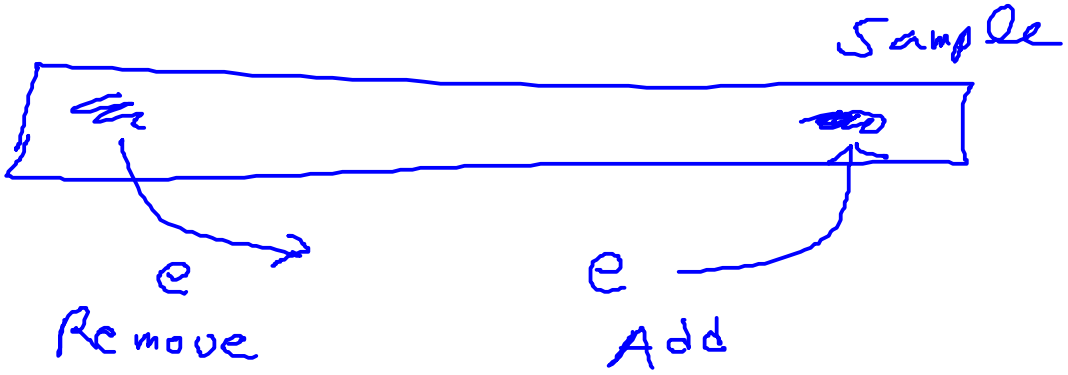
Lecture 16

Tuesday, March 06, 2012

How to distinguish metals and non-metals?

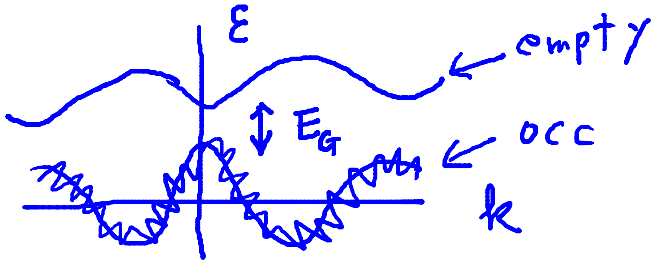
The defining characteristics of the metal and non-metal is the presence of an energy gap. What is the energy gap?



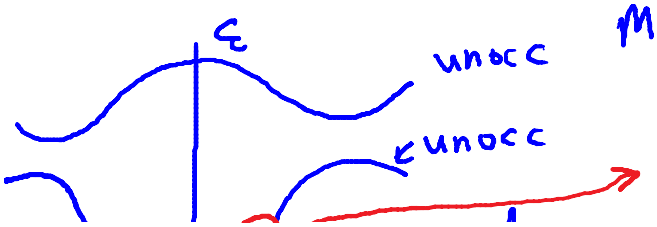
Consider the above procedure. You remove an electron from one part of the sample and then add it back to another part of the sample. Assume that these two parts are macroscopically separated so that they can be considered as two identical but separate samples. The energy gap is the minimum energy cost to accomplish this procedure.

This definition is very general, and applies to any materials, whether that material is described by the band theory (as we explored in last classes) or not.

(1) Within the band theory, the energy gap (E_G) is the energy separation between the top of the occupied state and the bottom of the un-occupied state.

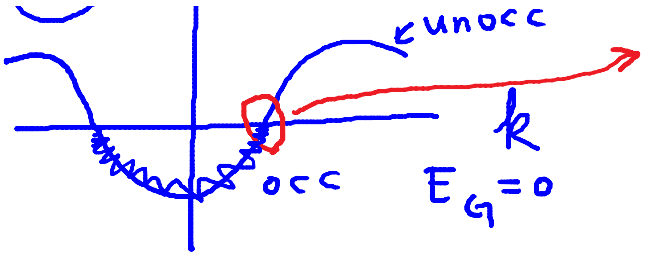


Insulator
Semiconductor



Metal

An unoccupied state is available right above the top of the occupied state. Thus, energy gap = 0



An unoccupied state is available right above the top of the occupied state. Thus, energy gap = 0.

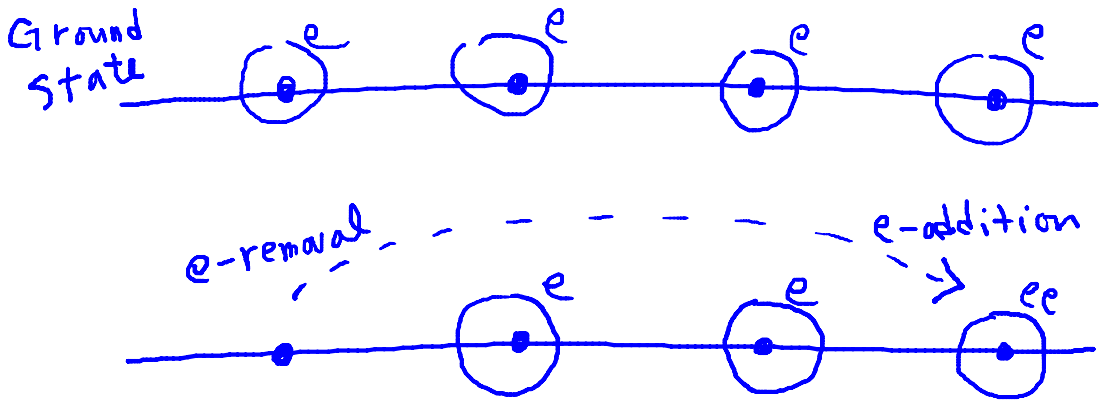
In the semiconductor field, the completely filled band above is called the **valence band**, while the completely empty band above it is called the **conduction band**.

(2) Now, as a different example, consider a H crystal. A 1D crystal. This is a hypothetical 1D crystal, but it is nevertheless a good example to consider. According to the band theory, this must be a metal, due to Wilson's rule. One electron per unit cell (per H), and thus we will necessarily be left with some partially filled band. Within the tight-binding theory that we did, the lowest lying 1s band is half filled, and all other bands (that we did not consider since they lie at high energies) are completely empty. And yet, the real H crystal that one makes in a lab (under moderate pressure) is an insulator. Why? It turns out that the electron-electron interaction energy that one ignores in the band theory is of more importance in this case. So, it is not correct to think of these electrons occupying a band. Instead it is correct to think of the ground state of a H crystal as something dramatically different -- a simple collection of H atoms! Each electron is bound to a proton, and there is no hopping of that electron to a neighboring site.

In order to realize that such a state must be a physical state can be realized from this reasoning. Consider any crystal, and imagine that you are free to expand the crystal so that the lattice constant can become arbitrarily large. When the lattice constant is very large (say 1 meter!), it is obvious that the physically correct picture is a simple collection of atoms, rather than electrons delocalized between atoms as in the band theory. So, any band theory will fail in the limit of large lattice constant. In the limit of large lattice constant, Coulomb interactions between electrons are much more important than band theory energy scale (hopping energy scale "t" which becomes very small).

Under natural conditions, some materials are in the regime where the band theory is a good description, while some other materials are in the regime where the band theory is not a good description. The H crystal above is an example of the latter. Materials such as MnO, CuO, MnF2, NiO, CeSi2, … (pretty much any material that involves 3d transition metal element or 4f rare earth element) are also examples of the latter, which can be called the "**strongly correlated electron materials**."

Back to the H crystal. The energy gap here is the so-called U, electron-electron repulsion. This is the concept of the **Mott insulator**.



The only difference between the ground state and the excited state (created by removing an electron and adding it to a different site) is the Coulomb interaction between two electrons on the same site, which we call "**Hubbard** U".

How to distinguish metals and non-metals (2)?

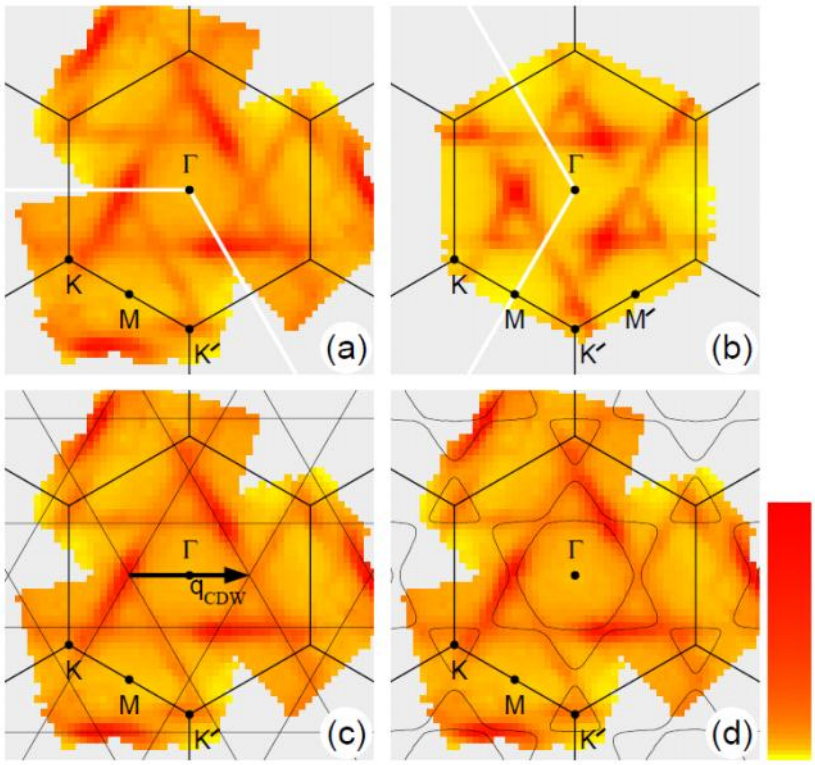
Related to the above general discussion, there is another method to distinguish metals from non-metals. **Metals have Fermi surface, while non-metals don't**. This statement is true for any materials, strongly correlated electron materials or not. It is however very easy to see this point for materials that are described well by the band theory, and so much discussion in the rest of this course (except magnetism and superconductivity) will focus on them.

Insulators and semi-conductors have no Fermi surface at zero temperature. Semi-metals do not have Fermi surface at zero temperature; instead they have Fermi "points", the limit of finite volume/area Fermi surfaces just before they disappear. In practice, if the Fermi surface is very small, then the material can be considered a semi-metal.

How do we know that the FS exists?

There is a rather direct way of observing a Fermi surface. This method is angle resolved photoelectron spectroscopy (ARPES), which is the method of choice for yours truly. This method uses the photoelectric effect, to excite electrons out of the specimen with high flux beams of ultraviolet light. The interpretation of the photoelectric effect (electron + photon → more energetic electron) mainly

involves only the energy conservation principle and the momentum conservation principle. Essentially, the electron that is taken out from the sample remembers the energy and the momentum it used to have inside the sample, and it is elementary for experimenters to compute them since the photon energy and the photon momentum involved in an ARPES experiment are very accurately known. The simplicity of the interpretation gives great power to the ARPES technique. From mid 90's, it has become quite routine to obtain "Fermi surface" maps, the intensity map at the chemical potential, where the high intensity region delineates the Fermi surface shape. The following is an example of such an early ARPES Fermi surface map, which was taken by yours truly for his Ph. D. thesis. The materials here are "Na/K purple bronzes" which have a strongly two dimensional electronic structure in a hexagonal unit cell. Shown here are the hexagonal BZ, ARPES maps, and theoretical predictions (c and d).



Other methods of probing the Fermi surface

ARPES works very well for layered materials with quasi-2D electronic structure. These are crystals that are built up by loosely coupled layers, and the electrons spend much of their time within the layer than hopping between layers. This class of materials include high temperature superconductors and graphene/graphite, and so ARPES continues to receive much attention.

However, the applicability of ARPES to more traditional 3D materials is weak.

What are other methods of probing the Fermi surface? Quantum oscillation measurements at high magnetic field is a very sharp measurement tool. Also, in recent days, the scanning tunneling spectroscopy has been used to infer the Fermi surface shape. Other effects such as Kohn anomaly (homework 4.5) can give information about the Fermi surface. None of these methods are as direct as ARPES in the sense that, while ARPES can give the shape of the Fermi surface without any band theory input, all these methods must use some prior knowledge from theory in order to interpret the results since they cannot obtain the absolute position of the Fermi surface and the exact shape of the Fermi surface. Nevertheless, these methods are very useful, and more advantageous to the ARPES technique in terms of applicability to 3D materials. Of these, we will review the quantum oscillation measurements below. Before doing that, though, we need to do establish some important facts about how the electron (or the hole) moves under an applied field.

The semi-classical equation of motion

In real situations, a mono-chromatic wave with a very sharply defined crystal momentum is more a fantasy than a reality. Even when a theoretical description prefers to use a plane wave or a infinitely sharply defined Bloch state, one is advised to always remember that physical situations always involve wave packets due to intrinsic effects such as finite lifetime and extrinsic effects such as experimental conditions imposed (finite sample size, smearing by instrumental resolutions, etc.).

Let us then consider an electron wave packet, which is built up with a distribution of wave vectors, say centered at $\vec{k}$ with a width scale $\Delta\vec{k}$. When the spatial extent of this wave packet $\sim 1/\Delta k$, determined by Heisenberg uncertainty principle, is much smaller than the length scale (the wave length or the decay length) of the applied field, then we have the semi-classical approximation.

In addition, we also generally assume that Δk is not too large. If Δk is too large, then the spatial extent of the wave packet can be as small as the lattice constant, in which case the band structure begins to blur out severely. So, we assume that Δk small compared to the wave vector scale for the energy dispersion itself. If follows from this assumption that the wave packet moves with the group velocity $\vec{v}_g = \frac{1}{\hbar}\frac{d\epsilon_k}{d\vec{k}}$.

In this case, the following semi-classical equation of motion applies.

$$\hbar \frac{d\vec{k}}{dt} = \vec{F}(\vec{r}, \vec{v}_g)$$

where $\vec{r}$ is the mean position, and $\vec{v}_g$ is the group velocity ($\partial\epsilon/\partial\vec{k}$), of the wave packet, and $\vec{F}$ is the classical force at the position $\vec{r}$ and the velocity $\vec{v}_g$. For a particle with charge q, in the presence of the $\vec{E}, \vec{B}$ fields:

$$\hbar \frac{d\vec{k}}{dt} = q\vec{E} + \frac{q}{c}\, \vec{v}_g \times \vec{B}$$

Let us see how this can be derived. Consider $H = \frac{1}{2m}\left(\vec{p} - \frac{q}{c}\vec{A}\right)^2 + q\phi$ where we assume that the length scales associated with $\vec{A}$ and ϕ (wave lengths) are much larger than $1/\Delta k$ of the wave packet. Otherwise, we leave $\vec{A}, \phi$ as unrestricted: e.g. they can be time dependent.

Consider a $T_{\vec{R}}$ operator, where $T_{\vec{R}} f(\vec{r}) = f(\vec{r} - \vec{R})$, i.e. $T_{\vec{R}}$ is the translation by $\vec{R}$. Our restriction on $\vec{R}$ is that it is a small lattice vector such that $R \ll$ length scales of the field.

We start from QM:

$$\frac{d\langle T_{\vec{R}}\rangle}{dt} = \frac{i}{\hbar}\langle\left[H, T_{\vec{R}}\right]\rangle$$

The left hand side here is trivial. $\langle T_{\vec{R}}\rangle \approx \langle e^{-i\vec{k}\cdot\vec{R}}\rangle = e^{-i\vec{k}\cdot\vec{R}}$.

So, $\frac{d\langle T_{\vec{R}}\rangle}{dt} \approx -i\dot{\vec{k}} \cdot \vec{R} e^{-i\vec{k}\cdot\vec{R}}$.

This is the leading order contribution, a term linear in $\vec{R}$. Our next task is to get the term of the same order on the right hand side.

Let us first consider the 2nd term in the Hamiltonian.

$$\frac{i}{\hbar}\langle\left[q\phi, T_{\vec{R}}\right]\rangle = \frac{i}{\hbar} q\langle\phi T_{\vec{R}} - T_{\vec{R}}\phi\rangle = \frac{iq}{\hbar}\langle\{\phi(\vec{r}) - \phi(\vec{r} - \vec{R})\}T_{\vec{R}}\rangle = \frac{iq}{\hbar}\vec{R}\cdot\langle\nabla\phi\, T_{\vec{R}}\rangle$$

$$\approx \frac{iq}{\hbar}\vec{R}\cdot\nabla\phi e^{-i\vec{k}\cdot\vec{R}}$$

[BEGIN]: Optional reading. The above is enough to show $\dot{\vec{k}} = q\vec{E}$.

Now, consider the first term in the Hamiltonian, which is more complicated.

$$\frac{i}{2m\hbar}\left\langle\left(\vec{p} - \frac{q}{c}\vec{A}\right)^2 T_{\vec{R}} - T_{\vec{R}}\left(\vec{p} - \frac{q}{c}\vec{A}\right)^2\right\rangle$$

$$= \frac{i}{2m\hbar}\left\langle\left\{\left(\vec{p} - \frac{q}{c}\vec{A}\right)^2 - \left(\vec{p} - \frac{q}{c}\vec{A}(\vec{r} - \vec{R})\right)^2\right\}T_{\vec{R}}\right\rangle$$

$$\approx \frac{i}{2m\hbar}\left\langle\left\{\left(\vec{p} - \frac{q}{c}\vec{A}\right)^2 - \left(\vec{p} - \frac{q}{c}\vec{A} + \frac{q}{c}(\vec{R}\cdot\nabla)\vec{A}\right)^2\right\}T_{\vec{R}}\right\rangle$$

$$\approx \frac{i}{2m\hbar}\left\langle\left\{-\frac{q}{c}\left(\vec{p}-\frac{q}{c}\vec{A}\right)\cdot(\vec{R}\cdot\nabla)\vec{A}-\frac{q}{c}\left((\vec{R}\cdot\nabla)\vec{A}\right)\cdot\left(\vec{p}-\frac{q}{c}\vec{A}\right)\right\}T_{\vec{R}}\right\rangle$$

$$\approx \frac{i}{2m\hbar}(-2)\frac{q}{c}\, m\vec{v}_g\cdot(\vec{R}\cdot\nabla)\vec{A}\, e^{-i\vec{k}\cdot\vec{R}}$$

$$= -\frac{iq}{\hbar c}\left\{\vec{R}\cdot\left(\vec{v}_g\times\vec{B}\right)+\vec{R}\cdot\left(\frac{d\vec{A}}{dt}-\frac{\partial\vec{A}}{\partial t}\right)\right\}e^{-i\vec{k}\cdot\vec{R}}$$

$\vec{p}-\frac{q}{c}\vec{A}\approx m\vec{v}_g$ for the wave packet

$\vec{v}\cdot(\vec{R}\cdot\nabla)\vec{A}=v_jR_i\partial_iA_j=\delta_{il}\delta_{jk}v_kR_l\partial_iA_j=(\epsilon_{mij}\epsilon_{mlk}+\delta_{ik}\delta_{jl})v_kR_l\partial_iA_j$

$=(\vec{R}\times\vec{v})\cdot(\nabla\times\vec{A})+\vec{R}\cdot(\vec{v}\cdot\nabla)\vec{A}=(\vec{R}\times\vec{v})\cdot\vec{B}+\vec{R}\cdot\left(\frac{d\vec{A}}{dt}-\frac{\partial\vec{A}}{\partial t}\right)=\vec{R}\cdot(\vec{v}\times\vec{B})+\vec{R}\cdot\left(\frac{d\vec{A}}{dt}-\frac{\partial\vec{A}}{\partial t}\right)$

So, collecting all terms:

$$-i\dot{\vec{k}}\cdot\vec{R}=\frac{iq}{\hbar}\vec{R}\cdot\nabla\phi-\frac{iq}{\hbar c}\left\{\vec{R}\cdot\left(\vec{v}_g\times\vec{B}\right)+\vec{R}\cdot\left(\frac{d\vec{A}}{dt}-\frac{\partial\vec{A}}{\partial t}\right)\right\}$$

Taking $\vec{R}$ to be $\vec{a},\vec{b},\vec{c}$, of the Bravais lattice, we have the vector identity:

$$\hbar\dot{\vec{k}}=-q\nabla\phi+\frac{q}{c}\left(\vec{v}_g\times\vec{B}+\frac{d\vec{A}}{dt}-\frac{\partial\vec{A}}{\partial t}\right)$$

With $\vec{E}=-\nabla\phi-\frac{1}{c}\frac{\partial\vec{A}}{\partial t}$, we have $\frac{d}{dt}\left(\hbar\vec{k}-\frac{q}{c}\vec{A}\right)=q\vec{E}+\frac{q}{c}\vec{v}_g\times\vec{B}$.

While this form is consistent with the classical limit ($\hbar\vec{k}$ is an average of the canonical momentum for each plane wave component of the wave packet, up to a constant reciprocal vector, and so $\frac{d}{dt}\left(\hbar\vec{k}-\frac{q}{c}\vec{A}\right)\approx\frac{d}{dt}\left(m\vec{v}_g\right)$), it is <u>not</u> $\frac{d}{dt}\left(\hbar\vec{k}\right)=\vec{F}\left(\vec{r},\vec{v}_g\right)=q\vec{E}+\frac{q}{c}\vec{v}_g\times\vec{B}$. Also, our derivation resulted in a gauge dependent form, which is clearly not desirable. The resolution of this problem is complicated, and is sketched in the following paragraph.

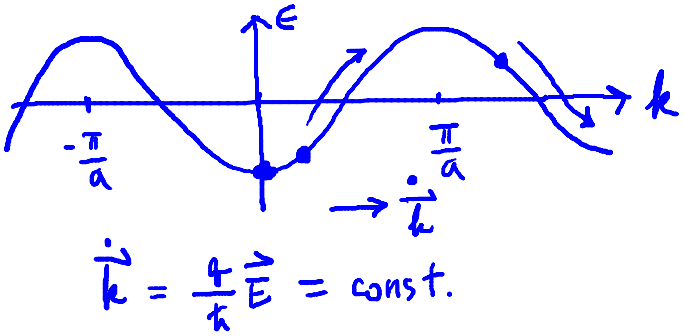
According to Zak, Phys. Rev. 168, 686 ('68), the semi-classical equation of motion should read as $\hbar\frac{d}{dt}\left[\vec{k}\right]=q\vec{E}+\frac{q}{c}\left[\frac{\partial\epsilon_n(\vec{k})}{\partial\vec{k}}\right]\times\vec{B}$ where [] means replacing $\vec{k}\to\vec{k}-i\frac{q}{2\hbar c}\vec{B}\times\frac{\partial}{\partial\vec{k}}$ (his derivation is for a constant $\vec{B}$ field and single band). $\vec{A}=\frac{1}{2}\vec{B}\times\vec{r}$ for a constant $\vec{B}$ field, and so $\vec{k}-i\frac{q}{2\hbar c}\vec{B}\times\frac{\partial}{\partial\vec{k}}$ can be thought of as related to $\vec{k}-\frac{q}{\hbar c}\vec{A}$ since one might expect $\vec{r}$ to be something like $i\frac{\partial}{\partial\vec{k}}$. As a matter of fact, Zak works in the representation that diagonalize both $T\left(\vec{R}\right)=e^{-i\vec{p}\cdot\vec{R}/\hbar}$ (translation operator in real space) and $T^*\left(\vec{G}\right)=e^{-i\vec{x}\cdot\vec{G}}$ (translation operator in momentum space). Since these two operators commute, it is possible to find the $\vec{k},\vec{q}$ representation, where $\vec{k}$ labels the eigenvalue of $T\left(\vec{R}\right)$, $e^{-i\vec{k}\cdot\vec{R}}$, and $\vec{q}$ labels the eigenvalue of $T^*\left(\vec{G}\right)$, $e^{-i\vec{q}\cdot\vec{G}}$. In this representation, he obtains $\vec{p}=-i\hbar\frac{\partial}{\partial\vec{q}}$ and $\vec{r}=i\frac{\partial}{\partial\vec{k}}+\vec{q}$. He further finds that, in the presence of the

magnetic field, the semi-classical equation $\hbar\frac{d}{dt}\vec{k} = q\vec{E} + \frac{q}{c}\, v_g(\vec{k}) \times \vec{B}$ has to be interpreted as $\vec{k}$ meaning $\left[\langle\vec{k}\rangle\right]$, where [] means the replacement of $\vec{k} \to \vec{k} - i\frac{q}{2\hbar c}\vec{B} \times \frac{\partial}{\partial\vec{k}}$. So, it is not really $\vec{k}$ when there is a magnetic field. However, this can be understood as a simple "re-mapping" or "re-labeling" of $\vec{k}$. Besides, the magnitude of the 2nd term is generally very small:
$\frac{q}{2\hbar c}Ba \sim \frac{e\hbar}{2mc}B\frac{mc^2}{(\hbar c)^2}a \sim \mu_B B\ \frac{1}{4\text{ eV Å}} \sim O(10^{-3})\ \text{Å}^{-1}$ (a is the lattice constant $\sim O\left(\frac{\partial}{\partial\vec{k}}\right)$, and μ_B is the Bohr magneton $= 6 \times 10^{-9}$ eV/gauss. The strongest B field that one can generate in a laboratory $\sim 10^6$ gauss; On the other hand, the term $\frac{q}{\hbar c}A \sim \frac{q}{\hbar c}Br$ is *not* necessarily small, since $r \gg a$.)
[END]: Optional reading.

Bloch Oscillation

Surprisingly, the semi-classical equation above predicts that if a partially filled band is subjected to a constant DC electric field, then an AC response may be obtained.

This is easy to see if there was one wave packet moving on a single band in one dimension.



According to the semi-classical EOM, **the wave vector simply changes at a constant rate**, under a constant $\vec{E}$ field, but without the $\vec{B}$ field. So, given a band sketched above, the wave packet will simply trace the dispersion relation, going up and coming down and repeating indefinitely, while the wave vector changes steadily. Why does this mean an AC current? The current density is $nq\vec{v}_g$ ($q = -e$ for the electron wave packet), where $n =$ number density, and $\vec{v}_g$ is the group velocity. Notice how the group velocity increases to a positive number, decreases to zero at $k = \pi/a$, and then reverses its direction, and then decreases further before increasing to 0 again ($k = 2\pi/a$). Thus the current is definitely an AC current. Given the band structure (which may not be a single cosine function in general), we expect that the AC current will contain harmonics of the $2\pi/T$,

where T is the period: the time it takes for $\Delta k = 2\pi/a$. Since the rate at which k changes is constant ($|qE|/\hbar$), $T = \frac{2\pi\hbar}{a|qE|}$. This is the so-called **Bloch oscillation**.

Each wave packet causes such an AC current, and the frequency of the current will not change when the same type of AC current is caused by a very many number of electron wave packets.

When would a Bloch oscillation be observed? $T \ll \tau$, where τ is the relaxation time. For a small T, one would like to have a large a (a nanoscale superstructure -- an artificial crystal -- rather than a natural crystal) and a large $|E|$ (semi-conductor instead of a metal). In semiconductors, τ is on the order of pico-seconds, and a superstructure with $a \sim 50$ Å and $E \sim 5 \times 10^4$ V/cm, giving $T \sim 0.2$ pico-seconds, will be enough.

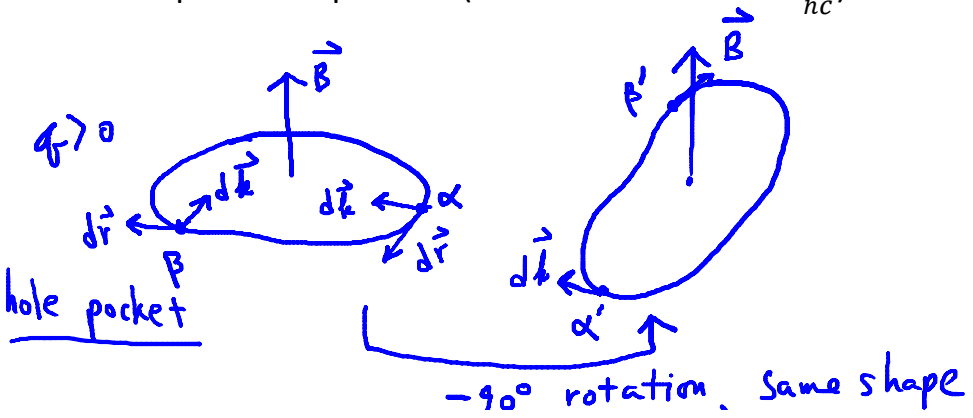
Motion of charged particle in a $\vec{B}$ field

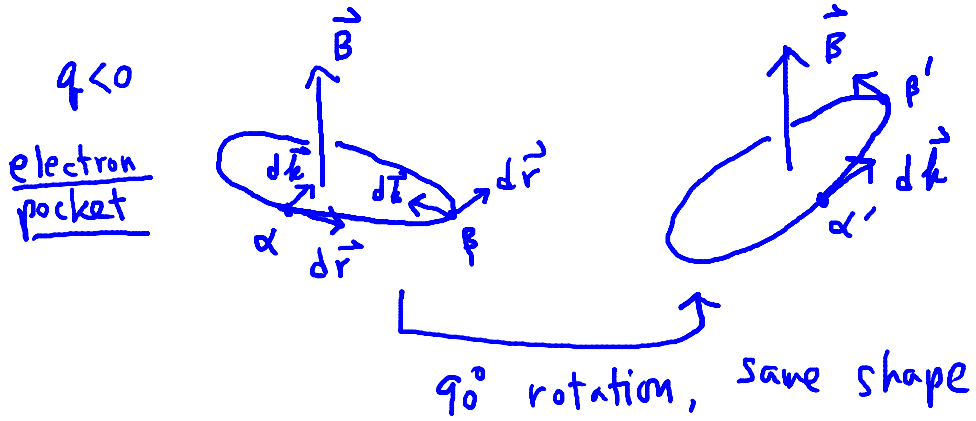
Let us consider a complementary situation when $\vec{E} = 0$ but $\vec{B} =$ constant. Classically, we know that this means a circular motion with the cyclotron frequency $\omega_c = \frac{qB}{mc}$. We shall now see what happens in the quantum case.

First, let us look at the equation of motion: $\hbar\dot{\vec{k}} = \frac{q}{c}\vec{v} \times \vec{B}$. Rewriting this $\hbar\frac{d\vec{k}}{dt} = \frac{q}{c}\frac{d\vec{r}}{dt} \times \vec{B}$, we get

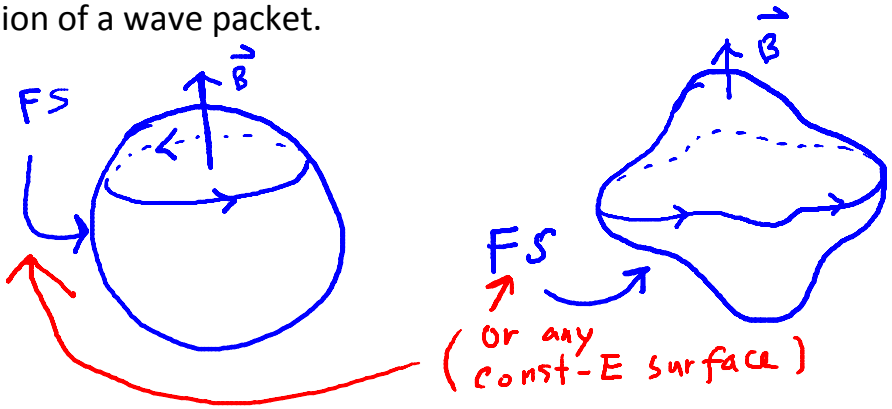
$$d\vec{k} = \frac{q}{\hbar c} d\vec{r} \times \vec{B}$$

What does this mean? Suppose $q > 0$. Let us say that there is a real space path that is followed by this wave packet ($d\vec{r}$). We consider a planar motion only, where the plane is perpendicular to the field. [The particle can have a finite constant velocity parallel to the field.] This equation means that the path of the particle in that plane is replicated (with a different scale $\frac{qB}{\hbar c}$).





It is important to realize that the shape of the orbit above is determined by the crossection of the constant energy surface. Why? (1) $\hbar\dot{\vec{k}} = \frac{q}{c}\vec{v} \times \vec{B}$ means that $\dot{\vec{k}} \cdot \vec{B} = 0$, meaning that the $\vec{k}$ component along the direction of $\vec{B}$ is constant (so the motion in $\vec{k}$ space is strictly two dimensional, while the motion in $\vec{r}$ space can be three dimensional with a constant velocity along the direction of $\vec{B}$). (2) $\frac{d\epsilon_{\vec{k}}}{dt} = \frac{\partial \epsilon_{\vec{k}}}{\partial \vec{k}} \cdot \dot{\vec{k}} = \vec{v} \cdot \dot{\vec{k}} = \frac{q}{c}\vec{v} \cdot \left(\vec{v} \times \vec{B}\right) = 0$. That is, the magnetic field doesn't do any work, as well-known. So a given band structure completely determines the motion of a wave packet.

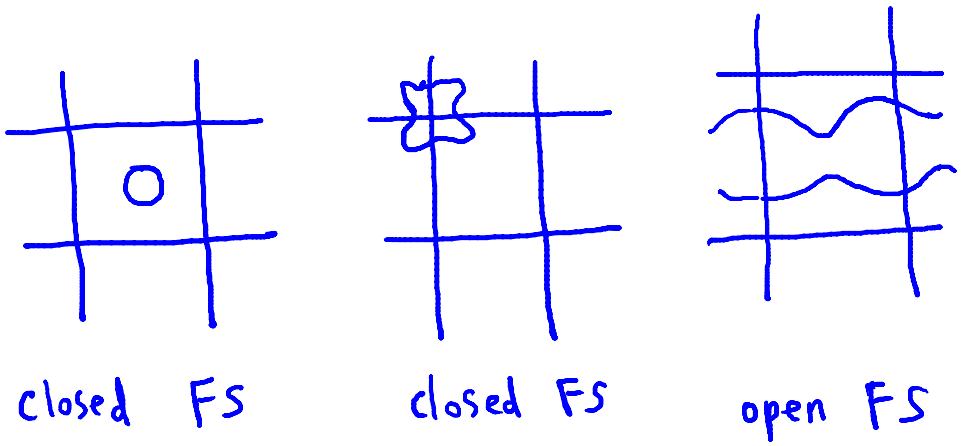


Holes and Electrons

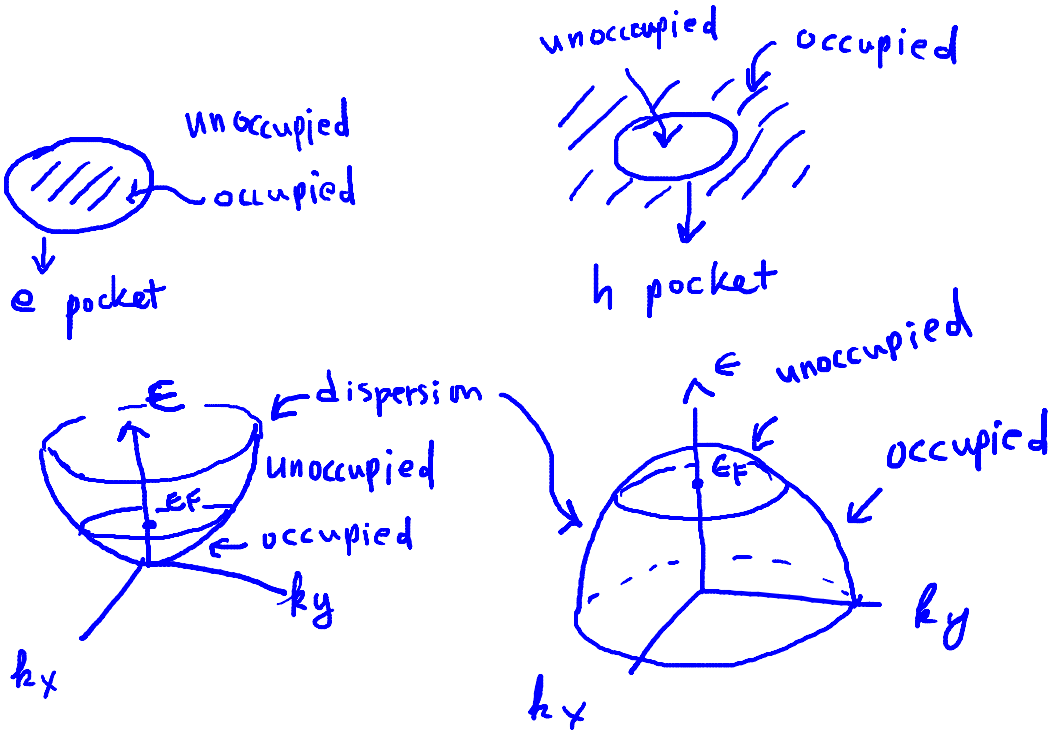
These days "topological effects" are in fashion. Topologically protected effects are robust, which is why they are important. Now, a very simple topological effect in solid state physics is the emergence of the hole as a physical entity in semi-conductors.

Let us recall how we defined the electron pocket and the hole pocket in relation to a Fermi surface. We assume that we have a closed Fermi surface, as an open Fermi surface leaves the electron or the hole nature poorly defined or

dependent on the direction.



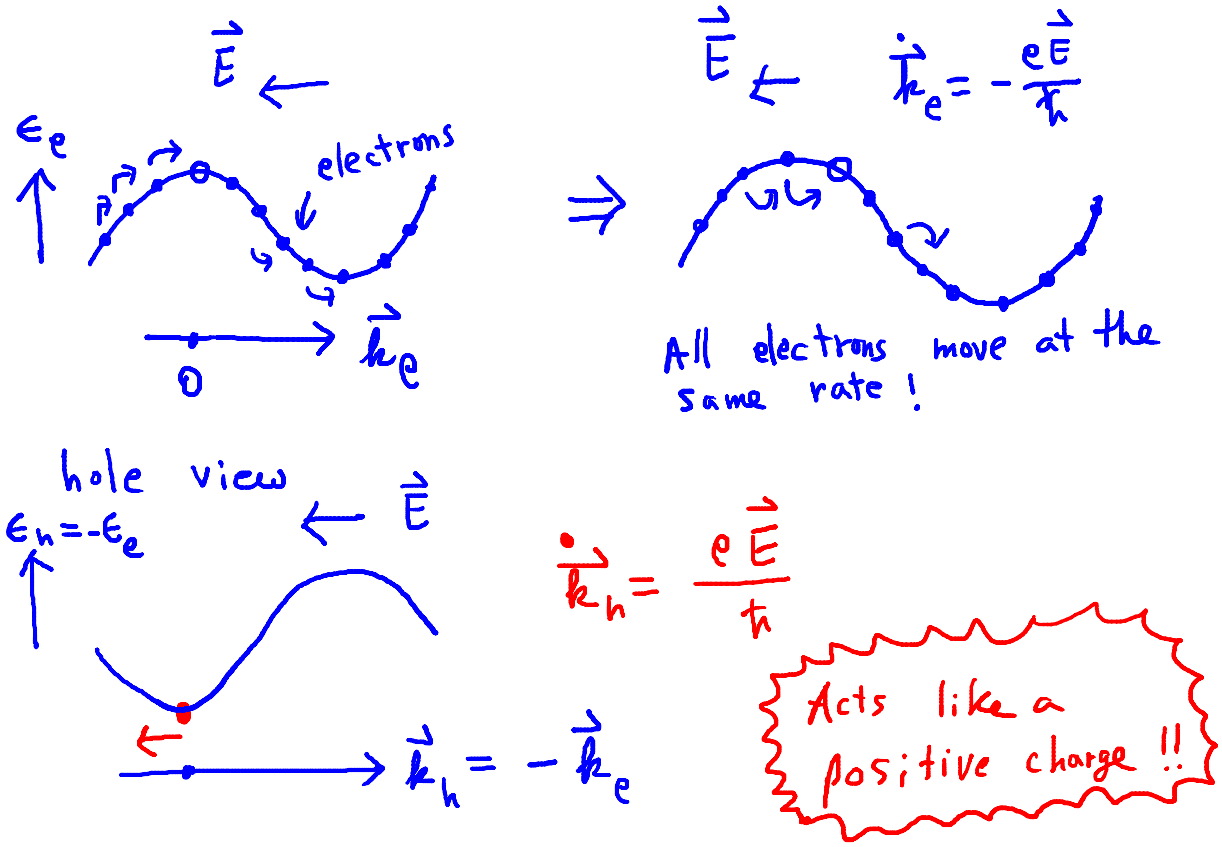
For a given closed Fermi surface, if the band dispersion that gives rise to that Fermi surface is occupied in the interior of the Fermi surface, then that is an electron pocket. If the interior is un-occupied, then that is a hole pocket.



Why is this -- e-pocket or a h-pocket -- a topological property? This is because, it is not possible to continuously deform an e-pocket to a h-pocket, and vice versa. (This topological argument is valid only in dimensions higher than one.)

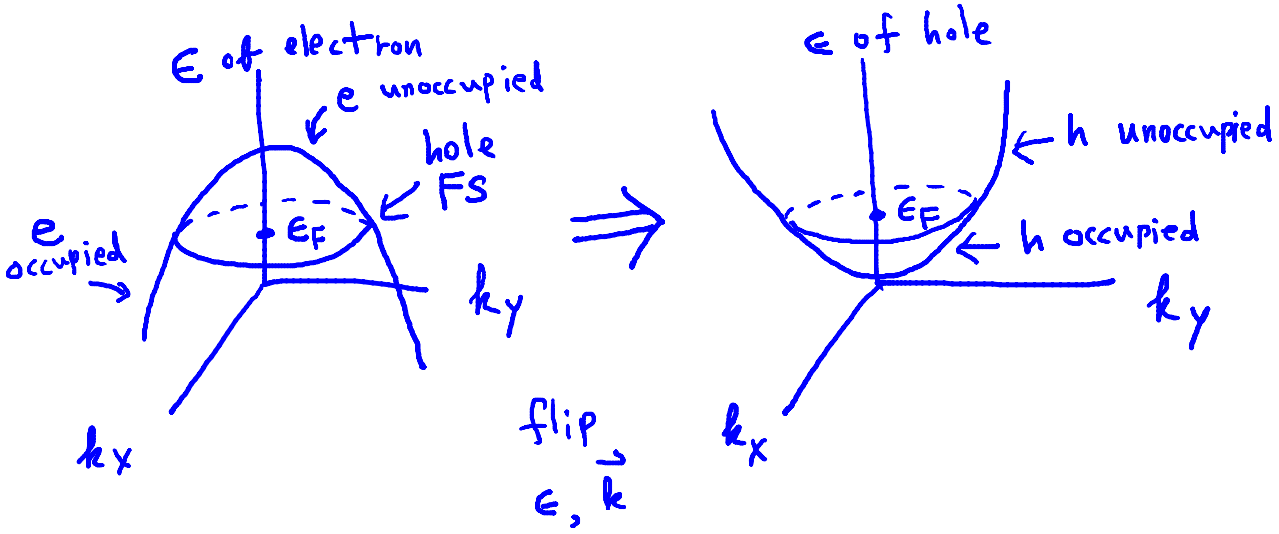
By studying the following diagrams carefully, one can grasp what we really mean by "hole." The most important fact is that the concept "hole" considers the completely filled band as the reference state, or the vacuum state if you like. Thus energy of the hole = (total energy of the N - 1 state) - (total energy of the N state), where the N state corresponds to the completed filled band. The wave

vector of the hole = (total wave vector of the N - 1 state) - (total wave vector of the N state). **So, to go from the electron diagram to the hole diagram, both the energy and the wave vector (momentum) need to be flipped in sign**. The same holds for the spin/orbital angular momentum. But, note that the group velocity remains the same.



Note, by the way, that the above diagram shows that in a completely filled band, each electron will move exactly the same way in $\vec{k}$ space. So, there is no net change. This is why materials with completely filled bands acts as though nothing happens (like no conductivity), since the overall state is not changing at all.

So it is not just the mere convenience that we call one FS an e-pocket or the other a h-pocket. They are fundamentally different. The transport phenomena arising from an electron pocket should be considered, as coming from electrons, while the transport phenomena arising from a hole pocket should be considered as holes, positively charged particles.



The hole may be defined as "the absence of an electron" but is a distinctly different particle, as measured by the probe sensitive to the charge of the particle. For instance, the Hall effect is such a measurement. Also, the relative sign of the heat conductivity and the electrical conductivity is another such measurement.