$ axis. One has to be careful when treating the spins quantum mechanically; for instance for spin 1/2 we have $S_z^2 = 1/4$, so the rotational symmetry is not broken at all. In the limit of large κ, the spin is forced to point maximally up or down; this reduces the Heisenberg model to the 3D Ising model.

Note. Classical spin waves in more detail. To avoid confusion we denote the classical magnetization field by $\mathbf{m}$. The Hamiltonian is

$$H = \int d\mathbf{r}\, \frac{1}{2}(\nabla\mathbf{m})^2 - \mathbf{h}\cdot\mathbf{m}, \quad (\nabla\mathbf{m})^2 = \partial_\alpha m_i \partial_\alpha m_i.$$

The corresponding equations of motion are

$$\frac{d\mathbf{m}}{dt} = \mathbf{m}\times\mathbf{h}_{\text{eff}}, \quad \mathbf{h}_{\text{eff}} = -\frac{\partial E}{\partial \mathbf{m}} = -\nabla^2\mathbf{m} + \mathbf{h}.$$

To investigate excitations, we linearize about the ground state $\mathbf{m}(x,t) = \mathbf{m}_0 + \delta\mathbf{m}(x,t)$ where $\mathbf{m}_0$ is parallel to $\mathbf{h}$, giving the equation of motion

$$\frac{d(\delta\mathbf{m})}{dt} = \delta\mathbf{m} \times \mathbf{h} - \mathbf{m}_0 \times \nabla^2\mathbf{m}.$$

This equation has plane wave solutions with dispersion relation

$$\omega(\mathbf{k}) = |\mathbf{h}| + \mathbf{k}^2.$$

When $\mathbf{h}$ is zero, we get a Nambu-Goldstone mode as expected.

Next, we consider domain walls.

- A (Weiss) domain is a region in a ferromagnet where the spins all point in the same direction. The formation of multiple domains is energetically favorable, because having all the spins point in the same direction yields a macroscopic magnetic field with substantial magnetic field energy. This energy isn't in the Heisenberg model, because the magnetic dipole-dipole force is tiny for individual spins, but it scales up faster than the exchange force penalty from domain walls.
- Generally, domain walls can occur whenever an order parameter space is disconnected. For example, if we define the order parameter for an antiferromagnet to be the magnetization of the even sites, then this parameter can flip sign at a domain wall. For antiferromagnets the formation of domain walls is not energetically favorable, but they can still be created as defects.
- The energetics of domain walls is rather complicated, depending on the size and detailed geometry of the sample, the strength of the interactions, and any disorder in the sample; for example, for a polycrystalline sample the domains may correspond to crystallites. These are small enough that they must contain a single domain, but multiple crystallites need not form a domain since there is no domain wall cost.
- In the Ising model, spins suddenly flip over when crossing domains. But more realistically, in the Heisenberg model, domain walls have nonzero width within which the spin rotates. They may be classified as either Bloch or Neel walls depending on the way the spin rotates. The size of a domain wall is determined by the ratio κ/J.
- In more detail, for a domain wall of length N, the energy per area is

  $$\delta E \sim NJS^2\frac{\pi^2}{2N^2} + N\kappa S^2$$

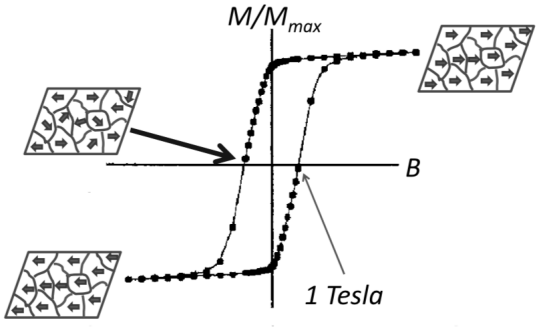
  which has a minimum for $N \sim \sqrt{J/\kappa}$. In real materials, often $N \sim 100$.

Example. Consider a cylinder of radius R and length L. If $R \ll L$, we have effectively two point charges, with charge $\pm q$ where $q \sim R^2$. The magnetic field energy for each is found by integrating E^2 with a short-distance cutoff of R, giving energy $q^2/R \sim R^3$. By adding a domain wall that splits the cylinder in half lengthwise, the charge q can be converted to a $\pm q/2$ dipole, which has negligible energy in comparison. The energy cost is RL, and the ratio of energy savings to energy cost is R^2/L. Hence for large enough L, the formation of domain walls is not favorable.

On the other hand, when $R \gg L$ we effectively have a parallel plate capacitor with energy R^2L. This energy can be almost completely eliminated by splitting the plates into regions of size L with alternating magnetization. This has cost $(L)(R^2/L^2)L = R^2$ where L is the height of the domain walls, R^2/L^2 is the number of domains, and L is the perimeter of each. The ratio of energy savings to energy cost is L, so for small enough L, the formation of domain walls is favorable.

Finally, we consider hysteresis.

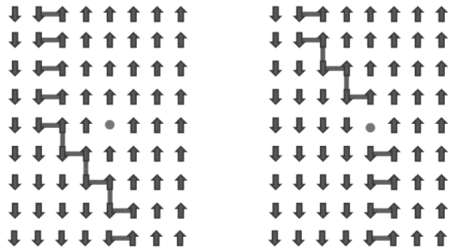
- We observe experimentally that ferromagnets display hysteresis: their current magnetization depends on their past history. When the external field is increased and then decreased, the magnetization follows a hysteresis curve.



  However, this wouldn't make sense in a perfect crystal because the energy cost of a domain is roughly constant; domains should continuously grow and shrink in response to applied fields. Hysteresis only occurs when domains are constrained in some way.

- Consider a polycrystalline sample. In the limit of large κ, the magnetization of a crystallite may point in either the $\pm\hat{\mathbf{z}}$ directions, say the $-\hat{\mathbf{z}}$ direction. When an external field along $\hat{\mathbf{z}}$ is turned on, the $\hat{\mathbf{z}}$ direction is favored, but the existing configuration is metastable. It requires a substantial field to remove this metastability. Once the domain flips, the field can be removed and it will remain in the $\hat{\mathbf{z}}$ state. Accounting for the different properties and environments of all the different domains yields a smooth hysteresis curve.

- Hysteresis also occurs for single-crystal samples with disorder, because domain walls can be "pinned" to defects, as shown below.



  Here every solid line denotes a neighboring pair of antialigned spins, and the energy cost is lower if the domain wall passes through a defect. Hence the presence of defects can impede the motion of domain walls, just as crystallite boundaries do, yielding hysteresis.

- Practically, hysteresis is used in magnetic disk drives. Here one wants a "hard" magnet which maintains its magnetization well in the absence of an external field. These are manufactured by introducing disorder to pin the domain walls.

Ferromagnetism can be understood using mean field theory.

- In Weiss mean field theory, or molecular mean field theory, one treats a single site exactly, and approximates the effects of all other sites with an expectation. This yields an expected magnetization for the single site, and the actual magnetization is solved by self-consistency.

- For example, we consider the ferromagnetic Ising model,

$$H = -\frac{1}{2}J\sum_{\langle i,j\rangle}\sigma_i\sigma_j + g\mu_B B\sum_j \sigma_j.$$

  Focusing on the Hamiltonian for a single site, and replacing the neighbors with an expectation,

$$H_i \sim (-Jz\langle\sigma\rangle + g\mu_B B)\,\sigma_i$$

  where the coordination number z is the number of neighbors of each site.

- Therefore, we have the self-consistency equation

$$\langle\sigma\rangle = -\frac{1}{2}\tanh(\beta(g\mu_B B - Jz\langle\sigma\rangle)/2).$$

  In particular, for zero external field, we have a phase transition when $\beta Jz/4 = 1$.

- We may also compute the paramagnetic susceptibility. For a small field B, we may expand the tanh linearly to find

$$\langle\sigma\rangle = -\frac{1}{4}\frac{\beta g\mu_B B}{1-\beta Jz/4}.$$

  Using $m = -g\mu_B\langle\sigma\rangle$, we find the zero-field susceptibility

$$\chi = \frac{\rho(g\mu_B)^2\mu_0/4}{k_B(T-T_c)} = \frac{\chi_{\text{Curie}}}{1-T_c/T}$$

  which diverges for $T \to T_c$ and approaches the result for Curie paramagnetism at high T.

- For an antiferromagnet, the mean field calculation is almost identical: the spins alternate, but this cancels the flipped sign of J, giving the same zero-field results. However, the susceptibility is instead

$$\chi = \frac{\chi_{\text{Curie}}}{1+T_c/T}$$

  which was the original experimental signature for antiferromagnetism.

- Mean field theory can be systematically improved by taking larger and larger "unit cells" and performing the calculation exactly within them. This is a predecessor to modern RG techniques.

Finally, we introduce the Hubbard model, one of the most important models in condensed matter.

- The Hubbard model is a tight binding model with hopping parameter t with a repulsive interaction between electrons on the same site,

$$H_{\text{int}} = \sum_i U n_{i\uparrow}n_{i\downarrow}$$

  which comes from Coulomb repulsion. There is a competition between this interaction, which prefers to align the spins, and the energy of the Fermi sea, which prefers to have an equal amount of each spin.

- We can understand this qualitatively by mean field theory. Here we will compute the kinetic energy as if the electrons were not interacting, and compute the interaction energy by averaging.

- To handle the interaction energy, we note that

$$n_{i\uparrow}n_{i\downarrow} = \frac{1}{4}(n_{i\uparrow}+n_{i\downarrow})^2 - \frac{1}{4}(n_{i\uparrow}-n_{i\downarrow})^2 \approx \frac{1}{4}\langle n_{i\uparrow}+n_{i\downarrow}\rangle^2 - \frac{1}{4}\langle n_{i\uparrow}-n_{i\downarrow}\rangle^2.$$

  For a unit cell of volume v, we have $M = (\mu_B/v)\langle n_{i\downarrow} - n_{i\uparrow}\rangle$. We assume the number of electrons is fixed, so the first term is a constant, giving

$$\langle H_{\text{int}}\rangle = -\frac{NU}{4}\left(\frac{Mv}{\mu_B}\right)^2 + \text{const.}$$

- Next we consider the shift in the Fermi sea. Working to lowest order, we let

$$E_{F,\uparrow} = E_F + \epsilon/2, \quad E_{F,\downarrow} = E_F - \epsilon/2$$

  and let the density of states per spin species per unit volume be $g(E_F)/2$. Then

$$N_\uparrow - N_\downarrow = \frac{g(E_F)}{2}\epsilon, \quad M = -\frac{g(E_F)\mu_B}{2}\epsilon.$$

  The shift in the Fermi sea energy is

$$\delta E = \int_{E_F}^{E_F+\epsilon/2} dE\,\frac{Eg(E)}{2} - \int_{E_F-\epsilon/2}^{E_F} dE\,\frac{Eg(E)}{2} = \frac{g(E_F)}{2}(\epsilon/2)^2 = \frac{g(E_F)}{2}\left(\frac{M}{\mu_B g(E_F)}\right)^2$$

  where we used our first result.

- Combining, the energy shift per unit volume is

$$\frac{\delta E_{\text{tot}}}{V} = \left(\frac{M}{\mu_B}\right)^2\left(\frac{1}{2g(E_F)} - \frac{vU}{4}\right)$$

  which results in a nonzero magnetization if $U > 2/g(E_F)v$, called the Stoner criterion.

- As argued earlier, if half of the states are occupied and U is large, then we get a Mott insulator in this model. It also turns out that the Mott insulator is an antiferromagnet. This is because in the antiferromagnetic state, the electrons can hop, while they cannot in the ferromagnetic state by Pauli exclusion. Treating the hopping t as a perturbation, this means that the antiferromagnetic state's energy is pushed down at second order.

5.3 Superconductivity

Superconductivity was discovered by Onnes in 1911, while attempting to test Drude theory at low temperatures. He instead found that mercury loses all electrical resistance below 4 K. Here we give a brief overview of the phenomenology of superconductivity.

- At atmospheric pressure, Hg, Nb, Pb, Sn, and V are all superconducting at a few Kelvin. Surprisingly, good conductors such as copper, silver, and gold display no signs of superconductivity at all. A few more elements become superconducting at extreme pressures, such as metallic hydrogen, which is predicted to be superconducting at room temperature. In the 1980s, more complicated alloys were discovered to be superconducting at around 100 K, ushering in the era of "high temperature superconductivity".

- Experimentally measuring zero resistance is a bit tricky, as any wires attached to the sample will themselves have resistance. This is circumvented using a four terminal approach: current is passed through the sample through two terminals, while a voltmeter is attached across two separate terminals. When the sample is superconducting, the voltage across the voltmeter's terminals is exactly zero, so the resistance of the voltmeter wires don't matter. Another, less direct method is the observation of persistent currents in the sample.

- In a classically perfect conductor, we have infinite conductivity, so the electric field vanishes inside the conductor. Then the emf along any loop vanishes, so Φ_B is constant through any loop, which implies $\mathbf{B}$ is constant in time.

- However, Meissner found that superconductors completely expel magnetic flux. The Meissner effect is much simpler to measure than vanishing resistivity, as it is an equilibrium property rather than a nonequilibrium transport property, so that it is taken as the modern definition of superconductivity. It is equivalent to the statement that a superconductor is a perfect diamagnet, so $\chi_m = -1$. Like all magnetic effects, it cannot be understood classically.

We now make some preliminary steps using a simple quasiclassical microscopic model.

- We model a superconductor classically with the Drude model in the limit $\tau \to \infty$. In this case the electrons simply accelerate freely without damping, $m\mathbf{a} = q\mathbf{E}$, implying

$$\dot{\mathbf{J}} = (n_s q^2/m)\mathbf{E}.$$

  This is different from the naive electrostatic model of a perfect conductor, where we simply set $\mathbf{E} = 0$ everywhere, but we will see this implies $\mathbf{E} = 0$ within the bulk anyway. Here n_s is the density of superconducting electrons, which we allow to be different from the density of electrons n.

- Taking the curl of both sides and applying Faraday's law,

$$\nabla \times \dot{\mathbf{J}} = \frac{n_s q^2}{m} \nabla \times \mathbf{E} = -\frac{n_s q^2}{m} \dot{\mathbf{B}}.$$

  Integrating both sides gives

$$\nabla \times \mathbf{J} = -\frac{n_s q^2}{m} (\mathbf{B} - \mathbf{B}_0)$$

  where $\mathbf{B}_0$ is a constant of integration.

- On the other hand, we know that $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$, so

$$\nabla \times \nabla \times \frac{\mathbf{B}}{\mu_0} = -\frac{n_s q^2}{m} (\mathbf{B} - \mathbf{B}_0).$$

  Simplifying the curl of curl gives

$$\frac{1}{\mu_0} \nabla^2 \mathbf{B} = \frac{n_s q^2}{m} (\mathbf{B} - \mathbf{B}_0).$$

  It's straightforward to check this forces the magnetic field to be uniform well inside a superconductor; we'll do this more explicitly below.

- So far this derivation is entirely classical; we may now account for the Meissner effect by setting $\mathbf{B}_0 = 0$, giving the London equation

$$\mathbf{B} = \lambda^2 \nabla^2 \mathbf{B}, \quad \lambda = \sqrt{\frac{m}{\mu_0 n_s q^2}}$$

  where λ is the London penetration depth. Similarly, we have $\mathbf{J} = \lambda^2 \nabla^2 \mathbf{J}$. These equations imply that magnetic fields and currents decay exponentially as we enter the superconductor. Hence, in the bulk both the electric and magnetic fields vanish.

- We may also arrive at the London equation by postulating

$$\mathbf{J} = -\frac{n_s q^2}{m} \mathbf{A}$$

  and taking the curl of both sides. Of course, this is a gauge-dependent statement; it is defined to hold in London gauge. In the time-independent case we have $\nabla \cdot \mathbf{J} = 0$ by charge conservation, so London gauge coincides with Coulomb gauge.

We can get a little more insight by thinking in terms of macroscopic wavefunctions, which we cover in much more detail in the notes on Condensed Matter.

- Given a single charged particle with wavefunction ψ, one cannot interpret the probability density and current as charge densities and currents. However, suppose the particle is a boson among many others. Since bosons tend to clump up in the same state, we can have macroscopically many particles with wavefunction ψ, which we call a "macroscopic wavefunction".

- The macroscopic wavefunction behaves much like a classical field. By the Born rule, its density and current $\mathbf{J}_p$ can be interpreted as electric densities and currents $\mathbf{J} = q\mathbf{J}_p$ in the semiclassical limit. We conventionally normalize it so that its total density is the number of particles N.

- This is how a classical electromagnetic field emerges from a quantum one; it also appears in superfluidity. The case of superconductivity is more subtle, because electrons are fermions. It turns out that phonons mediate an attractive force between electrons, leading to the formation of Cooper pairs, which are bosons.

- Writing $\psi = \rho^{1/2} e^{i\theta}$, the probability current is

$$\mathbf{J}_p = \frac{\rho}{m} \left(\nabla\theta - q\mathbf{A} \right).$$

  The probability current is related to the velocity of the electron by $\mathbf{J}_p = \rho \mathbf{v}$, so

$$m\mathbf{v} = \nabla\theta - q\mathbf{A}.$$

- In this picture, the superconducting current cannot dissipate because bosons tend to clump up into the same state; it is difficult to knock one out of the condensate. We may also derive the Meissner effect by taking the divergence of the above equation. In Coulomb gauge,

$$\nabla \cdot \mathbf{J}_p = \frac{\rho}{m} \nabla^2 \theta.$$

  In the steady state, the left-hand side vanishes, so θ must be constant. Hence

$$\mathbf{J}_p = -\frac{\rho q}{m} \mathbf{A}$$

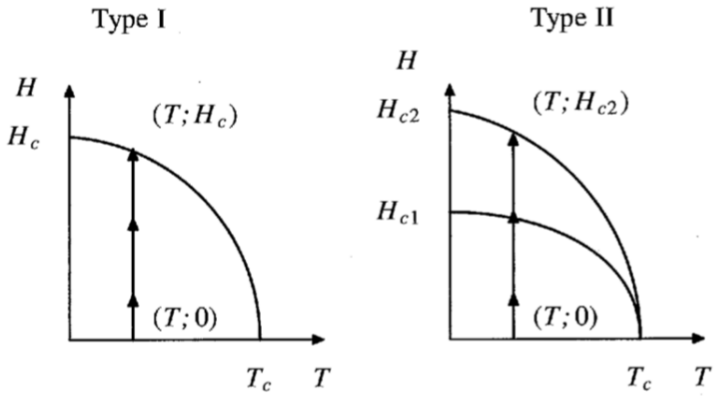
  which is equivalent to London's equation.

- Now consider a ring with thickness much greater than λ, through which a magnetic field passes. When it is cooled down, it becomes superconducting, which expels the magnetic field passing through the body of the ring. Superconducting currents holds the flux through the hole constant; if the magnetic field is subsequently turned off the flux is preserved.
- Inside the body of the ring and from the edges, we have $\mathbf{J}_p = 0$, so

$$\nabla\theta = q\mathbf{A}.$$

  Hence the fact that the wavefunction is single valued constrains the flux to be quantized in multiples of $\Phi_0 = h/q$. (Note that this argument doesn't apply to a single particle on a ring, because in that case there is no reason to set $\mathbf{J}_p = 0$.) This Φ_0 is reasonably large and readily observable, though one instead finds multiples of $h/2q$, because of the Cooper pairing.

Note. The quantization of flux is sometimes said to be a consequence of "single valuedness of the wavefunction", but this is a bit of an oversimplification. It is really a consequence of energy minimization: the energy cost of having the particles all leave the macroscopic coherent state exceeds the cost of enforcing quantized flux.



The same remarks hold for the Meissner effect. In a type I superconductor, the superconductivity is destroyed once the applied field $\mathbf{H}$ exceeds a critical value, allowing a flux to pass through the sample; this is simply because the energy cost of expelling the flux exceeds the energy savings of being in the superconducting state. In a type II superconductor, the magnetic flux instead becomes positive once $\mathbf{H}$ exceeds a lower critical value, due to the creation of vortex lines. These are non-superconducting regions which, by the above argument, carry a magnetic flux through them. Whether or not a superconductor is type I or type II again depends on energetics.

Note. An explicit vortex solution. A superconductor is characterized by three length scales: the coherence length ξ_0, the mean free path ℓ, and the penetration depth λ. We will see below that the ratio λ/ξ_0 determines whether a superconductor is type I or type II. A superconductor is said to be clean if $l \gg \xi_0$, though remarkably they may remain superconducting even if $l < \xi_0$.

We consider a vortex line along the z-axis in a clean superconductor. Using the Laplacian in cylindrical coordinates, the London equation is

$$\frac{d^2B_z}{dr^2} + \frac{1}{r}\frac{dB_z}{dr} - \frac{B_z}{\lambda^2} = 0.$$

This equation is only valid for $r > \xi_0$, so the region $r < \xi_0$ is interpreted as the vortex core, where the system is not superconducting. The solution to the differential equation is a Bessel function,

$$B_z(r) = \frac{\Phi_0}{2\pi\lambda^2}K_0(\lambda/r)$$

while in the vortex core B_z is constant instead. Using the asymptotics of the Bessel function,

$$B_z(r) = \frac{\Phi_0}{2\pi\lambda^2}\begin{cases}\log(r/\lambda) & r \ll \lambda, \\ e^{-r/\lambda}\sqrt{\pi\lambda/2r} & r \gg \lambda.\end{cases}$$

Of course, these limits are only relevant if $\xi_0 \ll \lambda$, which we assume here. In this case, Φ_0 is the total magnetic flux, neglecting the small amount of flux inside the vortex core, because

$$2\pi\int_{\xi_0}^{\infty} rB_z(r)\,dr = 2\pi\lambda^2\int_{\xi_0}^{\infty}\frac{d}{dr}\left(r\frac{dB_z}{dr}\right)dr = 2\pi\lambda^2\left(r\frac{dB_z}{dr}\right)_{r=\xi_0} = \Phi_0$$

where we used the London equation in reverse in the first step. To calculate the energy of the vortex core, note that by Ampere's law

$$\mathbf{J} = \frac{1}{\mu_0}\nabla\times\mathbf{B} = -\frac{1}{\mu_0}\frac{dB_z}{dr}.$$

The energy takes the form of kinetic energy, with $\mathbf{J} = nq\mathbf{v}$. Most of the energy comes from the region $r \lesssim \lambda$, and integrating gives an energy per length of

$$E \sim \frac{\Phi_0^2}{4\pi\mu_0\lambda^2}\log\frac{\lambda}{\xi_0}.$$

5.4 Ginzburg-Landau Theory

We can understand superconductivity in statistical mechanics by Ginzburg-Landau theory. First, we cover some thermodynamic preliminaries.

- The definition of magnetic work is notoriously tricky. For concreteness, consider a cylindrical sample inside a solenoid with N turns, length L, and current I. Then

  $$\mathbf{H} = \frac{N}{L}I\hat{\mathbf{z}}.$$

  The work done upon increasing the current infinitesimally is

  $$đW = -I\mathcal{E}\,dt = NI\,d\Phi$$

  where Φ is the flux through one loop.

- Therefore, we have

  $$đW = NAI\,dB = V\,\mathbf{H}\cdot d\mathbf{B} = \mu_0 V(\mathbf{H}\cdot d\mathbf{M} + \mathbf{H}\cdot d\mathbf{H}).$$

  The second term above is simply the work that would have to be done if the sample were not there at all; it is due to the self-inductance of the solenoid. As such, we choose to define the magnetic work (per unit volume) done on the sample as $\mu_0\,\mathbf{H}\cdot d\mathbf{M}$, though this is convention-dependent. For more discussion on this point, see the notes on Undergraduate Physics.

- In any case, the internal energy and free energies are defined as

  $$dU = T\,dS + \mu_0 V\,\mathbf{H}\cdot d\mathbf{M}, \quad F = U - TS, \quad G = U - TS - \mu_0 V\,\mathbf{H}\cdot\mathbf{M}.$$

  We will focus on the Gibbs free energy, with differential

  $$dG = -S\,dT - \mu_0 V\,\mathbf{M}\cdot d\mathbf{H}.$$

  This is the most convenient potential because T and $\mathbf{H}$ can be controlled experimentally.

- By integrating this expression, for a type I superconductor we have

$$G_s(T, H_c) - G_s(T, 0) = -\mu_0 V \int_0^{H_c} \mathbf{M} \cdot d\mathbf{H} = \frac{\mu_0}{2} H_c^2 V$$

where we used $\mathbf{B} = 0$. But at the critical field H_c, the superconductor is in equilibrium with the normal phase,

$$G_s(T, H_c) = G_n(T, H_c) \approx G_n(T, 0)$$

since $M \approx 0$ in the normal state.

- Therefore, at zero external field, the difference in Gibbs free energies between the superconducting and normal states is

$$G_s(T, 0) - G_n(T, 0) = F_s(T, 0) - F_n(T, 0) = -\frac{\mu_0}{2} H_c^2 V$$

which measures the condensation energy associated with entering the superconducting state. Numerically we find that this quantity is tiny, on the scale of μeV per atom, a puzzle that was explained by BCS theory.

- A type II superconductor is more complex, because the $\mathbf{M}\, d\mathbf{H}$ integral is nontrivial. Instead we conventionally define

$$F_s(T, 0) - F_n(T, 0) = -\frac{\mu_0}{2} H_c^2 V, \quad \frac{H_c^2}{2} \equiv \int_0^{H_{c2}} \mathbf{M} \cdot d\mathbf{H}.$$

The quantity H_c is called the thermodynamic critical field, though the only real transitions are at H_{c1} and H_{c2}.

- By a derivation analogous to the Clausius-Clapeyron equation, the latent heat of the superconducting/normal phase transition for a type I superconductor is

$$L = -\mu_0 H_c \frac{dH_c}{d \log T}.$$

This shows that the phase transition for a type I superconductor is discontinuous, except in the case of zero external field, where it is continuous. By contrast, the transitions for a type II superconductor are all continuous.

Now we introduce Landau-Ginzburg theory, only briefly covering aspects common to Landau theory in general; all these points are covered in detail in the notes on Statistical Field Theory.

- We describe the superconducting phase transition with a complex order parameter ψ. Here ψ may simply be viewed as a phenomenological, classical coarse-grained field which is nonzero in the superconducting state. With more work one can show that it is the macroscopic wavefunction corresponding to the condensate of Cooper pairs, bound pairs of electrons described by BCS theory. However, by virtue of universality, Landau-Ginzburg theory is more general than this; it successfully describes exotic superconductors where BCS theory does not apply.

- As a state, we neglect spatial variations and external fields, as well as fluctuations, working at the level of mean field theory. We postulate the free energy density is

$$f_s(T) = f_n(T) + a(T)|\psi|^2 + \frac{1}{2}b(T)|\psi|^4$$

as usual in Landau theory, where at the critical point we Taylor expand

$$a(T) = \dot{a}(T - T_c) + \ldots, \quad b(T) = b + \ldots.$$

There is thus a continuous phase transition at $T = T_c$, near which we have

$$|\psi| = \sqrt{\frac{\dot{a}}{b}(T_c - T)}\,\theta(T_c - T).$$

- Furthermore, the difference in free energy between the superconducting and normal states is

$$f_s(T) - f_n(T) = -\frac{\dot{a}^2(T - T_c)^2}{2b}$$

which yields a discontinuity in the specific heat, and gives the thermodynamic critical field

$$H_c = \frac{\dot{a}}{\sqrt{\mu_0 b}}\,(T_c - T).$$

Since Landau theory is based on a Taylor expansion near T_c, this is only valid near T_c.

- To account for spatial variations, we add a gradient term,

$$f_s(T) = f_n(T) + \frac{\hbar^2}{2m^*}|\nabla\psi|^2 + a(T)|\psi|^2 + \frac{1}{2}b(T)|\psi|^4.$$

Setting the variational derivative of the free energy to zero gives

$$-\frac{\hbar^2}{2m^*}\nabla^2\psi + \left(a + b|\psi|^2\right)\psi = 0$$

which has the form of a nonlinear Schrodinger equation.

- In one dimension, near an interface with $\psi(0) = 0$, we find the solution

$$\psi(x) = \psi_0 \tanh\frac{x}{\sqrt{2}\xi}$$

where ψ_0 is the field in the bulk. Hence variations in the field have the characteristic length

$$\xi = \sqrt{\frac{\hbar^2}{2m^*|a(T)|}}$$

which we identified as the coherence length above. Thus Landau theory predicts the coherence length diverges as $t^{-1/2}$ near the critical temperature.

- As usual in Landau theory, the situation becomes more complicated when we account for fluctuations. However, for most superconductors the Ginzburg criterion ensures the fluctuation effects are not visible until one is extremely close to the critical point, so that mean field theory is extremely accurate in practice. For high temperature superconductors, the coherence length is much smaller, so that one can see fluctuation effects within about 1 K of the critical temperature. The modified critical exponents match those predicted by statistical field theory.

Next, we introduce Landau-Ginzburg theory proper by coupling to a magnetic field.

- We use the standard minimal coupling procedure, except with a charge of $2e$ since ψ represents Cooper pairs. For later convenience, we'll also count the full magnetic field energy, so

$$f_s(T) = f_n(T) + \frac{\hbar^2}{2m^*}\left|\left(-i\nabla + \frac{2e}{\hbar}\mathbf{A}\right)\psi\right|^2 + a(T)|\psi|^2 + \frac{1}{2}b(T)|\psi|^4 + \frac{B^2}{2\mu_0}.$$

  This leads to the nonlinear Schrodinger equation

$$-\frac{\hbar^2}{2m^*}\left(\nabla + \frac{2ei}{\hbar}\mathbf{A}\right)^2\psi + \left(a + b|\psi|^2\right)\psi = 0.$$

- The current due to the magnetic field can be found by differentiating the free energy,

$$\mathbf{J}_s = -\frac{\delta F_s}{\delta \mathbf{A}(\mathbf{r})} = -\frac{2e\hbar i}{2m^*}(\psi^*\nabla\psi - \psi\nabla\psi^*) - \frac{(2e)^2}{m^*}|\psi|^2\mathbf{A}.$$

  The magnetic field is then sourced by the total current,

$$\nabla \times \mathbf{B} = \mu_0(\mathbf{J}_{\text{ext}} + \mathbf{J}_s).$$

- By construction, the free energy has the local $U(1)$ gauge symmetry

$$\psi(\mathbf{r}) \to \psi(\mathbf{r})e^{i\theta(\mathbf{r})}, \quad \mathbf{A}(\mathbf{r}) \to \mathbf{A}(\mathbf{r}) + \frac{\hbar}{2e}\nabla\theta.$$

  Hence the gauge-invariant measure of the phase gradient of the field is $\nabla\theta - (2e/\hbar)\mathbf{A}$, and indeed by expanding the free energy for a slowly varying phase, we find

$$F_s = F_s^0 + \rho_s \int d\mathbf{r}\left(\nabla\theta + \frac{2e}{\hbar}\mathbf{A}\right)^2, \quad \rho_s = \frac{\hbar^2}{2m^*}|\psi|^2.$$

  For concreteness, suppose there is no external field and gauge fix $\mathbf{A} = 0$. Then the phase of ψ is the same throughout space, even though this phase is modified by a *global* $U(1)$ transformation $\psi(\mathbf{r}) \to \psi(\mathbf{r})e^{i\theta}$. Hence the superconductor spontaneously breaks the global symmetry associated with a local gauge symmetry, a result sometimes incorrectly summarized as "spontaneous breaking of gauge symmetry". The same effect occurs in the Standard Model by the Higgs field.

- Now allowing for an external field, note that the current may also be expressed as

$$\mathbf{J}_s = -\frac{2e}{\hbar}\rho_s\left(\nabla\theta + \frac{2e}{\hbar}\mathbf{A}\right).$$

  We may choose a gauge where $\nabla\theta = 0$ in the ground state. Then

$$\mathbf{J}_s = -\rho_s\frac{(2e)^2}{\hbar^2}\mathbf{A}$$

  which is precisely the London equation; evidently this gauge is London gauge.

- To make the connection more explicit, we define

$$n_s = 2|\psi|^2, \quad m^* = 2m_e$$

in which case our two forms of the London equation coincide exactly. This further supports our interpretation of ψ, as m^* is the mass of a Cooper pair and $|\psi|^2$ is the density of Cooper pairs.

- By making these substitutions, we may express the London penetration depth in terms of the Landau-Ginzburg parameters,

$$\lambda(T) = \left(\frac{m_e b}{2\mu_0 e^2 |a(T)|}\right)^{1/2}.$$

The dimensionless ratio $\kappa = \lambda(T)/\xi(T)$ is thus temperature independent.

Example. Consider a superconducting ring of radius R, described by a field $\psi(\theta)$. We pass a flux Φ through the ring, and choose the gauge

$$A_\theta = \frac{\Phi}{2\pi R}.$$

The free energy is minimized for configurations where $\nabla\theta$ and $|\psi|$ are constant, so

$$\psi(\theta) = \psi_0 e^{in\theta}.$$

Plugging this back into the free energy, we find

$$F_s(T) = \frac{\hbar^2}{2m^* R^2} V|\psi|^2(\Phi - n\Phi_0)^2 + \frac{1}{2\mu_0}\int B^2\, d\mathbf{r} + \text{const}, \quad \Phi_0 = \frac{h}{2e}.$$

The second term can also be expressed in terms of the inductance of the ring; it is $LI^2/2 = \Phi^2/2L$. In general the first term will be much larger, so it is energetically favorable for the flux through the ring to be nearly a multiple of Φ_0. Hence we have derived flux quantization through energetic considerations. The fact that Φ_0 is indeed measured to be $h/2e$ rather than h/e provides experimental evidence for Cooper pairs. The macroscopically high potential barriers between different minima of the free energy are responsible for persistent currents.

Note. We can get a bit more insight by including the full dynamics of the electromagnetic field. In this case, the vev of ψ provides a mass term for the photon as in the Higgs mechanism,

$$\mathcal{L} \supset \frac{\hbar^2}{2m^*}\left|\frac{2e}{\hbar}\psi\right|^2 A^2 - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}.$$

The photon mass inside the superconductor is

$$m_A \sim \frac{e^2}{m^*}|\psi|^2 = \frac{e^2}{m^*}\frac{|a|}{b}$$

and this causes the Meissner effect, with electromagnetic fields decaying on the scale of the Compton wavelength $\lambda = 1/m$, matching our earlier expression for λ. In this perspective, the two length scales λ and ξ simply correspond to the inverse masses of the photon and Cooper pair.

Landau-Ginzburg theory allows one to treat superconductors with nontrivial spatial variation; we now use it to describe vortices.

- Abrikosov found that in a type II superconductor near H_{c2}, the vortices form a periodic structure called the Abrikosov flux lattice.

- Since the phase transition at H_{c2} is second order, the order parameter is small near H_{c2} and we may neglect the nonlinear term. Furthermore the magnetization is small, so $\mathbf{B} = \mu_0 \mathbf{H}$. We work in Landau gauge, $\mathbf{A} = xB\,\hat{\mathbf{y}}$, so the Landau-Ginzburg equation is

$$-\frac{\hbar^2}{2m^*}\left(\nabla^2 + \frac{4eBi}{\hbar}x\frac{\partial}{\partial y} - \frac{(2eB)^2}{\hbar^2}x^2\right)\psi = |a|\psi$$

  where we have neglected the cubic term because $|\psi|$ is small near H_{c2}.

- Without the nonlinear term, this is simply the Schrodinger equation for a particle in a magnetic field, with cyclotron frequency $\omega_c = 2eB/m^*$ and energy $|a|$. As shown in the lecture notes on Anyons, the solutions are Landau levels,

$$\psi(\mathbf{r}) = e^{i(k_y y + k_z z)} f(x - x_0), \quad x_0 = -\frac{\hbar k_y}{m\omega_c}$$

  where $f(x)$ is the n^{th} eigenstate of the harmonic oscillator with frequency ω_c, and the energy is

$$|a| = \left(n + \frac{1}{2}\right)\hbar\omega_c + \frac{\hbar^2 k_z^2}{2m^*}.$$

  Note that the Landau levels are not all degenerate, since we are working in three dimensions rather than the usual two.

- The important point is that the minimum possible value of the "energy" is $\hbar\omega_c/2$, which places a bound on $|a|$. Specifically, near the critical point,

$$\frac{1}{2}\hbar\omega_c = \dot{a}(T_c - T)$$

  where T is the superconducting transition temperature. Since ω_c depends on H, this may be inverted to an expression $H_{c2}(T)$,

$$\mu_0 H_{c2} = \frac{\Phi_0}{2\pi\xi(T)^2}.$$

  Since a vortex line carries flux Φ_0 and has characteristic radius ξ, the vortices are tightly packed near the superconducting transition.

- Comparing this to our expression for the thermodynamic critical field H_c earlier, we find

$$H_{c2} = \sqrt{2}\kappa H_c.$$

  Now, H_c was interpreted as the critical field below which a uniform nonzero ψ becomes favorable in a first order transition, while H_{c2} is the critical field below which flux lattice with small ψ becomes favorable in a second order transition. Hence if $H_c > H_{c2}$, the first order transition occurs first, so the superconductor is type I, while if $H_c < H_{c2}$, the superconductor is type II. The critical value of κ separating these two behaviors is $1/\sqrt{2}$.

- To show the configuration is indeed a flux lattice, note that the Landau levels of lowest energy have $n = 0$ and $k_z = 0$, so

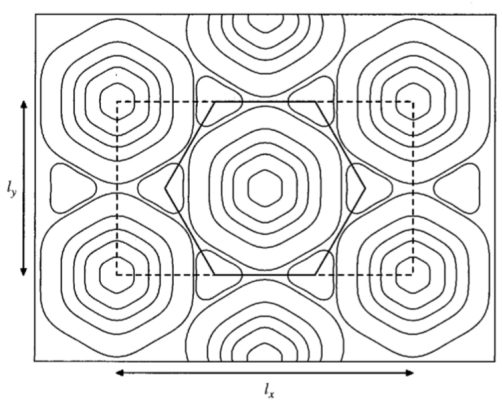
$$\psi(\mathbf{r}) = e^{ik_y y} e^{-(x-x_0)^2/\xi^2}.$$

  Taking periodic boundary conditions, the degenerate ground state has the form

$$\psi(\mathbf{r}) = \sum_m c_m e^{i(2\pi m y/l_y)} e^{-(x+m\Phi_0/Bl_y)^2/\xi^2}, \quad k_y = \frac{2\pi}{l_y} m, \quad m \in \mathbb{Z}.$$

  The state is periodic in the x direction as long as $c_{m+\nu} = c_m$ for some integer ν.

- The degeneracy is broken by the cubic term, so we look for the lowest energy configuration. Abrikosov considered the case $\nu = 1$, which leads to a square lattice. However, it turns out that there is a state of slightly lower energy with $\nu = 2$, which gives a triangular lattice.



  Curves of constant $|\psi|$ are shown for this lattice above.

- The flux lattice may be experimentally measured by sprinkling paramagnetic particles on the surface of the superconductor, as they are attracted to the regions of highest magnetic field. Alternatively, one can measure the field directly with a SQUID. Finally, since neutrons have a magnetic moment, they can scatter off the magnetic field, and the geometry of the flux lattice can be inferred from the diffraction pattern.

- So far we have only worked in a mean field approximation, neglecting any thermal fluctuations. Once these are added, the vortices behave dynamically, forming "vortex matter" which may be liquid or glassy. This presents a significant challenge for technological applications of high T_c superconductors. Such superconductors have very low H_{c1}, so vortices are always present, and the vortices flow like a liquid. Currents flowing in the superconductor cause the vortices to move, dissipating energy and providing a finite resistance.

6 Linear Response

6.1 Response Functions

We now consider how systems react to weak external influences, where the response is linear.

- Classically, a system with generalized coordinates $x_i(t)$ has the time evolution

$$\ddot{x_i} + g_i(\dot{x}, x) = 0.$$

  If an external influence is added, the equation of motion becomes

$$\ddot{x_i} + g_i(\dot{x}, x) = F_i(t)$$

  where the $F_i(t)$ are controlled externally, with a pre-specified time dependence.

- As a simple example, a mass on a spring has $g(\dot{x}, x) = \gamma\dot{x} + \omega_0^2 x$. Note that since friction is present, we aren't working with a Hamiltonian system.

- In the quantum case, the observables are Heisenberg operators $\mathcal{O}_i(t)$. Left alone, the Hamiltonian is $H(\mathcal{O})$. External influences add the term

$$H_s(t) = \phi_i(t)\mathcal{O}_i(t)$$

  where the $\phi_i(t)$ are sources under external control. For example, if $\mathcal{O}_i(t) = -\hat{x}$, then $\phi_i(t)$ is the force in the x–direction.

- We assume the system's response is linear,

$$\delta\langle\mathcal{O}_i(t)\rangle = \int dt'\, \chi_{ij}(t,t')\phi_j(t')$$

  where χ_{ij} called a response function or generalized susceptibility. Classically, quantum operators are replaced by dynamical variables, and the χ_{ij} are simply Green's functions.

- Assuming time translation invariance, we have

$$\chi_{ij}(t,t') = \chi_{ij}(t-t').$$

  Taking the Fourier transform,

$$\delta\langle\mathcal{O}_i(\omega)\rangle = \int dt'\, dt\, e^{i\omega t}\chi_{ij}(t-t')\phi_j(t') = \int dt'\, dt\, e^{i\omega(t-t')}\chi_{ij}(t-t')e^{i\omega t'}\phi_j(t') = \chi_{ij}(\omega)\phi_j(\omega).$$

  That is, the response is local in frequency space, as we expect from a linear theory.

It is useful to consider the real and imaginary parts of the response function. We drop the i and j indices, considering only a single source and response.

- If the source ϕ is real and $\mathcal{O}$ is Hermitian (so that $\langle\mathcal{O}\rangle$ is real), then $\chi(t)$ must also be real.

- We decompose the Fourier transform $\chi(\omega)$ into real and imaginary parts as

$$\chi(\omega) = \chi'(\omega) + i\chi''(\omega)$$

  where the χ' and χ'' functions are real. The reality condition implies that the Fourier transform of $\chi(-t)$ is $\chi^*(\omega)$.

- Now, the imaginary part is

$$\chi''(\omega) = -\frac{i}{2}\left(\chi(\omega) - \chi^*(\omega)\right) = -\frac{i}{2}\int dt\, e^{i\omega t}\left(\chi(t) - \chi(-t)\right).$$

  We see that the imaginary part of $\chi(\omega)$, called the dissipative/reactive part or the spectral function, comes from the part of $\chi(t)$ that is not invariant under time reversal. The spectral function "knows about the arrow of time", and it is odd in ω.

- The real part, also called the reactive part, is

$$\chi'(\omega) = \frac{1}{2}(\chi(\omega) + \chi^*(\omega)) = \frac{1}{2}\int dt\, e^{i\omega t}\left(\chi(t) + \chi(-t)\right).$$

  This part doesn't care about the arrow of time, and it is even in ω.

- With multiple sources, the appropriate definitions of χ' and χ'' are the Hermitian and anti-Hermitian parts of the matrix $\chi_{ij}(\omega)$,

$$\chi'_{ij}(\omega) = \frac{1}{2}(\chi_{ij}(\omega) + \chi^*_{ji}(\omega)), \quad \chi''_{ij}(\omega) = -\frac{i}{2}\left(\chi_{ij}(\omega) - \chi^*_{ji}(\omega)\right).$$

- Causality is the condition

$$\chi(t) = 0 \text{ for all } t < 0.$$

  Imposing causality is essentially choosing the causal/retarded Green's function.

- Now consider the Fourier transform

$$\chi(t) = \int_{-\infty}^{\infty} đ\omega\, e^{-i\omega t}\chi(\omega)$$

  where ω is now regarded as complex. When $t < 0$, the right-hand side is a contour integral which can be closed in the upper-half plane. The value of the integral is the sum of the residues of $\chi(\omega)$ inside the contour, weighted by $e^{-i\omega t}$. However, since the integral is zero for all $t < 0$, $\chi(\omega)$ can have no poles in the upper-half plane. Hence causality is related to analyticity.

6.2 Kramers-Kronig

The real and imaginary parts of an analytic function are related by complex analysis.

- Let $\rho(\omega)$ be a meromorphic function (i.e. analytic except at isolated poles). Define

$$f(\omega) = \frac{1}{i\pi}\int_a^b \frac{\rho(\omega')}{\omega' - \omega} d\omega'.$$

  When $\omega \in [a, b]$, the integral diverges at $\omega = \omega'$.

- We may regularize the integral by shifting the pole to $\omega' \pm i\epsilon$, or equivalently, shifting ω to $\omega \mp i\epsilon$. However, these two choices go around the pole in different directions, so their difference picks up the residue of the pole,

$$\frac{f(\omega + i\epsilon) - f(\omega - i\epsilon)}{2} = \rho(\omega).$$

  This implies that $f(\omega)$ is discontinuous across the real axis for all $\omega \in [a, b]$. If ρ is everywhere analytic, this is a branch cut.

- The principal value of the integral comes from averaging these two possibilities,

$$\frac{f(\omega+i\epsilon)+f(\omega-i\epsilon)}{2}=\frac{1}{i\pi}\mathcal{P}\int_a^b\frac{\rho(\omega')}{\omega'-\omega}d\omega'.$$

- To get some intuition, note that shifting the pole up/down gives integrands

$$\frac{1}{\omega'-(\omega\pm i\epsilon)}=\frac{\omega'-\omega}{(\omega'-\omega)^2+\epsilon^2}\pm\frac{i\epsilon}{(\omega'-\omega)^2+\epsilon^2}.$$

  Taking the principle value integral is equivalent to working with the average of these integrands, and hence the real part, which is a symmetric regularization of $1/(\omega'-\omega)$. This makes it clear that our definition of the principal value is the same as the usual definition for real integrals,

$$\mathcal{P}\int_a^c f(x)\,dx=\lim_{\epsilon\to0^+}\int_a^{b-\epsilon}f(x)\,dx+\int_{b+\epsilon}^c f(x)\,dx$$

  where $x=b$ is the location of a singularity. More generally, for a singular contour integral the principal value is defined by omitting a disk about each pole.

- If we instead subtract the two integrands, we get the imaginary part, which limits to $\delta(\omega-\omega')$ as $\epsilon\to0$, recovering our first result.

- Note that none of the statements above required analyticity; they also apply for any sufficiently smooth real function.

We now apply our above results to our response function $\chi(\omega)$, with $(a,b)\to(-\infty,\infty)$. We only demand that $\chi(z)$ falls off faster than $1/|z|$ at infinity.

- The fact that $\chi(\omega)$ is analytic in the upper-half plane implies that $f(\omega-i\epsilon)=0$, since the integral is a contour integral that may be closed in the upper-half plane, containing no poles.

- Combining our two main results above, we have

$$\chi(\omega)=\frac{1}{i\pi}\mathcal{P}\int_{-\infty}^{\infty}d\omega'\frac{\chi(\omega')}{\omega'-\omega}.$$

  Taking the real and imaginary parts, we have the Kramers-Kronig relations

$$\operatorname{Re}\chi(\omega)=\mathcal{P}\int_{-\infty}^{\infty}\frac{d\omega'}{\pi}\frac{\operatorname{Im}\chi(\omega')}{\omega'-\omega},\quad\operatorname{Im}\chi(\omega)=-\mathcal{P}\int_{-\infty}^{\infty}\frac{d\omega'}{\pi}\frac{\operatorname{Re}\chi(\omega')}{\omega'-\omega}$$

  which relate the real and imaginary parts of the response function. If one of the two is known, the entire response function can be found.

- Now let $\rho(\omega)=\operatorname{Im}\chi(\omega)$. This is not an analytic function, so we no longer have $f(\omega-i\epsilon)=0$, but it is smooth enough for our other results to hold. Then we have

$$\int_{-\infty}^{\infty}\frac{d\omega'}{i\pi}\frac{\operatorname{Im}\chi(\omega')}{\omega-\omega-i\epsilon}=\operatorname{Im}\chi(\omega)+\mathcal{P}\int_{-\infty}^{\infty}\frac{d\omega'}{i\pi}\frac{\operatorname{Im}\chi(\omega')}{\omega'-\omega-i\epsilon}.$$

  Applying a Kramers-Kronig relation, we find

$$\chi(\omega)=\int_{-\infty}^{\infty}\frac{d\omega'}{\pi}\frac{\operatorname{Im}\chi(\omega')}{\omega'-\omega-i\epsilon}$$

  which shows explicitly how the dissipative part of the response function determines everything.

- The DC susceptibility is the response of a system to a DC field,

$$\chi = \lim_{\omega\to 0}\chi(\omega) = \left.\frac{\partial\langle\mathcal{O}\rangle}{\partial\phi}\right|_{\omega=0}.$$

  Using our above result, we immediately have

$$\chi = \int_{-\infty}^{\infty}\frac{d\omega'}{\pi}\frac{\operatorname{Im}\chi(\omega')}{\omega'-i\epsilon}$$

  Then if we know how much the system absorbs at all energies, we know its response at zero frequency. This is an example of a "thermodynamic sum rule".

Example. The damped harmonic oscillator. The equation of motion is

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = F(t).$$

The response function is the Green's function of the system,

$$x(t) = \int_{-\infty}^{\infty} dt'\,\chi(t-t')F(t').$$

To find the response function, we take the Fourier transform of the above for

$$x(t) = \int đ\omega dt'\, e^{-i\omega(t-t')}\chi(\omega)F(t').$$

Applying $\partial_t^2 + \gamma\partial_t + \omega_0^2$ to both sides, using $\partial_t \leftrightarrow -i\omega$, we have

$$F(t) = \int đ\omega dt'\, e^{-i\omega(t-t')}(-\omega^2 - i\gamma\omega + \omega_0^2)\chi(\omega)F(t').$$

For this to be true, the $d\omega$ integral must give a delta function, so we must have

$$\chi(\omega) = \frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}$$

on the real axis, which is extended to the complex plane in the usual way. This is the desired response function, i.e. the retarded Green's function.

- The susceptibility is $\chi(\omega = 0) = 1/\omega_0^2$. Indeed, $x = F/\omega_0^2$ as expected.
- The poles are located at

$$\omega_* = -\frac{i\gamma}{2} \pm \sqrt{\omega_0^2 - \gamma^2/4}$$

  and these are indeed located outside the upper-half plane.

- The reactive part of the response function is

$$\operatorname{Re}\chi(\omega) = \frac{\omega_0^2 - \omega^2}{(\omega_0^2-\omega^2)^2 + \gamma^2\omega^2}.$$

  This measures the system's 'in phase' response to a given frequency; note that it is zero at resonance since the response there is $\pi/2$ out of phase.

- The dissipative part of the response function is

$$\operatorname{Im}\chi(\omega) = \frac{\omega\gamma}{(\omega_0^2-\omega)^2+\gamma^2\omega^2}$$

  which peaks at resonance, as expected.

- In the limit $\gamma \to 0$, the poles approach the real axis at $\omega = \pm\omega_0$, sitting just below it. The dissipative part of the response function tends to two delta functions at $\omega = \pm\omega_0$.

Example. The undamped harmonic oscillator. If we carry out the same logic as before, we find

$$\chi(\omega) = \frac{1}{\omega_0^2-\omega^2}$$

with the time-domain response function being

$$\chi(t) = \int đ\omega\, \frac{e^{-i\omega t}}{\omega_0^2-\omega^2}.$$

This integral is not well-defined, due to the divergences at $\omega = \pm\omega_0$, so the expression for $\chi(t)$ is ambiguous. This is simply because we have a choice between retarded and advanced Green's functions. If we integrate above the poles (or equivalently push the poles downward) we get the retarded Green's function

$$\chi_r(\omega) = \frac{1}{\omega_0^2-(\omega+i\epsilon)^2} = \frac{1}{\omega_0^2-\omega^2-2i\omega\epsilon}.$$

Comparing this to our damped harmonic oscillator result, we see that in the time domain, the $i\epsilon$ adds damping to make the Green's function go to zero at $t \to \infty$. The advanced Green's function

$$\chi_a(\omega) = \frac{1}{\omega_0^2-(\omega-i\epsilon)^2}$$

instead has poles in the upper-half plane, and has 'anti-damping' to go to zero at $t \to -\infty$. Taking the principal value gives the average of the two, which corresponds to time-symmetric damping.

Note that in the case of finite damping, the time-domain Green's function is unambiguous: only the retarded Green's function exists, since the advanced Green's function blows up. In this case the presence or absence of the $i\epsilon$ makes no difference.

Example. We can explicitly link $\operatorname{Im}\chi(\omega)$ to energy dissipation. The dissipation rate dW/dt is

$$\frac{dW}{dt} = F(t)\dot{x}(t) = F(t)\frac{d}{dt}(\chi \star F)(t)$$

where $\star$ is convolution. Fourier transforming, the time derivative becomes $-i\omega$, giving

$$\frac{dW}{dt} = \int đ\omega đ\omega'\,(-i\omega)e^{-i(\omega+\omega')t}\chi(\omega)F(\omega)F(\omega').$$

Now we drive the system with a force of the form

$$F(t) = F_0\cos\Omega t \quad \leftrightarrow \quad F(\omega) = 2\pi F_0(\delta(\omega-\Omega)+\delta(\omega+\Omega)).$$

Note that we can't use the common complex exponential form $F_0 e^{-i\Omega t}$ because our equation isn't linear in F. (The exponential trick only works when everything is linear, because $\operatorname{Re} z \operatorname{Re} w \neq \operatorname{Re} zw$.) On a deeper level, the assumption that F was real was our starting point above, without which our nice analytic results fail.

Plugging this in, we find

$$\frac{dW}{dt} = -iF_0^2\Omega(\chi(\Omega)e^{-i\Omega t} - \chi(-\Omega)e^{i\Omega t})(e^{-i\Omega t} + e^{i\Omega t}).$$

Taking a time average, we have

$$\overline{\frac{dW}{dt}} = -iF_0^2\Omega(\chi(\Omega) - \chi(-\Omega)) = 2F_0^2\Omega \operatorname{Im}\chi(\Omega)$$

where we used the fact that $\operatorname{Re}\chi(\omega)$ is real and $\operatorname{Im}\chi(\omega)$ is odd. Then the rate of energy dissipation is proportional to $\operatorname{Im}\chi(\omega)$, as expected.

Note. We should be more careful with the terminology above. We have really calculated the rate at which the system absorbs energy from the driving force. This happens to be equal to the rate at which the system dissipates energy in the steady state, where here we mean dissipation as irreversible energy loss into the environment.

Now consider the quantum case. If we consider a small quantum system, such as a two-state system, it doesn't make sense to say it dissipates in the thermodynamic sense; we really mean that it just absorbs energy. However, in condensed matter, we are often concerned with very large (i.e. $N \sim 10^{23}$) quantum systems, so it does make sense to think about absorption as thermodynamic dissipation. The energy gets "lost" in the many degrees of freedom in the system, and never returns for entropic reasons.

6.3 The Kubo Formula

We now return to quantum mechanics and prove the Kubo formula.

- We work in interaction picture, so the operators $\mathcal{O}_i$ evolve by the undriven Hamiltonian, and the states evolve by the external perturbation,

  $$U(t, t_0) = T\exp\left(-i\int_{t_0}^{t} H_s(t')\,dt'\right).$$

  We assume that in the distant past $t_0 \to -\infty$ the state is ρ_0, so

  $$\rho(t) = U(t)\rho_0 U^{-1}(t), \quad U(t) = U(t, -\infty).$$

- To first order, the expectation of $\mathcal{O}_i$ in the presence of sources $\phi(t)$ is

  $$\langle\mathcal{O}_i(t)\rangle|_\phi = \operatorname{tr}\rho(t)\mathcal{O}_i(t) = \operatorname{tr}\rho_0 U^{-1}(t)\mathcal{O}_i(t)U(t) \approx \operatorname{tr}\rho_0\left(\mathcal{O}_i(t) + i\int_{-\infty}^{t} dt'\,[H_s(t'), \mathcal{O}_i(t')]\right).$$

  The first term is just the expectation value of $\mathcal{O}_i$ in the presence of no sources, so

  $$\delta\langle\mathcal{O}_i\rangle = i\int_{-\infty}^{t} dt'\,\langle[\mathcal{O}_j(t'), \mathcal{O}_i(t)]\rangle\phi_j(t')$$

  where the expectation value is taken with respect to ρ_0.

- Comparing this to the definition of the response function, we have the Kubo formula

$$\chi_{ij}(t-t') = -i\theta(t-t')\langle[\mathcal{O}_i(t), \mathcal{O}_j(t')]\rangle.$$

Equivalently, the susceptibility in frequency space is

$$\chi_{ij}(\omega) = -i\int_0^\infty dt\, e^{i\omega t}\langle[\mathcal{O}_i(t), \mathcal{O}_j(0)]\rangle.$$

Causality is built in because we imposed boundary conditions in the past.

- We can get a bit more insight by going to spectral representation. Suppose we begin in the thermal state $\rho_0 = e^{-\beta H}$. Then

$$\chi_{ij}(\omega) = -i\int_0^\infty dt\, e^{i\omega t}\,\mathrm{tr}\, e^{-\beta H}[\mathcal{O}_i(t), \mathcal{O}_j(0)].$$

To simplify this, we insert copies of the identity in the unperturbed energy eigenbasis, noting that $\mathcal{O}_i(t) = U^{-1}\mathcal{O}_i(0)U$ where $U = e^{-iHt}$. Then we have

$$\chi_{ij}(\omega) = -i\int_0^\infty dt\, e^{i\omega t}\sum_{mn} e^{-E_m\beta}\left[(\mathcal{O}_i)_{mn}(\mathcal{O}_j)_{nm}e^{i(E_m-E_n)t} - (\mathcal{O}_j)_{mn}(\mathcal{O}_i)_{nm}e^{i(E_n-E_m)t}\right]$$

where the matrix elements are in the energy basis.

- This integral does not converge, so we add an $i\epsilon$ damping, to find

$$\chi_{ij}(\omega+i\epsilon) = \sum_{mn}\frac{(\mathcal{O}_i)_{mn}(\mathcal{O}_j)_{nm}}{\omega + E_m - E_n + i\epsilon}(e^{-E_m\beta} - e^{-E_n\beta}).$$

The response function has poles just below the real axis, as expected by causality. Each pole in the response function adds a delta function to the dissipation, as we saw in the $\gamma \to 0$ limit of the harmonic oscillator, reflecting the fact that the system can absorb energy by performing a transition. For a quantum field theory the poles may merge into a branch cut.

Example. Revisiting dissipation. The rate of energy absorption is

$$\frac{dW}{dt} = \frac{d}{dt}\,\mathrm{tr}\,\rho H = \mathrm{tr}(\dot\rho H + \rho\dot H).$$

Since this is a physical observable, it may be computed in any picture, so we switch to Schrodinger picture. Then $i\dot\rho = [H, \rho]$, so the first term is zero, and $\dot H = \mathcal{O}\dot\phi(t)$, where we're working with one source for simplicity. Therefore

$$\frac{dW}{dt} = \mathrm{tr}(\rho\mathcal{O}\dot\phi) = \langle\mathcal{O}\rangle|_\phi\dot\phi = (\langle\mathcal{O}\rangle|_{\phi=0} + \delta\langle\mathcal{O}\rangle)\,\dot\phi.$$

Taking a periodic source $\phi(t) = \phi_0\cos(\Omega t)$ and time averaging, the first term vanishes, leaving

$$\overline{\frac{dW}{dt}} = \frac{1}{T}\int_0^T dt\int dt'\,\chi(t-t')\phi(t')\dot\phi(t).$$

By manipulations similar to the classical case, we find

$$\overline{\frac{dW}{dt}} = 2\Omega|\phi_0|^2\,\mathrm{Im}\,\chi(\Omega)$$

which is a nearly identical result. For a large quantum system, thermodynamics requires $\overline{dW}/dt \geq 0$, so $\Omega\,\mathrm{Im}\,\chi(\Omega)$ must be nonnegative. More generally, $\Omega\,\mathrm{Im}\,\chi_{ij}(\Omega)$ must be positive semidefinite.

Example. The fluctuation-dissipation theorem. The simplest kind of correlation function we could consider is the two-point correlator

$$S_{ij}(t) = \langle \mathcal{O}_i(t)\mathcal{O}_j(0)\rangle$$

which roughly measures the correlation of fluctuations in $\mathcal{O}_i$ and $\mathcal{O}_j$. As in the previous example, we work in a thermal ensemble $\rho_0 = e^{-\beta H}$. Expanding the definition, we have

$$S_{ij}(t) = \operatorname{tr} e^{-\beta H}\mathcal{O}_i(t)\mathcal{O}_j(0) = \operatorname{tr} e^{-\beta H}\mathcal{O}_i(t)e^{\beta H}e^{-\beta H}\mathcal{O}_j(0) = \operatorname{tr} \mathcal{O}_i(t+i\beta)e^{-\beta H}\mathcal{O}_j(0)$$

where we regarded $e^{-\beta H}$ as generating time translation in imaginary time. Then

$$S_{ij}(t) = S_{ji}(-t - i\beta)$$

where we used time-translation invariance. This result is more intuitive in Fourier space, where

$$S_{ij}(\omega) = S_{ji}(-\omega)e^{\beta\omega}.$$

Physically, this is the statement that detailed balance holds in thermal equilibrium.

We can now use these identities to relate the two-point correlator to the imaginary part of the response function. Starting with

$$\chi''_{ij}(t) = -\frac{i}{2}\left(\chi_{ij}(t) - \chi_{ji}(-t)\right)$$

and applying the Kubo formula and time translation invariance, we find

$$\chi''_{ij}(t) = -\frac{1}{2}(S_{ij}(t) - S_{ji}(-t)).$$

Taking the Fourier transform and applying our S_{ij} identities gives the fluctuation-dissipation theorem,

$$\chi''_{ij}(\omega) = -\frac{1}{2}(1 - e^{-\beta\omega})S_{ij}(\omega).$$

This result links fluctuations in equilibrium, captured by $S_{ij}(\omega)$, to the dissipation rate near equilibrium, captured by $\chi''_{ij}(\omega)$. The physical meaning is clearer if we invert the expression, for

$$S_{ij}(\omega) = -2(n_B(\omega) + 1)\chi''_{ij}(\omega), \quad n_B(\omega) = (e^{\beta\omega} - 1)^{-1}.$$

We see that there are two separate contributions to the fluctuations; the $n_B(\omega)$ is due to thermal effects, and the extra $+1$ can be thought of as due to inherently quantum fluctuations. In the classical limit, the first term dominates, giving

$$S_{ij}(\omega) = -\frac{2kT}{\omega}\chi''_{ij}(\omega)$$

which is the classical fluctuation-dissipation theorem.