Contents

EP431 Semiconductor Physics : Course content

- Putting many atoms together, use 1D array of delta functions.
- Bloch's theorem & band structure.
- k.p + effective mass, effective g-factor,
- equation of motion in a band, Concept of holes: what is the correct k vector of a hole, equation of motion, group velocity etc.
- Methods of band structure calculation: Tight binding, Wannier functions, Pseudo potential and OPW
- Band structure of common semiconductors (Si, Ge, GaAs)
- doping in semiconductors, carrier density, Mott transition
- Some transport, scattering mechanisms & optics in bulk semiconductors
- pn junction, Heterostructures, Poisson Schrodinger band structure calculation,
- MOSFET & 2DEGs: Quantum Hall Effect (if time permits...)

Evaluation

All quizzes and exams will be problem-solving based. You must be able to calculate algebraic/numerical answers to problems. There will be little or no descriptive questions. You can use one formula sheet and calculators during exams/quizzes.

- Quiz 1 ≈10-15%
- Midsem ≈30%
- Quiz 2 ≈ 10-15%
- Endsem ≈30-40%

Audit Policy

Treat class quizzes, midsem and endsem as "take home" exams. Submit within pre-decided times. You need to score at least 50%. Sitting through the lectures is NOT audit. You must demonstrate that you can solve problems.

References

- Chapters on semiconductor physics in the two "Solid State Physics" books by C. Kittel, and Ashcroft & Mermin.
- Semiconductor Physics, K. Seeger (more like a Handbook)
- Semiconductor Physics, Yu & Cordona
- Physics of Low Dimensional Semiconductors, J H Davies (for the heterostructure/ Quantum Hall aspects)
- References to review papers etc as discussed during lectures.

Expected Background

The course is run as one of the 400-level "honours course". It is expected that the those taking the course are able to do Quantum Mechanics at the level of QM-2 and Statistical physics (first course) as taught in the Physics department.

Chapter 1

Putting atoms together : a toy model

Typically a small number of atoms come together to form molecules. A large number of atoms come together to form a solid. In the process the energy levels of each isolated atom undergoes important changes, irrespective of whether the solid so formed is crystalline or not. This is a very fundamental aspect and we will try to understand this using a "toy model".

1.1 The 1d delta function

We simplify things drastically and model an atom by a 1d delta function potential. We then ask for the bound states.

$$V(x) = -|V_0| L \delta(x) \tag{1.1}$$

There is a reason we write the δ function potential this way. We can then think of this as the limiting case of a potential well whose strength is V_0 and extent is L. It will also allow us to set the strength of the potential realistically, to be similar to that of an atom. *e.g.* We can take $V_0 \sim 10$ electron-volts and $L \sim 1$Å, keeping hydrogen atom in mind. Writing the strength this way is better for keeping units and dimensions correct. Can you see why? We need to solve the Schrödinger equation for the bound states, that means solve

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\psi = -|E|\psi \tag{1.2}$$

where the solutions satisfy

$$\psi(x \to \pm\infty) = 0 \tag{1.3}$$

At all points ψ must be of the form

$$\psi(x) = Ae^{kx} + Be^{-kx} \tag{1.4}$$

$$k^2 = \frac{2m|E|}{\hbar^2} \tag{1.5}$$

To satisfy condition 1.3 the wavefunctions must be as shown in fig 1.1.

Integrating the Schrödinger equation once we obtain the condition the derivative $\dfrac{d\psi}{dx}$ must satisfy at the location of the delta function. For an infinite discontinuity (like a delta function) the left and right derivatives are not equal, which is clear from the cusp in the fig 1.1. Integrating it again gives the condition that $\psi(x)$ itself must be continuous. (We will be using these two conditions repeatedly, so make sure you understand how they come about.)

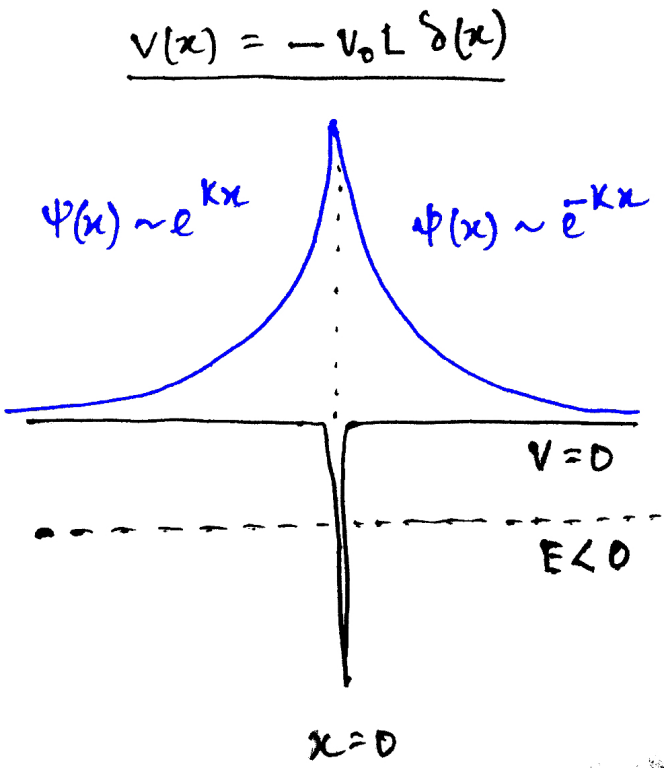


Figure 1.1: The wavefunction around a delta function for a bound state. Of course it is possible to have solutions with $E > 0$, these will not be bound, in the sense the particle may be found infinitely far from the point of attraction with significant probability. These are also called scattering states.

$$\psi(x = 0^+) = \psi(x = 0^+) \tag{1.6}$$

$$\left.\frac{d\psi}{dx}\right|_{0^+} - \left.\frac{d\psi}{dx}\right|_{0^-} = \frac{2m}{\hbar^2}(-V_0 L)\psi(0) \tag{1.7}$$

$$= -\beta\psi(x = 0^+)$$

$$\text{where} \qquad \beta = \frac{2m}{\hbar^2}|V_0 L| \tag{1.8}$$

Applying the two conditions gives the solution for the allowed value of k and hence E

$$2k = \beta \tag{1.9}$$

$$\therefore \quad E = -\frac{mL^2 {V_0}^2}{2\hbar^2} \tag{1.10}$$

There is always only one bound state in this problem. We state without formal proof that in 2d and 3d this is not the case. A potential would need to have a minimum strength before it can support a bound state. In 1d any attractive potential, however weak has at least one bound state.

1.1.1 Two delta functions : the molecule

We now want to put two attractive points separated by a distance a, as shown in figure 1.2 .So we have

$$V(x) = -V_0 L\left(\delta(x) + \delta(x - a)\right) \tag{1.11}$$

Note the following points and justify them

- For $x < 0$ we must have $\psi(x) = A_0 e^{kx}$

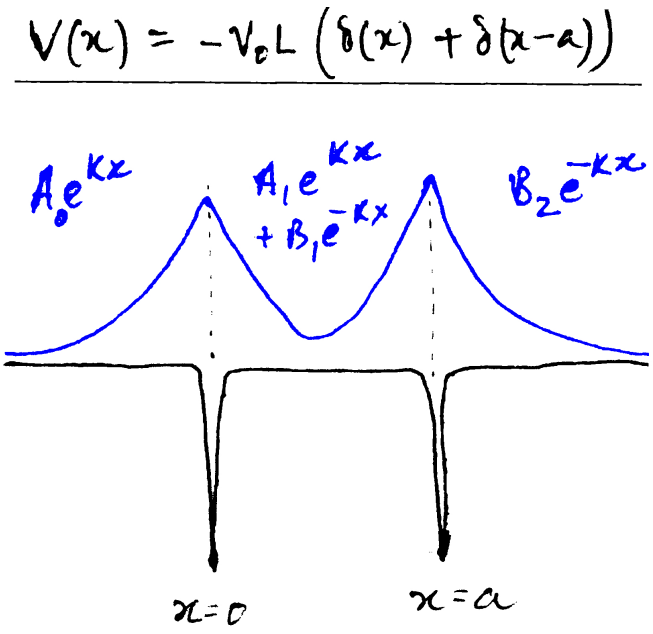


Figure 1.2: The wavefunction around two delta functions for a bound state. In the middle one can have both $+k$ and $-k$ solutions. The figure shows only one possible solution for $\psi(x)$. How does the other one look like?

- For $0 < x < a$ we can have $\psi(x) = A_1 e^{kx} + B_1 e^{-kx}$
- For $x > a$ we must have $\psi(x) = B_2 e^{-kx}$
- We have five unknowns A_0, A_1, B_1, B_2, k.
- What are the five independent equations?
- Apply the conditions on ψ and $\dfrac{d\psi}{dx}$ at $x = 0$ and $x = a$. Normalisation provides the other one.
- However in these problems it is not always essential to normalise the wavefunction. In that case we can simply set $A_0 = 1$ and solve for the rest. Often one wants to know the bound state energy only.
- As an exercise formulate the problem. You can get two solutions for k. Prove that in such cases one solution lies above the single well solution and the other lies above.

In figure 1.3 we have shown what happens to the energy levels as the two "atoms" are brought closer to each other.

1.1.2 Many delta functions : chain of atoms or a "solid"

we will solve the more general problem for N attractive centers

$$V(x) = -V_0 L \sum_{n=0}^{N-1} \delta(x - na) \tag{1.12}$$

There are a number of ways to solve this problem. We will do it in a slightly unconventional way. Later on we will revisit this type of problems many times and exploit the translational symmetry.

- Notice that the solution to the wavefunction in any region is of the form $A_n e^{kx} + B_n e^{kx}$
- Two relations connect the solutions for regions n and $n + 1$ at the point $x = na$, where $n = 0, 1, ..., N - 1$

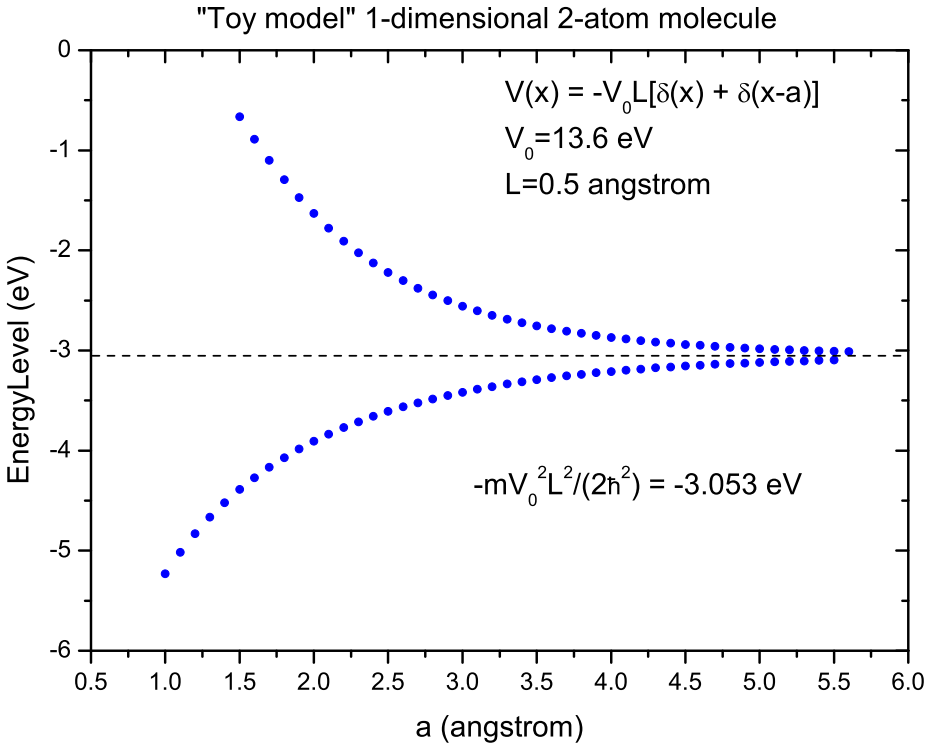


Figure 1.3: The bound state energy level splits into two as the two atoms come closer. The lower one corresponds to the "bonding" solution, where the electron density is higher in the middle. The other "anti-bonding" solution has the electron density going to zero in the middle.

- The continuity of the wavefucntion requires

$$A_n e^{kna} + B_n e^{-kna} = A_{n+1} e^{kna} + B_{n+1} e^{-kna} \tag{1.13}$$

- The discontinuity of the derivative requires

$$A_{n+1} k e^{kna} - B_{n+1} k e^{-kna} - \left(A_n k e^{kna} - B_n k e^{-kna} \right) = -\beta \left(A_n e^{kna} + B_n e^{-kna} \right) \tag{1.14}$$

The two equations 1.13 and 1.14 relate (A_{n+1}, B_{n+1}) with (A_n, B_n) via linear equations. They can be written quite neatly in a matrix form.

$$\begin{pmatrix} e^{kna} & e^{-kna} \\ k e^{kna} & -k e^{-kna} \end{pmatrix} \begin{pmatrix} A_{n+1} \\ B_{n+1} \end{pmatrix} = \begin{pmatrix} e^{kna} & e^{-kna} \\ (k-\beta) e^{kna} & -(k+\beta) e^{-kna} \end{pmatrix} \begin{pmatrix} A_n \\ B_n \end{pmatrix} \tag{1.15}$$

The inverse of 2×2 matrices are easy to calculate and we get an explicit relation

$$\begin{pmatrix} A_{n+1} \\ \\ B_{n+1} \end{pmatrix} = \begin{pmatrix} 1 - \dfrac{\beta}{2k} & -\dfrac{\beta}{2k} e^{-2kna} \\ \dfrac{\beta}{2k} e^{2kna} & 1 + \dfrac{\beta}{2k} \end{pmatrix} \begin{pmatrix} A_n \\ \\ B_n \end{pmatrix} \tag{1.16}$$

- Such matrices which relate the right hand side variables with the left hand side ones are often called *transfer matrices*. They are particularly useful in solving 1d problems.

- we can now keep multiplying such matrices and relate (A_N,B_N) with (A_0,B_0)

First do this for the 2 delta-fn problem from the last section. We multiply the two transfer matrices for $x = 0$ and $x = a$ must have

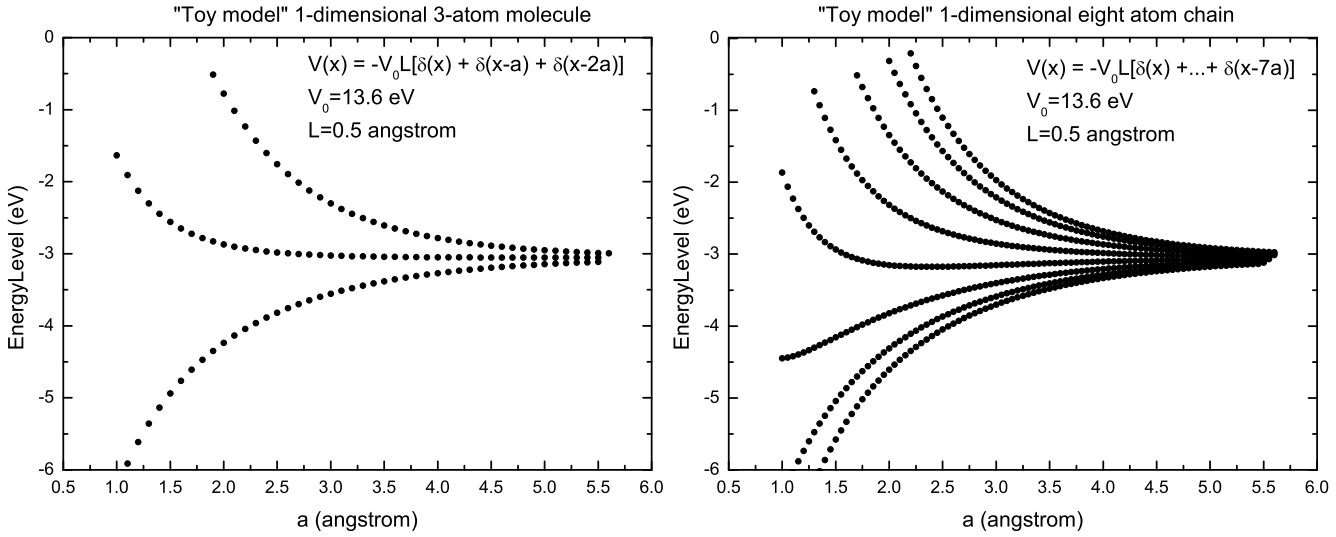


Figure 1.4: The bound state energy level splits into N states as the two atoms come closer.

$$\begin{pmatrix} 0 \\ B_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1-\frac{\beta}{2k} & -\frac{\beta}{2k}e^{-2ka} \\ \frac{\beta}{2k}e^{2ka} & 1+\frac{\beta}{2k} \end{pmatrix}}_{x=a} \underbrace{\begin{pmatrix} 1-\frac{\beta}{2k} & -\frac{\beta}{2k} \\ \frac{\beta}{2k} & 1+\frac{\beta}{2k} \end{pmatrix}}_{x=0} \begin{pmatrix} A_0 \\ 0 \end{pmatrix} \tag{1.17}$$

$$= \begin{pmatrix} \left(1-\frac{\beta}{2k}\right)^2 - \left(\frac{\beta}{2k}\right)^2 e^{-2ka} & -\left(1-\frac{\beta}{2k}\right)\frac{\beta}{2k} - \left(1+\frac{\beta}{2k}\right)\frac{\beta}{2k}e^{-2ka} \\ \frac{\beta}{2k}\left(1-\frac{\beta}{2k}\right)e^{2ka} + \left(1+\frac{\beta}{2k}\right)\frac{\beta}{2k} & \left(1+\frac{\beta}{2k}\right)^2 - \left(\frac{\beta}{2k}\right)^2 e^{2ka} \end{pmatrix} \begin{pmatrix} A_0 \\ 0 \end{pmatrix}$$

The equation can only be satisfied if the $(1,1)$ element of the full transfer matrix is zero. Hence the implicit equation for k is

$$\left(1-\frac{\beta}{2k}\right)^2 - \left(\frac{\beta}{2k}\right)^2 e^{-2ka} = 0 \tag{1.18}$$

- We see that if a is very large then the exponential is very small. Hence $\frac{\beta}{2k} \approx 1$ which is what we got for the single well case. This of course must hold for consistency.
- Because of the square term we can see that two solutions will emerge. You should be able to show that one of them is less than the $k = \beta/2$ solution in energy, the other (if it exists) is greater.

$$\frac{2k}{\beta} - 1 = \pm e^{-ka} \tag{1.19}$$

- It is possible to solve for the hydrogen molecule problem exactly, using the full Coulomb potential from each nucleus. However it is quite complicated algebraically.

We can now address the N-atom problem. We see that the problem will involve the N^{th} power of $\frac{\beta}{2k}$ in the $(1,1)$ element of the full transfer matrix. We can expect N solutions. Only those which lie below zero are bound state solutions. For $N > 2$ it is not possible to do the solution analytically since it involves a mix of polynomials and exponentials. However it is straightforward to plot the function and read off all the zero crossings numerically. This is how the data for the figure 1.4 and 1.5 have been generated.
In figure 1.5, we have stack-plotted all the energy values for a certain N with N in the x-axis.

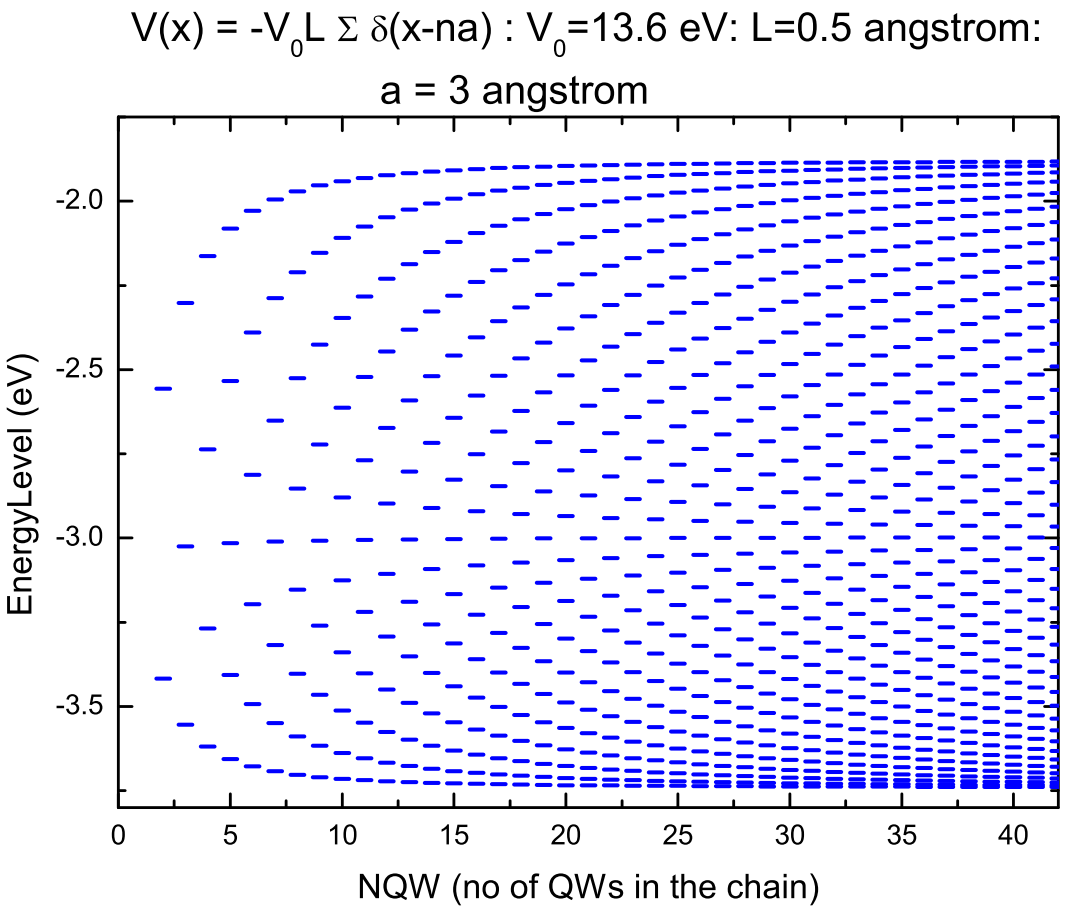


Figure 1.5: The distribution of the energies for the N-atom problem with $a = 3$Å. Study the features carefully.

- Notice how the bound state has spread out into a "band". The width of the band saturates once N reaches a value $10 - 20$. After that the density of points within the band increases and in general it will have N levels.
- However as N increases the spacing between this discrete values becomes very small. For realistic values on $N{\sim}10^{23}\text{cm}^{-3}$ in a solid, that many states will lie within a few eV of energy. The gap between two successive levels is so small that we can treat them as a continuous function for all practical purposes.
- You may have noticed a concentration of the energy levels towards the upper and lower edge of the band - this is a typical feature of 1-dimensionality, it will not be so in 2 & 3 dimensions.

The "bare-bones" approach that we just took has severe limitations.

- It is difficult to extend this to 2 and 3 dimension.
- We have made no use of the translational symmetry and gone for a brute-force numerical solution.
- We have not treated a system with more than one atomic energy level.
- However the basic picture that has emerged is very robust. In fact even if not all the potential wells are of equal strength, *i.e.* there is a distribution of V_0, the generic result will still hold.
- Now, you can think about one more possibility. Suppose the chain of atoms was made into a loop by tying the two free ends together. How do you think the solutions will be modified?

1.2 Problems

1. Suppose you had a 1d quantum well with depth $V_0 = 13.6$eV and width $L = 1$Å. How many bound states are possible? How does the energy of the lowest state compare with the bound state for a delta function potential $V(x) = -|V_0|L\delta(x)$. ?

2. For the potential $V(x) = -|V_0|L(\delta(x) + \delta(x - a))$, solve for the two possible eigenfunctions and sketch them . Derive a condition for which the higher energy state becomes unbound.

3. Modify the code given at the end and generate numerically the density of states $D(E)$ and plot it vs E.

4. Finally, this is somewhat open ended...Can you guess the limitations of the code? How can you improve it? Suppose the strengths of the delta function are not all same, but they are chosen from a random distribution (with a given mean and width). How do you think the answer would be modified?

1.3 Code to generate the transfer matrix

You can play with it and see how things evolve...

```
/*
Program to calculate bound state energies of an array of N delta-fn QWs.
Using a transfer matric method but not invoking Bloch's theorem

This version prints out the number of QWs, and the allowed energy levels
in the system.

The boundary condition is "free". It is a chain with open ends and not a
loop. For N > 2 it would be good to have an option for a chain/loop
and see how the difference evolves/vanishes.

The code can also solve for the bound state energies E<0, as a function
of the lattice spacing.
*/

#include <stdio.h>
#include <stdlib.h>
#include <math.h>
double hbar, m_electron;

int main()
{
  double V0,L,k,beta,a, da,beta_a, betaby2k, ka, ka_min, ka_max, dka,E_bound_in_eV,twokann, exptwokann;
  double m11, m12, m21, m22, tmp11, tmp12,tmp21,tmp22, mold11,mold12,mold21,mold22;
  double lastval;
  int NQW, nn, flag;
  FILE *fpout;


  /* all quantities in SI, unless otherwise written */
  hbar=1.05e-34;
  m_electron=9.1e-31;

  fpout=fopen("output.dat","w");

  V0=13.6*1.6e-19; /* 13.6 eV */
  L=0.5e-10;         /* 0.5 angstrom */
  beta= (2*m_electron/(hbar*hbar))*V0*L;


  a=3e-10; /* 3 angstroms lattice */

  for(NQW=2; NQW<=100; NQW++)
    {

    beta_a=beta*a;
    printf("NQW=%d\n",NQW);

  ka_min=0.5;
  ka_max=5;
  dka=1e-5;

  flag=0;

  for(ka=ka_min; ka<ka_max;ka+=dka)
     {
```

```
       nn=0;
       betaby2k=beta_a/(2*ka);

       mold11=1.0-betaby2k;
       mold12=-1.0*betaby2k;  /*  nn = 0, 1,2,3...etc  where x=nn*a */
       mold21=betaby2k;       /*  for nn=0,exp(2*ka*nn)=1 */
       mold22=1.0+betaby2k;



       for(nn=1; nn<NQW; nn++)
          {
           twokann=2*ka*((double)nn); /* minimise the number of exponentials within the loop */
           exptwokann=exp(twokann);
           m11=1.0-betaby2k;
           m12=-1.0*betaby2k/exptwokann;  /** nn = 0, 1,2,3...etc  where x=nn*a */
           m21=betaby2k*exptwokann;
           m22=1.0+betaby2k;


           tmp11=m11*mold11 + m12*mold21;   /* multiply the matrices */
           tmp12=m11*mold12 + m12*mold22;
           tmp21=m21*mold11 + m22*mold21;
           tmp22=m21*mold12 + m22*mold22;


           mold11=tmp11;  /* store the new values */
           mold12=tmp12;
           mold21=tmp21;
           mold22=tmp22;
          }



      if(flag){ /* check for zero-crossings of the function */
               if ( ((lastval >0) && (mold11 <0)) || ((lastval < 0) && (mold11 >0)))
                  { E_bound_in_eV= -1.0*((hbar*hbar/(2*m_electron))*(ka/a)*(ka/a))/(1.6e-19);
                   fprintf(fpout,"%d\t%e\t%e\n",NQW,a*1e10,E_bound_in_eV); /* distance in angstrom */
                  }
               }


      lastval=mold11;
      flag=1;
      }
    }

  fclose(fpout);
  return 0;
}
```

Chapter 2

Reciprocal Lattice : in a "nutshell"

Suppose we have a periodic function. In 1D this means

$$f(x+L) = f(x) \tag{2.1}$$

where L is a fixed length - the period of the function. We know that the Fourier series of such a function will contain "frequencies" $n\frac{2\pi}{L}$ where n is any integer. In 3D this implies the existence of *three* linearly independent vectors $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, $\boldsymbol{a_3}$, such that

$$f(\boldsymbol{r} + n_1\boldsymbol{a_1} + n_2\boldsymbol{a_2} + n_3\boldsymbol{a_3}) = f(\boldsymbol{r}) \tag{2.2}$$

This means any translation vector of the form

$$\boldsymbol{R_n} = n_1\boldsymbol{a_1} + n_2\boldsymbol{a_2} + n_3\boldsymbol{a_3} \tag{2.3}$$

leaves the function invariant. $\boldsymbol{n}$ is a shorthand for for all possible (positive & negative) integer set (n_1,n_2, n_3). Which Fourier components will exist?

$$F(\boldsymbol{k}) = \int_{\substack{\text{all}\\ \text{space}}} d^3\boldsymbol{r}\ f(\boldsymbol{r})e^{i\boldsymbol{k}.\boldsymbol{r}} \tag{2.4}$$

The answer can be deduced as follows

$$\begin{aligned}
F(\boldsymbol{k}) &= \int d^3\boldsymbol{r}\ f(\boldsymbol{r})e^{i\boldsymbol{k}.\boldsymbol{r}} \\
&= \int d^3\boldsymbol{r}\ f(\boldsymbol{r}+\boldsymbol{R_n})e^{i\boldsymbol{k}.\boldsymbol{r}} \quad \forall\ \boldsymbol{n} \\
&= e^{-i\boldsymbol{k}.\boldsymbol{R_n}} \int d^3\boldsymbol{r}\ f(\boldsymbol{r}+\boldsymbol{R_n})e^{i\boldsymbol{k}.(\boldsymbol{r}+\boldsymbol{R_n})} \\
&= e^{-i\boldsymbol{k}.\boldsymbol{R_n}}\ F(\boldsymbol{k}) \\
\therefore e^{-i\boldsymbol{k}.\boldsymbol{R_n}} &= 1 \\
\implies \boldsymbol{k}.\boldsymbol{R_n} &= 2\pi \times \text{integer}
\end{aligned} \tag{2.5}$$

The condition can be satisfied if we can find a set of three vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{b}_3$, such that

$$\boldsymbol{a}_i.\boldsymbol{b}_j = 2\pi\delta_{ij} \tag{2.6}$$

Then any vector of the form

$$\boldsymbol{G_m} = m_1\boldsymbol{b_1} + m_2\boldsymbol{b_2} + m_3\boldsymbol{b_3} \tag{2.7}$$

will satisfy the condition 2.5 and exist in the set.
To complete the argument we need to show that the condition 2.6 is sufficient to determine the set of vectors $\boldsymbol{b}_i$ in any dimension.

- For example in 3 dimension, each vector has 3 components so we would need 3×3 equations to determine the full set $\boldsymbol{b_1}$,$\boldsymbol{b_2}$,$\boldsymbol{b_3}$.
- Since i, j both take values 1,2,3- the condition 2.6 indeed generates 3×3 equations.
- We can see that the same argument would work in any dimension - we need $n \times n$ equations and they are provided by 2.6
- However these linear equations must be independent and solvable -this is guaranteed if the determinant

$$\begin{vmatrix} a_{11} & a_{12} & ..... & a_{1n} \\ a_{21} & a_{22} & ..... & a_{2n} \\ .. & .. & .. & .. \\ a_{n1} & a_{n2} & ..... & a_{nn} \end{vmatrix} \neq 0 \tag{2.8}$$

  You should prove this for yourself.
- PROBLEM: We have N^2 equations, but the determinant is only $N \times N$. Why is this sufficient?
- Again in 3d it leads to the familiar condition $\boldsymbol{a_1}.(\boldsymbol{a_2} \times \boldsymbol{a_3}) \neq 0$, meaning that the three vector should not be co-planer. Another way to express the same result is that the three vectors should enclose some "volume". In any dimension the determinant written above generates the analogue of volume enclosed by the vectors. In 2d the definition generates the cross-product which is the area of the parallelogram formed by the two vectors.
- PROBLEM: Show that the expression for the reciprocal vectors would be the following in 3D: (This is the form most textbooks will give you)

$$\begin{aligned} \boldsymbol{b_1} &= 2\pi \frac{\boldsymbol{a_2} \times \boldsymbol{a_3}}{\boldsymbol{a_1}.(\boldsymbol{a_2} \times \boldsymbol{a_3})} \\ \boldsymbol{b_2} &= 2\pi \frac{\boldsymbol{a_3} \times \boldsymbol{a_1}}{\boldsymbol{a_1}.(\boldsymbol{a_2} \times \boldsymbol{a_3})} \\ \boldsymbol{b_3} &= 2\pi \frac{\boldsymbol{a_1} \times \boldsymbol{a_2}}{\boldsymbol{a_1}.(\boldsymbol{a_2} \times \boldsymbol{a_3})} \end{aligned} \tag{2.9}$$

- What does $\boldsymbol{a_1}.(\boldsymbol{a_2} \times \boldsymbol{a_3})$ imply? Work out the relevant expression for two dimensions.
- find a relation between the volumes $\boldsymbol{a_1}.(\boldsymbol{a_2} \times \boldsymbol{a_3})$ and the reciprocal unit cell volume $\boldsymbol{b_1}.(\boldsymbol{b_2} \times \boldsymbol{b_3})$
- PROBLEM: Consider a finite chain with N particles of mass m with spacing a, such that the density may be written as

$$\rho(x) = \sum_{n=0}^{N-1} m\delta(x - na)$$

  Calculate the fourier transform of its density $\rho(k)$. What is the half width of each peak? How does your result differ from that of an infinite chain?
- PROBLEM: A lattice is built up of many atoms arranged in a periodic manner. Consider a function $f(\boldsymbol{r})$ that gives the electron density contributed by each atom (say). The full density at a point is the sum total of all the contributions from all the atoms.

$$\rho(\boldsymbol{r}) = \sum f(\boldsymbol{r} - \boldsymbol{R_n})$$

  Show that the fourier transform of this is a product of two distinct terms, irrespective of whether the sum is finite or infinite. What is the physical significance of each term?

In general we will use the symbol $\boldsymbol{R}$ for the real space vectors and the symbol $\boldsymbol{G}$ for the reciprocal or Fourier space vectors.

Chapter 3

Electrons in a periodic potential: Bloch's theorem

We know that the lattice sites form a repetitive array, so the potential felt by an electron moving in it will be periodic as well, this means we can write potential as a Fourier sum over all reciprocal lattice vectors.

3.1 Derivation of the theorem

$$\begin{aligned} H &= T + V \\ &= T + \sum_{\boldsymbol{G}} v_{\boldsymbol{G}} e^{i\boldsymbol{G}.\boldsymbol{r}} \end{aligned} \tag{3.1}$$

Let's assume we know all the components $v_{\boldsymbol{G}}$ and want to solve for the the eigenvalues and eigenfunctions. We know that if the potential was zero, then the solutions would have been free particle like, now as a consequence of the existence of the potential, many plane waves states $|\boldsymbol{k}\rangle$ must be mixed, so

$$|\psi\rangle = \sum_{\boldsymbol{k}} c_{\boldsymbol{k}} |\boldsymbol{k}\rangle \tag{3.2}$$

where

$$\langle \boldsymbol{r}|\boldsymbol{k}\rangle = \frac{1}{\sqrt{L^3}} e^{i\boldsymbol{k}.\boldsymbol{r}} \tag{3.3}$$

To solve for $c_{\boldsymbol{k}}$ is the obvious target. We let H act on the *ket* $|\psi\rangle$, which we assume to be an eigenstate of H, and then left multiply with the state *bra* $\langle \boldsymbol{k'}|$.

$$\begin{aligned} \langle \boldsymbol{k'}|H|\psi\rangle &= \sum_{\boldsymbol{k}} c_{\boldsymbol{k}} \langle \boldsymbol{k'}|T|\boldsymbol{k}\rangle + \sum_{\boldsymbol{k}} c_{\boldsymbol{k}} \langle \boldsymbol{k'}|V|\boldsymbol{k}\rangle \\ \therefore \quad E c_{\boldsymbol{k'}} &= \frac{\hbar^2}{2m} k'^2 c_{\boldsymbol{k'}} + \sum_{\boldsymbol{k}} \sum_{\boldsymbol{G}} c_{\boldsymbol{k}} v_{\boldsymbol{G}} \delta_{\boldsymbol{k'},\boldsymbol{k}+\boldsymbol{G}} \\ \therefore \quad 0 &= \left(\frac{\hbar^2}{2m} k'^2 - E \right) c_{\boldsymbol{k'}} + \sum_{\boldsymbol{G}} c_{\boldsymbol{k'}-\boldsymbol{G}} v_{\boldsymbol{G}} \end{aligned} \tag{3.4}$$

The sum is over the reciprocal lattice vectors $\boldsymbol{G}$. This has the form of an eigenvalue equation. But we have only one equation for a large number of unknowns -all the $c_{\boldsymbol{k}-\boldsymbol{G}}$. In principle there can be an infinite number of them.
Now notice a few important points

1. Eqn 3.4 connects a state $\boldsymbol{k}$ with states that can be reached by reciprocal lattice translations. *i.e.* The potential only mixes the states $|\boldsymbol{k}\rangle$, $|\boldsymbol{k}-\boldsymbol{G_1}\rangle$, $|\boldsymbol{k}-\boldsymbol{G_2}\rangle$ and so on.

2. But there are some states which cannot be taken to one another by reciprocal lattice translations: take any two states in the first Brillouin zone, that are not on the zone boundary.

3. The previous two points together imply that $\boldsymbol{k}$ from the first Brillouin zone can be used to label a wavefunction that will contain $|\boldsymbol{k}\rangle$, $|\boldsymbol{k}-\boldsymbol{G_1}\rangle$, $|\boldsymbol{k}-\boldsymbol{G_2}\rangle$.... These wavefunctions are the Bloch wavefunctions.

Now we complete the solution. In the process of getting to eqn. 3.4 we could have left multiplied with another state $\langle \boldsymbol{k'}-\boldsymbol{K}|$ where $\boldsymbol{K}$ is any reciprocal lattice vector. Then we would get the equation

$$\left(\frac{\hbar^2}{2m}(\boldsymbol{k'}-\boldsymbol{K})^2 - E\right) c_{\boldsymbol{k'}-\boldsymbol{K}} + \sum_{\boldsymbol{G}} c_{\boldsymbol{k'}-\boldsymbol{K}-\boldsymbol{G}} v_{\boldsymbol{G}} = 0 \tag{3.5}$$

Hence

$$\left(\frac{\hbar^2}{2m}(\boldsymbol{k'}-\boldsymbol{K})^2 - E\right) c_{\boldsymbol{k'}-\boldsymbol{K}} + \sum_{\boldsymbol{G'}} c_{\boldsymbol{k'}-\boldsymbol{G'}} v_{\boldsymbol{G'}-\boldsymbol{K}} \quad = \quad 0$$

where we have written $\boldsymbol{G'} = \boldsymbol{K}+\boldsymbol{G}$, remembering that $\boldsymbol{k'}$ and $\boldsymbol{K}$ are fixed vectors for a particular row. The primes are now unnecessary in the summation indices.

$$\left(\frac{\hbar^2}{2m}(\boldsymbol{k}-\boldsymbol{K})^2 - E\right) c_{\boldsymbol{k}-\boldsymbol{K}} + \sum_{\boldsymbol{G}} c_{\boldsymbol{k}-\boldsymbol{G}} v_{\boldsymbol{G}-\boldsymbol{K}} = 0$$

$\boldsymbol{G}$ and $\boldsymbol{K}$ run over the same set of (reciprocal lattice) vectors. Now we see the following:

1. There will be as many solutions as the dimension of the matrix.

2. Each $\boldsymbol{k}$ vector (not a reciprocal lattice vector) in the first Brillouin zone gives a matrix for us to solve. The $E(\boldsymbol{k})$ relation thus has as many branches as the dimension of the matrix.

3. Further we can substitute $\boldsymbol{k}+\boldsymbol{G}$ for $\boldsymbol{k}$ and show that the set of equations 3.5 remain the same. You should work this out as a problem.

4. An important consequence of this is that $E(\boldsymbol{k}) = E(\boldsymbol{k}+\boldsymbol{G})$. We only need to solve the eigenvalue equations for $\boldsymbol{k}$ in the first Brillouin zone.

We can see that the wavefunction must have the form

$$\begin{aligned}
|\psi_{\boldsymbol{k}}\rangle &= \sum_{\boldsymbol{G}} c_{\boldsymbol{G}} |\boldsymbol{k}-\boldsymbol{G}\rangle \\
\langle \boldsymbol{r}|\psi_{\boldsymbol{k}}\rangle \equiv \psi_{\boldsymbol{k}}(\boldsymbol{r}) &= \sum_{\boldsymbol{G}} c_{\boldsymbol{G}} e^{i(\boldsymbol{k}-\boldsymbol{G}).\boldsymbol{r}} \\
&= e^{i\boldsymbol{k}.\boldsymbol{r}} \sum_{\boldsymbol{G}} c_{\boldsymbol{G}} e^{-i\boldsymbol{G}.\boldsymbol{r}} \\
&= e^{i\boldsymbol{k}.\boldsymbol{r}} u(\boldsymbol{r}) \qquad (3.6) \\
\therefore \quad \psi_{\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R}) &= e^{i\boldsymbol{k}.(\boldsymbol{r}+\boldsymbol{R})} u(\boldsymbol{r}+\boldsymbol{R}) \\
&= e^{i\boldsymbol{k}.(\boldsymbol{r}+\boldsymbol{R})} \sum_{\boldsymbol{G}} c_{\boldsymbol{G}} e^{-i\boldsymbol{G}.\boldsymbol{r}} e^{-i\boldsymbol{G}.\boldsymbol{R}} \qquad (3.7) \\
&= e^{i\boldsymbol{k}.\boldsymbol{R}} \psi(\boldsymbol{r}) \qquad \left(\because e^{-i\boldsymbol{G}.\boldsymbol{R}} = 1\right)
\end{aligned}$$

Since where $\boldsymbol{k}$ is in the first Brillouin zone. This is Bloch's theorem in many equivalent forms. The function $u(\boldsymbol{r})$ has the symmetry of the direct lattice.

3.1.1 Translation invariance and Bloch's theorem

The Hamiltonian of the lattice is invariant under a translation by any lattice vector $\boldsymbol{R}$. We construct the translation operator in terms of the momentum operator $\boldsymbol{p}$ as

$$\begin{aligned} T(\boldsymbol{R}) &= e^{i\boldsymbol{R}.\boldsymbol{p}/\hbar} && (3.8)\\ T(\boldsymbol{R})^{-1} H T(\boldsymbol{R}) &= H \\ \therefore [T(\boldsymbol{R}), H] &= 0 && (3.9) \end{aligned}$$

Since $\boldsymbol{p}$ is hermitian, $T(\boldsymbol{R})$ is unitary. The commutator tells us that $T(\boldsymbol{R})$ and H has simultaneous eigenfunctions. So if $|\Psi\rangle$ is an eigenfunction of H, it should also satisfy the two following conditions

$$\begin{aligned} T(\boldsymbol{R})\psi(\boldsymbol{r}) &= c(\boldsymbol{R})\psi(\boldsymbol{r}) && (3.10)\\ T(\boldsymbol{R})\psi(\boldsymbol{r}) &= \psi(\boldsymbol{r}+\boldsymbol{R}) && (3.11) \end{aligned}$$

Now what is $c(\boldsymbol{R})$?

$$\begin{aligned} \langle \boldsymbol{k}|T(\boldsymbol{R})|\psi\rangle &= \langle\psi|T^{\dagger}(\boldsymbol{R})|\boldsymbol{k}\rangle^* \\ &= \left(\int d^3\boldsymbol{r}\ \psi^*(\boldsymbol{r}).e^{-i\boldsymbol{R}.\boldsymbol{p}/\hbar}.e^{i\boldsymbol{k}.\boldsymbol{r}}\right)^* \\ &= \left(\int d^3\boldsymbol{r}\ \psi^*(\boldsymbol{r}).e^{i\boldsymbol{k}.(\boldsymbol{r}-\boldsymbol{R})}\right)^* \\ &= e^{i\boldsymbol{k}.\boldsymbol{R}}\langle\boldsymbol{k}|\psi\rangle && (3.12) \end{aligned}$$

Then left multiply 3.10 with $\langle\boldsymbol{k}|$ and put together the result with 3.12. This leads to

$$\left(c(\boldsymbol{R}) - e^{i\boldsymbol{k}.\boldsymbol{R}}\right)\langle\boldsymbol{k}|\psi\rangle = 0 \qquad (3.13)$$

Using this and eqn 3.11 you should be able to reproduce the properties of the Bloch functions.

3.1.2 Significance of $\boldsymbol{k}$

The vector $\boldsymbol{k}$ looks like momentum ($\hbar\boldsymbol{k}$), but it is not. To see this, let us calculate the momentum of a particle in a Bloch state. We need the expectation value of the momentum operator $\boldsymbol{p} = -i\hbar\nabla$

$$\langle\boldsymbol{p}\rangle = \langle\psi_{\boldsymbol{k}}| - i\hbar\nabla|\psi_{\boldsymbol{k}}\rangle = \hbar\boldsymbol{k} + \hbar\sum_{\boldsymbol{G}}\boldsymbol{G}|c_{\boldsymbol{k}+\boldsymbol{G}}|^2 \qquad (3.14)$$

PROBLEM : Prove eqn. 3.14.

To distinguish $\hbar\boldsymbol{k}$ from momentum, we will call it the *crystal momentum*. The significance of this will be clear when we write the equations of semiclassical dynamics of the electron.

3.1.3 Origin of the band gap: solving the matrix equation

Let us for the moment consider a simplified "toy" case to appreciate an (perhaps the most) important consequence of a periodic potential.

1. We consider a 1-dimensional case (chain with a period a)

2. We consider a periodic potential

$$V(x) = 2V_0 \cos\frac{2\pi x}{a} = V_0 e^{i\frac{2\pi}{a}x} + V_0 e^{-i\frac{2\pi}{a}x} \tag{3.15}$$

which means that only two (reciprocal lattice) components of the potential contribute:

$$V_{\frac{2\pi}{a}} = V_{-\frac{2\pi}{a}} = V_0 \tag{3.16}$$

So as to write out the matrix of eqn 3.5 we need to order all the reciprocal vectors according to some rule (which we can frame for ourselves). In the 1D case we can write all the RLVs as:

$$V_n = n.\frac{2\pi}{a} \qquad (\text{n} = ... -2, -1, 0, 1, 2, ...) \tag{3.17}$$

So the matrix resulting from the eqn. 3.6 is a tridiagonal matrix, a *part* of which looks like (assuming $\hbar^2/2m = 1$) for simplicity

$$\begin{vmatrix}
\left(k+3\frac{2\pi}{a}\right)^2 - E & V_{\frac{2\pi}{a}} & 0 & 0 & 0 & 0 & 0 \\
V_{-\frac{2\pi}{a}} & \left(k+2\frac{2\pi}{a}\right)^2 - E & V_{\frac{2\pi}{a}} & 0 & 0 & 0 & 0 \\
0 & V_{-\frac{2\pi}{a}} & \left(k+\frac{2\pi}{a}\right)^2 - E & V_{\frac{2\pi}{a}} & 0 & 0 & 0 \\
0 & 0 & V_{-\frac{2\pi}{a}} & k^2 - E & V_{\frac{2\pi}{a}} & 0 & 0 \\
0 & 0 & 0 & V_{-\frac{2\pi}{a}} & \left(k-\frac{2\pi}{a}\right)^2 - E & V_{\frac{2\pi}{a}} & 0 \\
0 & 0 & 0 & 0 & V_{-\frac{2\pi}{a}} & \left(k-2\frac{2\pi}{a}\right)^2 - E & V_{\frac{2\pi}{a}} \\
0 & 0 & 0 & 0 & 0 & V_{-\frac{2\pi}{a}} & \left(k-3\frac{2\pi}{a}\right)^2 - E
\end{vmatrix} = 0 \tag{3.18}$$

In the notation of eqn. 3.6, $\boldsymbol{G}$ varies along a row, $\boldsymbol{K}$ is kept fixed. For every $\boldsymbol{K}$ we have a new row. If we had more components of the potential (*e.g.*$V_{\frac{4\pi}{a}}$,$V_{-\frac{4\pi}{a}}$) then another band would appear symmetrically about the diagonal line. The matrix is necessary symmetrical and "band diagonal".

Now if in the matrix we did the replacement $k \to k + n\frac{2\pi}{a}$, what happens? Notice that k only occurs in the diagonal. The off diagonal terms would not change at all. Convince yourself that the effect of the substitution is simply sliding the entire diagonal by a few units along the diagonal. The eigenvalues and eignevectors would remain same. Write out a small section of the matrix, do the replacement and convince yourself!

Finally, can you work out the reason for concentrating close to the region $k, k \pm \frac{\pi}{a}, k \pm \frac{2\pi}{a}$? Why was it all right to leave out the regions far from this?

Zero potential case : the "empty lattice"

Now let's consider the apparently trivial case (but there is purpose!) where all the potential components are zero. Then the eigenvalues must be the free electron values

$$E = \frac{\hbar^2}{2m}\left(k - n\frac{2\pi}{a}\right)^2 \qquad \text{with} \quad -\frac{\pi}{a} < k < \frac{\pi}{a} \tag{3.19}$$

It might be a little counter-intuitive to see how the plot with shifted parabolas look. We are used to seeing one continuous parabola ($E \propto k^2$) as the dispersion, because in the free particle case there is no lattice. But if we now shift any vector from outside the first Brillouin zone into the first zone by reciprocal vector translations then this is how it will look. This way of plotting it is called the "reduced zone" scheme and is the most commonly used way of plotting bandstructure. If the lattice is 2d or 3d (Kittel gives an example) then these plots can look quite striking, but it is perfectly logical way of plotting it. We will see soon that the effect of a small non-zero potential would be to smooth the many sharp corners and crossings in the plot as seen in Fig. 3.1.

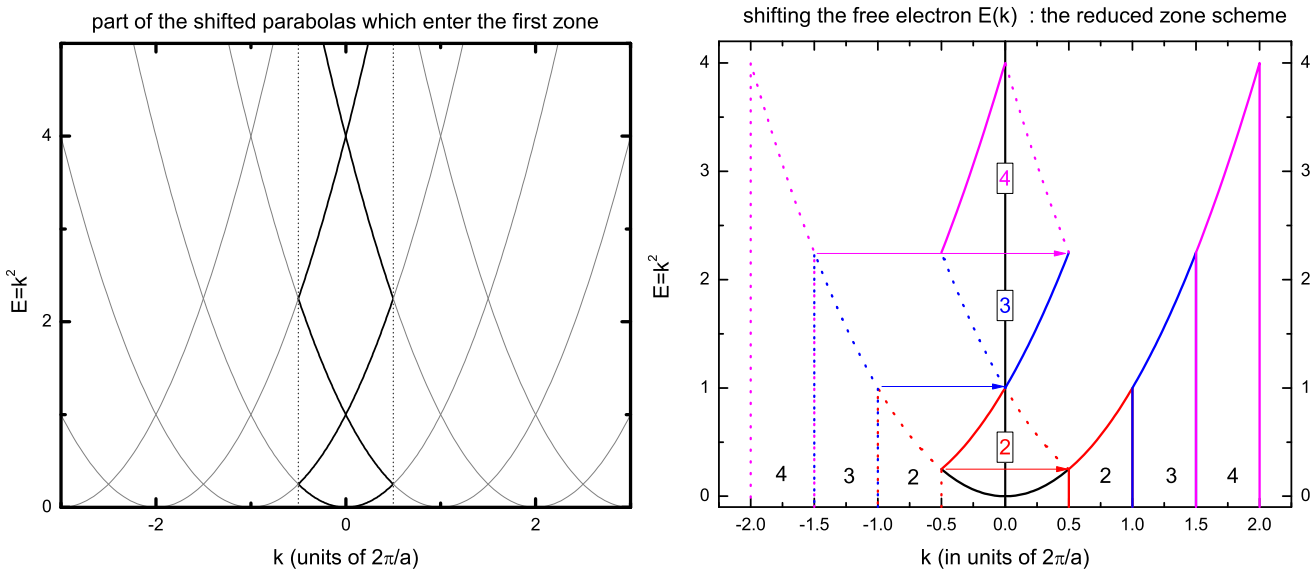


Figure 3.1: (left)Free electron $E(k)$ in the reduced zone scheme. The equations $E = \left(k - n.\frac{2\pi}{a}\right)^2$, and the part which enters the first Brillouin zone. The part in the first zone is in black, the outside parts are grayed out. (right) Another similar plot , showing how the same can be obtained from $E = k^2$ by shifting different parts by different reciprocal lattice translations.

Turning on a small potential

One might think that a way to solve the problem with a small periodic potential would be to use perturbation, starting with the free electron eignestates. This however is not in general useful. The main reason is that the free electron system is highly degenerate. Many of the degeneracies (*e.g.* between $\boldsymbol{k}$ and $-\boldsymbol{k}$) are not lifted by the periodic potential to first order. Second order degenerate perturbation is algebraically very messy! Even if the problem is done that way, it will not give us the physical insight offered by Bloch's theorem.

However, it turns out that there is a particular case where first order perturbation does break the degeneracy. This happens when the two degenerate states are separated (in a vector sense) by a reciprocal lattice vector.

We can see that at the point where the shifted parabolas cross, two eignevalues are degenerate. The first case happens at $k = \frac{\pi}{a}$. Here the $E = k^2$ and $E = \left(k - \frac{2\pi}{a}\right)^2$ branches give the same value if $V = 0$. What happens if we turn on a small V? First we just solve for the eigenvalues of the matrix equation 3.18. The plot shows a crucial prediction of the Bloch equation. If we have a potential with a Fourier component ($v_{\boldsymbol{G}}$) then a gap opens at $\boldsymbol{G}/2$ These are precisely the zone boundaries. Note that the sharp crossings of Fig. 3.1 have been "rounded off". It turns out that this is a generic feature of turning on a small interaction potential and you will see this in many situations.

PROBLEM : A quick (though bit handwaving) way of seeing what happens when a small potential is "turned on" is to solve the 2×2part of the matrix eqn 3.18:

$$\begin{vmatrix} k^2 - E & V_0 \\ V_0 & \left(k - \frac{2\pi}{a}\right)^2 - E \end{vmatrix} = 0 \tag{3.20}$$

where $V_{-\frac{2\pi}{a}} = V_{\frac{2\pi}{a}} = V_0$. Now solve for the eignevalues, this should give the behaviour near $k = \pi/a$. You should be able to show that the two eigenvalues at $k = \pi/a$ differ by $2V_0$.

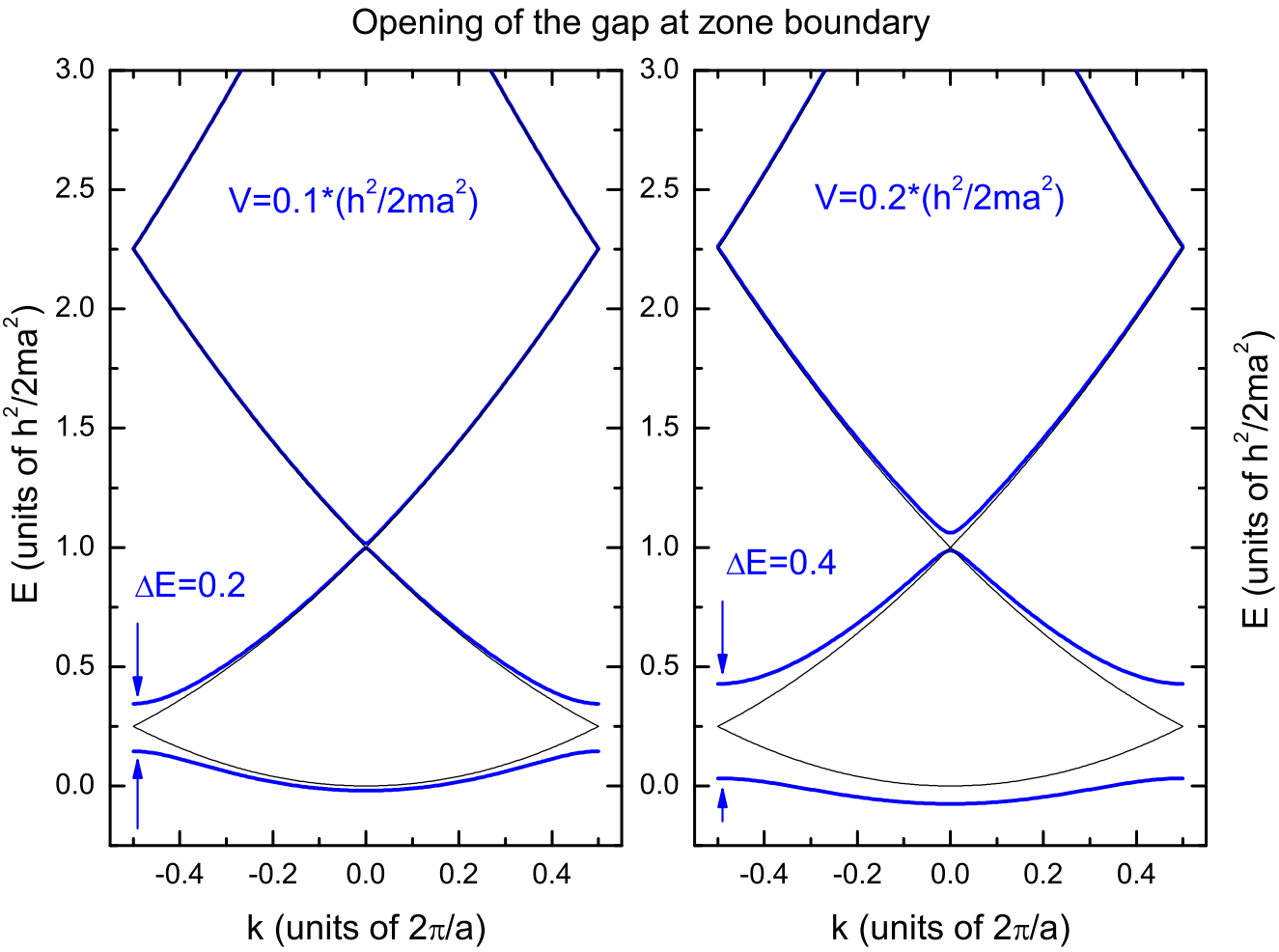


Figure 3.2: $E(k)$ in the reduced zone scheme, with a single non-zero component of the potential as in the equation 3.18. The energies have been scaled with $E = h^2/2ma^2$, which is 4 times the kinetic energy at the zone boundary, to make the plot. Similarly k has been scaled by $2\pi/a$. The black lines show the free electron result with no potential turned on. Note that the potential also affects higher bands but much less..

Brillouin zone boundary and the intersection of the parabolas

For a certain $\boldsymbol{G}$ two parabolas given by $E = \boldsymbol{k}^2$ and $E = (\boldsymbol{k} - \boldsymbol{G})^2$ will intersect at a point given by

$$\frac{\boldsymbol{G}}{2}.\left(\frac{\boldsymbol{G}}{2} - \boldsymbol{k}\right) = 0 \tag{3.21}$$

You should be able to show that this implies that the tip of the vector $\boldsymbol{k}$ lies on (line or plane depending on 2D/3D) the perpendicular bisector of the vector $\boldsymbol{G}$ drawn from the same origin. This is precisely how we construct the first Brillouin zone of the reciprocal lattice. Thus the lifting of the degeneracies begins at the zone boundary. This is a very important generic feature.

Extended and repeated zone schemes

Exactly the same data can be plotted in some different ways. The repeated zone scheme simply means that we emphasize the periodicity of the solution that $E(\boldsymbol{k}) = E(\boldsymbol{k} + \boldsymbol{G})$. The extended zone scheme shows very clearly the deviation from the free electron parabola. The data used to plot Fig. 3.2 has been plotted in these two schemes in the next figure, Fig .3.3.

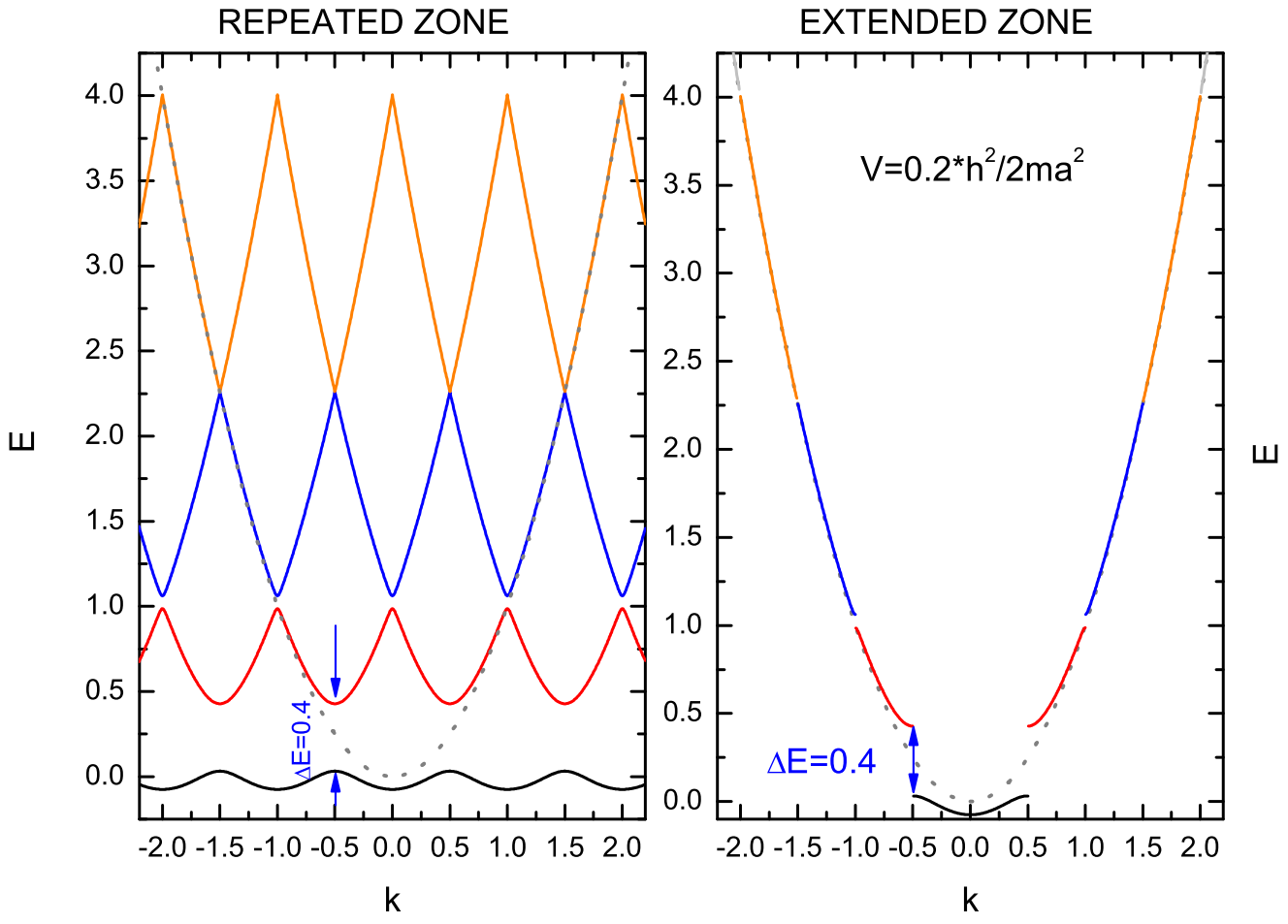


Figure 3.3: $E(k)$ in the repeated and extended zone schemes. The data used is exactly the same as in Fig. 3.2.

Band structure as a perturbation problem

Some of the unperturbed state kets $|\boldsymbol{k}\rangle$ will now be mixed. The point to note is that only those states of the form $|\boldsymbol{k}+\boldsymbol{G}\rangle$ will mix with $|\boldsymbol{k}\rangle$. We write the perturbed state kets to *first* order and perturbed energies to *second* order.

$$\begin{aligned} |\psi_{\boldsymbol{k}}\rangle &= |\boldsymbol{k}\rangle + \sum_{\boldsymbol{G}\neq\boldsymbol{0}} \frac{\langle \boldsymbol{k}|V|\boldsymbol{k}+\boldsymbol{G}\rangle}{E^0(\boldsymbol{k}) - E^0(\boldsymbol{k}+\boldsymbol{G})}|\boldsymbol{k}+\boldsymbol{G}\rangle \\ &= \sum_{\boldsymbol{G}\neq\boldsymbol{0}} \frac{v_{\boldsymbol{G}}}{E^0(\boldsymbol{k}) - E^0(\boldsymbol{k}+\boldsymbol{G})}|\boldsymbol{k}+\boldsymbol{G}\rangle \qquad (3.22) \\ E(\boldsymbol{k}) &= E^0(\boldsymbol{k}) + \cancel{\langle \boldsymbol{k}|V|\boldsymbol{k}\rangle} + \sum_{\boldsymbol{G}\neq\boldsymbol{0}} \frac{|\langle \boldsymbol{k}|V|\boldsymbol{k}+\boldsymbol{G}\rangle|^2}{E^0(\boldsymbol{k}) - E^0(\boldsymbol{k}+\boldsymbol{G})} \\ &= E^0(\boldsymbol{k}) + \sum_{\boldsymbol{G}\neq\boldsymbol{0}} \frac{|v_{\boldsymbol{G}}|^2}{E^0(\boldsymbol{k}) - E^0(\boldsymbol{k}+\boldsymbol{G})} \qquad (3.23) \end{aligned}$$

PROBLEM : We have dropped the first order energy shift in eqn. 3.23. Why are we justified in doing this? When will this be incorrect?

---

1. We do not need to make the sum in eqn. 3.22 run over all $\boldsymbol{k'} \neq \boldsymbol{k}$ but only some of them because $\langle \boldsymbol{k}|V|\boldsymbol{k'}\rangle = 0$ unless $\boldsymbol{k'} = \boldsymbol{k}+\boldsymbol{G}$. The consequence of all this is that the first order result gives us something of the Bloch form.
2. But if $E^0(\boldsymbol{k}) - E^0(\boldsymbol{k}+\boldsymbol{G}) = 0$ then the denominator will vanish and this will not work. This happens when two states are degenerate. This is precisely where two parabolas in Fig. 3.1 intersect. These are the zone boundaries.
3. At the zone boundaries, the energy shift is no longer of second order, this is then a degenerate first order problem. The solution (see any quantum mechanics text) for the energy shifts are to be obtained by taking the eigenvalues of the matrix V_{ij} between the degenerate states. This is precisely what we did when we solved for (in a previous problem):

$$\begin{vmatrix} E^0(\boldsymbol{k}) - E & \langle \boldsymbol{k}|V|\boldsymbol{k}+\boldsymbol{G}\rangle \\ \langle \boldsymbol{k}+\boldsymbol{G}|V|\boldsymbol{k}\rangle & E^0(\boldsymbol{k}+\boldsymbol{G}) - E \end{vmatrix} = 0 \qquad (3.24)$$

3.1.4 The Kronig-Penny model

A simple model of a periodic potential is shown in the figure. The well-barrier-well sequence is more realistic than a single fourier component. It is a very useful "toy model" for understanding how the energy bands and band gaps in a solid arises.

PROBLEM : Revise the single electron in a finite quantum well. Consider a potential given by

$$\begin{aligned} V &= 0 \quad \text{for} \quad -w < x < 0 \\ &= V_0 \qquad \text{otherwise} \end{aligned}$$

Write the wavefunctions for bound states, piecewise as follows

$$\begin{aligned} \Psi_1 &= Ae^{i\alpha x} + Be^{-i\alpha x} \quad \text{for} \quad -w < x < 0 \\ \Psi_2 &= Ce^{-\beta x} \qquad\qquad \text{for} \quad x > 0 \\ \Psi_3 &= De^{\beta x} \qquad\qquad \text{for} \quad x < -w \qquad (3.25) \end{aligned}$$

where α and β are given by:

$$\begin{aligned} \alpha^2 &= \frac{2mE}{\hbar^2} \\ \beta^2 &= \frac{2m(V_0 - E)}{\hbar^2} \end{aligned} \tag{3.26}$$

The wavefn and its derivative must be continuous at the two boundaries of the potential well. Show that the solutions are obtained by solving the set of linear equations:

$$\begin{pmatrix} 1 & 1 & -1 & 0 \\ i\alpha & -i\alpha & \beta & 0 \\ e^{-i\alpha w} & e^{i\alpha w} & 0 & -e^{-\beta w} \\ i\alpha e^{-i\alpha w} & -i\alpha e^{i\alpha w} & 0 & -\beta e^{-\beta w} \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} = 0 \tag{3.27}$$

There are 5 unknowns, but one of the coefficients can be chosen arbitrarily and normalisation will take care of it. Setting the determinant to zero you can show that the wavevector α can be obtained from one of the two conditions

$$\begin{aligned} \tan\frac{\alpha w}{2} &= \frac{\beta}{\alpha} \\ \tan\frac{\alpha w}{2} &= -\frac{\alpha}{\beta} \end{aligned} \tag{3.28}$$

Once k is determined calculate E and then the coefficients A, B, C, D. By arbitrarily setting $C = 1$ The *unnormalised* wavefunction can be written as

$$\begin{aligned} \Psi_1 &= \cos\alpha x - \frac{\beta}{\alpha}\sin\alpha x && \text{for } [-w, 0] \\ \Psi_2 &= e^{-\beta x} && \text{for } [0, \infty] \\ \Psi_3 &= \left(\cos\alpha w + \frac{\beta}{\alpha}\sin\alpha w\right) e^{\beta(w+x)} && \text{for } [-\infty, -w] \end{aligned} \tag{3.29}$$

We now put many such potential wells in an array separated by b units. So we have a repeating structure with period $a = b + w$. In the barrier regions the wavevector cannot be real, because $E < V_0$. However the exponential can have both decaying and growing parts, because the barrier region does not tend to infinity. There is no chance for one exponential to blow up. We can now write the wavefunction in well (Ψ_w) and barrier (Ψ_b) as :

$$\Psi_w = Ae^{i\alpha x} + Be^{-i\alpha x} \tag{3.30}$$

$$\Psi_b = Ce^{\beta x} + De^{-\beta x} \tag{3.31}$$

where α and β are defined by 3.26 as before. Because of the periodicity of the structure, the wavefunction must satisfy the Bloch condition. Hence we can write , with k as crystal momentum:

$$\begin{aligned} \Psi_w(0) &= \Psi_b(0) \\ \left.\frac{d\Psi_w}{dx}\right|_{x=0} &= \left.\frac{d\Psi_b}{dx}\right|_{x=0} \\ \Psi_b(b) &= e^{ika}\Psi_w(-w) \\ \left.\frac{d\Psi_b}{dx}\right|_{x=b} &= e^{ika}\left.\frac{d\Psi_w}{dx}\right|_{x=-w} \end{aligned} \tag{3.32}$$

Notice how the Bloch condition helped us to write the last two equations of the set. We now have four equations connecting A, B, C, D and k.
The four equations are :

$$\begin{aligned} A+B &= C+D \\ Ai\alpha - Bi\alpha &= C\beta - D\beta \\ Ce^{\beta b} + De^{-\beta b} &= e^{ika}\left[Ae^{-i\alpha w} + Be^{i\alpha w}\right] \\ C\beta e^{\beta b} - D\beta e^{-\beta b} &= e^{ika}\left[Ai\alpha e^{-i\alpha w} - Bi\alpha e^{i\alpha w}\right] \end{aligned} \tag{3.33}$$

This leads to the following consistency condition:

$$\left(\frac{\beta^2-\alpha^2}{2\alpha\beta}\right)\sinh\beta b\sin\alpha w + \cosh\beta b\cos\alpha w = \cos ka \tag{3.34}$$

What are the allowed values of k

Let's assume the structure is a loop with length L such that $L = Na$ Because we impose periodic boundary conditions we must have

$$\Psi(x+L) = \Psi(x+Na) = \Psi(x) \tag{3.35}$$

But Bloch condition requires

$$\Psi(x+L) = e^{ikL}\Psi(x) \tag{3.36}$$

These two together imply

$$\begin{aligned} e^{ikL} &= 1 \\ \therefore kL &= 2n\pi \\ \therefore k &= \frac{2n\pi}{L} \\ \therefore k &= \frac{n}{N}\frac{2\pi}{a} \end{aligned} \tag{3.37}$$

In any interval $2\pi/a$ there are thus N states, where N is the number of unit cells in the structure. Later on we will prove it for 2D and 3D also for any lattice.
The equation 3.34 can be solved numerically, following the process:

- Notice that the solution $E(k)$ is periodic in k. So we restrict k within one period.
- Choose a k, in the range $-\pi/a < k < \pi/a$. This fixes the RHS.
- The only unknown in the LHS is the energy E, because both α and β are determined by E.
- Allow α to vary from 0 to $\sqrt{2mV_0/\hbar^2}$. Pick up all the solutions. There are a finite number of them only.
- Calculate the corresponding E values and plot them vs k (not α).
- This is the band structure.

Here's an example
And a couple more, where we have increased the separation, keeping all other parameters same.
And then:

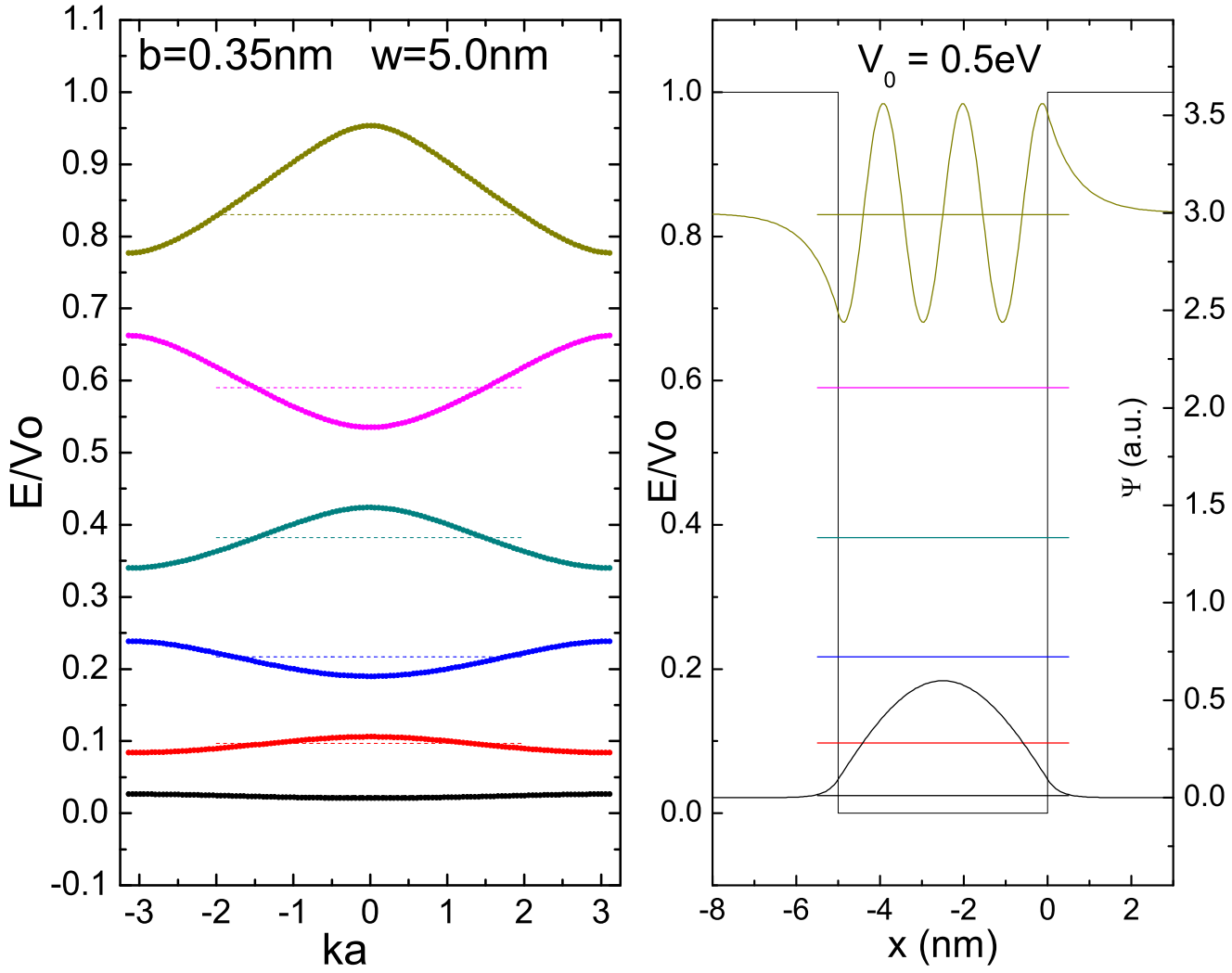


Figure 3.4: Notice how a single energy level has spread out into a band. The deeper levels are not very spread out, the higher (loosely bound) levels are spread more. This indeed is the crux of band structure. The more the possible overlap between wavefunctions at neighbouring sites, more will be the spread of the band. The right hand figure shows the single potential solutions for eigenvalues and eignefunctions. The left hand figure is a plot of the eigenvalues (allowed energies) when a series of them are laid out on a chain.

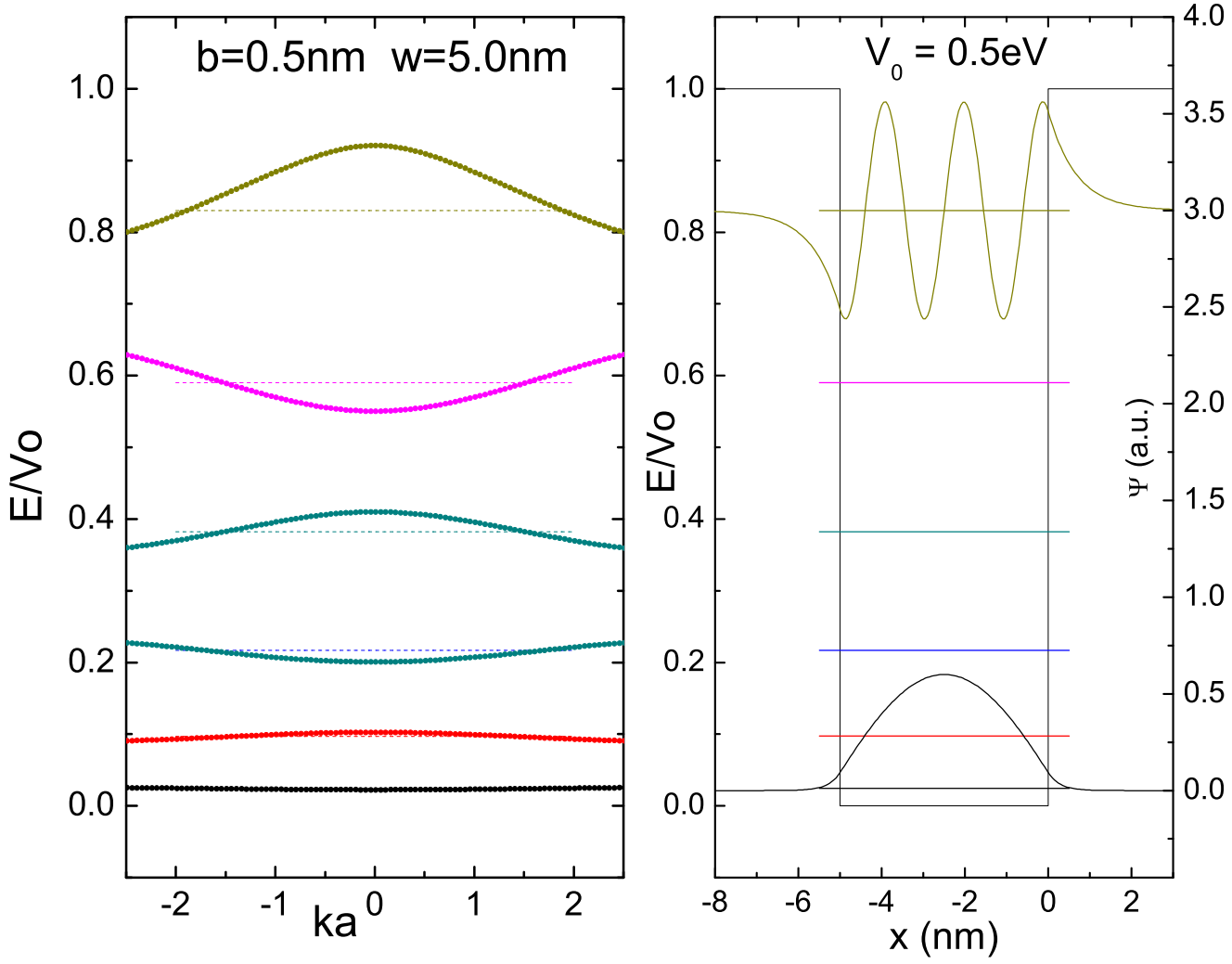


Figure 3.5: b=0.5nm, w= 5nm, V_0=0.5eV. Notice that the bands have become narrower

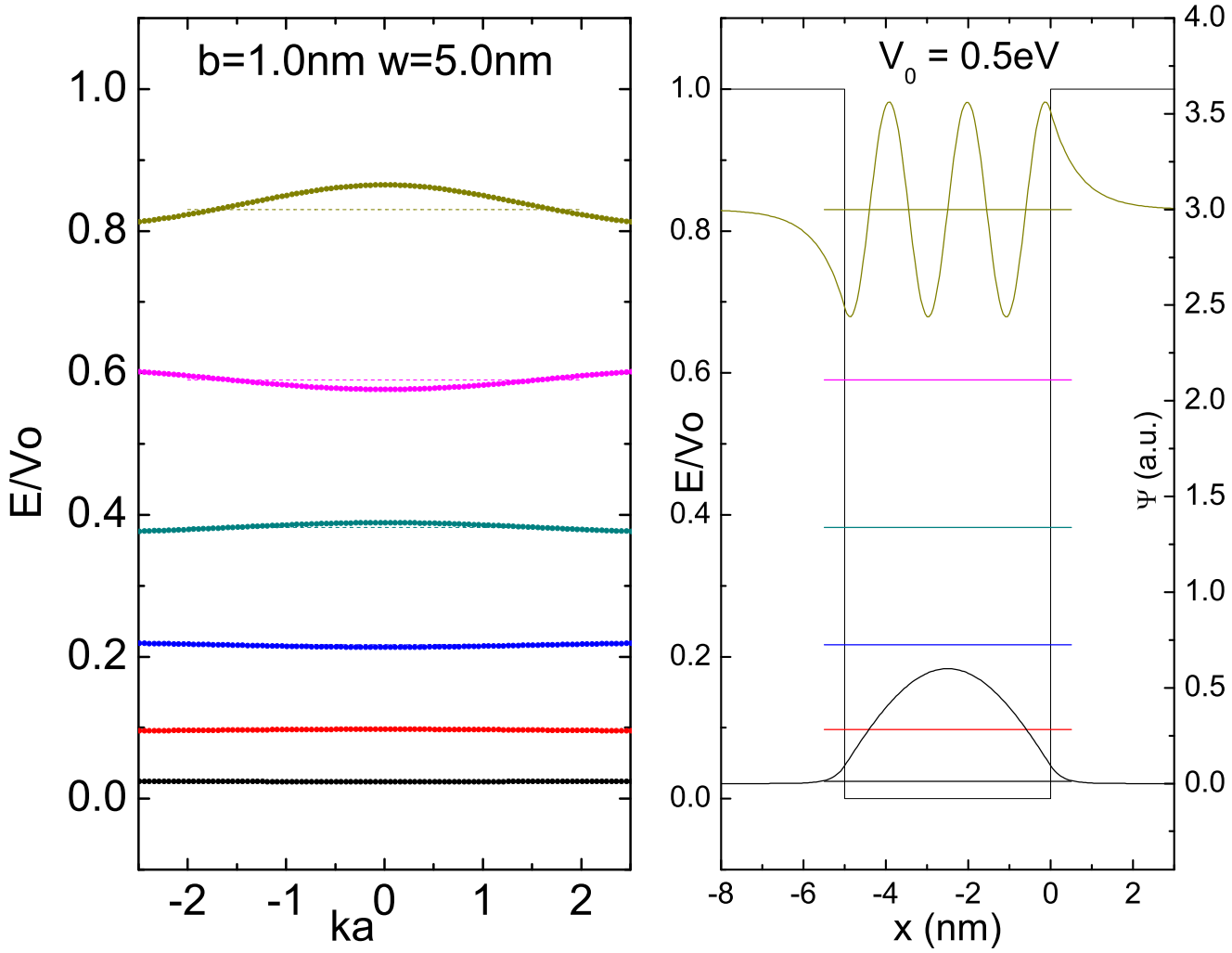


Figure 3.6: b=1nm, w= 5nm, V_0=0.5eV. The bands are getting even more narrow.

3.1.5 The Band gap: classification of conducting, non-conducting and semiconducting substances

We have spoken about the importance of the density of states at the Fermi level before, while calculating quantities like specific heat, susceptibility, conductivity etc. Now since the energy spectrum has gaps in it, it is possible to have the Fermi level in qualitatively different positions.

Fermi level in a continuous band : Metals

If the electrons fill a band partially, then the $D(E_F)$ will in general be large. It will be easy to accelerate an electron with a small electric field. Responses which depend on $D(E_F)$ - like thermal and electrical conductivities will also be large.

Fermi level in a large bandgap: Insulators

If the band gap is large (few electron volts) and the available electrons fill up the lower band fully, then the Fermi level is forced to lie in a gap. $D(E_F) = 0$ and the material cannot respond to small electric fields. This is an insulator.

Fermi level in a small bandgap: Semiconductors

By small we mean typically not more than ~ 2 eV. The distinction between semiconductors and insulators is a qualitative one. In semiconductors like Silicon, Germanium, Gallium Arsenide there are some carriers in the upper band due to thermal excitations at room temperature. Large bandgap materials can behave like semiconductors at high temperatures.

3.2 Motion of an electron in a band: Bloch oscillation, group velocity and effective mass

How do the concepts of position, velocity etc carry over to the Bloch electrons? Recall that a Bloch state $|\Psi\rangle = e^{ikx}|u_k\rangle$ is not in general an eigenfunction of the momentum operator : $p = -i\hbar\frac{\partial}{\partial x}$.

3.2.1 Effect of an electric field

Consider a (Bloch state) electron subjected to an electric field $\boldsymbol{F}$. We denote the electric field with $\boldsymbol{F}$ to avoid confusion with the energy E. For a free particle, with charge q, we would expect the driving equation to be

$$\frac{d\boldsymbol{p}}{dt} = \hbar\frac{d\boldsymbol{k}}{dt} = q\boldsymbol{F} \tag{3.38}$$

The electric or magnetic field enters the Schrodinger equation via the potentials ($\boldsymbol{A}$,V), the most obvious (but not the only) way to introduce an electric field would be $V(\boldsymbol{r}) = -\boldsymbol{F}.\boldsymbol{r}$ The problem with this is that then the total potential (and hence the hamiltonian) would no longer have the symmetry $V(\boldsymbol{r}+\boldsymbol{R}) = V(\boldsymbol{r})$. However since

$$\boldsymbol{E} = -\frac{\partial \boldsymbol{A}}{\partial t} - \nabla V \tag{3.39}$$

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \tag{3.40}$$

We introduce the extra E using a time dependent $A = qFt$. So the Hamiltonian including $A(t)$ becomes:

$$H\psi = \left[\frac{1}{2m}(p + qFt)^2 + qV(x)\right]\psi = E(t)\psi \tag{3.41}$$

At anytime t, it has the requisite translational symmetry and its eigenstates should be Bloch states. The eigenvalues would however become time dependent. For simplicity we write x instead of the vector $\boldsymbol{r}$ and

so on - but the arguments work equally well in 2 or 3 dimensions.

To solve this we assume a solution in the form,

$$\psi = e^{i\lambda(t)x}\phi \tag{3.42}$$

At this point ϕ is not a known function. Now notice that

$$\begin{aligned} (p + qFt)\, e^{i\lambda(t)x}\phi &= \hbar\lambda e^{i\lambda(t)x}\phi + e^{i\lambda(t)x}p\phi + qFt\ e^{i\lambda(t)x}\phi & (3.43)\\ &= (\hbar\lambda + qFt)\, e^{i\lambda(t)x}\phi + e^{i\lambda(t)x}p\phi & (3.44)\end{aligned}$$

If we chose $\lambda(t)$ such that

$$\hbar\lambda(t) + qFt = 0 \tag{3.45}$$

then the first term will vanish. By applying the same operation twice we get

$$e^{i\lambda x}\left[\frac{1}{2m}p^2 + qV(x)\right]\phi = e^{i\lambda x}E(t)\phi \tag{3.46}$$

Notice how the explicit time dependence of the hamiltonian has been canceled. ϕ must be of the Bloch form where the k parameter can be time dependent *i.e.* $k(t)$.

$$\begin{aligned} \psi &= e^{i\lambda x}.e^{ik(t)x}u_k(x)\\ \text{and}\quad \psi(x+L) &= \psi(x)\\ \text{since}\ \ u_k(x+L) &= u_k(x)\\ \therefore e^{i[\lambda(t)L+k(t)L]} &= 1\\ \therefore -\frac{qFt}{\hbar} + k(t) &= \frac{2n\pi}{L} & (3.47)\\ \therefore \frac{dk}{dt} &= \frac{qF}{\hbar} & (3.48)\end{aligned}$$

The Bloch state evolves into another Bloch state with a different k value. In this case $\hbar k$ behaves as if it was the momentum of the particle. This however is not correct as we have seen earlier. The equation 3.48 can be generalized to 2 or 3d easily. More importantly in presence of a magnetic field it can be shown (we state this without proof) that

$$\hbar\frac{d\boldsymbol{k}}{dt} = q\left(\boldsymbol{F} + \boldsymbol{v}\times\boldsymbol{B}\right) \tag{3.49}$$

where $\boldsymbol{v} = \dfrac{1}{\hbar}\nabla_{\boldsymbol{k}}E(\boldsymbol{k})$, the group velocity of the particles. This we will prove next, but before that let's ask one question.

Why is this simple result very remarkable?

The force that acts on the electrons in a solid is a combination of the electric field of the ions and the field applied from outside - for example by connecting a battery (of typically a few volts) across a metal wire or a piece of the semiconductor. The field created by the ions is several thousand times larger than the field resulting from the potential drop due to the battery. A simple estimate of the periodic potential from ions is 10-100V across 1Å. The battery may create a field of $\sim$ 10V/millimeter, *which is not even a millionth of the ionic field*. Yet it *appears* that the evolution of the $\boldsymbol{k}$ vector is only due to the external field. The "real" momentum of the electron evolves due to the total force on it, but the evolution of the $\boldsymbol{k}$ vector neatly factors out - giving the impression that it *is* the momentum. The index $\boldsymbol{k}$ that we introduced in the Bloch wavefunction, is called the crystal momentum. We emphasize that it is not the real momentum. Infact *the momentum transferred by an external electric field to a piece of wire is exactly zero*, because the wire is as a whole electrically neutral and an electric field cannot impart any momentum to an object that is overall electrically neutral.

The group velocity of Bloch electrons:

$$\begin{aligned} p|\Psi\rangle &= \hbar k|\Psi\rangle + e^{ikx}(-i\hbar)\frac{\mathrm{d}}{\mathrm{d}x}|u_k\rangle \\ \therefore (p-\hbar k)\,|\Psi\rangle &= e^{ikx}p|u_k\rangle \\ \therefore p|\Psi\rangle &= e^{ikx}\,(p+\hbar k)\,|u_k\rangle \end{aligned} \tag{3.50}$$

It is left as a little exercise to show that this implies the following result

$$\begin{aligned} \left(\frac{p^2}{2m}+V\right)|\Psi\rangle &= E(k)|\Psi\rangle \\ H(k)|u_k\rangle = \left(\frac{(p+\hbar k)^2}{2m}+V\right)|u_k\rangle &= E(k)|u_k\rangle \\ \frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{(p+\hbar k)^2}{2m}+V\right)|u_k\rangle &= \frac{\mathrm{d}}{\mathrm{d}k}E(k)|u_k\rangle \\ \frac{\hbar}{m}\,(p+\hbar k)\,|u_k\rangle + H(k)\frac{\mathrm{d}}{\mathrm{d}k}|u_k\rangle &= \frac{\mathrm{d}E}{\mathrm{d}k}|u_k\rangle + E\frac{\mathrm{d}}{\mathrm{d}k}|u_k\rangle \\ \frac{\hbar}{m}\langle u_k|p+\hbar k|u_k\rangle + \cancel{\langle u_k|H(k)\frac{\mathrm{d}}{\mathrm{d}k}|u_k\rangle} &= \frac{\mathrm{d}E}{\mathrm{d}k}\langle u_k|u_k\rangle + \cancel{E\langle u_k|\frac{\mathrm{d}}{\mathrm{d}k}|u_k\rangle} \qquad (3.51) \\ \therefore \langle\Psi|\frac{p}{m}|\Psi\rangle &= \frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}k} \qquad (3.52) \end{aligned}$$

Bloch oscillation

We now know the following

- The $E(k)$ relation is periodic in k. Let's consider a case where $E = E_0 - 2\gamma\cos ka$.
- An electric field (F) changes the k vector such that $dk/dt = -eF/\hbar$.
- Now suppose there is no scattering, what is the equation of motion of the electron?

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}k} \\ &= \frac{2\gamma a}{\hbar}\sin k(t)a \\ &= \frac{2\gamma a}{\hbar}\sin\left(k(0)a - \frac{eaFt}{\hbar}\right) \\ \therefore x(t) &= \frac{2\gamma a}{\hbar}\int_0^t \mathrm{d}t\ \sin\left(k(0)a - \frac{eaFt}{\hbar}\right) \end{aligned} \tag{3.53}$$

We therefore see that the electron will *oscillate* with a frequency $\omega_{Bloch} = eaF/\hbar$, a very striking prediction of quantum mechanics that is believed to be correct but has only been partially verified. Putting $a = 1$nm, $F = 10$ V/cm we find that the frequency would be $\sim 10^9$Hz.

PROBLEM : How is the amplitude of Bloch oscillation related to the other quantities in the problem? For a bandwidth of 1 ev, how does this length compare to typical mean free paths in a very clean metal at low temperature - say 500 microns?

Effective mass

Mass is the constant of proportionality between the applied force, $\boldsymbol{F}$, and the acceleration produced $\boldsymbol{a}$. If we carry over this idea to the motion of electrons in a band, we get

$$\begin{aligned} \boldsymbol{a} &= \frac{d}{dt}\boldsymbol{v} \\ &= \frac{d}{dt}\frac{1}{\hbar}\nabla_{\boldsymbol{k}}E(\boldsymbol{k}) \end{aligned}$$

Writing it out componentwise and remembering that v_x can be a function of *all* components k_x, k_y and k_z we get:

$$\begin{aligned} a_i &= \frac{\mathrm{d}}{\mathrm{dt}}v_i(k_j) \\ &= \frac{\partial v_i}{\partial k_j}\frac{\mathrm{d}k_j}{\mathrm{dt}} \\ &= \frac{1}{\hbar}\frac{\partial^2 E}{\partial k_i \partial k_j}\frac{F_j}{\hbar} \end{aligned} \tag{3.54}$$

The sum over the repeated index j is implied. The structure of the eqn. 3.54 tells us that we can write an inverse effective mass matrix

$$\mathrm{M}^{-1}_{\mathrm{ij}} = \frac{1}{\hbar^2}\frac{\partial^2 E}{\partial k_i \partial k_j} \tag{3.55}$$

Calculating the effective mass in a band: $k.p$ method

Using what we did for calculating the group velocity of Bloch electrons, you can show the following

---

PROBLEM : A Bloch state is a solution of the Schrodinger equation in a periodic potential, $V(\boldsymbol{r})$, of the crystal. It can be written as

$$\Psi_{n\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}.\boldsymbol{r}}u_{n,\boldsymbol{k}}(\boldsymbol{r})$$

where $u_{n,\boldsymbol{k}}(\boldsymbol{r}) = u_{n,\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R})$ for any direct lattice vector $\boldsymbol{R}$. Given that $\Psi_{n\boldsymbol{k}}(\boldsymbol{r})$ is a solution of

$$\left[\frac{1}{2m}\boldsymbol{p}^2 + V(\boldsymbol{r})\right]\Psi_{n,\boldsymbol{k}}(\boldsymbol{r}) = E_{n,\boldsymbol{k}}\Psi_{n,\boldsymbol{k}}(\boldsymbol{r})$$

Show that the function $u_{n,\boldsymbol{k}}(\boldsymbol{r})$ satisfies

$$\left[\frac{1}{2m}\left(\boldsymbol{p}+\hbar\boldsymbol{k}\right)^2 + V(\boldsymbol{r})\right]u_{n,\boldsymbol{k}}(\boldsymbol{r}) = E_{n,\boldsymbol{k}}u_{n,\boldsymbol{k}}(\boldsymbol{r})$$

The equations are so far exact. It follows that we can use this result to formulate band structure as perturbation problem around $\boldsymbol{k}=0$ in the following way. Expanding the result of the problem

$$\left[\frac{\boldsymbol{p}^2}{2m} + V(\boldsymbol{r}) + \frac{\hbar}{m}\boldsymbol{k}.\boldsymbol{p} + \frac{\hbar^2k^2}{2m}\right]|u_{n,\boldsymbol{k}}(\boldsymbol{r})\rangle = E_{n,\boldsymbol{k}}|u_{n,\boldsymbol{k}}(\boldsymbol{r})\rangle$$

We treat the $\boldsymbol{k}$-independent part as H_0 and the $\boldsymbol{k}$-dependent part as the perturbation such that

$$H_0 = \frac{\boldsymbol{p}^2}{2m} + V(\boldsymbol{r}) \tag{3.56}$$

$$H_{\boldsymbol{k}} = \frac{\hbar}{m}\boldsymbol{k}.\boldsymbol{p} + \frac{\hbar^2k^2}{2m} \tag{3.57}$$

The eigenfunctions of H_0 are $|u_{n,\mathbf{0}}(\boldsymbol{r})\rangle$, which must form a complete set, using which we can express the $|u_{n,\boldsymbol{k}}(\boldsymbol{r})\rangle$, where n is the band index. This is a very useful fact which at once allows us to write till second order (notice the indices and their meaning carefully, for convenience we now write $E_{n,\boldsymbol{k}}$ as $E_n(\boldsymbol{k})$)

$$E_n(\boldsymbol{k}) = E_n(\mathbf{0}) + \langle u_{n,\mathbf{0}}|H_{\boldsymbol{k}}|u_{n,\mathbf{0}}\rangle + \sum_{n\neq m} \frac{|\langle u_{n,\mathbf{0}}|H_{\boldsymbol{k}}|u_{m,\mathbf{0}}\rangle|^2}{E_n(0) - E_m(0)} \tag{3.58}$$

Now

$$\begin{aligned} \langle u_{n,\mathbf{0}}|H_{\boldsymbol{k}}|u_{n,\mathbf{0}}\rangle &= \hbar\boldsymbol{k}.\langle u_{n,\mathbf{0}}|\frac{\boldsymbol{p}}{m}|u_{n,\mathbf{0}}\rangle + \frac{\hbar^2k^2}{2m} \\ &= \boldsymbol{k}.\,\nabla_{\boldsymbol{k}}E_n(\boldsymbol{k})|_{\boldsymbol{k}=0} + \frac{\hbar^2k^2}{2m} \\ &= \frac{\hbar^2k^2}{2m} \end{aligned} \tag{3.59}$$

At $\boldsymbol{k} = 0$ this is trivially true, but provided we have an extremum at $\boldsymbol{k} = \boldsymbol{k}_0$ then also this will work. Now for the second order part , the matrix elements are taken between states such that $n\neq m$, so a number like $\frac{\hbar^2k^2}{2m}$ cannot contribute anything. So we get:

$$\sum_{n\neq m} \frac{|\langle u_{n,\mathbf{0}}|H_{\boldsymbol{k}}|u_{m,\mathbf{0}}\rangle|^2}{E_n(0) - E_m(0)} = \frac{\hbar^2}{m^2}\sum_{n\neq m} \frac{\langle u_{n,\mathbf{0}}|\boldsymbol{k}.\boldsymbol{p}|u_{m,\mathbf{0}}\rangle\langle u_{m,\mathbf{0}}|\boldsymbol{k}.\boldsymbol{p}|u_{n,\mathbf{0}}\rangle}{E_n(0) - E_m(0)} \tag{3.60}$$

So we see how the presence of other bands changes the free electron dispersion explicitly by writing:

$$E_n(\boldsymbol{k}) = E_n(\mathbf{0}) + \frac{\hbar^2k^2}{2m} + \frac{\hbar^2}{m^2}\sum_{n\neq m} \frac{\langle u_{n,\mathbf{0}}|\boldsymbol{k}.\boldsymbol{p}|u_{m,\mathbf{0}}\rangle\langle u_{m,\mathbf{0}}|\boldsymbol{k}.\boldsymbol{p}|u_{n,\mathbf{0}}\rangle}{E_n(0) - E_m(0)} \tag{3.61}$$

We can now see how differentiating eqn. 3.61 twice (w.r.t k_i and k_j) that the effective mass matrix is

$$\mathrm{M}^{-1}_{ij} = \frac{1}{\hbar^2}\frac{\partial^2E_n}{\partial k_i\partial k_j} = \frac{1}{m}\delta_{ij} + \frac{2}{m^2}\sum_{n\neq m} \frac{\langle u_{n,\mathbf{0}}|p_i|u_{m,\mathbf{0}}\rangle\langle u_{m,\mathbf{0}}|p_j|u_{n,\mathbf{0}}\rangle}{E_n(0) - E_m(0)} \tag{3.62}$$

If the bandgaps are large then the effective mass differs very little from the free electron mass. At least qualitatively, you can show why narrow bandgap semiconductors have small effective masses, using eqn. 3.62

A similar expansion is also useful around any extremum (where the group velocity vanishes). For an expansion around $\boldsymbol{k}_0$, we write $\boldsymbol{k} = \boldsymbol{k}_0 + \delta\boldsymbol{k}$ and

$$\begin{aligned} H &= \frac{1}{2m}\left(\boldsymbol{p} + \hbar\boldsymbol{k}\right)^2 + V \\ &= \underbrace{\left[\frac{1}{2m}\left(\boldsymbol{p} + \hbar\boldsymbol{k}_0\right)^2 + V\right]}_{H_{\boldsymbol{k}_0}} + \underbrace{\frac{\hbar\delta\boldsymbol{k}}{m}.\left(\boldsymbol{p} + \hbar\boldsymbol{k}_0\right) + \frac{\hbar^2}{2m}\delta\boldsymbol{k}.\delta\boldsymbol{k}}_{H_{\delta\boldsymbol{k}}} \end{aligned} \tag{3.63}$$

We would then have

$$E(\boldsymbol{k}) = E(\boldsymbol{k}_0) + \langle u_{n,\boldsymbol{k}_0}|H_{\delta\boldsymbol{k}}|u_{n,\boldsymbol{k}_0}\rangle + \sum_{n\neq m} \frac{|\langle u_{n,\boldsymbol{k}_0}|H_{\delta\boldsymbol{k}}|u_{m,\boldsymbol{k}_0}\rangle|^2}{E_n(0) - E_m(0)} \tag{3.64}$$

The first order correction is

$$\begin{aligned}\langle u_{n,\boldsymbol{k}_0}|H_{\boldsymbol{\delta}k}|u_{n,\boldsymbol{k}_0}\rangle &= \frac{\hbar\delta\boldsymbol{k}}{m}.\langle u_{n,\boldsymbol{k}_0}|\boldsymbol{p}+\hbar\boldsymbol{k}_0|u_{n,\boldsymbol{k}_0}\rangle + \frac{\hbar^2}{2m}\delta\boldsymbol{k}.\delta\boldsymbol{k} \\ &= \hbar(\boldsymbol{k}-\boldsymbol{k}_0).\langle\Psi_{\boldsymbol{k}_0}|\frac{\boldsymbol{p}}{m}|\Psi_{\boldsymbol{k}_0}\rangle + \frac{\hbar^2}{2m}(\boldsymbol{k}-\boldsymbol{k}_0)^2 \\ &= (\boldsymbol{k}-\boldsymbol{k}_0).\nabla_{\boldsymbol{k}}E(\boldsymbol{k})|_{\boldsymbol{k}_0} + \frac{\hbar^2}{2m}(\boldsymbol{k}-\boldsymbol{k}_0)^2 \end{aligned} \tag{3.65}$$

If the gradient vanishes (extremum) then there is no first order correction. The form of the second order term too remains similar.

The expression given here is quantitatively correct if the maxima of the valence band and the minima of the conduction band occur at the same $\boldsymbol{k}$, it is not necessary for the extrema to be at $\boldsymbol{k}=0$. But in some of the most common semiconductors we encounter, like Si,Ge, this is not correct. They are indirect gap material.

For silicon the conduction band minima lie on the six equivalent Δ -lines along $\langle 100\rangle$ -directions and occur at about 0.85% of the way to the zone boundary. These are the well-known, equivalent ellipsoidal constant energy valleys. The highest point of the valence band is still at $\boldsymbol{k}=0$ or the zone center.

For Ge, the CB has eight minima at the zone boundary itself in the $\langle 111\rangle$ direction. The valence band maximum is at $\boldsymbol{k}=0$

If we plot the electron effective mass vs band gap of some semiconductors, we will see that Si and Ge are outliers. See fig 3.7

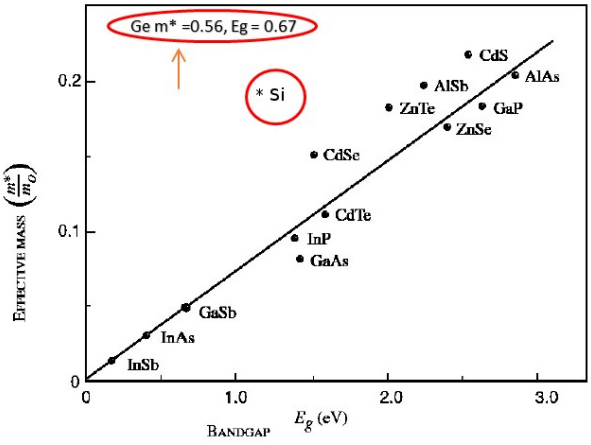


Figure 3.7: The figure shows effective mass of various semiconductors plotted against their bandgap. The indirect gap Si and Ge are outliers. Infact the effective mass of electrons in conduction band of Ge is 0.56, and E_g=0.66 eV, which is out of scale of the plot. The figure is taken from *Semicondcutor Optoelectronics* by J. Singh and slightly modified.

For these materials, we need to extend the calculation till large k values, covering the full zone. So simple results upto k^2 order will not give the correct picture. In fact the band gap of Ge is 0.67 eV (at 300K) and for Si the gap is 1.1 eV. For these materials a simple comparison of band gaps and effective masses would not work. There are techniques of extending these kind of calculations over the full zone, but they are

necessarily quite heavily numerical and do not give simple algebraic (analytic) results. See for example: Band structures of Ge and InAs: A 20 band ***k.p*** model S. Ben Radhia et al. *Journal of Applied Physics* **2**, 4422 (2002) for an example. The figure 3.8 shows an example where you can see the location of the various minima.

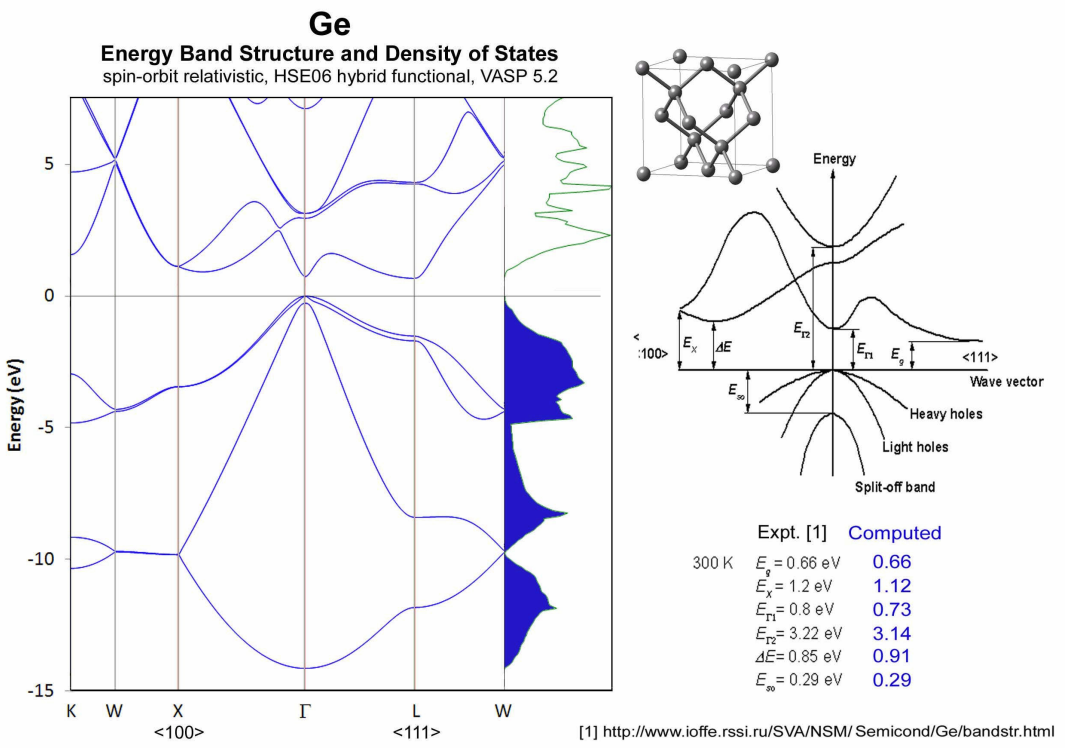


Figure 3.8: An example of a calculated band structure. The methods used are much more involved than what we have discussed so far. However notice the similarity of the lower (deeper) bands with what we got from a simple quantum well array. The higher ones are of interest as far as the semiconducting behaviour is concerned. The zero level demarcates the filled from the unfilled states in the undoped material. The figure is from *www.materialsdesign.com*

The location of the electron pockets in the first Brillouin zone (six for Si, eight for Ge) are shown in figure 3.9 The inset shows the names given to various corner points and mid points of the faces. There are eight hexagonal faces and six square faces in the structure, usually called a " truncated octahedron".

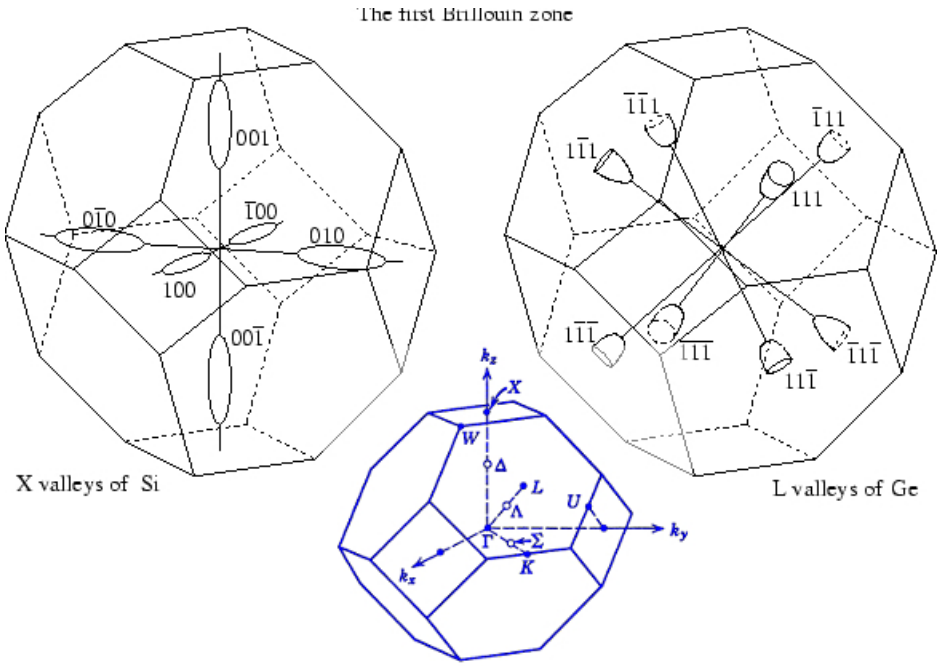


Figure 3.9: The location of the electron pockets in Si and Ge. The figure is from *Physical Modeling of Electron Transport in Strained Silicon and Silicon-Germanium*, Sergey Smirnov, *PhD Thesis, University of Vienna,* 2003.

Chapter 4

Bandstructure I : Tight binding or Linear Combination of Atomic Orbitals (LCAO)

In the last chapter we considered the problem of one electron in the periodic potential of the lattice and solved for its energy eigenvalues. It is possible (and useful) to look at the problem from another point of view. We consider that we are building up a solid atom by atom, like building up a molecule.

4.1 Diatomic molecule and Linear chain of atoms

4.1.1 Diatomic molecule

As an initial problem let's consider building up a molecule form two atoms that are not necessarily identical. When they are far apart then the wavefunctions must be same as the wavefunctions of the isolated atoms - we call the atoms a and b. So the Hamiltonian of the system must be

$$H_{ab} = T + V_a + V_b \tag{4.1}$$

Our basis set is going to be the states $|a\rangle$ $|b\rangle$, centered on atom a and atom b respectively, when they are very far apart. So that the basis set satisfies

$$\begin{aligned} (T + V_a)|a\rangle &= E_0^a|a\rangle & (4.2)\\ (T + V_b)|b\rangle &= E_0^b|b\rangle & (4.3) \end{aligned}$$

The Linear Combination of Atomic Orbitals method means looking for solutions of eqn 4.1 of the form

$$|\psi\rangle = \alpha|a\rangle + \beta|b\rangle \tag{4.4}$$

If eqn. 4.4 is a solution then we must have (E is the unknown eigenvalue we want to solve for)

$$\begin{aligned} \langle a|H_{ab}|\psi\rangle &= E\langle a|\psi\rangle \\ \langle b|H_{ab}|\psi\rangle &= E\langle b|\psi\rangle & (4.5) \end{aligned}$$

As it stands the set of eqn. 4.5 is exact, but to proceed we need to understand the physical significance of each term and approximate them reasonably.

$$\langle a|b\rangle \approx 0 \tag{4.6}$$

This means that there is negligible overlap between the atomic orbitals.

$$\langle a|T + V_a + V_b|b\rangle \equiv t \tag{4.7}$$

Under the action of the Hamiltonian the state $|a\rangle$ and $|b\rangle$ can mix a little bit. We will come across this type of a term many times in future. A term of this type is called a hopping term. It is important to understand why we claim that that the hopping term (eqn. 4.7) can be larger than the direct overlap term 4.6. See the fig. 4.1 and study it carefully. You should be able to reason out why we could ignore the expression in eq. 4.6 but retain the hopping term.

$$\langle a|H_{ab}|a\rangle = \langle a|T + V_a|a\rangle + \langle a|V_b|a\rangle = E_0^a + \langle a|V_b|a\rangle \equiv \tilde{E_0^a} \tag{4.8}$$

You should now be able to appreciate the physical significance of the approximation (known as the tight binding approximation) show that eqn. 4.5 leads (assuming both the atoms are identical, it does not mean $V_a = V_b$, because they are still centered at different points, though the functional forms will be similar.

$$\begin{pmatrix} \tilde{E_0} - E & t \\ t^* & \tilde{E_0} - E \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0 \tag{4.9}$$

The solution is obtained by setting the determinant to zero.

$$E = \tilde{E_0} \pm |t| \tag{4.10}$$

$$|\psi\rangle = \frac{1}{\sqrt{2}}\left(|a\rangle \mp |b\rangle\right) \tag{4.11}$$

PROBLEM : Complete the algebra leading to eqn. 4.10 and eqn. 4.11. The lower energy state is called the bonding state (in chemistry) and the higher energy state is called the antibonding state. Which state has the higher electron density at the mid-point between the two atoms?

4.1.2 Linear chain of atoms with nearest neighbour interaction

We now extend the ideas of tight binding with one hopping term to a linear chain of atoms,each spaced by a units. We will always consider a chain that has its ends joined together. Periodic boundary conditions are then obviously easy to apply. This means that the $N + 1^{th}$ atom is same as the 1^{st} atom.

The hamiltonian is then

$$H = T + V_1 + V_2 + \ldots + V_N \tag{4.12}$$

Remember that although there are N sites/atoms/potential wells, there is only one particle co-ordinate. We are solving for single particle eigenstates in the potential created by all the atoms.

Single orbital on a site

Since this is a periodic potential, we use Bloch's theorem in combination with the tight binding idea of using wavefunctions ($|\phi_n\rangle$) localized at the n^{th} atomic site as our starting point. The wavefunction is then

$$|\psi_k\rangle = \frac{1}{\sqrt{N}} \sum_n e^{ikna} |\phi_n\rangle \tag{4.13}$$

$$\therefore \langle r|\psi_k\rangle = \frac{1}{\sqrt{N}} \sum_n e^{ikna} \langle r|\phi_n\rangle \tag{4.14}$$

$$\psi_k(r) = \frac{1}{\sqrt{N}} \sum_n e^{ikna} \phi(r - na) \tag{4.15}$$

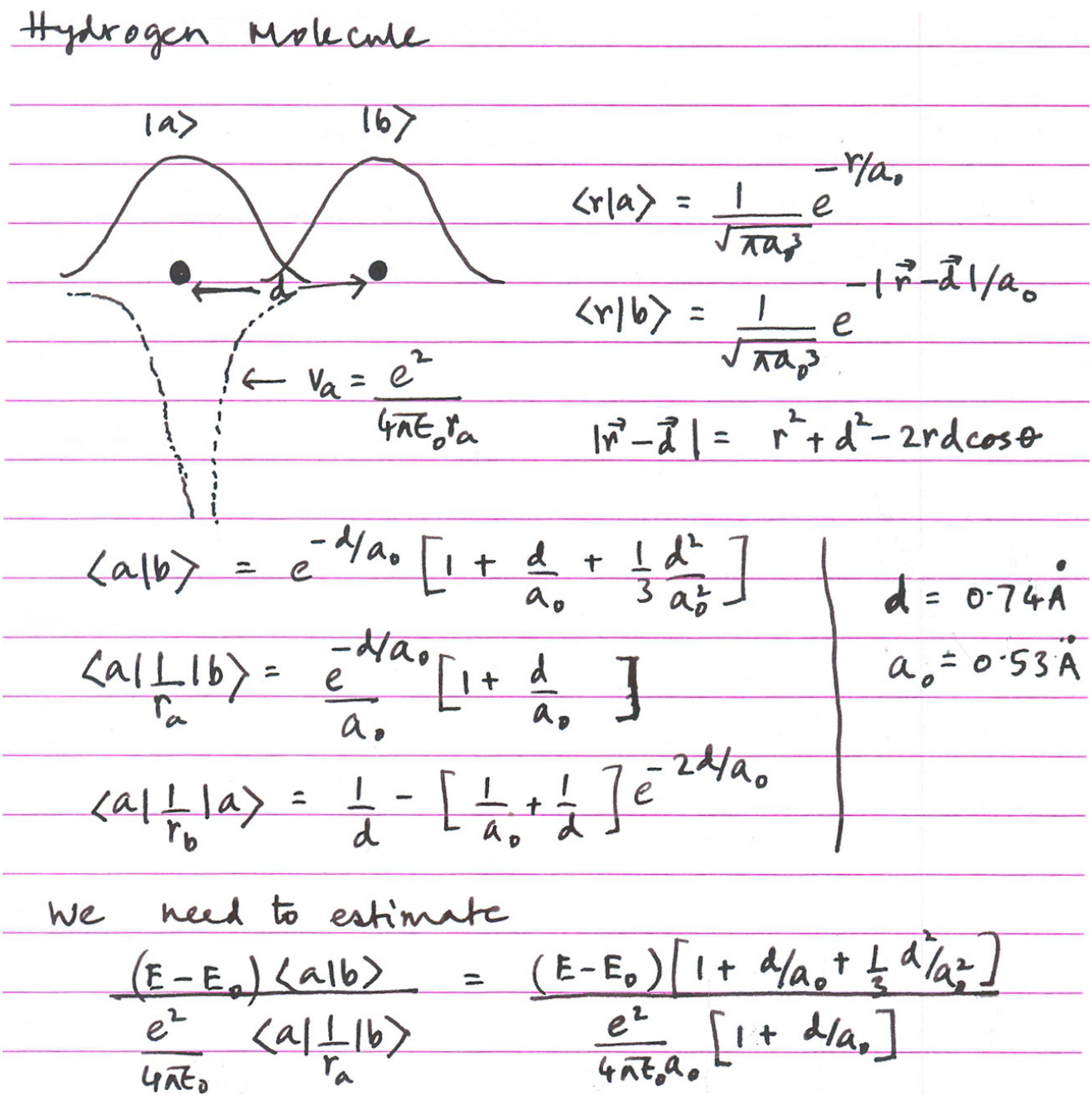


Figure 4.1: The overlap of the two 1s orbitals in Hydrogen molecule. Some of the intermediate steps are left for you to fill in - you should be able to estimate the term we retained and the term we dropped. Though the wavefunctions are specific to the H_2 molecule, the general conclusion would be true for any two tightly bound states, separated by a not too large an amount. Of course if we keep increasing the separation, then the hopping term would also go to zero.

where $|\phi_n\rangle$ is the wavefunction localized on the n^{th} atom, satisfying

$$(T + V_n)\,|\phi_n\rangle = E_0|\phi_n\rangle \tag{4.16}$$

$$\begin{aligned}\langle\phi_n|H|\phi_n\rangle &= E^0 + \langle\phi_n|V_{n-1}|\phi_n\rangle + \langle\phi_n|V_{n+1}|\phi_n\rangle \\ &= \tilde{E}_0\end{aligned} \tag{4.17}$$

PROBLEM : Show that the wavefunction 4.13

1. satisfies the Bloch criteria $\psi(r) = e^{ikr}u_k(r)$, where $u_k(r + na) = u_k(r)$
2. is correctly normalized provided a certain assumption is made. What is the assumption?

The periodic boundary condition requires that the values of k be quantized. However you can see that if N is large then the quantization gets more and more finely spaced and k becomes continuous in the large N limit.

$$e^{ik(N+1)a} = e^{ika} \tag{4.18}$$

$$\therefore kNa = 2m\pi \tag{4.19}$$

$$k = \frac{2\pi}{a}\frac{m}{N} \tag{4.20}$$

The problem is now surprisingly straightforward, because there are no unknown parameters in eqn. 4.13, all we need to do is take the expectation value

$$\begin{aligned}E(k) &= \langle\psi_k|H|\psi_k\rangle \\ &= \langle\psi_k|T + V_1 + V_2 + V_3 + \ldots + V_N|\psi_k\rangle && (4.21)\\ &= \frac{1}{N}\sum_{n,m} e^{-ikna}e^{ikma}\langle\phi_n|H|\phi_m\rangle && (4.22)\\ &= \frac{1}{N}\sum_{n=m}\langle\phi_n|H|\phi_n\rangle + \frac{1}{N}\sum_{n=m\pm1} e^{ik(m-n)a}\langle\phi_n|H|\phi_m\rangle + \cancel{\frac{1}{N}\sum_{|n-m|>1}\ldots} && (4.23)\\ &\approx \tilde{E}_0 + \left(te^{ika} + te^{-ika}\right) && (4.24)\\ \therefore E(k) &= \tilde{E}_0 + 2t\cos ka && (4.25)\end{aligned}$$

We have solved the band structure, t has the same significance (nearest neighbour hopping) that we discussed in eqn. 4.7.
Two points to note

1. The bandwidth is proportional to the hopping term.
2. It is surprisingly easy to generalize the result to 2 and 3 dimensions, because *all* we need to do is sum over the nearest neighbours! If they are symmetrically located then the bandwidth would simply be $2zt$ where z is the number of nearest neighbours or the co-ordination number of the lattice.

PROBLEM : Calculate the group velocity of a particle at the bottom of the band and at the corner ($k = \pm\pi/a$). Show that there is a point of inflection (where the second derivative changes sign) somewhere between $k = 0$ and $k = \pm\pi/a$.

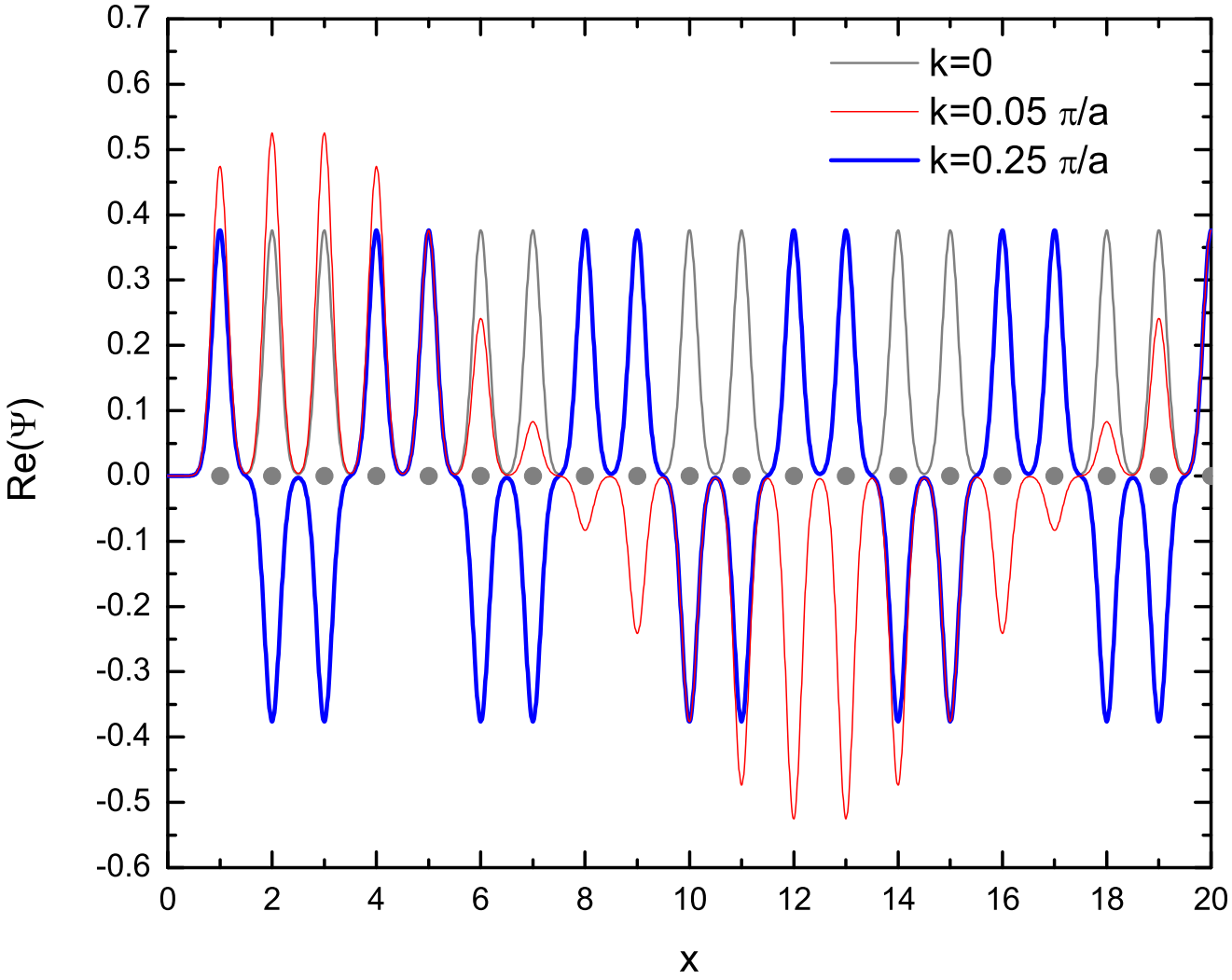


Figure 4.2: The dots are the atomic sites, the $k = 0$ wavefunction shows what the atomic states are like. The other two show what the linear combination of those wavefunctions, as given by Bloch's theorem , would look like. See eqn. 4.13.

How does the Bloch function look?

Here's a plot of how the functions look. The $k = 0$ wavefunction is shown for reference, because that has the maximum resemblance with the "atomic" wavefunctions. Here we assumed that the atomic wavefunction is a gaussian. See Fig. 4.2.

Generalising to 2 and 3 dimensions: with 1 oribital per site

The generalisation is easy. To handle 2 and 3d lattices we need to write the wavefunction as

$$|\psi_{\boldsymbol{k}}\rangle = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{R}} e^{i\boldsymbol{k}.\boldsymbol{R}} |\phi_{\boldsymbol{R}}\rangle \tag{4.26}$$

$$\langle \boldsymbol{r}|\psi_{\boldsymbol{k}}\rangle = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{R}} e^{i\boldsymbol{k}.\boldsymbol{R}} \phi(\boldsymbol{r}-\boldsymbol{R}) \tag{4.27}$$

Where the sum runs over all direct lattice vectors $\boldsymbol{R}$ and $|\phi_{\boldsymbol{R}}\rangle$ is the atomic state centered at $\boldsymbol{R}$. While taking the expectation value of energy we will group the series of terms into three and ignore the interaction between sites which are not nearest neighbours or next-nearest-neighbours:

$$H = T + V_1 + V_2 + V_3 + \ldots + V_N \tag{4.28}$$

$$E(\boldsymbol{k}) = \langle \psi_{\boldsymbol{k}} | H | \psi_{\boldsymbol{k}} \rangle \tag{4.29}$$

$$= \frac{1}{N} \sum_{\boldsymbol{R}=\boldsymbol{R'}} \langle \phi_{\boldsymbol{R'}} | H | \phi_{\boldsymbol{R}} \rangle + \frac{1}{N} \sum_{\substack{\boldsymbol{R},\boldsymbol{R'} \\ \text{nearest neighbours}}} e^{i\boldsymbol{k}.(\boldsymbol{R}-\boldsymbol{R'})} \langle \phi_{\boldsymbol{R'}} | H | \phi_{\boldsymbol{R}} \rangle + \frac{1}{N} \sum_{\substack{\boldsymbol{R},\boldsymbol{R'} \\ \text{further than nearest neighbours}}} \ldots \tag{4.30}$$

$$\approx E_0 + \sum_{\substack{\text{nearest} \\ \text{neigh-} \\ \text{bours}}} e^{i\boldsymbol{k}.\boldsymbol{R}} t_{\boldsymbol{R}} \tag{4.31}$$

Since all sites are identical, it is sufficient to sum over the nearest neighbours of the site at $\boldsymbol{R} = 0$.

PROBLEM : Consider a 2-d rectangular lattice with sides a and b.

1. Show that following eqn. 4.31 the bandstructure would be of the form

$$E(k_x, k_y) = E_0 - 2t_1 \cos(ak_x) - 2t_2 \cos(bk_y) \tag{4.32}$$

2. What is the reciprocal lattice? Draw the first Brillouin zone.

3. Plot the constant energy contours, assuming $t_1 > t_2 > 0$ and $a < b$. Why is this physically reasonable?

4. Plot some constant energy contours. How do the contours look for small k? How do the shapes change at slightly larger k? Do all constant energy contours close within the first Brillouin zone?

Similarly for 3d lattices like BCC with 8 nearest neighbours and FCC with 12 neighbours can be summed up.

PROBLEM : Tight-binding bandstructure with a single orbital per site gives on BCC and FCC

1. For Body Centered Cubic lattice write down the co-ordinates of the nearest neighbours of $(0, 0, 0)$

2. Then show, with 8 nearest neighbour hopping terms and a as the side of the cube

$$E(k_x, k_y, k_z) = E_0 + 8t \cos\frac{k_x a}{2} \cos\frac{k_y a}{2} \cos\frac{k_x a}{2} \tag{4.33}$$

3. For Face Centered Cubic lattice write down the co-ordinates of the nearest neighbours of $(0, 0, 0)$

4. Then show, with 12 nearest neighbour hopping terms and a as the side of the cube:

$$E(k_x, k_y, k_z) = E_0 + 4t \left[\cos\frac{k_x a}{2} \cos\frac{k_y a}{2} + \cos\frac{k_y a}{2} \cos\frac{k_z a}{2} + \cos\frac{k_z a}{2} \cos\frac{k_x a}{2} \right] \tag{4.34}$$

Counting the number of states in $\boldsymbol{k}$ space

Geometrically, periodic boundary condition in 1d means, putting all the lattice points on a ring. In 2d it means putting them on a "torus", in 3d it would be some hypersurface that we can only define mathematically.

Now think of a lattice with lattice vectors $\boldsymbol{a_1}$,$\boldsymbol{a_2}$, $\boldsymbol{a_3}$ and corresponding reciprocal lattice vectors $\boldsymbol{b_1}$, $\boldsymbol{b_2}$,$\boldsymbol{b_3}$. Let's take the $N = N_1N_2N_3$ as the number of unit cells in the crystal. Note that it doesn't imply that we are taking a cubic/rectangular volume, only. In general (algebraically, in any dimension) periodic (or Born von Karman) boundary conditions means that we require

$$\begin{aligned}
\psi(\boldsymbol{r} + N_i\boldsymbol{a_i}) &= \psi(\boldsymbol{r}) && (4.35)\\
\therefore e^{iN_i\boldsymbol{k}.\boldsymbol{a_i}} &= 1 \qquad \text{(Bloch's theorem)} && (4.36)\\
\therefore \boldsymbol{k} &= \sum_i \frac{m_i}{N_i}\boldsymbol{b_i} \qquad \text{for integer } m && (4.37)\\
& && (4.38)
\end{aligned}$$

The volume of allowed $\boldsymbol{k}$-space per point is then:

$$\begin{aligned}
\Delta\boldsymbol{k} &= \frac{\boldsymbol{b_1}}{N_1}.\frac{\boldsymbol{b_2}}{N_2} \times \frac{\boldsymbol{b_3}}{N_1} && (4.39)\\
&= \frac{1}{N}\frac{(2\pi)^3}{v_{unit\ cell}} && (4.40)\\
&= \frac{(2\pi)^3}{V_{crystal}} && (4.41)
\end{aligned}$$

This means:

1. Whenever we need to sum over all states we can thus interchange discrete summation and continuous integration by the following rule

$$\sum_{\boldsymbol{k}} (\ldots) \to \int \frac{V}{(2\pi)^3} \mathrm{d}^3\boldsymbol{k}\,(\ldots) \tag{4.42}$$

2. Since the $\boldsymbol{k}$-space density is uniform, we can start from here and use the $E(\boldsymbol{k})$ relation to convert this into density of states in energy $D(E)$.

3. We need to multiply this by 2 for spin 1/2 particles like electrons. In general by $(2s + 1)$ if the particle has spin s, because each k-state can accommodate one particle with spin $-1/2$, one with spin $1/2$ etc.

The density of states in energy

We want to write an expression for the number of states between between E to $E + \delta E$. We will call this function $D(E)$.

In 1d it is trivial. We choose an interval k to $k + \delta k$ and the corresponding interval in energy $E(k)$ to $E(k + \delta k) = E + \delta E$. The number of states in these two intervals must be equal, so:

$$\begin{aligned}
D(E)\delta E &= D(k)\delta k && (4.43)\\
&= \frac{L}{2\pi}\frac{\delta k}{\delta E} && (4.44)\\
&= \frac{L}{2\pi}\frac{1}{\mathrm{d}E/\mathrm{d}k} && (4.45)
\end{aligned}$$

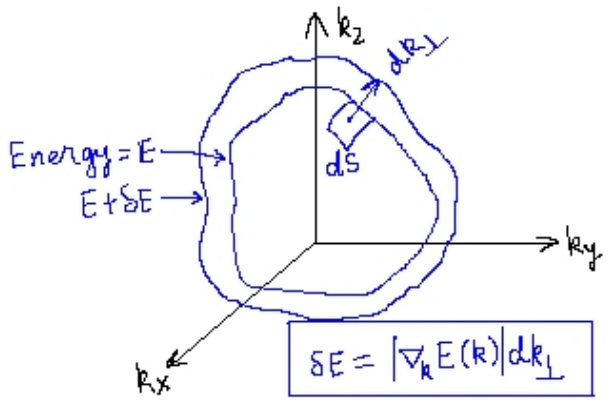


Figure 4.3: Calculation of the k-space area/volume between two equal energy contours differing slightly in energy.

Because there is only one component of k the derivative is simple.

Now, in 2d and 3d we proceed as follows, follow the logic carefully, this would reappear many times in different places.

1. We need to count all the states that lie between the two constant energy contours E and $E + \mathrm{d}E$. These can have quite complex shapes depending on the $E(k_x, k_y, k_z)$ relation.

2. To do this we calculate the k-space volume enclosed by the two contours and multiply the number with $V/(2\pi)^3$. Our volume element here is $\mathrm{d}S\mathrm{d}k_\perp$.

3. The normal to an "equipotential" is given by the gradient. Hence the normal to $E(k_x, k_y, k_z) =$ constant will be given by $\nabla_{\boldsymbol{k}} E(k_x, k_y, k_z)$ So:

$$\begin{aligned} \delta E &= |\nabla_{\boldsymbol{k}} E(k_x, k_y, k_z)| \mathrm{d}k_\perp && (4.46) \\ \therefore D(E)\delta E &= \frac{V}{(2\pi)^3} \int_S \mathrm{d}S \frac{\delta E}{|\nabla_{\boldsymbol{k}} E(k_x, k_y, k_z)|} && (4.47) \\ \therefore D(E) &= (2s+1) \frac{V}{(2\pi)^3} \int_S \frac{\mathrm{d}S}{|\nabla_{\boldsymbol{k}} E(k_x, k_y, k_z)|} && (4.48) \end{aligned}$$

   In the last step we have included the spin-degeneracy. V is the sample volume. If we want density of states per unit volume, obviously this will be dropped.

4. Points where the group velocity vanishes can give rise to singularities in $D(E)$, but these will be integrable. If the gradient vanishes then we should be able to expand E around this point as

$$E(k_x, k_y, k_z) = E_0 + \frac{\hbar^2}{2m_x} k_x^2 + \frac{\hbar^2}{2m_y} k_y^2 + \frac{\hbar^2}{2m_z} k_z^2 \qquad (4.49)$$

   If all the coefficients are positive, then this is a band minima, if all are negative it is a band maxima and if they are mixed it is a saddle point. Points where the density of states or its derivative is singular are called *van Hove singularities*.

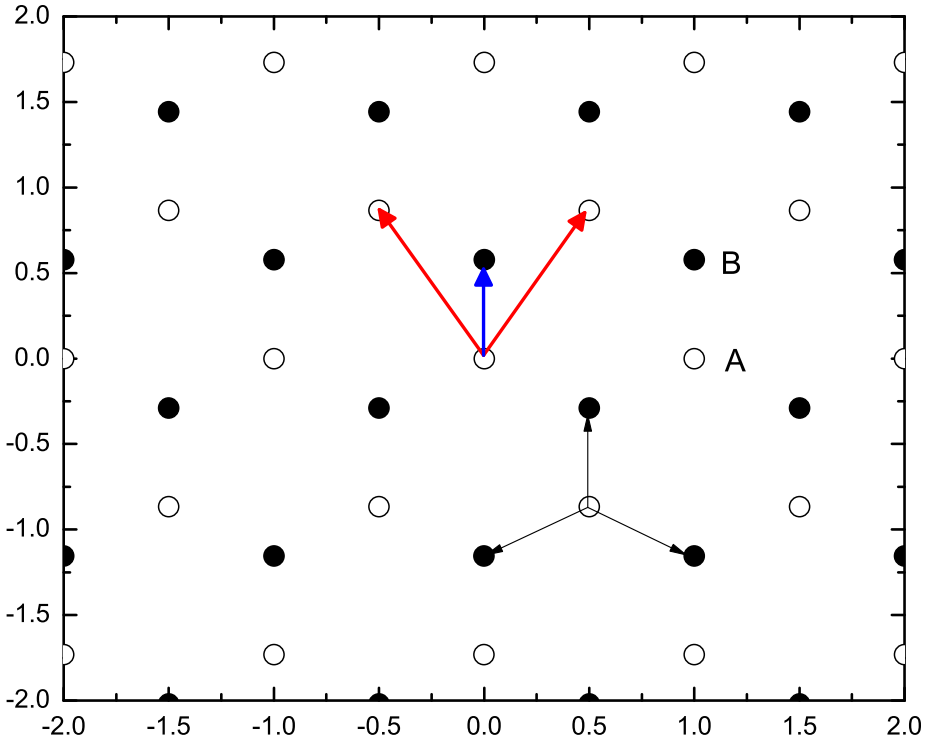


Figure 4.4: Lattice vectors of Graphene. All the atoms are Carbon, but the two types of sites mean that the Bloch functions cannot be written down straightway.

4.1.3 More than 1 orbital per site

At the beginning of the chapter we calculated the wavefunction/energy levels of a diatomic molecule. We had to solve for the eigenvalues and then get the coefficients of the atomic states which made up the molecular wavefunctions. But no such procedure was needed when we solved the linear chain. Why? The reason is that the symmetry (Bloch's theorem) told us what the coefficients would be. We now ask, what if there are two atoms a and b per lattice site (the basis can of course have more) or two orbitals on the same atom (like a $2s$ and $2p$ orbital or some s and d orbitals..). In these cases we still begin with the atomic wavefunctions, but Bloch's theorem cannot tell us how much of the wavefunction of site a and site b to take. We must solve for those.

Consider the example of a single sheet from graphite (graphene). The triangular lattice has a two atom basis. We take the following as lattice and vectors:

$$\begin{aligned} \boldsymbol{a_1} &= \frac{a}{2}(1, \sqrt{3}) \\ \boldsymbol{a_2} &= \frac{a}{2}(-1, \sqrt{3}) \end{aligned} \tag{4.50}$$

The two point basis is composed of :

$$\begin{aligned} \text{Type A atoms} &: (0,0) \\ \text{Type B atoms} &: \boldsymbol{d} = a(0, \frac{1}{\sqrt{3}}) \end{aligned} \tag{4.51}$$

Notice that the nearest neighbours of A are B type atoms. Thus the largest hopping terms would occur between A-B overlaps.

Which orbitals to consider

First, it is known that Graphite is a "flaky" material, meaning that it tends to peel off in layers, this indicates that it consists of sheets held together somewhat loosely. In fact the inplane nearest neighbour C-C distance is 1.42Å, whereas the interplaner distance is about 3.4Å. Also we know that the strongest σ bonds are all in plane. So we would start by assuming that we can treat each "sheet" as a two dimensional entity.

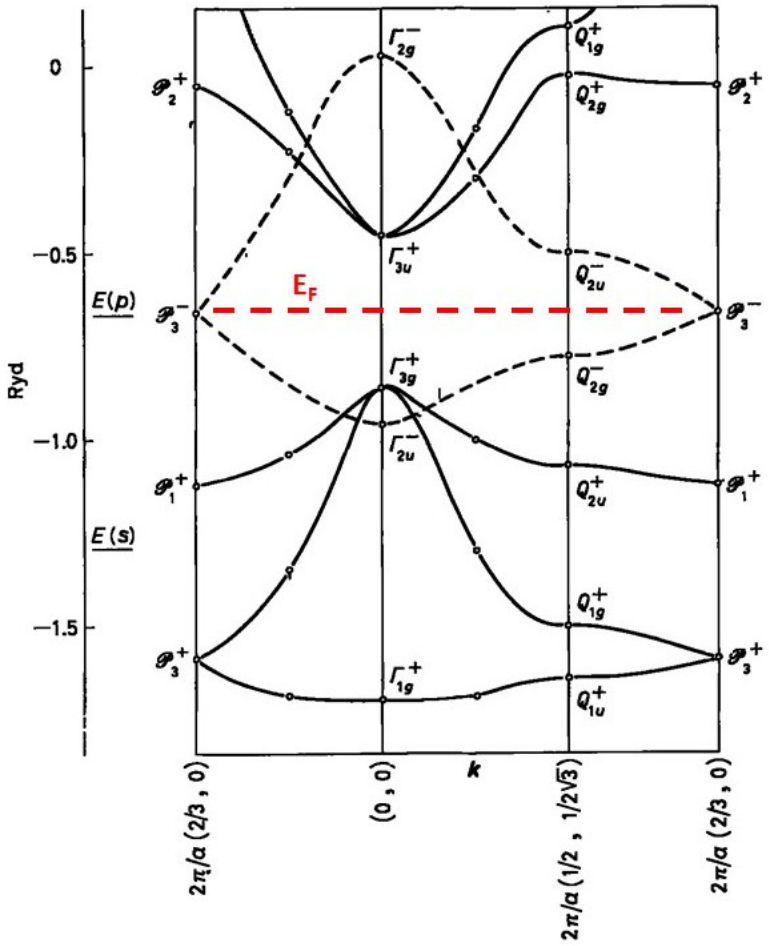


Figure 4.5: Notice the overall band structure of graphene. The figure is taken from "Band structure and optical properties of graphite and of the layer compounds GaS and GaSe", F. Bassani, G. Pastori Parravicini, *Nuovo Cimento* **B50**, pp 95128 (July 1967).

Carbon atom has electronic configuration $1s^2 2s^2 2p^2$. The binding energies of the "core" $1s$ state is -21.4 Ry, the $2s$ state is at -1.29 Ry and the $2p$ state is at -0.66 Ry. (Rydberg = 13.6 eV). This means that the $1s$ level will in practice have no dispersion at all. The $2s$, $2p_x$, $2p_y$ and $2p_z$ states can now be treated together.

It would seem then we need to consider an 8×8 problem, because there are four atomic orbitals and two sets of Bloch functions originating from the two atoms per unit cell. However this 8×8 problem (or the hamiltoninan matrix) neatly factors out into one 6×6 and one 2×2 block. The 6-block consists of the the states from s, p_x and p_y states and the bonding orbitals it generates lies lower in energy than the p_z states. They do not "mix" due to different symmetries of the wavefunctions with respect to reflection on the plane of the sheet *i.e.* think of the plane of the sheet as a mirror.

If the sample has N unit cells it has $2N$ atoms. The number of available states in each band is N *for each spin orientation*, hence $2N$ in total. There are eight electrons (we are neglecting two $1s$ electrons here, they are much deeper in energy and effectively dispersionless) from two atoms per cell, We have to accommodate $8N$ electrons. So we expect the lowest four bands to be fully occupied. These turn out be not too intertwinned (see fig 4.5 making the separation of filled and unfilled states easy to visualize.

It turns out that the Fermi energy will lie just above the filled band from the p_z state. We can thus hope to look at the 2×2 block arising out of the p_z state and hope to understand a bit of what happens near the Fermi level.

The 2×2 problem

If $|\phi_A\rangle$ and $|\phi_B\rangle$ are states centered at A and B. So we form a set of two Bloch functions and make a linear combination with unknown (to be solved for) coefficients α and β

$$\begin{aligned} |\psi_{\boldsymbol{k},A}\rangle &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}} e^{i\boldsymbol{k}.\boldsymbol{R}}|\phi_A\rangle && (4.52)\\ |\psi_{\boldsymbol{k},B}\rangle &= \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}} e^{i\boldsymbol{k}.\boldsymbol{R}}|\phi_B\rangle && (4.53)\\ |\psi_{\boldsymbol{k}}\rangle &= \alpha|\psi_{\boldsymbol{k},A}\rangle + \beta|\psi_{\boldsymbol{k},B}\rangle && (4.54)\end{aligned}$$

Now we have:

$$\begin{aligned} H|\psi_{\boldsymbol{k}}\rangle &= E|\psi_{\boldsymbol{k}}\rangle \\ \langle\psi_{\boldsymbol{k}A}|H|\psi_{\boldsymbol{k}}\rangle &= E\langle\psi_{\boldsymbol{k}A}|\psi_{\boldsymbol{k}}\rangle \\ \langle\psi_{\boldsymbol{k}B}|H|\psi_{\boldsymbol{k}}\rangle &= E\langle\psi_{\boldsymbol{k}B}|\psi_{\boldsymbol{k}}\rangle && (4.55)\end{aligned}$$

The idea can be extended to more complex basis sets. For this we need to the following (the calculation/justification is left as an exercise).

$$\begin{aligned} \langle\psi_{\boldsymbol{k}A}|\psi_{\boldsymbol{k}A}\rangle &= 1 \\ \langle\psi_{\boldsymbol{k}A}|\psi_{\boldsymbol{k}B}\rangle &\approx 0 && (4.56)\\ \langle\psi_{\boldsymbol{k}A}|H|\psi_{\boldsymbol{k}A}\rangle &= \tilde{E}_0 \\ \langle\psi_{\boldsymbol{k}A}|H|\psi_{\boldsymbol{k}B}\rangle &= \sum_{\substack{\text{nearest}\\ \text{neighbours}}} e^{i\boldsymbol{k}.\boldsymbol{R}}\langle\phi_A|H|\phi_B\rangle && (4.57)\\ &= \left(e^{i\boldsymbol{k}.\boldsymbol{0}} + e^{-i\boldsymbol{k}.\boldsymbol{a_1}} + e^{-i\boldsymbol{k}.\boldsymbol{a_2}}\right)\langle\phi(\boldsymbol{r})|H|\phi(\boldsymbol{r}-\boldsymbol{d})\rangle && (4.58)\\ &= \left(1 + 2\cos\left(\frac{k_x a}{2}\right)e^{-i\frac{\sqrt{3}}{2}k_y a}\right)t && (4.59)\\ &= F(k_x,k_y)t && (4.60)\end{aligned}$$

where t is again the nearest neighbour hopping amplitude. The hopping term occurs between the an atom and its 3 neighbours, shown by black arrows in Fig. 4.4. If we considered next nearest neighbours as well, these would have come from the terms in $\langle\psi_{\boldsymbol{k}A}|H|\psi_{\boldsymbol{k}A}\rangle$. (Justify this as an exercise).

With these in place the set of eqns. 4.55 gives the eigenvalue equation:

$$\begin{pmatrix} \tilde{E}_0 & tF(k_x,k_y) \\ tF^*(k_x,k_y) & \tilde{E}_0 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = E\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad (4.61)$$

PROBLEM :
Show that:

1. the eigenvalues of matrix 4.61 are given by

$$E(k_x, k_y) = \tilde{E}_0 \pm t|F(k_x, k_y)| \tag{4.62}$$

$$\text{where} \quad |F|^2 = 1 + 4\cos^2\frac{k_x a}{2} + 4\cos\frac{k_x a}{2}\cos\frac{\sqrt{3}}{2}k_y a \tag{4.63}$$

2. The reciprocal lattice vectors of the graphene lattice are given by:

$$\begin{aligned} \boldsymbol{b_1} &= \frac{2\pi}{a}\left(1, \frac{1}{\sqrt{3}}\right) \\ \boldsymbol{b_2} &= \frac{2\pi}{a}\left(-1, \frac{1}{\sqrt{3}}\right) \end{aligned} \tag{4.64}$$

3. Calculate the co-ordinates of the six points where the two bands touch.

Two branches have now appeared, this is a common feature in problems where the lattice has a two basis points, pretty much similar situations occur for electron energy bands as well as phonon bands (we will see later).

Now we identify the first Brillouin zone of the triangular lattice and plot the energy eignenvalues in that zone.

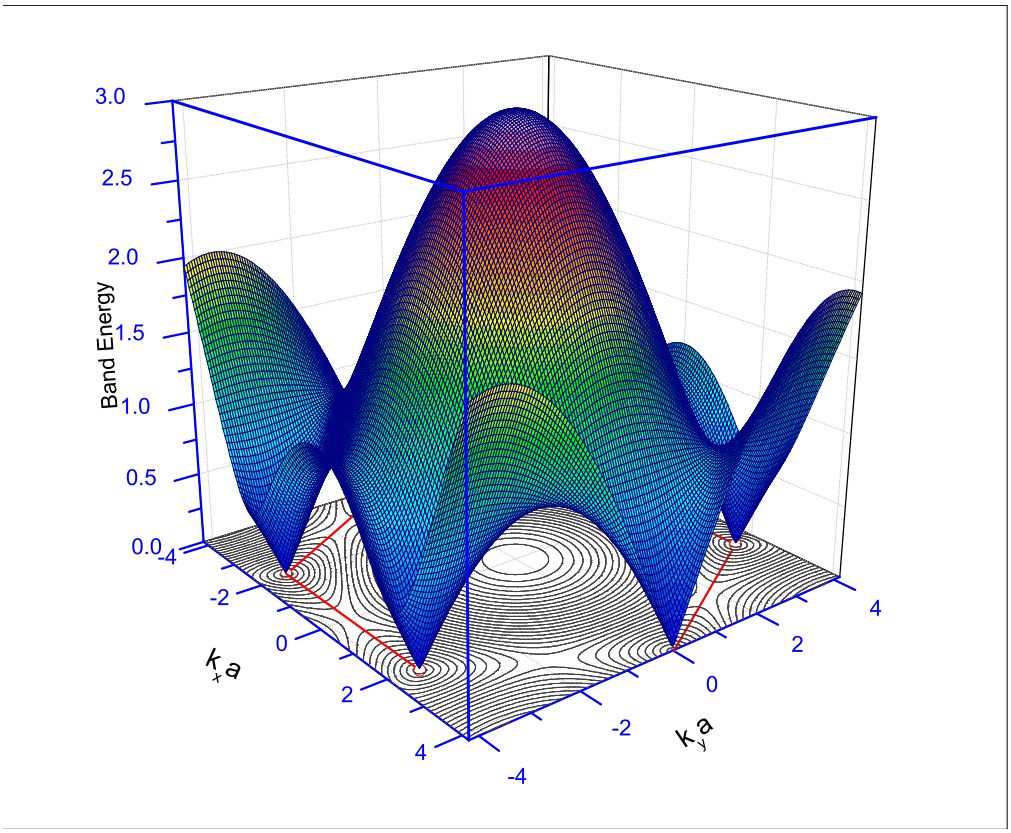


Figure 4.6: Notice that at the six corner points the upper and lower bands touch. The dispersion relation near those points is linear and hence the electrons near those points behave very differently from electrons in most other common substances which have a parabolic dispersion. Also the "touch" implies that we have a zero bandgap material.

4.2 The graphene problem in more detail

Let's now do the tight binding problem in a little bit more detail:

$$\begin{pmatrix} H_{AA} & H_{AB} \\ H_{BA} & H_{BB} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = E \begin{pmatrix} S_{AA} & S_{AB} \\ S_{BA} & S_{BB} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{4.65}$$

$$\tag{4.66}$$

Is the general expression we had discussed earlier.
The coefficients α and β are the unknowns in the problem now. They denote the contribution of the A and B sublattice to the single electron (Bloch) wavefunction spread over the entire lattice. Their values will depend on the hopping term t and the sum-over-sites term $F(\boldsymbol{k})$, but we will need to define a few more quantities. The nearest neighbour interactions are between the AB sites and is contained in the H_{AB} terms. The next nearest neighbour interaction is between AA and BB sites. We expect six identical terms, see the figure

$$\begin{aligned} H_{AA} &= E_A^0 + \sum_{nnn} e^{i\boldsymbol{k}.\boldsymbol{R}_{nnn}} \underbrace{\langle \phi_A(r)|H|\phi_A(r-a)\rangle}_{t'} \\ &= E_A^0 + t' \left(\underbrace{e^{i\boldsymbol{k}.\boldsymbol{a}_1} + e^{i\boldsymbol{k}.\boldsymbol{a}_2} + e^{-i\boldsymbol{k}.\boldsymbol{a}_1} + e^{-i\boldsymbol{k}.\boldsymbol{a}_2} + e^{i\boldsymbol{k}.(\boldsymbol{a}_1-\boldsymbol{a}_2)} + e^{i\boldsymbol{k}.(\boldsymbol{a}_1-\boldsymbol{a}_2)}}_{F_1(\boldsymbol{k})} \right) \end{aligned} \tag{4.67}$$

and

$$\begin{aligned} S_{AA} &= S_{BB} = 1 \\ S_{AB} &= \sum_{nn} e^{i\boldsymbol{k}.\boldsymbol{R}_{nn}} \underbrace{\langle \phi_A(r)|\phi_B(r-d)\rangle}_{s} \end{aligned} \tag{4.68}$$

$$S_{AB} = s \left(\underbrace{1 + e^{-i\boldsymbol{k}.\boldsymbol{a}_1} + e^{-i\boldsymbol{k}.\boldsymbol{a}_2}}_{F(\boldsymbol{k}) \quad \text{as before}} \right) \tag{4.69}$$

Notice that

$$F_1(\boldsymbol{k}) = |F(\boldsymbol{k})|^2 - 3 \tag{4.70}$$

So the more detailed equation now reads:

$$\begin{pmatrix} E_A + t'F_1(\boldsymbol{k}) - E & (t - sE)F(\boldsymbol{k}) \\ (t - sE)F^*(\boldsymbol{k}) & E_B + t'F_1(\boldsymbol{k}) - E \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0 \tag{4.71}$$

The effect of t' and non-identical atoms on A and B sites

Notice that the nnn term (t') results in a redefinition of the diagonal term and will not change the form of the equation. We can further see that if the A and B site energies are not equal then the diagonal terms will not be equal. A practical example is hexagonal boron nitride (h-BN), where the sites consist of two chemically different species. We have, allowing for the possibility that A and B atoms are not identical, the following:
Defining

$$\begin{aligned} E_A + t'F_1(\boldsymbol{k}) &= E_0 + \frac{\mu}{2} \\ E_B + t'F_1(\boldsymbol{k}) &= E_0 - \frac{\mu}{2} \end{aligned} \tag{4.72}$$

$$\tag{4.73}$$

and setting our zero level of energy to E_0

$$\begin{vmatrix} \frac{\mu}{2} - E & (t - sE)F(\boldsymbol{k}) \\ (t - sE)F^*(\boldsymbol{k}) & -\frac{\mu}{2} - E \end{vmatrix} = 0 \tag{4.74}$$

It tells us that the two bands will not touch if $\mu \neq 0$. So if the two atoms in the lattice are chemically different then there would be an energy gap. This is why monolayer graphene does not have an energy gap but hexagonal BN does.

Now we can solve equation 4.74 with identical atoms and the on-site energy as our reference point. We get

$$E(\boldsymbol{k}) = \frac{t' F_1(\boldsymbol{k}) \pm t|F(\boldsymbol{k})|}{1 \pm s|F(\boldsymbol{k})|} \tag{4.75}$$

The two solutions are identical (bands will touch) when $F(\boldsymbol{k}) = 0$ and hence $F_1 = -3$, but the curvature of the upper and lower bands will not be identical.

4.2.1 The solution near the six minimas

Let us denote the corner points where the bands touch by $\boldsymbol{K}$, and the small deviation as $\boldsymbol{q}$. So $\boldsymbol{k} = \boldsymbol{K} + \boldsymbol{q}$. Since the diagonal entries in equation 4.71 only depend on $|F(\boldsymbol{k})|^2$, if we expand $F(\boldsymbol{k})$ to first order in deviation from the corner points we will have only off-diagonal terms to retain.

$$\begin{aligned} F(\boldsymbol{k}) &= 1 + e^{-i\boldsymbol{k}.\boldsymbol{a}_1} + e^{-i\boldsymbol{k}.\boldsymbol{a}_2} \\ &= 1 + e^{-i\boldsymbol{K}.\boldsymbol{a}_1} e^{-i\boldsymbol{q}.\boldsymbol{a}_1} + e^{-i\boldsymbol{K}.\boldsymbol{a}_2} e^{-i\boldsymbol{q}.\boldsymbol{a}_2} \\ &= 1 + e^{-i\boldsymbol{K}.\boldsymbol{a}_1}\left[1 + (i\boldsymbol{q}.\boldsymbol{a}_1) - \frac{(\boldsymbol{q}.\boldsymbol{a}_1)^2}{2!} + ...\right] + e^{-i\boldsymbol{K}.\boldsymbol{a}_2}\left[1 + (i\boldsymbol{q}.\boldsymbol{a}_2) - \frac{(\boldsymbol{q}.\boldsymbol{a}_2)^2}{2!} + ...\right] \\ &= \underbrace{1 + e^{-i\boldsymbol{K}.\boldsymbol{a}_1} + e^{-i\boldsymbol{K}.\boldsymbol{a}_2}}_{F(\boldsymbol{K})=0} \underbrace{- \; i \; \left[e^{-i\boldsymbol{K}.\boldsymbol{a}_1}(\boldsymbol{q}.\boldsymbol{a}_1) + e^{-i\boldsymbol{K}.\boldsymbol{a}_2}(\boldsymbol{q}.\boldsymbol{a}_1)\right]}_{\text{first order}} \; + \; \mathcal{O}(|\boldsymbol{q}|^2 a^2) \end{aligned} \tag{4.76}$$

To write the Hamiltonian matrix correct to first order, we construct the following table first

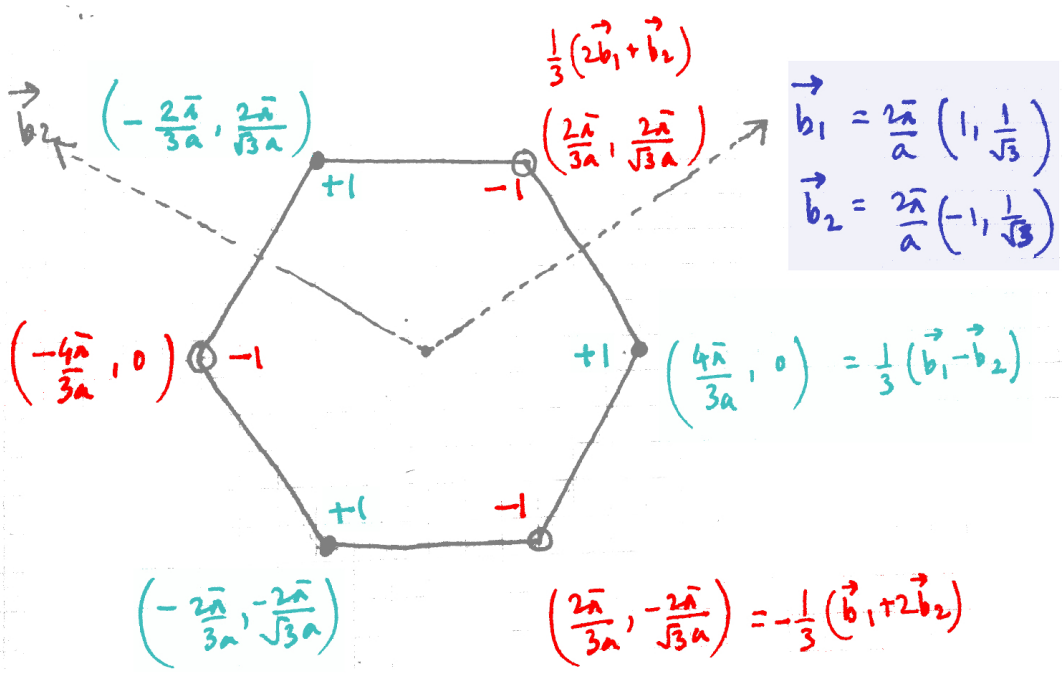


Figure 4.7: The figure shows the reciprocal lattice vector of graphene, the First Brillouin zone and the co-ordinate of the corner points

BZ corner pt	RLV combination	$e^{-i\boldsymbol{K}.\boldsymbol{a}_1}$	$e^{-i\boldsymbol{K}.\boldsymbol{a}_2}$	$F(\boldsymbol{K})$	First order terms	Point type
$+\frac{4\pi}{3a}$, 0	$\frac{1}{3}(\boldsymbol{b}_1 - \boldsymbol{b}_2)$	$e^{-i2\pi/3}$	$e^{i2\pi/3}$	0	$-\frac{a\sqrt{3}}{2}(q_x - iq_y)$	$K(\xi = +1)$
$+\frac{2\pi}{3a}$, $+\frac{2\pi}{\sqrt{3}a}$	$\frac{1}{3}(2\boldsymbol{b}_1 + \boldsymbol{b}_2)$	$e^{i2\pi/3}$	$e^{-i2\pi/3}$	0	$\frac{a\sqrt{3}}{2}(q_x + iq_y)$	$K'(\xi = -1)$
$-\frac{2\pi}{3a}$, $+\frac{2\pi}{\sqrt{3}a}$	$\frac{1}{3}(\boldsymbol{b}_1 + 2\boldsymbol{b}_2)$	$e^{-i2\pi/3}$	$e^{i2\pi/3}$	0	$-\frac{a\sqrt{3}}{2}(q_x - iq_y)$	$K(\xi = +1)$
$-\frac{4\pi}{3a}$, 0	$-\frac{1}{3}(\boldsymbol{b}_1 - \boldsymbol{b}_2)$	$e^{i2\pi/3}$	$e^{-i2\pi/3}$	0	$\frac{a\sqrt{3}}{2}(q_x + iq_y)$	$K'(\xi = -1)$
$-\frac{2\pi}{3a}$, $-\frac{2\pi}{\sqrt{3}a}$	$-\frac{1}{3}(2\boldsymbol{b}_1 + \boldsymbol{b}_2)$	$e^{-i2\pi/3}$	$e^{i2\pi/3}$	0	$-\frac{a\sqrt{3}}{2}(q_x - iq_y)$	$K(\xi = +1)$
$+\frac{2\pi}{3a}$, $-\frac{2\pi}{\sqrt{3}a}$	$-\frac{1}{3}(\boldsymbol{b}_1 + 2\boldsymbol{b}_2)$	$e^{i2\pi/3}$	$e^{-i2\pi/3}$	0	$\frac{a\sqrt{3}}{2}(q_x + iq_y)$	$K'(\xi = -1)$

Notice how the pattern alternates (these are the alternating K and K' type points). We also know from explicit calculation of the overlap integrals that the hopping parameter $t \approx -3$eV. Taking the negative

value of this in account we replace t by $-|t|$. The hamiltonian matrix becomes

$$\begin{pmatrix} 0 & tF(\boldsymbol{k}) \\ tF^*(\boldsymbol{k}) & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{|t|a\sqrt{3}}{2}(q_x - iq_y) \\ \frac{|t|a\sqrt{3}}{2}(q_x + iq_y) & 0 \end{pmatrix} \tag{4.77}$$

or

$$\begin{pmatrix} 0 & tF(\boldsymbol{k}) \\ tF^*(\boldsymbol{k}) & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{|t|a\sqrt{3}}{2}(q_x + iq_y) \\ -\frac{|t|a\sqrt{3}}{2}(q_x - iq_y) & 0 \end{pmatrix} \tag{4.78}$$

They can be brought together by using the Pauli matrices and a symbol (say ξ) that takes value $+1$ or -1. Notice also that the quantity $\frac{|t|a\sqrt{3}}{2}$ has the dimension of velocity$\times\ \hbar$. We define

$$v_F = \frac{|t|a\sqrt{3}}{2\hbar} = \frac{3|t|d}{2\hbar} \tag{4.79}$$

Here a is *not* the nearest neighbour distance and we can get rid of the $\sqrt{3}$ by defining the same quantity in terms of d, which *is* the nearest neighbour distance 1.42Å. It will turn out that this velocity plays the role of Fermi velocity. The $E(\boldsymbol{k})$ relation near the corner points does not allow us to define mass and hence $v = \frac{\hbar k}{m}$ type of definitions would not obviously work. We bring together the K and K' type behaviour by writing the 2×2 matrix using the Pauli matrices:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{4.80}$$

$$\sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{4.81}$$

$$H = \xi\hbar v_F(\sigma_x q_x + \xi\sigma_y q_y) \tag{4.82}$$

This hamiltonian acts on the column vector (spinor) $\begin{pmatrix} \alpha_+ \\ \beta_+ \end{pmatrix}$, or $\begin{pmatrix} \alpha_- \\ \beta_- \end{pmatrix}$. The coefficients of the two sublattice contributions near the *inequivalent* $\xi = +1$ and $\xi = -1$ points are not identical. The subscripts denote this distinction.

It would be useful if we could somehow remove ξ from inside the bracketed part, then the $\boldsymbol{\sigma}.\boldsymbol{q}$ kind of term would emerge naturally. This is possible if we invert the order of writing α and β near the K' points. So the four component column vector would read :

$$\Psi = \begin{pmatrix} \alpha_+ \\ \beta_+ \\ \beta_- \\ \alpha_- \end{pmatrix} \tag{4.83}$$

The entire set of equations can now be written as :

$$\hbar v_F \begin{pmatrix} 0 & q_x - iq_y & 0 & 0 \\ q_x + iq_y & 0 & 0 & 0 \\ 0 & 0 & 0 & -(q_x - iq_y) \\ 0 & 0 & -(q_x + iq_y) & 0 \end{pmatrix} \begin{pmatrix} \alpha_+ \\ \beta_+ \\ \beta_- \\ \alpha_- \end{pmatrix} = E \begin{pmatrix} \alpha_+ \\ \beta_+ \\ \beta_- \\ \alpha_- \end{pmatrix} \tag{4.84}$$

or more compactly as a Dirac equation with zero mass term and the speed of light replaced by v_F

$$\left[\hbar v_F \begin{pmatrix} \boldsymbol{\sigma}.\boldsymbol{q} & 0 \\ 0 & -\boldsymbol{\sigma}.\boldsymbol{q} \end{pmatrix} + \underbrace{I_4 m_0 c^2}_{m_0=0}\right] \Psi = E\Psi \tag{4.85}$$

The properties of the low energy electrons ($q << 1/a$) around the BZ corner points, can be obtained by solving this equation. We shall return to this later. However, remember that once the electrons gain enough energy, the first order approximation that led to this equation would no longer be valid.

4.3 Measuring the effective mass: Cyclotron resonance

Since the effective mass is the curvature of the $E(\boldsymbol{k})$ relation, matching this to experimental data is an important part of refining band structure calculations. How do we measure this? We use the fact that in magnetic field the electron orbits have a frequency of rotation. One way is to measure the resonant frequencies in a magnetic field. We will see that this frequency is related to effective mass in a particular direction. $\boldsymbol{M}$ is 3×3 matrix or the effective mass tensor. It is symmetric. The equation of motion of an electron in a band is:

$$\boldsymbol{M}\frac{d}{dt}\boldsymbol{v} = -e\boldsymbol{v} \times \boldsymbol{B} \tag{4.86}$$

If we look for oscillatory solutions, we must have

$$\boldsymbol{v} = \tilde{\boldsymbol{v}}_0 e^{i\omega t} \tag{4.87}$$

Let's direct the magnetic field along z axis, so that $\boldsymbol{B} = B_0\hat{\boldsymbol{z}}$.

PROBLEM : Show that the last two equations imply

$$\begin{vmatrix} M_{xx} & M_{xy} + i\frac{eB_0}{\omega} & M_{xz} \\ M_{xy} - i\frac{eB_0}{\omega} & M_{yy} & M_{yz} \\ M_{zx} & M_{zy} & M_{zz} \end{vmatrix} = 0 \tag{4.88}$$

And expanding the determinant gives

$$\frac{\det M}{M_{zz}} = \frac{e^2 B_0^2}{\omega^2} \tag{4.89}$$

Notice that if we rotate the direction of the magnetic field, we can bring another effective mass into focus.

Chapter 5

Carrier densities and dopants

References:

1. Chapter 5 (4^{th} edition), *Solid State Electronic Devices*, B. G. Streetman
2. Chapter 3 *Semiconductor Physics*, K Seeger
3. Greg Snider's homepage has the tool used to calculate band structures.
   See <www.nd.edu/∼gsnider>

Our target is to answer the following questions :

- How many carriers are there in the bands?
- How many dopants ionize? Where is the Fermi level? What is the driving equation?
- How can we qualitatively sketch the bending of the bands near a surface, metal-semiconductor contact, p-n junctions and heterointerfaces?
- Finally, what (self consistent) equations relate the charge densities and band profiles?

5.1 Carrier concentration and doping

At $T = 0$ in a pure semiconductor, the conduction band is empty and the valence band is full. A completely full or a completely empty band cannot carry current. We will see soon that under these circumstances the Fermi energy lies in the gap between the valence and conduction band. The density of states at the Fermi level is zero. The semiconductor is an insulator at this point.

In reality there is no qualitative distinction between semiconductors and insulators. The distinction is that the bandgap of an insulator is large - *e.g.* Silicon oxide has $E_g \sim 9$eV, Diamond has $E_g \sim 5$eV and so on. The bandgap of typical semiconductors is in the range of nearly zero to 3-4 eV. At very high temperatures, if an insulator hasn't already melted, it will act as a semiconductor.

Carriers in a semiconductor's bands come from two sources:

1. Thermally excited electrons in conduction band and the corresponding vacancies left behind in the valence band.
2. Some *suitable* foreign atoms called dopants which can put some electrons in CB or capture some electrons from VB. Sometimes crystal defects can also play the role of foreign atoms.

Consider a group V atom like Phosphorous replacing an atom of group IV Silicon in the lattice. It has one more electron compared to Si. We keep aside the question about how to get the P atom to replace the Si for the time being - but that is not a trivial question. A "dopant" will not work as a dopant if it does not sit in the right place. It is possible for a P atom to somehow go in as an "interstitial", that will not work. Also the same atom may act as an acceptor or a donor in some cases. For example if Si is incorporated in GaAs lattice, replacing a Ga atom, it will act as a donor. If it replaces an As atom it will act as an acceptor. You can figure out the reason.

The crucial fact is that the binding energy of that remaining electron becomes very low. We give a very simplified model - usually called the "hydrogenic impurity model". We assume that the outermost electron in P behaves as if it is tied to a hypothetical nucleus - that is the P^+ ion core. The binding energy and Bohr radius of an H atom (1s state) is

$$E = -\frac{me^4}{8\varepsilon_0^2 h^2} \tag{5.1}$$

$$a_B = 4\pi\varepsilon_0 \frac{\hbar^2}{me^2} \tag{5.2}$$

Now we make two crucial claims. Inside the "medium" the free electron mass would be modified such that $m \to m_{eff}$ and $\varepsilon_0 \to \varepsilon_0\varepsilon_r$. Typically $\epsilon_r \sim 10-15$ for most semiconductor lattices and $m_{eff} \sim 0.1m$. That means the binding energy would reduce by a factor of ~ 1000 and the Bohr radius would increase by a factor of about ~ 100. So instead of $E = 13.6$eV the binding energy will be a few 1-10 meV, the Bohr radius will increase from 0.5Å to ~ 50Å. This means that the electron will be exploring something of the order of a 10×10×10 lattice units. This in retrospect justifies the use of the lattice dielectric which is a quantity meaningful only if averaged over sum volume of the lattice. Also the fact that the electron gets spread over a large area, means that replacing the free electron mass with the band effective mass can be justified. If the binding energy drops to a few meV, it is clear that at room temperature ($k_BT = 25$meV) these can be almost fully ionised. The order of magnitude of these numbers ensure that semiconductors can be useful at room temperature.

How can we make the arguments better for using the band effective mass? For a direct gap semiconductor we proceed as follows: We need to treat the extra potential introduced by the impurity atom as a "perturbation" and then solve for the wave function.

$$H = H_0 - \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{e^2}{r} \tag{5.3}$$

Here the potential seen by the electron can be split into two generic parts - the fast varying potential of the host lattice and the much more slowly varying "extra" potential brought about by the dopant/impurity atom's core. We will call these $V_{lattice}$ and V_{slow} respectively. We know the solution of the hamiltonian with the fast varying $V_{lattice}$ part, which are the Bloch functions.

where $H_0 = T + V = \left[-\frac{\hbar^2}{2m}\nabla^2 + V_{lattice}(\boldsymbol{r})\right]$ is the KE + fast varying lattice periodic potential part. $m = m_0$ is the *free electron mass*. The relevant solution to this must be formed out of the eigenfunctions of H_0, the Bloch functions (mostly from near the bottom of the band), $C_n(\boldsymbol{k})$ are the linear coefficients associated with them.

$$\Psi = \int_{BZ} \frac{d\boldsymbol{k}}{(2\pi)^3} C_n(\boldsymbol{k})\ \phi_n(\boldsymbol{k}, \boldsymbol{r}) \tag{5.4}$$

where $\phi_n(\boldsymbol{k})$ are the Bloch functions of the relevant band. They satisfy

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{lattice}\right]\phi_n(\boldsymbol{k}, \boldsymbol{r}) = E_n(\boldsymbol{k})\phi_n(\boldsymbol{k}, \boldsymbol{r}) \tag{5.5}$$

Ψ satisfies

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{lattice} + V_{slow}\right]\Psi = E\Psi \tag{5.6}$$

$E_n(\boldsymbol{k})$ are known, E is the unknown we need to solve for. For notational simplicity we assume that the wavefunctions we are dealing with are confined to one band and drop the band index n. In general there can be a sum over n. We also drop the $\boldsymbol{r}$ from the arguments of the ϕ. But remember that the spatical derivative will act only on those.

$$\begin{aligned}
E\Psi &= E\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k}) \\
&= -\frac{\hbar^2}{2m}\nabla^2\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k}) \;+\; (V_{lattice}+V_{slow})\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k}) \\
&= \int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\left[-\frac{\hbar^2}{2m}\nabla^2\phi(\boldsymbol{k})\right] \;+\; (V_{lattice}+V_{slow})\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k}) \\
&= \int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\left[(E(\boldsymbol{k})-V_{lattice})\phi(\boldsymbol{k})\right] \;+\; (V_{lattice}+V_{slow})\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k})
\end{aligned}$$

Notice that the $V_{lattice}$ term will cancel. So far the calculation has been exact and we have not dropped anything. We have now got rid of the second derivative which can produce fast variations (recall the derivatives make a function "rougher", integration generally makes them "smoother". It is important to keep these in mind while making approximations. A derivative acting on a function which has small value but a jump, can suddenly produce a large term. Throwing away a term on which a derivative acts, before the derivative has been "allowed to act" on it, while making an approximation is something one should not do.

We will now make the crucial approximation that this is an isotropic parabolic band, near a minima. This is a practical (though not fully general) situation.

$$\begin{aligned}
E\Psi &= \int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\left[E(\boldsymbol{k})\right]\phi(\boldsymbol{k}) \;+\; V_{slow}\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k}) \\
&= \int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\left[E_0+\frac{\hbar^2}{2m_{eff}}\boldsymbol{k}^2\right]\phi(\boldsymbol{k}) \;+\; V_{slow}\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k}) \\
\therefore (E-E_0)\Psi &= \int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\left[\frac{\hbar^2}{2m_{eff}}\boldsymbol{k}^2\right]\phi(\boldsymbol{k}) \;+\; V_{slow}\int_{BZ}\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\phi(\boldsymbol{k})
\end{aligned}$$

Notice the way m_{eff} enters into the equation, via the dispersion relation. It is *not put in by hand.*

Recall that Bloch functions are like

$$\phi_n(\boldsymbol{k},\boldsymbol{r}) = u_n(\boldsymbol{k},\boldsymbol{r})e^{i\boldsymbol{k}.\boldsymbol{r}} \tag{5.7}$$

Now we claim that the following approximation should work.

$$\Psi \approx u_n(0,\boldsymbol{r})\underbrace{\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\;e^{i\boldsymbol{k}.\boldsymbol{r}}}_{F(\boldsymbol{r})} \tag{5.8}$$

Which physically means that we are assuming that relatively few states from near the extrema are required to build the total wavefunction. This is not inconsistent with the previous approximation that the dispersion is isotropic and parabolic in our region of interest. So we have

$$(E-E_0)u(0,\boldsymbol{r})\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})e^{i\boldsymbol{k}.\boldsymbol{r}} = u(0,\boldsymbol{r})\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})\frac{\hbar^2}{2m_{eff}}\boldsymbol{k}^2e^{i\boldsymbol{k}.\boldsymbol{r}} + V_{slow}u(0,\boldsymbol{r})\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})e^{i\boldsymbol{k}.\boldsymbol{r}}$$

$$(E-E_0)\underbrace{\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})e^{i\boldsymbol{k}.\boldsymbol{r}}}_{F(\boldsymbol{r})} = -\frac{\hbar^2}{2m_{eff}}\nabla^2\underbrace{\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})e^{i\boldsymbol{k}.\boldsymbol{r}}}_{F(\boldsymbol{r})} + V_{slow}\underbrace{\int\frac{d\boldsymbol{k}}{(2\pi)^3}C(\boldsymbol{k})e^{i\boldsymbol{k}.\boldsymbol{r}}}_{F(\boldsymbol{r})}$$

$$\therefore\left[-\frac{\hbar^2}{2m_{eff}}\nabla^2 + V_{slow}\right]F(\boldsymbol{r}) = (E-E_0)F(\boldsymbol{r}) \tag{5.9}$$

For a impurity Coulomb potential, it leads to the *hydrogen atom like-*

$$\left[-\frac{\hbar^2}{2m_{eff}}\nabla^2 + \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{e^2}{r}\right]F(\boldsymbol{r}) = (E-E_0)F(\boldsymbol{r}) \tag{5.10}$$

This forms the basis of the hydrogenic impurity model. It also tells us when that works and what are its limitations. You can try to construct the equation if the band is not isotropic, as an exercise. The $F(r)$ function that we introduced is not really a wavefunction, in the sense that it is not the solution of the original Schrodinger equation. The wavefunction is Ψ, which we defined at the beginning and can be constructed if we know $F(r)$.
The justification of a similar model for indirect (and multiple valley) semiconductors like Si is more involved (this was given by Luttinger). Also the simple form of the equation will need modification if the band is not isotropic. The summary of the result is that for direct gap single valley (GaAs) the shallow donor level cluster around one number. See the figure 5.1. For indirect gap Si, the clustering does not happen so well. We will in general take this numbers as experimentally determined parameters.

5.2 A few useful numbers about Si, Ge and GaAs

5.3 Fermi Level in an intrinsic (undoped) semiconductor

If the material is undoped, then all the electrons in the conduction band (CB) must have been thermally excited from the valence band (VB). This fact is sufficient to tell us where the (E_f) should be. The electron and hole densities must be,

$$n = \int_{E_C}^{\infty}\mathrm{d}E\ D(E)f(E) \tag{5.11}$$

$$p = \int_{-\infty}^{E_V}\mathrm{d}E\ D(E)\left(1-f(E)\right) \tag{5.12}$$

Let us assume that the dispersion relations are very simple

$$E_e(k) = E_C + \frac{\hbar^2k^2}{2m_e} \tag{5.13}$$

$$E_h(k) = E_V - \frac{\hbar^2k^2}{2m_h} \tag{5.14}$$

where E_C, E_V denote the bottom and the top of the conduction and valence bands respectively

Table 5.1: List of commonly used parameters of Silicon, Germanium and Gallium Arsenide

	Silicon	Germanium	Gallium Arsenide
Atoms cm^{-3}	5.0×10^{22}	4.4×10^{22}	4.4×10^{22}
Crystal structure	Diamond	Diamond	Zincblende
Density (gm cm^{-3})	2.33	5.33	5.32
Dielectric constant	11.9	16	13.1
Electron affinity (eV)	4.1	4.0	4.1
Effective density of states in conduction band at 300K: $N_c(cm^{-3})$	2.8×10^{19}	1.04×10^{19}	4.7×10^{17}
Effective density of states in valence band at 300K: $N_v(cm^{-3})$	1.04×10^{19}	6.0×10^{18}	7.0×10^{18}
Band gap at 300K (eV)	1.12	0.66	1.42
Intrinsic carrier concentration at 300K: $n(cm^{-3})$	1.5×10^{10}	2.4×10^{13}	1.8×10^{6}
electron effective mass : in units of m_0, the free electron mass	0.98, 0.19	1.64, 0.082	0.067
hole effective mass : in units of m_0, the free electron mass	0.16, 0.49	0.044, 0.28	0.082, 0.45
Intrinsic electron mobility at 300K ($cm^2V^{-1}s^{-1}$)	1350	3900	8500
Intrinsic hole mobility at 300K ($cm^2V^{-1}s^{-1}$)	480	1900	400
Electron diffusion coefficient at 300K (cm^2s^{-1})	35	100	220
Hole diffusion coefficient at 300K (cm^2s^{-1})	12.5	50	10

The density of states (in 3D), including spin degeneracy is then given by:

$$D(E) = \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_C)^{1/2} \qquad \text{for} \qquad E > E_C \tag{5.15}$$

$$D(E) = \frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2}(E_V-E)^{1/2} \qquad \text{for} \qquad E < E_V \tag{5.16}$$

PROBLEM : Calculate the density of states in 2D and 1D for parabolic bands. Explain why it is allright to take one of the limits to be infinity in equations 5.11 & 5.12, even though all bands have finite extents.

Now to evaluate equations 5.11 & 5.12 we proceed as :

$$\begin{aligned} n &= \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int_{E_C}^{\infty} \mathrm{d}E\ (E-E_C)^{1/2}\frac{1}{e^{\beta(E-E_F)}+1} \\ &= \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\frac{1}{\beta^{3/2}}\int_0^{\infty}\mathrm{d}u\ \frac{u^{1/2}}{e^u e^{-\beta(E_F-E_C)}+1} \qquad \text{where} \qquad u=\beta(E-E_C) \\ &= 2\left(\frac{2\pi m_e k_B T}{h^2}\right)^{3/2}\left(\frac{2}{\sqrt{\pi}}\int_0^{\infty}\mathrm{d}u\ \frac{u^{1/2}}{e^u e^{\beta(E_C-E_F)}+1}\right) \end{aligned} \tag{5.17}$$

$$\tag{5.18}$$

Now we identify the integral within the brackets as a Fermi-Dirac integral, defined as :

$$F_j(z) = \frac{1}{\Gamma(j+1)}\int_0^{\infty}\mathrm{d}x\frac{x^j}{e^z e^x+1} \tag{5.19}$$

Further the "effective density of states" in the conduction band is defined as

$$N_C = 2\left(\frac{2\pi m_e k_B T}{h^2}\right)^{3/2} \tag{5.20}$$

Note however that the dimension of N_C is not the same as $D(E)$. With the definitions eqn 5.19 & 5.20, eqn 5.17 then reduces to

$$n = N_C F_{1/2}\left(\frac{E_C - E_F}{k_B T}\right) \tag{5.21}$$

You can prove that the number of holes is given by

$$p = N_V F_{1/2}\left(\frac{E_F - E_V}{k_B T}\right) \tag{5.22}$$

The Fermi-Dirac integrals appear often in physics. They are tabulated as "special functions". We can show that if E_F is reasonably *below* E_C, such that $\frac{E_F - E_C}{k_B T} < -4$ the integral is very closely approximated by $e^{\frac{E_F - E_C}{k_B T}}$. This is called the non-degenerate regime where the electron and hole densities are given by

$$n = N_C e^{\beta(E_F - E_C)} \tag{5.23}$$
$$p = N_V e^{\beta(E_V - E_F)} \tag{5.24}$$

For charge neutrality we must have $n_i = p_i$ for undoped (intrinsic) semiconductors only. Multiplying eqns 5.23 & 5.24

$$\begin{aligned} n_i^2 &= n_i p_i = N_C N_V e^{-\beta(E_C - E_V)} \\ \therefore\ n_i &= \sqrt{N_C N_V}\ \ e^{-\beta E_g/2} \end{aligned} \tag{5.25}$$

Clearly the intrinsic carrier density falls rapidly with increasing band gap. We can compare the result with the data in table 5.1.

Note however that the product of the carrier densities is independent of the location of the Fermi level even when $n \neq p$. This is a very important fact and allows us to write

$$np = n_i^2 \tag{5.26}$$

even when the source of the charges are dopants. In such cases (we will see in the next section) the Fermi level moves away from its intrinsic position. The electron and hole densities can them become vastly unequal - but they do so in such a way that the np product still remains the same. We now solve eqns 5.23 & 5.24 for E_f and get the intrinsic Fermi level (E_{fi})

$$E_{fi} = \frac{E_C + E_V}{2} + \frac{3}{4} k_B T \ln \frac{m_h}{m_e} \tag{5.27}$$

PROBLEM : Show that the deviation of the electron density (n) from intrinsic density (n_i) and the deviation of the Fermi level (E_f) from the intrinsic Fermi level (E_{fi}) are related as :

$$n = n_i e^{\beta(E_f - E_{fi})} \tag{5.28}$$

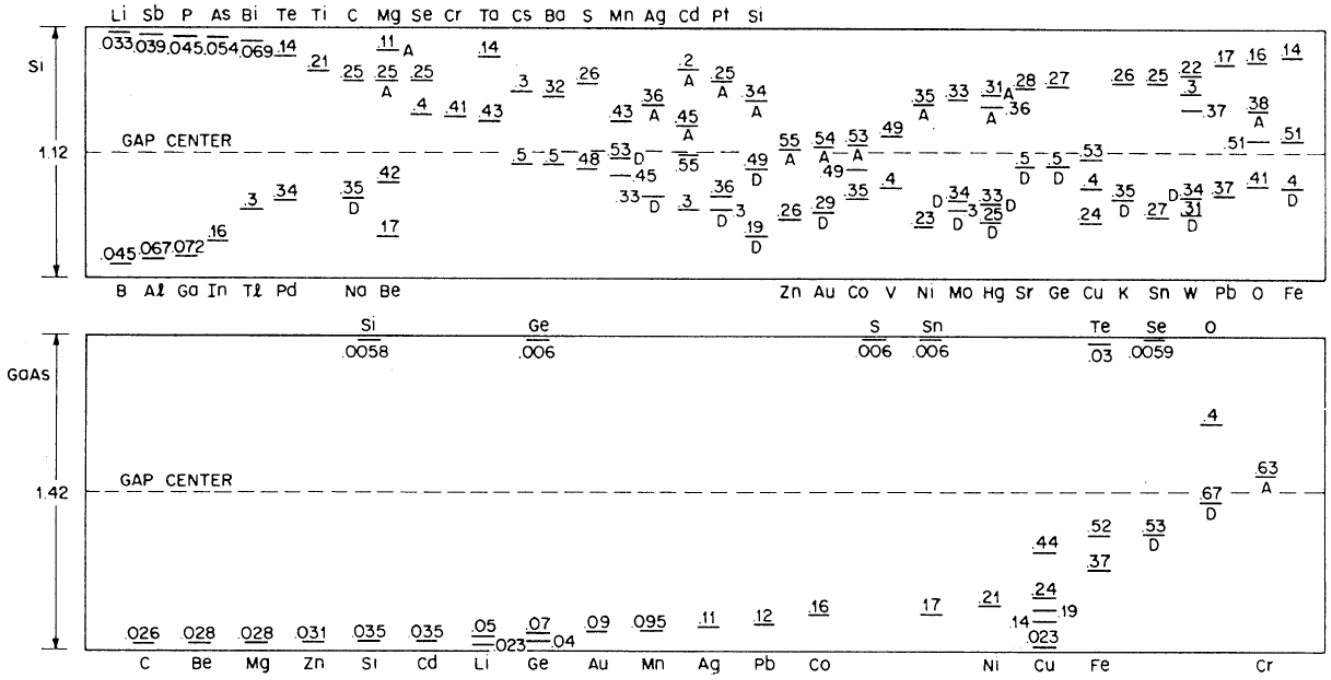


Figure 5.1: Impurity levels in Silicon and Gallium Arsenide (taken from the book by S.M. Sze). Notice that the shallowest levels tend to cluster somewhat around a value.

5.4 Fermi level in a doped semiconductor

We now come to the more practical situation, where there are dopants and ask: where is the Fermi level? If there are dopants then n and p are no longer equal. In fact the number of carriers supplied by ionised dopants can be several orders larger than the intrinsic carrier densities.

The fundamental point is that all the atoms of the host lattice and the dopants were initially neutral. But inside the semiconductor there are now four sources of charge :

- Negatively charged electrons in the conduction band (n)
- Unoccupied (positively charged) ionised donor atoms (N_d^+)
- Negatively charged ionised acceptor atoms (N_a^-)
- Unoccupied states (holes) in the valence band

The sum total of all these four must continue to be zero. To start with the valence band was full and the conduction band was empty (intrinsic semiconductor), then we put in neutral donor atoms (capable of giving out an electron) and neutral acceptor atoms (capable of capturing an electron). So the sum total must remain zero.

Thus if we can write down the carrier concentrations in the conduction and valence band and calculate the fraction of dopants which are ionised (as a function of E_f) then we can have an equation where E_f is the only unknown. This is how one determines the location of E_f

$$n + N_a^- = p + N_d^+ \tag{5.29}$$

We know how to write n and p as a function of E_f using equations 5.23 and 5.24. Now we ask, what is the probability that a donor will ionise. This question is an interesting exercise in statistical physics. The donor site (*e.g* Phosphorous in Silicon) can exist in four states

1. It may lose its electron (charge = +1 , energy = 0)

2. It may be occupied by a spin up electron (charge = 0, energy = E_D)

3. It may be occupied by a spin down electron (charge = 0, energy = E_D)

4. It may be occupied by one spin up and one spin down electron (charge = -1 , energy = $2E_D + U$ where "U" is the large repulsive energy cost of putting two electrons on the same site, making the state very improbable.)

The dopant densities are not very large compared to the density of atoms of the host lattice. It is rarely more than 1 in 10^3 to 10^4. So we can treat each dopant atom in isolation and the electron can be localised on the atom.[1] Each dopant can exchange electrons with the "sea" of conduction band electrons. It is thus in equilibrium with a larger system and can exchange particles with it - thus its temperature and chemical potential must be the same as that of the larger system.

So we write the partition function as (with μ, the chemical potential set as E_f)

$$\begin{aligned} Z_G &= \sum_{E,N} e^{-\beta(E-\mu N)} \\ &= e^{-\beta(0-0)} + e^{-\beta(E_D-E_f)} + e^{-\beta(E_D-E_f)} + e^{-\beta(2E_D+U-2E_f)} \\ &\approx 1 + 2e^{-\beta(E_D-E_f)} \end{aligned} \tag{5.30}$$

The mean occupancy (probability that the dopant is *not* ionised) is then,

$$\begin{aligned} 1 - \frac{N_D^+}{N_D} &= 0.P(0) + 1.P(\uparrow) + 1.P(\downarrow) + 2.P(\uparrow\downarrow) \\ &= \frac{2e^{-\beta(E_D-E_f)}}{Z_G} \\ &= \frac{2e^{-\beta(E_D-E_f)}}{1 + 2e^{-\beta(E_D-E_f)}} \\ &= \frac{1}{\frac{1}{2}e^{\beta(E_D-E_f)} + 1} \end{aligned} \tag{5.31}$$

Note that this is not simply a Fermi-Dirac distribution. The fraction of ionised donors is

$$\frac{N_D^+}{N_D} = \frac{1}{1 + 2e^{-\beta(E_D-E_F)}} \tag{5.32}$$

The similar expression for the fraction of ionised (negatively charged) acceptors is

$$\frac{N_A^-}{N_A} = \frac{1}{1 + 4e^{-\beta(E_F-E_A)}} \tag{5.33}$$

The factor 4 is a result of the fact that the electron sitting on the acceptor could have come from four possible places - spin up/down from heavy hole band, spin up/down from light hole band. The split off band does not come into the picture because it is too far down.

One type of dopant only

If we neglect the valence band and the acceptors (which can be justified if only donors are present), combining eqns 5.23 and 5.32 we get

[1] However if the dopant densities *are* very high then the dopant states will not be localised. This condition is called "Mott transition". It is among the most studied problems in semiconductor physics.

$$N_C e^{\beta(E_f - E_C)} = \frac{N_D}{1 + 2e^{-\beta(E_D - E_f)}} \tag{5.34}$$

E_f is the only unknown in eqn 5.34 and can be solved (numerically if required).

PROBLEM : Show that the carrier density can now be obtained by solving the following equation: (which is in turn obtained by using eqn 5.34)

$$n^2 + nN_C\frac{e^{-\beta\Delta}}{2} - N_D N_C\frac{e^{-\beta\Delta}}{2} = 0 \tag{5.35}$$

where $\Delta = E_C - E_D$.
The fermi level can be obtained by solving

$$x^2 + x\frac{e^{-\beta\Delta}}{2} - \frac{N_D}{N_C}\frac{e^{-\beta\Delta}}{2} = 0 \tag{5.36}$$

where $x = e^{\beta(E_f - E_C)}$

If you put $N_D = 0$ in either of the two equations you would get an unphysical answer. Why is this so?

PROBLEM :
In a system with N_D donors, N_D acceptors, N_D^+ donors and N_A^- acceptors are ionised. Each donor (acceptor) level has a degeneracy of g_D (g_A). There are n electrons in the conduction band and p holes in the valence band. (In general $g_D = 2$, but g_A may be different from 2.). Then

$$N_D^+ = \frac{N_D}{(g_D n/N_C)\exp\beta(E_C - E_D) + 1} \tag{5.37}$$

And the corresponding result for the acceptors:

$$N_A^- = \frac{N_A}{(g_A p/N_V)\exp\beta(E_A - E_V) + 1} \tag{5.38}$$

Here N_C and N_V are the conduction and valence band effective density of states which have been defined earlier. Notice that the fermi energy does not appear in these relations.

A semiconductor may be doped with both (acceptors and donors) types of dopants. In a situation where there are a large number of donors and a few acceptors ($i.e\ N_D >> N_A$), how would eqn 5.35 (the previous problem) be modified?

5.4.1 Thermal ionisation (Saha equation) of the dopant system

It is instructive to calculate the fraction of ionised dopants in another way. We can think of the problem as a thermal ionization of bound states - in a way that is very similar to the method of calculating the ratio of ionised to unionised atoms (of a certain species) in a hot plasma. We want to find the "chemical equilibrium point" of the reaction:

$$\text{atom} \ + \ \text{ionisation energy} \leftrightarrow \text{ionised atom} \ + \ \text{electron} \tag{5.39}$$

A certain fraction of atoms will exist in the dissociated state and a certain fraction will remain in the undissociated state. The fraction which minimises the free energy of the entire system (at a certain temperature) will be the equilibrium point.

Taking this approach we can calculate the ratio N_D^+/N_D by minimising the free energy of the entire system of free electrons and the dopants. First we write the free energy so that the free electron concentration n is the only variable.

$$\begin{aligned} F_{system} &= F_{electrons} + F_{dopants} \\ F_{electrons} &= -kT\ln\frac{z^n}{n!} \text{ where for a single electron} \\ z &= \sum_{all\ \ states} e^{-\beta E} \\ &= V\frac{2}{h^3}\int \mathrm{d}^3\mathbf{p}\ \ e^{-\frac{\beta p^2}{2m}} \\ &= 2V\left(\frac{2\pi mkT}{h^2}\right)^{3/2} \end{aligned} \tag{5.40}$$

Now since $N_D - n$ dopant sites are occupied we have for the internal energy (U) and entropy (S)

$$U = -\Delta(N_D - n) \tag{5.41}$$

$$S = k\ln\left(2^{N_D-n}\frac{N_D!}{n!(N_D-n)!}\right) \tag{5.42}$$

$$F_{dopants} = U - TS \tag{5.43}$$

PROBLEM : Minimise $F_{system} = F_{electrons} + F_{dopants}$ w.r.t n, using Stirling's approximation for factorials as needed and show that you get exactly the same result as eqn 5.35. This is essentially a variant of the "Saha ionisation" equation, applied to a situation where the atoms and ions are not mobile, but only the electrons are.

5.4.2 General method of solving for the Fermi level

Consider a situation where a semiconductor is doped with N_D donors and N_A acceptors. We want the general solution for the location of E_F and all the carrier densities, ionisation probabilities.
Since the semiconductor is overall neutral we have using the charge neutrality condition

$$n + N_A^- = p + N_D^+ \tag{5.44}$$

$$N_C F_{1/2}\left(\frac{E_C - E_F}{k_B T}\right) + \frac{N_A}{1 + g_A e^{\beta(E_A - E_F)}} = N_V F_{1/2}\left(\frac{E_F - E_V}{k_B T}\right) + \frac{N_D}{1 + g_D e^{\beta(E_F - E_D)}} \tag{5.45}$$

- here $g_A = 4$ and $g_D = 2$ are the acceptor and donor degeneracies. E_A and E_D are the acceptor and donor levels.
- Since E_F is the only unknown here, we can plot the LHS and RHS by treating E_F as an independent variable. The position where they intersect must be the solution. Figure ... illustrates the situation.
- Notice that E_F is temperature dependent.
- Once E_F is determined all the quantities can be determined. In general this cannot be done analytically. An example on numerical solution is given in Fig 5.2, 5.3.

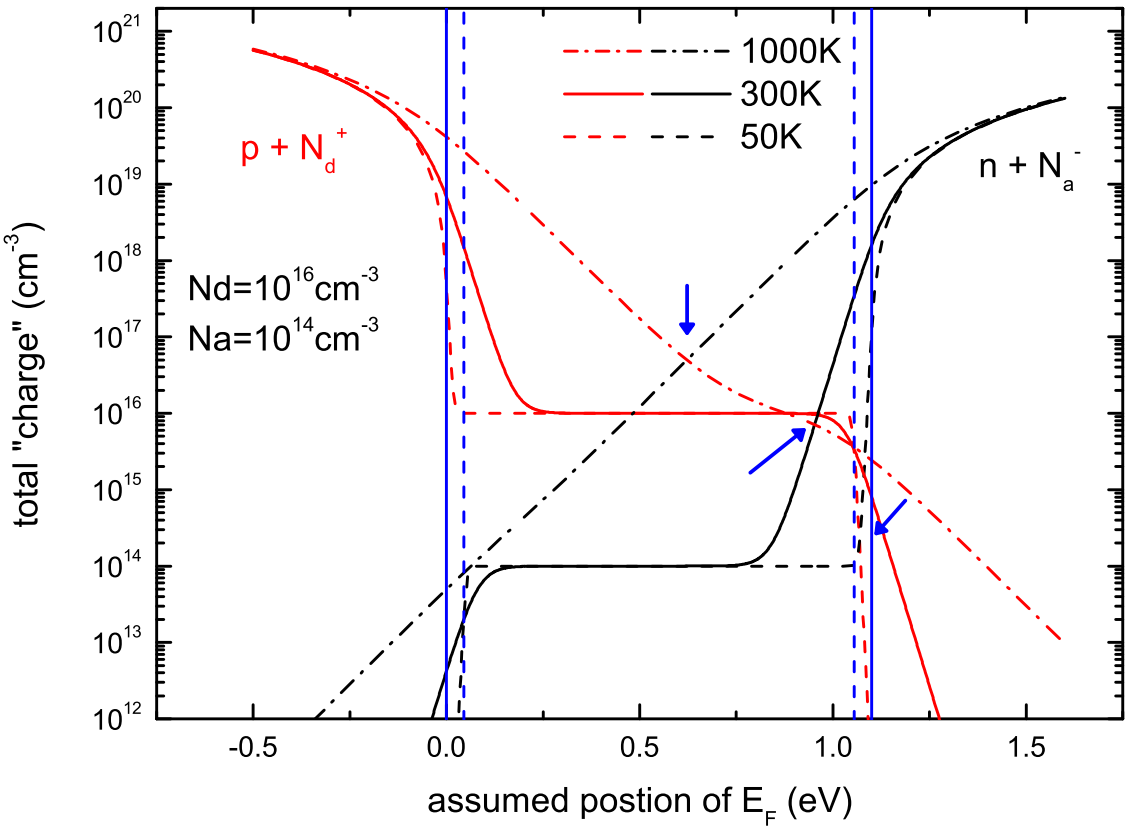


Figure 5.2: The LHS and RHS of eqn 5.45 are plotted for different temperatures. The intersection point is the solution for E_F.

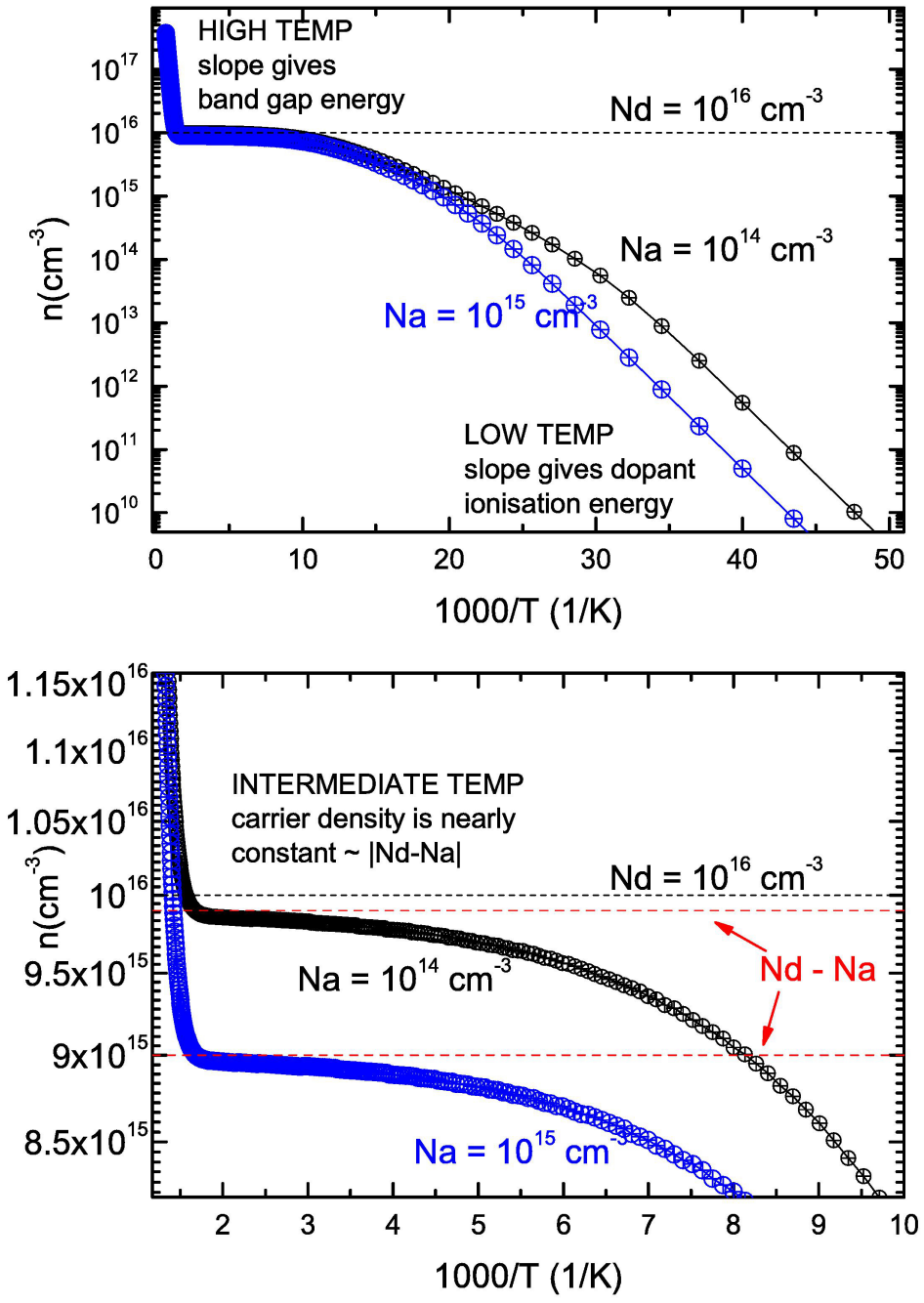


Figure 5.3: Figure shows the resulting carrier densities obtained from $n(T)$. Notice the various regimes.

5.5 The concept of a hole

We have seen earlier how an electron in a Bloch state behaves. In semiconductors we often encounter situations where a band is almost completely full but has a few vacant states. This happens because the Fermi level in an undoped semiconductor typically lies between the top of a filled band and the bottom of an unfilled band. We make the following observations and develop the method of dealing with this in steps. In what follows we denote the electron's wavevector by $\boldsymbol{k}_e$.

Remember that a hole is NOT a particle with a positive charge! It is one of the major conceptual mistakes.

1. The sum total of all the wavevectors in a band is

$$\sum_{all\ states} \boldsymbol{k} = 0$$

$$\therefore \sum_{all\ states} \boldsymbol{k} = -\boldsymbol{k}_e \tag{5.46}$$

$$\tag{5.47}$$

if one state $\boldsymbol{k}_e$ is vacant. This "missing" wavevector is assigned to a "hole" and we say that the wavevector of the hole state is

$$\boldsymbol{k}_h = -\boldsymbol{k}_e \tag{5.48}$$

PROBLEM : Using this condition, explain why a hole state would tend to move in a k-space direction opposite to that of an electron when an electric field is applied? At no stage should you invoke anything to do with a positive charge.

2.

$$\sum_{all\ states} \boldsymbol{v}_g = 0 \tag{5.49}$$

Since $\boldsymbol{v}_g = \frac{1}{\hbar}\boldsymbol{\nabla} E(\boldsymbol{k})$ and $E(\boldsymbol{k}+\boldsymbol{G}) = E(\boldsymbol{k})$ the result follows. We are essentially integrating the derivative of a periodic function over one period. This must be zero. This result is a *crucial* one as well.

3. Since one electron is missing from the band the energy of a hole should be

$$E_h = -E_e \tag{5.50}$$

4. Using equation 5.48 and eqn. 5.50 we arrive at the conclusion

$$\boldsymbol{v}_h = \boldsymbol{v}_e \tag{5.51}$$

The hole state moves with the same group velocity as the electron state.

5. The equation of motion (Lorentz force) for an electron state can be written down and *converted* to an equation for the hole as follows

$$\frac{d\boldsymbol{k}_e}{dt} = -|e|\,(\boldsymbol{E} + \boldsymbol{v}_e \times \boldsymbol{B})$$

$$\therefore -\frac{d\boldsymbol{k}_h}{dt} = -|e|\,(\boldsymbol{E} + \boldsymbol{v}_h \times \boldsymbol{B})$$

$$\therefore \frac{d\boldsymbol{k}_h}{dt} = |e|\,(\boldsymbol{E} + \boldsymbol{v}_h \times \boldsymbol{B}) \tag{5.52}$$

The hole state evolves as if it contains a particle of charge $|e|$. *This emerges from the equation of motion and was not put in by hand. There is no justification of assigning a positive charge to a*

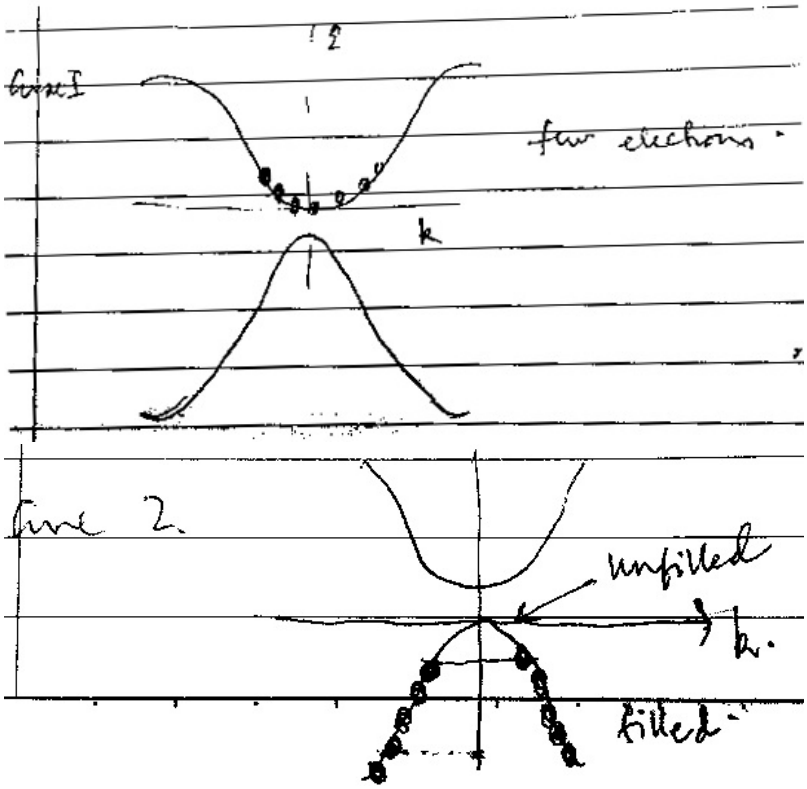


Figure 5.4:

hole a priori. However the equation does predict that the Lorentz force on the holes is oppositely directed (as compared to an electron). This is what leads to the opposite sign for the Hall voltage for a system of holes. It also implies that cyclotron orbits would have an opposite sense of circulation.

6. Now consider a simple situation with a few electrons as shown in fig 5.4 (top).

   An electric field $\boldsymbol{E}$ is applied and the scattering time is τ. All the $\boldsymbol{k}$ states shift a little bit

$$\boldsymbol{k}_e \to \boldsymbol{k}_e + \underbrace{\frac{-|e|\tau \boldsymbol{E}}{\hbar}}_{\delta \boldsymbol{k}} \tag{5.53}$$

   If the region we are dealing with is parabolic, then

$$\boldsymbol{v}_g = \frac{1}{\hbar}\boldsymbol{\nabla} E = \frac{\hbar}{m_{eff}}\boldsymbol{k} \tag{5.54}$$

   Due to the electric field

$$\boldsymbol{v}_g \to \boldsymbol{v}_g^0 + \underbrace{\frac{\hbar}{m_{eff}}\delta \boldsymbol{k}}_{\delta \boldsymbol{v}_g} \tag{5.55}$$

Hence

$$\boldsymbol{j} = -|e| \sum_{\substack{\text{occ} \\ \text{states}}} \boldsymbol{v}_g \tag{5.56}$$

$$= \underbrace{-|e| \sum_{\substack{\text{occ} \\ \text{states}}} \boldsymbol{v}_g^0}_{=0 \text{ in equilibrium}} \quad -|e| \sum_{\substack{\text{occ} \\ \text{states}}} \frac{\hbar}{m_{eff}} \left(\frac{-|e|\tau \boldsymbol{E}}{\hbar} \right) \tag{5.57}$$

$$= n_{occ}|e| \underbrace{\overbrace{\left(\frac{|e|\tau}{m_{eff}} \right)}^{\text{mobility}} \boldsymbol{E}}_{\text{drift velocity : } v_d} \tag{5.58}$$

$$= n_{occ} e \boldsymbol{v}_d \tag{5.59}$$

The commonly used expression. The drift velocity is typically a few cm to m/s. In a copper wire electrons do not drift much faster than human beings can run.

Now consider a situation where only a few states are unoccupied (fig 5.4 (bottom)). In this case the expression for the current is more conveniently evaluated as

$$\boldsymbol{j} = -|e| \sum_{\substack{\text{occ} \\ \text{states}}} \boldsymbol{v}_g = |e| \sum_{\substack{\text{unocc} \\ \text{states}}} \boldsymbol{v}_g \tag{5.60}$$

using the condition 5.49. The sum over occupied states is not easy to do because it will span over a large extent of the band over which the same parabolic approximation will not work. But all the *unoccupied* states are close to the top extremum and a sum over these states can be done exactly the way we did the previous sum. Clearly the end result would be

$$\boldsymbol{j} = n_{unocc} e \boldsymbol{v}_d \tag{5.61}$$

However the effective mass must be evaluated for these states near the top of the band. The sum runs over unoccupied states and looks as if the current is caused by the unoccupied states (holes) with positive effective mass. Notice the crucial role played by condition 5.49. Without that condition we would not have got the simplification.

7. In a situation where one band has a few electrons and another band has a few "holes" we can get the total current by summing over the electrons for the upper band and the holes for the lower band. As long as we do not mix up the picture within the same band, the result will be correct. In a situation where there are n electrons and p holes, we would have

$$j = (n|e|\mu_e + p|e|\mu_p)E \tag{5.62}$$

the quantity in the bracket is called the conductivity. Convince yourself that the signs are all correct!

5.6 Hall effect

How do we know the number of electrons and (also the sign of the charge carrier) in a conducting solid? It is interesting to read the line of thinking that led to the "Hall effect" experiment (See Ashcroft/Mermin) but we will just give the summary here.

Consider a bar shaped metal/semiconductor (of known thickness) with crossed electric and mangetic fields as shown. As the electrons acquire a drift velocity due to the electric field, they must also feel the Lorenz force.

$$\mathbf{F} = -|e|\mathbf{v}_d \times \mathbf{B} \tag{5.63}$$

Since the current cannot flow out of the sides, we must assume that an electric field (E_y) arises in body of the sample that balances this force. This electric field (or the integral of this field, as a transverse voltage) is what we can measure. Since

$$\begin{aligned} V_H &= w.E_y = w.v_{dx}B \\ &= w.\frac{j_x}{ne}B \\ &= \frac{B}{ne}\frac{1}{t}I_x \end{aligned} \tag{5.64}$$

A remarkably simple result, with n as the only unknown because B, I_x, t and V_H are measured or set by the experimenter.

Doing it a little better...

We need to concentrate on the deviation in momentum from the equilibrium, it is implied now that we are talking about momentum in the sense of an average deviation from equilibrium

$$\Delta\mathbf{p}(t) = \langle \mathbf{p}(t) - \mathbf{p}_0 \rangle = m\mathbf{v}_d \tag{5.65}$$

We write out the transport equation with the magnetic field present as :

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -e\left(\mathbf{E} + \mathbf{v}_d \times \mathbf{B}\right) - \frac{\mathbf{p}}{\tau} \tag{5.66}$$

Let us assume that $\mathbf{E}$ is in the $x-y$ plane and $\mathbf{B}$ is along $\hat{\mathbf{z}}$, which is a very common configuration: Using the definition of current from eqn. 5.59 and writing out the components we get for the steady state

$$\begin{aligned} 0 &= -eE_x - \frac{eB}{m}p_y - \frac{p_x}{\tau} \\ 0 &= -eE_y + \frac{eB}{m}p_x - \frac{p_y}{\tau} \end{aligned} \tag{5.67}$$

Defining the cyclotron frequency

$$\omega_c = \frac{eB}{m} \tag{5.68}$$

and the zero field conductivity and resistivity as

$$\sigma_0 \equiv \frac{1}{\rho_0} = \frac{ne^2\tau}{m} \tag{5.69}$$

The set of eqns. 5.67 can be written as :

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \rho_0 \begin{pmatrix} 1 & \omega_c\tau \\ -\omega_c\tau & 1 \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} = \begin{pmatrix} \rho_0 & \frac{B}{ne} \\ -\frac{B}{ne} & \rho_0 \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} \tag{5.70}$$

PROBLEM : Eqn. 5.70 implies that $\mathbf{j}$ and $\mathbf{E}$ are not parallel to each other in presence of a magnetic field. The angle, δ, between these two vectors is called the "Hall angle". Show that

$$\tan\delta = \omega_c\tau \tag{5.71}$$

Notice that cleaner (large τ) the substance is, easier it is for the electrons to complete a rotation without suffering large collisions. The effect of the magnetic field will be significant if the charged particle can

complete one full rotation without getting scattered. A pure material will have a larger Hall angle than an impure one.

We need to remember that the quantity $\omega_c\tau = \frac{e\tau}{m}B = \mu B$. What is the typical value of this at "low" field? With $\tau \sim 10^{-14}$ sec and $B \sim 0.1$ Tesla we get $\mu B = \frac{1.6\times10^{-19}\ 0.1\ 10^{-14}}{9.1\times10^{-31}} \sim 10^{-2} - 10^{-3}$. This is important, since it means that quite often we can use $1 + \mu^2 B^2 \approx 1$ to pretty good accuracy at moderate fields. Indeed, the effective mass will differ from the free electron mass that we have put in here - but that will not drastically affect our order of magnitude estimate. This allows us to rewrite and invert the equation 5.70.

$$\begin{aligned}
\begin{pmatrix} E_x \\ E_y \end{pmatrix} &= \rho_0 \begin{pmatrix} 1 & \mu B \\ -\mu B & 1 \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} \\
\therefore \begin{pmatrix} j_x \\ j_y \end{pmatrix} &= \frac{\sigma_0}{1+\mu^2 B^2} \begin{pmatrix} 1 & -\mu B \\ \mu B & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \\
&\approx \sigma_0 \begin{pmatrix} 1 & -\mu B \\ \mu B & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix}
\end{aligned} \tag{5.72}$$

We discussed earlier that the sign of Hall voltage for the hole type conductivity would be opposite. This means that the off-diagonal elements of the matrix 5.70 will come with opposite sign.

Now it often happens that there are some electrons and some holes conducting simultaneously. It is possible that there are two electron bands or two hole bands also. We will see how to write the Hall effect equations to a two band case where there are n electrons and p holes. The generalisation to a many-band case is conceptually straightforward after that - though algebraically quite messy...

We write $\boldsymbol{j} = \boldsymbol{j}_n + \boldsymbol{j}_p$ and then add the two components. The electric and magnetic fields seen by the charge carriers must be the same hence (the notation should be self-explanatory)

$$\begin{aligned}
\begin{pmatrix} j_x \\ j_y \end{pmatrix} &= \left[\sigma_n \begin{pmatrix} 1 & -\mu_n B \\ \mu_n B & 1 \end{pmatrix} + \sigma_p \begin{pmatrix} 1 & \mu_p B \\ -\mu_p B & 1 \end{pmatrix}\right] \begin{pmatrix} E_x \\ E_y \end{pmatrix} \\
&= \begin{pmatrix} \sigma_n + \sigma_p & -\sigma_n\mu_n B + \sigma_p\mu_p B \\ \sigma_n\mu_n B - \sigma_p\mu_p B & \sigma_n + \sigma_p \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix}
\end{aligned} \tag{5.73}$$

Ultimately we need $\frac{E_y}{j_x}$ with j_y set to zero. To invert the last relation we require the determinant

$$\Delta = (\sigma_n + \sigma_p)^2 + (\sigma_n\mu_n - \sigma_p\mu_p)^2 B^2 \tag{5.74}$$

We have

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} \sigma_n + \sigma_p & \sigma_n\mu_n B - \sigma_p\mu_p B \\ -\sigma_n\mu_n B + \sigma_p\mu_p B & \sigma_n + \sigma_p \end{pmatrix} \begin{pmatrix} j_x \\ 0 \end{pmatrix} \tag{5.75}$$

Hence, The Hall resistivity

$$\begin{aligned}
\frac{E_y}{j_x} &= -\frac{(\sigma_n\mu_n - \sigma_p\mu_p)B}{\Delta} \\
&\approx -\frac{1}{e}\frac{n\mu_n^2 - p\mu_p^2}{(n\mu_n + p\mu_p)^2}B
\end{aligned} \tag{5.76}$$

- We have used $\sigma_n = ne\mu_n$, $\sigma_p = pe\mu_p$ and $\mu_n^2 B^2 \approx 0$, $\mu_p^2 B^2 \approx 0$. Some intermediate steps are left out.

- The method for generelisation to a multi-band case should be obvious after this.
- Notice that in a case where there are both electrons and holes, the sign of the Hall voltage would depend upon the relative mobilities of the two carriers.
- In almost all cases we get $\mu_n > \mu_p$, so with $n = p$, the Hall voltage will not be zero.

5.7 Mott transition

References:

1. The transition to the metallic state, N.F. Mott, *Philosophical Magazine*, **6**:62, 287-309
2. The transition to the metallic State, P.P. Edwards and M.J. Sienko, *Accounts of Chemical Research*, **15**, 87-93 (1982)

We consider the following thought experiment. Suppose we have a collection of atoms with one loosely bound electron (like *s* electrons) arranged on a cubic lattice (say) with a very large lattice constant. Now we start shrinking the lattice. One might argue that this can be achieved to some extent by applying pressure to some real material. However, we will make the connection with reality a little later. We want to know whether this material will conduct electric current at $T = 0$. This depends on whether it has free electrons in the conduction band in the limit of $T \to 0$.

Common sense would say that if these atoms are too far away from each other then the atoms effectively do not see each other and there is no way an electron can move from one atom to another. In a tight binding sense, the overlap integral would be zero. The question is at what separation (and how) does the system switch to being a band metal?

This question has very deep ramifications but let us first try to frame a strategy to solve it.

Band theory alone would not help. Tight binding would just suggest that the bandwidth would go to zero exponentially with distance. If the atoms are very far apart what is the difficulty in conduction? Well for an electron to move from one site to another there would be times when two of these sit on top of one another. This would cost a lot of Coulomb repulsion because the wavefunctions are localised on each site as long as the *s* electrons are bound to the atoms. But if the binding energy goes to zero then all electrons are delocalised and the question of large repulsion due to confinement in a small volume (one atomic site) does not arise. Recall that we gave a similar argument to rule out double occupancy for donor sites in a semiconductor.

The potential between a lattice site and the outer electron will be the Coulomb potential modified by the screening due to the lattice and the other free electrons already present in the conduction band. This may then be written in the Thomas-Fermi screening approximation as

$$V(r) = -\frac{1}{4\pi\epsilon_0}\frac{e^2}{\kappa r}e^{-q_{TF}r} \tag{5.77}$$

where q_{TF} is the Thomas Fermi wavevector that depends on how many free electrons are already there.

$$q_{TF}^2 = \frac{me^2}{\pi\epsilon_0\kappa\hbar^2}n^{1/3} \tag{5.78}$$

In 3D not every potential has a bound state. There needs to be a minimum "strength" of a potential - before it will develop a bound state. We know that in the limit $q \to 0$ the screened Coulomb (or Yukawa) potential has bound states - it just becomes the bare Coulomb. On the other hand if q_{TF} is very large the potential drops too fast to develop a bound state. It turns out that the condition for *no bound state* is :

$$q_{TF} > \frac{me^2}{4\pi\epsilon_0\kappa\hbar^2} \tag{5.79}$$

Notice that the combination on the rhs is precisely the inverse of effective (hydrogenic) Bohr radius (a_B). The relation is

$$a_B = \frac{4\pi\epsilon_0\kappa\hbar^2}{me^2} \tag{5.80}$$

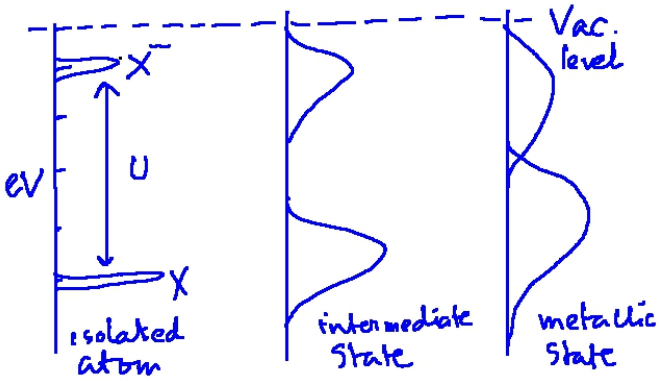


Figure 5.5: The Coulomb repulsion energy U when atoms are far away is too large for transport to take place. But the bandwidth increases as the interatomic separation decreases. There comes a point when the overlap integrals are large enough and transport is possible from site to site.

This allows us, to conclude using 5.79 and 5.78 that if

$$n^{1/3}a_B > 0.25 \tag{5.81}$$

then there is no bound state. Figure 5.5 shows another way of thinking of this. In the isolated atom the state with one extra electron lies very high in energy (by an amount U) for reasons given before. Now with this extra input we can calculate the spreading of both the levels, assuming some realistic wavefunctions of both the states (neutral and the one with an extra electron). As the atoms come closer at some point the band width may become sufficiently large so that the two bands overlap. The electron can then move seamlessly from from one site to the next. This is the metallic state.
Let us now see the connection of this with dopants in semiconductors. The dopants do not form a regular lattice, but their number can be controlled. For the moment let us forget about order with the following handwaving justification. The bandwidth depends (in a tight binding sense) on the co-ordination number and the nearest neighbour overlap integral. So we have some justification of ignoring what might be happening to the sites far away and just take the average density of sites.

The host semiconductor does the job of keeping the dopants on place and provide a background dielectric constant. The effective Bohr radius and effective mass that we use are the ones appropriate for the host semiconductor.

We can control the number of dopants that we put in. If they are all ionised then the electron density would just be same as the doping density. It is this number that enters eqn. 5.81.
An experimental example is shown in Fig. 5.6. In general metallic state implies two things:

1. The conductivity tends to a finite value as $T \to 0$

2. The conductivity increases with decreasing temperature

Microscopically this implies that the electrons are delocalised. Insulating state would mean just the opposite. An important point here is that ultimately temperature plays no role in the problem because we are talking about two possible states (metallic or insulating) in the limit of zero temperature. Thus this kind of phase transition is distinct from the other thermal phase transitions (like water freezing to ice) that we are familiar with.
How well does this hold for doped semiconductors? See the Figure 5.7. It agrees remarkably well, inspite of so many assumptions that we made. It tells us that the basic idea of screening by conduction electrons making the Coulomb repulsion cost going to zero, is a very robust one.

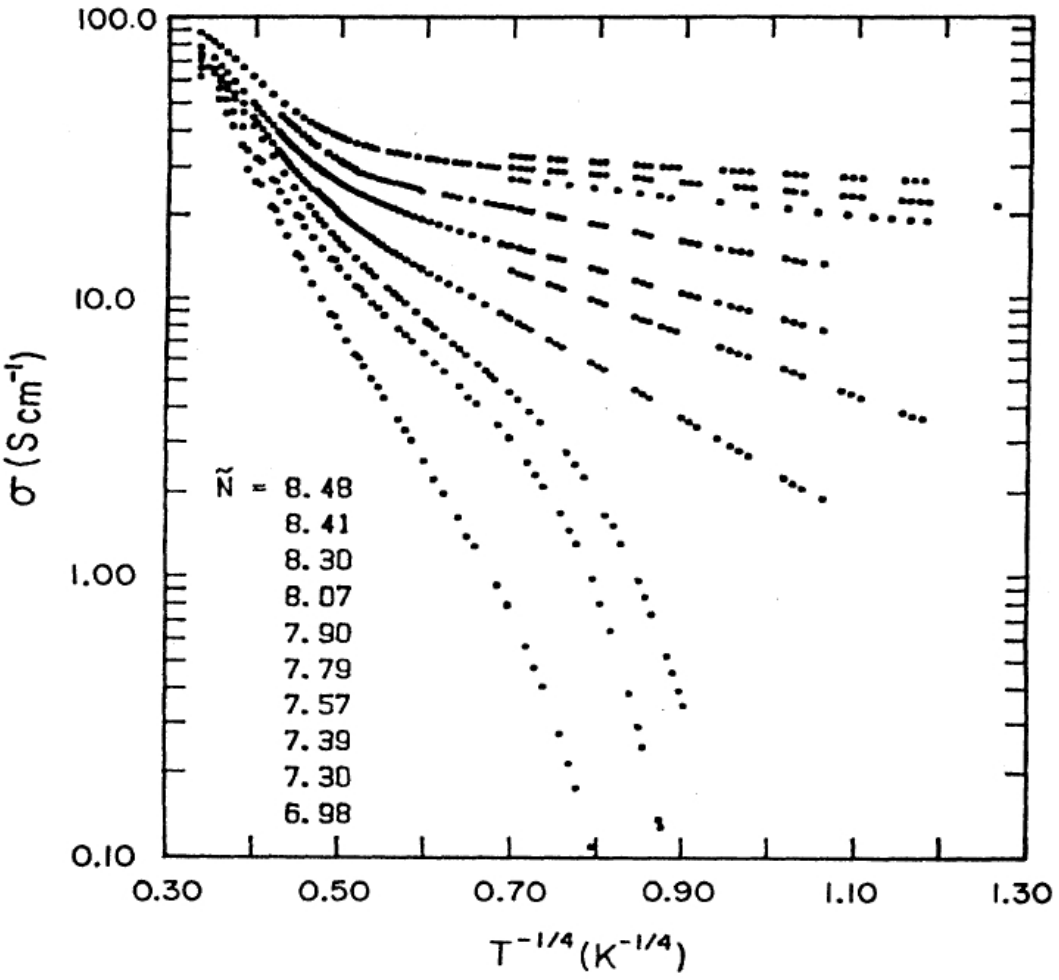


Figure 5.6: A plot of $\log \sigma(N,T)$ vs $T^{-1/4}$ for Arsenic donor doped Silicon. Notice how the behaviour changes at higher doping. At low T, $\sigma(T)$ tends to a finite value rather than dropping sharply to zero. The doping levels are in units of 10^{18}cm^{-3}. The data is taken from a paper, "dc conductivity of arsenic doped silicon near the metal insulator transition", by W.N. Shafarman, D.W. Koon and T.G. Castner, *Physical Review B*, **40**, 1216-1231 (1989). This data is on the insulating side, but at the highest doping levels the temperature dependence has almost started flattening off.

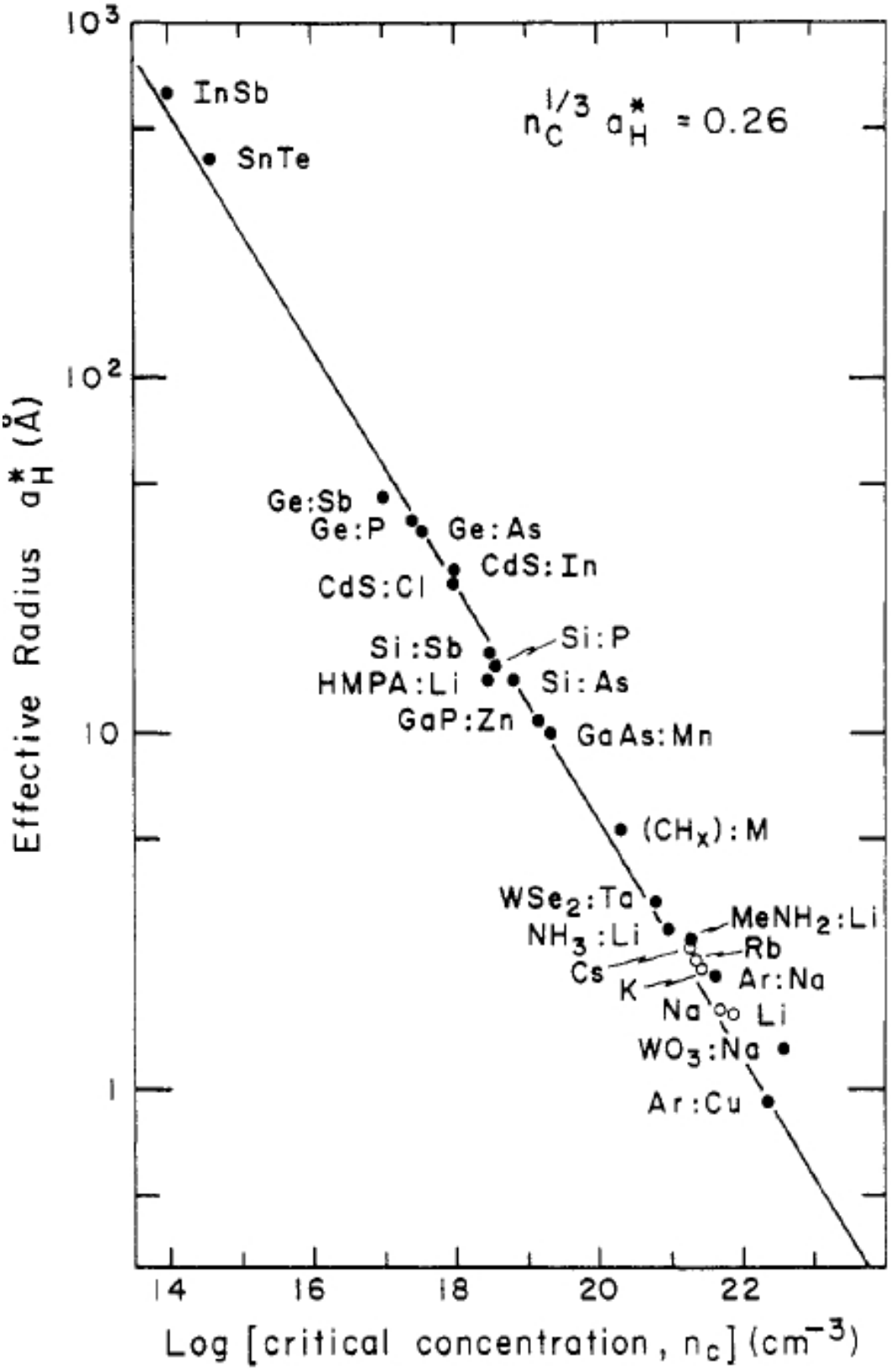


Figure 5.7: At what doping density does the transition occur? In this plot the straight line is corresponds to the Mott formula 5.81. The data points are experimental numbers at which the transition has been seen to occur. Notice the remarkable agreement for a number of material. The data is taken from a paper "The Transition to the Metallic State", P.P. Edwards and M.J. Sienko, *Accounts of Chemical Research*, **15**, 87-93 (1982).

These arguments were originally given by Neville Mott starting from late 1940s. There are several review articles that you can get, since a lot of interesting phenomena happens due to the competing effects of electron-electron interaction, disorder and dimensionality.

Chapter 6

Band-bending and junctions in semiconductors

References:

1. Chapter 5 (4^{th} edition), *Solid State Electronic Devices*, B. G. Streetman
2. Chapter 3 *Semiconductor Physics*, K Seeger
3. Greg Snider's homepage has the tool used to calculate band structures.
   See <www.nd.edu/∼gsnider>

6.1 Metal-semiconductor junctions

We now know the following important things :

1. How to calculate the charge density, if we know the location of E_f. Ignoring the holes and acceptors for the time being to keep the number of terms to a minimum, we have

$$n(x) - N_D^+(x) = N_C e^{\beta(E_f(x)-E_C(x))} - N_D \frac{1}{1+2e^{-\beta(E_D(x)-E_f(x))}} \tag{6.1}$$

$$\rho(x) = -|e|(n(x) - N_D^+(x)) \tag{6.2}$$

2. The charge density is related to the electrostatic potential (V) as

$$\nabla^2 V(x) = -\frac{\rho(x)}{\epsilon_r \epsilon_0} \tag{6.3}$$

3. The scalar potential is essentially the bottom of the conduction band.
4. In equilibrium E_f is constant, recall that current flow requires a gradient in the electrochemical potential or Fermi level.

6.1.1 Situations with no current flow

Now let's see how we can put this in practice - a (somewhat idealised) metal in contact with a semiconductor (see fig). The work function of a metal (ϕ_m in our discussion) is the energy an electron sitting at the Fermi level of the metal needs to escape from inside the metal to outside (vacuum level). [1] Similarly

[1] This can be typically about 4-5 eV, it depends a lot on which crystal face we are considering and how clean the surface is. Since we are going to ignore these aspects to highlight the basic concept, our discussion is a bit idealised here.

ϕ_s is the work function of the semiconductor in question. The two objects are brought in contact, so that they can exchange electrons. If $|\phi_s| < |\phi_m|$, then transferring an electron from the semiconductor to the metal is energetically favourable.

There is a little complication though - in a semiconductor the Fermi level is often in a gap - thus no electron may acctualy be right at the Fermi level. To account for this we define the electron affinity (χ) of the semiconductor as the energy difference between the vacuum level and the bottom of the conduction band of the semiconductor sufficiently deep inside.

When the two objects touch the conduction band and the Fermi level of the metal would be separated by $\phi_B = \phi_m - \chi$. Since deep inside the semiconductor (where the surface should have no effect) the bottom of the band and the Fermi level must continue to be separated by $\phi_s - \chi$, this dictates that the total drop of the conduction band of the semiconductor must be $\phi_m - \phi_s$

Charge separation must give rise to an extra electrostatic potential - and it is reasonable to expect that the bands would start bending in a way that would result in a barrier preventing the flow after some time. At this point the metal and the semiconductor's Fermi level must be identical. Applying the set of conditions that we just talked about produces the band diagram shown in Fig. 6.1.

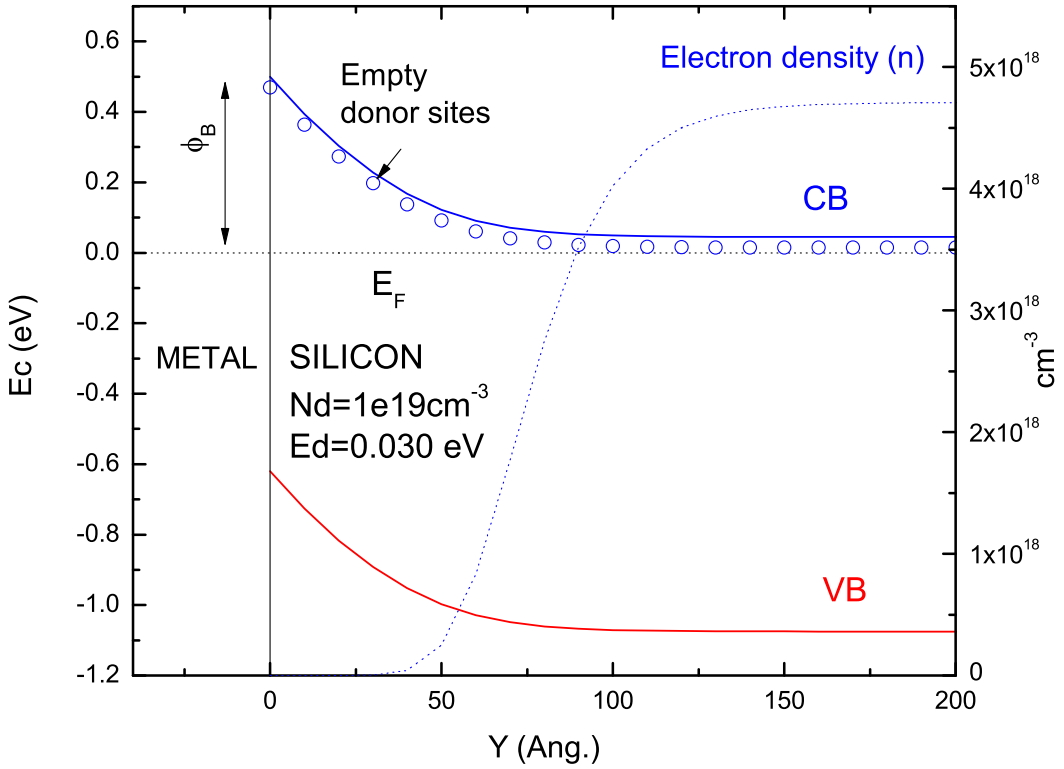


Figure 6.1: The band bending near the surface of a n-type semiconductor to metal contact, where $|\phi_s| < |\phi_m|$. The calculation has been done using a program written by Greg Snider (Notre Dam University)

Some charge has moved from the semiconductor to the metal. This charge came from the dopants sitting considerably above E_f. If a dopant site is pushed above E_f then it must be charged, because the electron cannot reside at a site sitting much above E_f. The bands in the metal didn't have to bend a lot to accomodate this extra charge, because the density of states of a metal near E_f is very large.

To (numerically) solve eqns 6.1, 6.2 and 6.3 we can proceed as follows.

1. Since E_f is constant, everything can be measured relative to E_f, by setting $E_f = 0$.

2. The gradient of the scalar potential is the same as the gradient of the conduction band bottom (E_C).

3. We make a guess for $E_C(x)$ and use this to calculate the expected charge density by using eqns 6.1 and 6.2.

4. This calculated charge density should give a new guess for the potential via Poisson's equation (eqn 6.3).

5. We use this potential and go back to step 3.

6. The iterative process can continue till the change in two successive iterations becomes very small (our convergence criteria)

7. But all these equations are differential equations - they need proper boundary conditions. In the calculation of Fig. 6.1, we set the slope $dE_C/dx = 0$ deep inside the material and $E_C = \phi_B$ at the other end. Choosing the correct boundary condition depends on the physical situation.

PROBLEM : Try to draw the band diagram of the metal-semiconductor contact when $|\phi_s| > |\phi_m|$ and the semiconductor is p-type. Where will the semiconductor accommodate the electrons flowing in?

Why doesn't the depletion zone extend to the metal as well?

We now apply the same process to a p-n junction, see Fig. 6.2.

6.1.2 How realistic are these calculations?

We remarked at the beginning of this section that there are some idealisations. The work function of a metal in reality depends on which crystal face we are using, how clean it is etc. This means that if we deposit a thin film of a metal (say gold on silicon) on a semiconductor, we can't really take the values for a crystal of gold and clean silicon and predict what the barrier will be. Also the density of states near the surface of a semiconductor is modified by the presence of surface states - which ultimately mean that the Schottky barrier needs to determined experimentally. However the band diagram of the barrier that we drew and the principles for solving the band-bending are sufficiently generic.

6.1.3 When is a contact not a "Schottky" ?

In the previous section we considered the work function of the metal to be larger $|\phi_s| < |\phi_m|$ and the semiconductor to be n-doped. As a consequence some electrons flowed from the semiconductor to the metal. What if $|\phi_s| > |\phi_m|$. Clearly the band bending must be different, because if electrons flow into the semiconductor then the dopant states and the conduction band cannot remain much above the Fermi-level. But the dopants (if they drop below E_f must hold on to their own electrons (they must be occupied). and the conduction band would have to accommodate the electrons. Under these conditions no depletion zone can form and hence there should be no Schottky barrier.

Real ohmic contacts

In reality ohmic contacts on a semiconductor are made by depositing an alloy that often contains one noble metal (Gold) and another element that can act as a dopant. For example an alloy of Gold-Germanium is commonly used to make ohmic contacts to n-type Gallium Arsenide. After depositing the metal the sample is generally annealed (heated to a high temperature) very rapidly so that the Germanium diffuses into the surface, heavily dopes the region around it allowing the Gold to make a contact with no barrier. The microscopic mechanisms of ohmic contact formation are not very simple and you would find a good deal of research work happening on these.

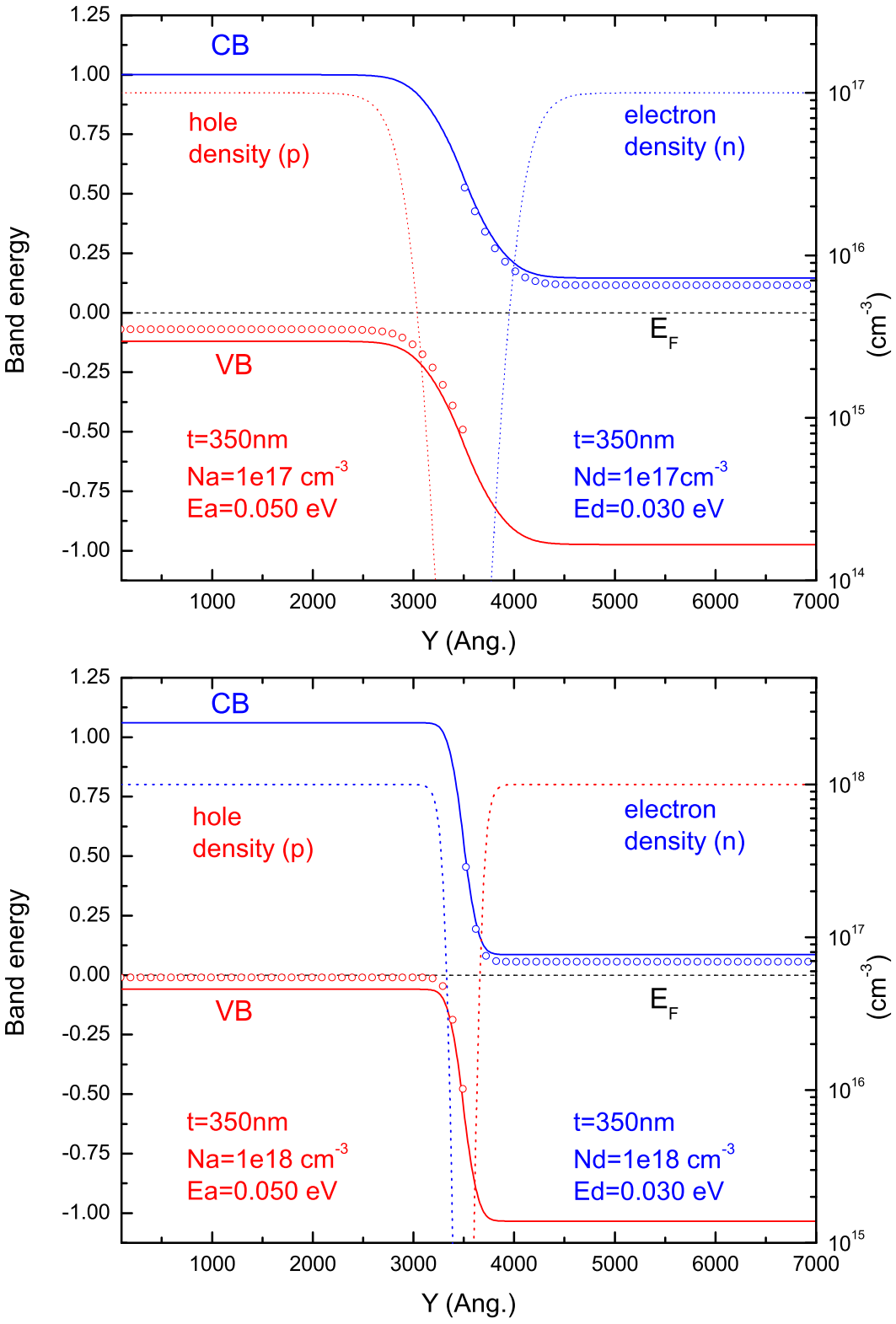


Figure 6.2: The band bending near a p-n junction. Note that the junction becomes sharper at a higher doping level. The Fermi energy also moves closer to the dopant level at higher doping concentration. The calculation has been done using a program written by Greg Snider (Notre Dam University)

Gold-Beryllium alloy can be used to contact p-type Gallium Arsenide.

Gold-Antimony alloy can be used to contact n-type Silicon...and so on.

6.1.4 Situations with varying E_f: what more is needed?

We noted earlier that the current is related to the gradient of the electrochemical potential (in a 1-dimensional case) as

$$j = -n(x)\mu\frac{\mathrm{d}E_f(x)}{\mathrm{d}x} \tag{6.4}$$

So we now have three variables to deal with - other than the charge density and the profile of the bottom of the band. We also need to calculate the the profile of $E_f(x)$.

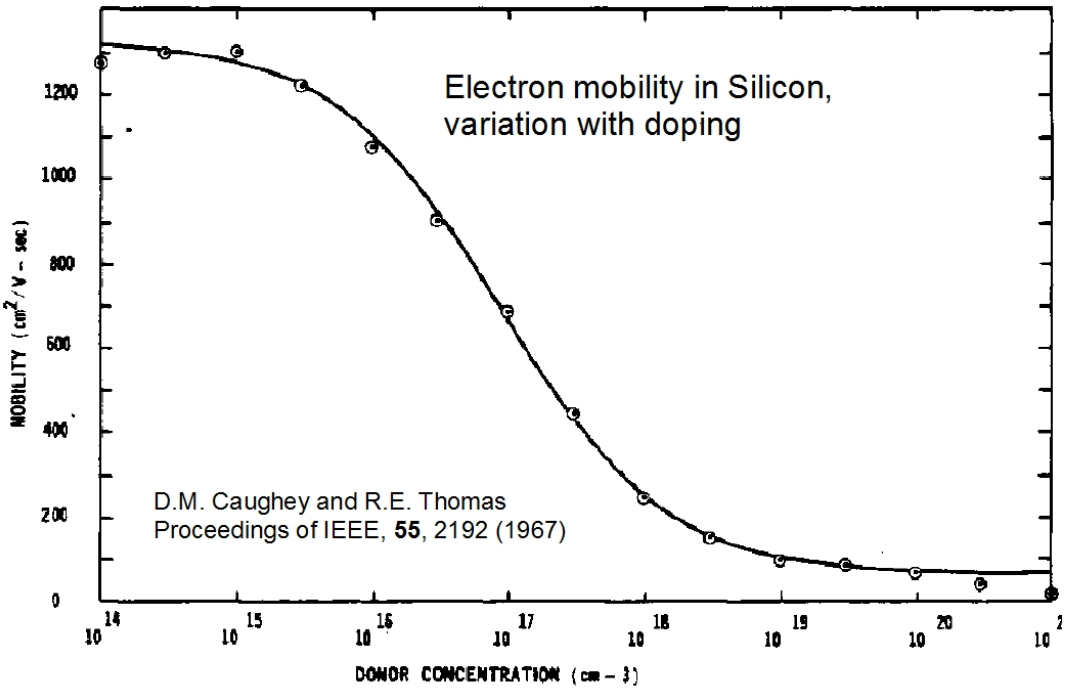
We expect μ to be a function of n.

In general the product $n\mu$ would increase with increasing carrier density, it doesn't necessarily imply that μ will be larger at higher densities. However at least in a 1-dimensional situation it is easy to see that wherever $n\mu$ is large, $\frac{\mathrm{d}\mu}{\mathrm{d}x}$ must be small. This reminds us of what to expect if we apply a voltage across a string of resistances (in series). The largest voltage drop must occur across the largest resistance, because the current through each of them is constant.

When the current flow is very small we can approximate the situation by saying that all the drop in E_f must be across the most resistive region (like a barrier) if we can identify one. This is however an approximation to get around the fact that the variation of mu with n is in general a hard and very system dependent problem. An empirical approach is shown in Fig 6.3.

Mobility model

Empirical relations and estimates can be used to get mobility as a function of carrier density from experimental data, here's an example



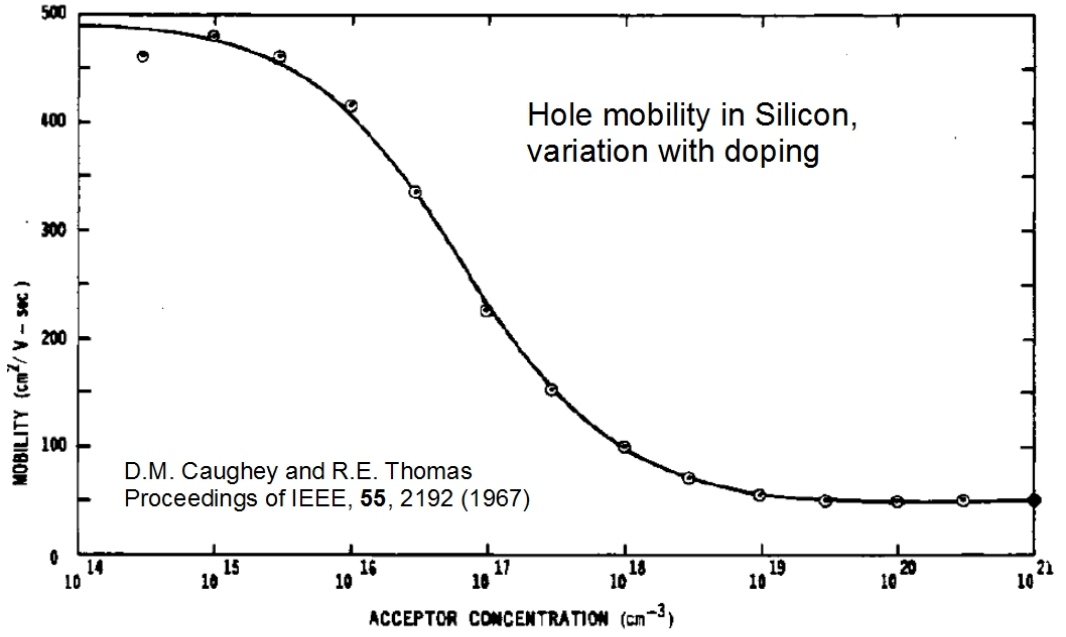


Figure 6.3: The variation of mobility with doping and carrier density can be empirically modelled from experimental data and used to solve the current equation numerically.

PROBLEM : Band bending at the p-n junction. The total drop in the profile of the bands shown in Fig. 6.2 can be calculated in two different ways. First let us see the method given in most text books. The flow of charge through the junction can be thought to have a drift (forced by the electric field) and a diffusion (forced by density gradient) component - at equilibrium, when the electrochemical potential in constant, these two components must add upto zero. So we get for no electron current

$$\begin{aligned} J_{drift} + J_{diffusion} &= 0 \\ -ne\mu\frac{dV}{dx} - De\frac{dn}{dx} &= 0 \end{aligned} \tag{6.5}$$

We have used the standard relation between current, diffusion constant and density gradient. (A full justification of this set of equations require the Boltzmann transport formulation, which we haven't done.) Then solve the differential equation using $D/\mu = kT/e$ and the assumption that all dopants are ionised. So that the electron density on the n-side is $n = N_D$ and on the p-side it is $n = n_i^2/N_A$. You should get the result for the total change in electrostatic potential as one moves from one side of the junction to the other. The electron bands are higher on the p-side.

$$\Delta V = \frac{kT}{e}\ln\frac{N_A N_D}{n_i^2} \tag{6.6}$$

Now think of the same in another way. Let us not mention diffusion constant at all, but use the fact that the electrochemical potential (E_f) is constant. Here the free energy of the electrons can be written, including the electrostatic potential as

$$F = -k_B T \ln\frac{z^n}{n!} + neV \tag{6.7}$$

where the electron density $n(x)$ is a function of position. And

$$z = 2\Omega\left(\frac{2\pi m k_B T}{h^2}\right)^{3/2} \tag{6.8}$$

is the partition function of a single free electron moving in the conduction band. Ω is the volume which should drop out of the calculation. Since we assume full ionisation we van neglect the entropy contribution coming from possible number of ways to distribute the bound electrons among the dopants.
Differentiating this w.r.t. n to get the electrochemical potential, first write

$$E_f(x) = \frac{\partial F}{\partial n(x)} \tag{6.9}$$

And then show that setting $E_f(x) =$ constant leads to exactly the same condition as before. Convince yourself that in both cases the approximations that we made are acctually identical. They are both consequences of Boltzmann statistics applied to the free electron gas in the conduction band.

The reverse and forward biased metal-semiconductor junction

In Fig. 6.1 we plotted the band diagram of a metal-semiconductor junction with no voltage applied (E_f = constant). No current flows at equilibrium. The tunnel rates from both sides balance each other. Now imagine that the electron energies on the metal side is raised by connecting the metal the negative terminal of a battery. The drop in the electrochemical potential must happen predominantly over the depletion region. See Fig. 6.4

- Electrons which try to cross over from the metal to the semiconductor still see almost the same barrier. The current that can pass through is $I_{m\to s} = AT^2 e^{-\phi_B/kT}$, where A is a constant.

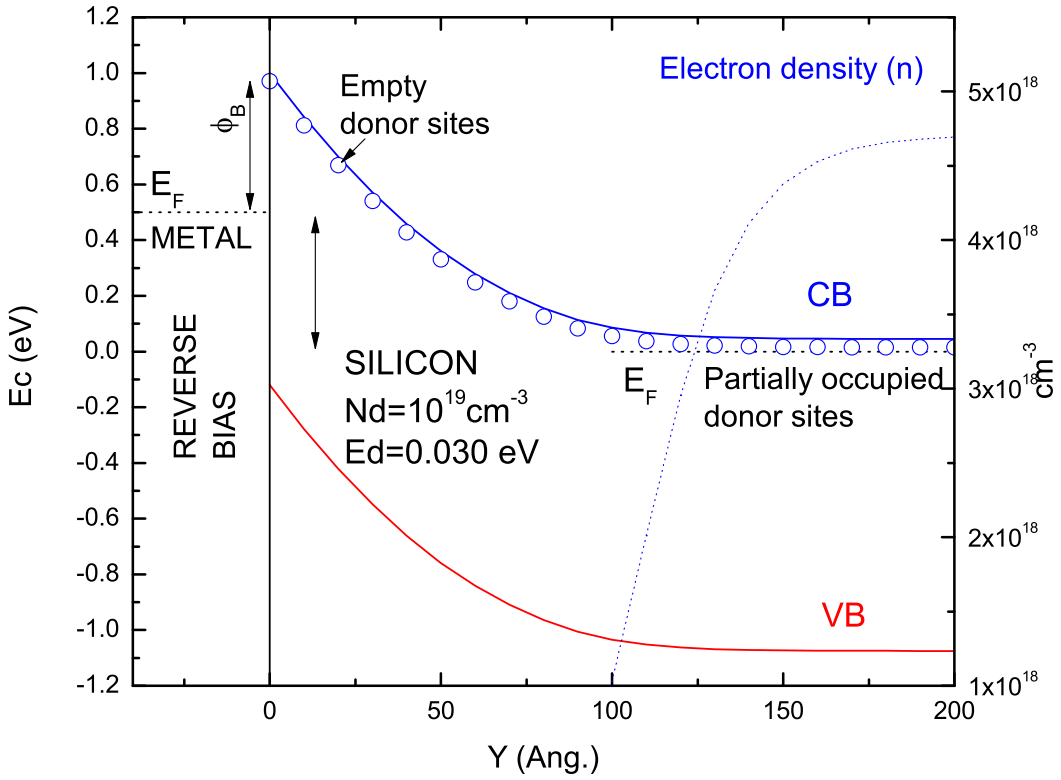


Figure 6.4: Approximate band bending near a reverse biased metal-semiconductor junction. The calculation has been done using a program written by Greg Snider (Notre Dam University)

- But those which try to cross from the semiconductor to the metal now see a higher barrier. Typically tunnelling probability through a barrier would drop exponentially with the height of the barrier. $I_{s\to m} = AT^2 e^{-(\phi_B+V)/kT}$, where V is the voltage bias on the semiconductor w.r.t. the metal. In this case $V > 0$. Remember that positive voltage bias lowers electron energies.
- Thus only the reverse saturation current now flows

What happens when the electron energies in the metal are lowered? See Fig. 6.5

- Electrons which try to cross over from the metal to the semiconductor still see almost the same barrier. $I_{m\to s}$ remains the same.
- But those which try to cross from the semiconductor to the metal now see a lower barrier. Typically tunnelling probability through a barrier would increase exponentially as the height of the barrier is lowered. $I_{s\to m} = AT^2 e^{-(\phi_B-V)/kT}$
- A large number of electrons can now flow from the semiconductor to the metal. We take the difference of the left going and the right going currents to get the total current which is the well known diode equation : $I = I_0(e^{eV/kT} - 1)$, where I_0, the reverse saturation current, is determined by the height of the Schottky barrier.

The Richardson formula : $T^2 e^{-W/kT}$

The calculation of how many electrons can make it through the barrier at a metal semiconductor interface is very similar to the way we calculate how many electrons a hot filament can emit. The electrons in a metal can be effectively pictured as being inside a box with the outside world (vacuum level) at a height W above the Fermi energy of the metal and U above the bottom of the conduction band of the metal. All energies are measured from the bottom of the conduction band, which is usual for the free electrons. E_F

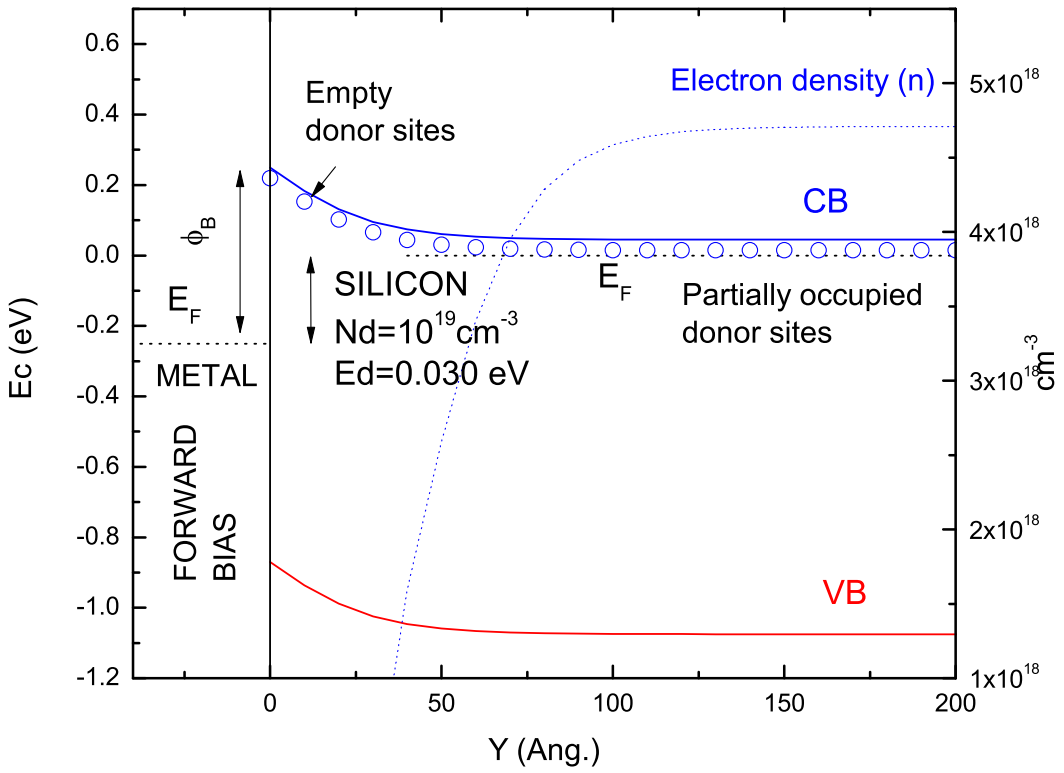


Figure 6.5: Approximate band bending near a forward biased metal-semiconductor junction. The calculation has been done using a program written by Greg Snider (Notre Dam University)

is Fermi level and hence $W = U - E_F$. This is the work function barrier which keeps the free electrons from jumping out of the "box". Now consider the situation shown in the figure.

- The electrons are in random motion and some of them will hit the boundary. Do they have enough energy to come out? Notice that in the drawing the potential barrier is infinitely thick. In such a case there is no quantum mechanical tunnelling. To be able to come out of the metal, the x component of the electron's velocity should satisfy

$$\frac{mv_x^2}{2} > U \tag{6.10}$$

- How many such electrons will hit an area A (normal to the boundary) in time Δt? This is a very common calculation in kinetic theory. The total charge coming out should be

$$Q = e\frac{2m^3}{h^3}\int_{\sqrt{2U/m}}^{\infty} dv_x \int_{-\infty}^{\infty} dv_y \int_{-\infty}^{\infty} dv_z (Av_x\Delta t) f(E) \tag{6.11}$$

The factor $\dfrac{2m^3}{h^3}$ comes from the density of states ($k = mv/\hbar$). In general $f(E)$ will be the Fermi distribution, but for high temperatures we can approximate it with the Boltzmann distribution and hence

$$f(E) \approx e^{-\beta\left(mv_x^2+mv_y^2+mv_z^2-E_F\right)} \tag{6.12}$$

The integral can be done explicitly and gives

$$J = \frac{Q}{A\Delta t} = \frac{4\pi em}{h^3}\left(kT^2\right)e^{-W/kT} \tag{6.13}$$

The same reasoning is valid for a metal semiconductor interface though the work function will be different, much less than the 4.5 eV or so, typical of a tungsten filament.

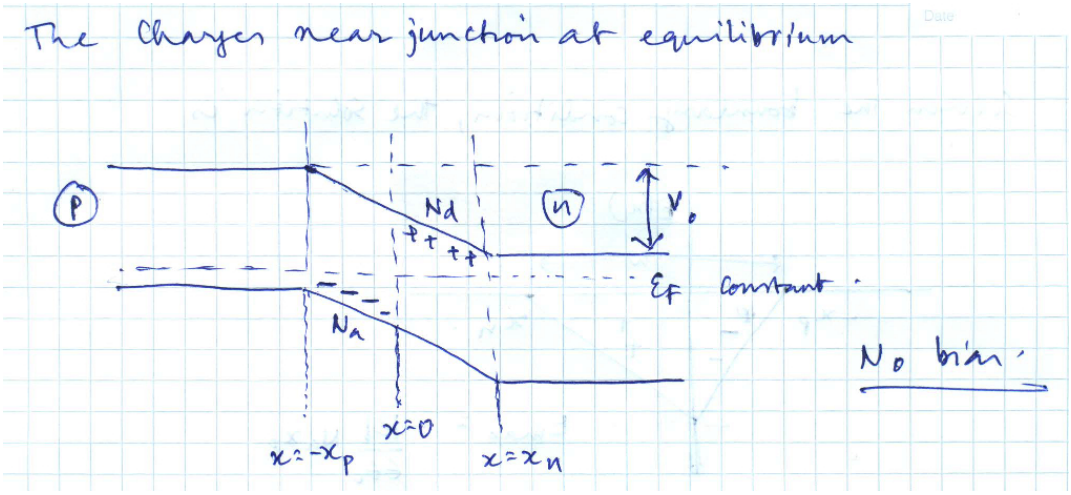


Figure 6.6: Linearised version of the *pn* junction.

6.2 The p-n junction in detail

To calculate the important quantities at a p-n junction analytically, we need to make some simplifying assumptions. The first thing we do is approximate the band structure near the junction with piecewise straight lines. We assume that the "junction" lies between the extents $-x_p < x < x_n$ with a sharp boundary at $x = 0$, where the p region ends and the n region begins. Such sharp junctions are *possible*. It is also possible not to have them so sharp. The variation of the doping is something that we can control to a good extent.

Consider a surface bounded by $x = -x_p$ and $x = x_n$. Since the bands become flat just outside these, the electric field becomes zero. So we must have

$$\epsilon_0 \int \boldsymbol{E}.\boldsymbol{dS} = -Q_{encl} = -\left(-|e|N_a x_p + |e|N_d x_n\right) \tag{6.14}$$

$$N_a x_p = N_d x_n \tag{6.15}$$

The electric fields are given by:

$$-x_p < x < 0 \quad \frac{dE}{dx} = -\frac{|e|N_a}{\epsilon_0 \epsilon_r} \quad E(-x_p) = 0 \tag{6.16}$$

$$0 < x < x_n \quad \frac{dE}{dx} = \frac{|e|N_d}{\epsilon_0 \epsilon_r} \quad E(x_n) = 0 \tag{6.17}$$

The solution must be two straight lines meeting at $x = 0$, since the electric field must have the same value there, see fig 6.7
The total change in the electrostatic potential in crossing the junction from $-x_p$ to x_n is V_0, say.

$$\int_{-x_p}^{x_n} E(x)dx = -V_0 \tag{6.18}$$

This must be equal in magnitude to the area of the triangle in figure 6.7.

$$\frac{1}{2}(x_p + x_n)\frac{|e|}{\epsilon_0 \epsilon_r} N_a x_p = V_0 \tag{6.19}$$

$$\frac{1}{2}(x_p + x_n)\frac{|e|}{\epsilon_0 \epsilon_r} N_d x_n = V_0 \tag{6.20}$$

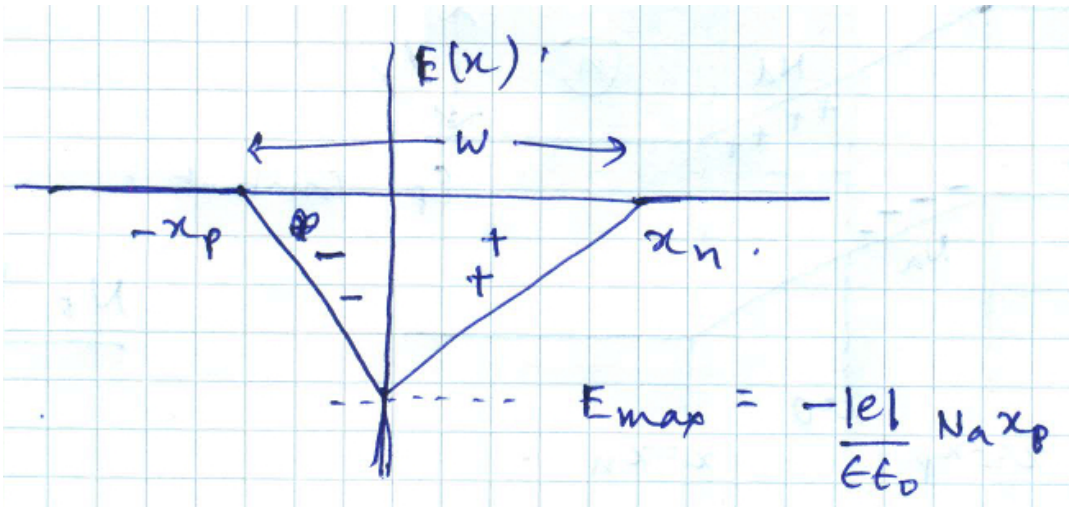


Figure 6.7: Electric field at the pn junction.

we call $w = x_p + x_n$ as the depletion width. The two previous equations can be solved for w. This gives

$$w^2 = \frac{\epsilon_0 \epsilon_r}{|e|} V_0 \left(\frac{1}{N_a} + \frac{1}{N_d} \right) \tag{6.21}$$

Since the change in the scalar potential becomes $V_0 - |V_f|$ when there is forward bias V_f and $V_0 + |V_r|$ when there is a reverse bias V_r, the expression can be generalised to

$$w = \sqrt{\frac{\epsilon_0 \epsilon_r}{|e|} \left(\frac{1}{N_a} + \frac{1}{N_d} \right) (V_0 - V_f)} \tag{6.22}$$

The width of the depletion region increases if there is a reverse bias, and decreases if there is a forward bias. We now need to determine V_0, so that can calculate x_p, x_n and V_0 all in terms of the doping levels alone.

6.2.1 Drift and Diffusion currents in equilibrium

In equilibrium, no current flows across the junction, but there is a strong variation of the electric field as well as the carrier densities. This means that the drift and the diffusion components will both exist, but they will cancel each other. Here n is the electron density at that point, μ_n is the electron mobility, D_n is the diffusion co-efficient. The electric current due to electron flow can be written as

$$J_n = n|e|\mu_n E + D_n |e| \frac{dn}{dx} \tag{6.23}$$

Notice the sign of the second term. If $dn/dx > 0$, it means that the concentration of electrons is increasing in the positive direction of x. Then the electron particle flow will be towards $-x$ due to diffusion. So the electric current is in $+x$ direction. If we had written the similar equation with holes it would have been:

$$J_p = p|e|\mu_p E - D_p |e| \frac{dp}{dx} \tag{6.24}$$

If the concentration of holes is increasing in the positive direction of x, then the hole particle flow will be towards $-x$ due to diffusion. So the electric current is in $-x$ direction.
Notice the difference in sign of the diffusion current. The electric current due to drift is always in the direction of the electric field. However if we had written the equation for particle currents, the sign of the drift term would have been different.
Let us take the electron equation and proceed:

$$\begin{aligned} J_n &= n|e|\mu E + D|e|\frac{dn}{dx} = 0 \\ \therefore n\mu\frac{dV}{dx} &= D\frac{dn}{dx} \\ \int_{-x_p}^{x_n} dV &= \frac{D}{\mu}\int_{-x_p}^{x_n}\frac{dn}{n} \end{aligned}$$

Since $n{\sim}N_d$ on the n side and $n{\sim}\frac{n_i^2}{N_a}$ on the p side, we can complete the integration, giving

$$V_0 = \frac{kT}{e}\ln\frac{N_aN_d}{n_i^2} \tag{6.25}$$

Here we have calculated the potential of the n-side w.r.t. the p-side, *i.e.* $V_0 = V(x_n) - V(x_p)$. The quantity is $+ve$. This means that the bands are lower on the n side, which is correct.
We could have of course calculated the same using $J_p = 0$ as well. The final result would have been the same.

It is important to remember the typical values of this voltage $V_0 \approx 0.7$volts for Silicon, the value changes relatively little becuase the doping concentrations occur inside a logarithm.

Similarly the barrier width or the depletion region is typically between 100 nm to $\sim$ 1 micron.

You cannot measure this electrostatic potential by placing a voltmeter across the junction, the voltmeter will measure zero although the electric field is about 1 volt/micron and hence about 1 million volts across 1 meter. A voltmeter will measure the electrochemical potential difference which is precisely zero. In cases where the conduction band and the electrochemical potential both have the same slope (typical situation in a metal wire, with no thermal gradient), the electric field and the slope of the electrochemical potential (Fermi level) are identical. In these cases the voltmeter will tell you the electrostatic potential difference, because it is identical to the electrochemical potential difference. Electrostatic potential differences across a junction, if they are to be measured, has to be inferred from current voltage characteristics or some optical transitions.

There are many other equivalent forms one can write the three equations. An useful result is an expression for the maximum electric field. You can prove the following easily:

$$E_{max} = -\frac{|e|}{\epsilon_0\epsilon_r}N_ax_p \tag{6.26}$$

$$= -\frac{|e|}{\epsilon_0\epsilon_r}\frac{N_aN_d}{N_a+N_d}w \tag{6.27}$$

$$|E| = \sqrt{\frac{|e|}{\epsilon_0\epsilon_r}\frac{N_aN_d}{N_a+N_d}(V_0 - V_f)} \tag{6.28}$$

6.2.2 Minority carrier injection

What happens when we create a slightly non-equilibrium situation by injecting a few holes in an n-type region? We will deal with a situation of "low level injection". This means that the hole concentration (in the n type region) has gone above its equilibrium concentration ($p \approx \frac{n_i^2}{N_d}$), but not so high as to approach anywhere near the electron concentration. So $p << n$. The n and p concentration can differ by orders of magnitude.

Ratio of drift and diffusion currents

The Einstein relation ($D/\mu = kT/e$) tells us the ratio of the two coefficients, but not the actual magnitude of the two components of the current. We need to know, what is $\frac{J_{diff}}{J_{drift}}$? Clearly, if the density is uniform, (bottom of CB and E_F are parallel), then there is no diffusion. However there are situations, particularly in semiconductors when this is certainly not true. We analyse the *magnitude of the ratio* as follows, using the expressions for the currents from eqns 6.23 and 6.24.

$$\begin{aligned} \frac{J_{diff}}{J_{drift}} &= \frac{D}{\mu}\frac{1}{E}\frac{1}{p}\frac{dp}{dx} \\ &= \frac{kT}{|e|}\frac{1}{E}\frac{d\ln p}{dx} \end{aligned} \tag{6.29}$$

Suppose a lot of holes are introduced into an n region. The continuity equation in this case requires the generation (G) and recombination (R) rates in place, since the holes will ultimately recombine if placed in an n-region. This rate is simply taken to be proportional to the deviation from the equilibrium density ($\Delta p = p - p_0$) with a relaxation time τ.

$$\frac{\partial J}{\partial x} + \frac{\partial p}{\partial t} = G - R \tag{6.30}$$

$$\text{with} \quad J = -D\frac{\partial p}{\partial x} \quad \text{and} \quad R = \frac{\Delta p}{\tau} \tag{6.31}$$

$$\text{steady state} \quad \frac{\partial p}{\partial t} = 0 \tag{6.32}$$

$$\therefore \frac{d^2\Delta p}{dx^2} = \frac{\Delta p}{D\tau} = \underbrace{\frac{\Delta p}{L_p{}^2}}_{\text{diffusion length}} \tag{6.33}$$

$$\therefore \Delta p = \Delta p(0)e^{-x/L_p} \tag{6.34}$$

Which is a simple exponential decay with a length scale L_p or L_n.

Notice that to simplify the equation, we ignored the drift component of the minority (hole) current in writing the second step. Was this consistent? For regions outside the depletion zone E is quite small it may be $\sim 10V/m$, the experimentally measured hole diffusion lengths are several microns, a typical value may be L_p∼10micron. With $kT/e = 25$meV we get

$$\frac{J_{diff}}{J_{drift}} = \frac{kT}{e}\frac{1}{EL_p} \approx 250 \tag{6.35}$$

The very important conclusion we arrive at is *the minority current is mostly diffusive*. A more general analysis of this can be done, there is in fact a classic experiment which addresses the measurement of the speed and recombination rate of the minority carriers - due to Haynes and Schokley.

The calculation of the current through a pn junction requires the concept we just developed. It is important that typical values of the junction width w is much smaller than the minority diffusion length. $w << L_p, L_n$. So very little loss of the holes injected from the p side, or electrons injected from the n side, occurs within the junction itself. This understanding is *very* important!

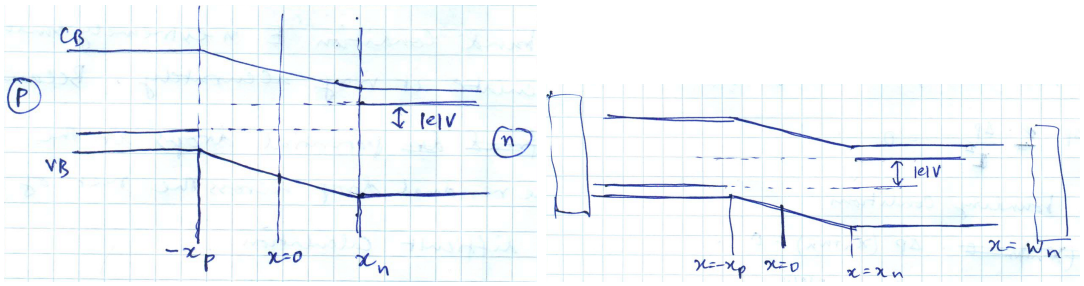


Figure 6.8: (Left) Linearised version of the *pn* junction, with a voltage bias. The bias shown here is forward bias, it means that the barrier is lowered. Notice that the −ve voltage applied to the n-side has raised the band, the +ve voltage on the p-side has lowered the bands. So the net effect is to reduce the barrier. The opposite of this will happen if the voltages are reversed. The n-side will go down and the p-side will go up, making the barrier height larger. (Right) The contacts are just areas where the carrier densities have relaxed back to their equilibrium values. Here we will assume that they are sufficiently far, meaning $w >> L_p, L_n$.

6.2.3 The current through a voltage biased pn junction

In equilibrium (zero bias) the drift and the diffusion terms cancel each other exactly. However if we calculate the magnitude of these terms at the junction, then it turns out that both are very large numbers individually. Consider the value near $x = 0$

$$\begin{aligned} E &\sim \frac{1V}{1\mu} \sim 10^6 \text{V.m}^{-1} \\ n &\sim 10^{17}\text{cm}^{-3} \\ \mu &\sim 10 - 1000\text{cm}^2\text{V}^{-1}\text{s}^{-1} \\ \therefore J_{drift} &\sim 10^4\text{A.cm}^{-2} \end{aligned} \tag{6.36}$$

In reality the current density through a diode are a lot less, it is 1 - 10^{-2}A.cm^{-2}. This means that two large numbers actually cancel out substantially, leaving a small number as the result. In such cases the calculation can very easily go wrong. On the other hand if we move deep into the n and p regions, the situation is almost like a metal (with low carrier density). Here $J_{diff} \sim 0$, because no major density gradient exists, but the electric field is also very small, since the bands are nearly flat. So here, unless we calculate the band slopes very accurately, we cannot get the correct electric field.

- To get the full current at any point we must add the electron and hole components of the current.
- The correct analytic way is to look at the edges of the depletion region, along with an assumption that very little recombination occurs within the depletion region. The correctness of this assumption comes from measured values of diffusion lengths and depletion widths.
- Holes from p-side, will appear on the n-side. But once they are on the n-side they are accounted for by diffusive component alone, because they are the minority carriers.
- The same argument can be given for electrons.
- Hence the total current is given by :

$$J_{total} = J_{n,diff}(-x_p) + J_{p,diff}(x_n) \tag{6.37}$$

  Notice the subscripts and where they are being evaluated very carefully. This is a key conceptual point.

- Although it is difficult to write the exact electrochemical potential at each point, we do know the difference of E_F between the two sides, this is what we measure or set. So the target would be to express everything in terms of the difference in E_F between the two sides.

- But this excess hole density must relax to its very low equilibrium value on the n doped side at $x = w_n (p \sim \frac{{n_i}^2}{N_d})$. This determines how the density must be relaxing. But if we know how the (excess) hole density is relaxing, it will allow us to calculate the diffusion current (due to hole concentration gradient) on the n doped side. If $w_n >> L_p$, then the calculation is trivial (exponential decay)

  The excess minority carrier density (for the holes on n side) can be defined as

$$\begin{aligned} \Delta p(x > x_n) &= p(x) - p(x = w_n) \\ \Delta n(x < -x_p) &= n(x) - n(x = -w_p) \end{aligned} \tag{6.38}$$

$$\Delta p(x) = \Delta p(x_n) e^{-\frac{x - x_n}{L_p}} \quad \text{for} \quad x > x_n \tag{6.39}$$

$$\Delta n(x) = \Delta n(-x_p) e^{\frac{x + x_p}{L_n}} \quad \text{for} \quad x < -x_p \tag{6.40}$$

- So now by using eqn 6.37 and the standard expression for diffusion current, we can evaluate the total current. This should be

$$\begin{aligned} J_{total} &= J_{n,diff}(-x_p) + J_{p,diff}(x_n) \\ &= |e| \frac{D_p}{L_p} \Delta p(x_n) + |e| \frac{D_n}{L_n} \Delta n(-x_p) \end{aligned} \tag{6.41}$$

  Even if we do not make the assumption $w >> L_p, L_n$, it can be done analytically, it will be a combination of two exponentials that will satisfy the correct boundary conditions. You can prove (as an exercise) that this will be

$$\Delta p(x) = \Delta p(x_n) \frac{\sinh\left[\frac{w_n - (x - x_n)}{L_p}\right]}{\sinh \frac{w_n}{L_p}} \quad \text{for} \quad x > x_n \tag{6.42}$$

$$\Delta n(x) = \Delta n(x_p) \frac{\sinh\left[\frac{w_p - (x + x_p)}{L_n}\right]}{\sinh \frac{w_p}{L_n}} \quad \text{for} \quad x < -x_p \tag{6.43}$$

- But the key point is to express the injected (excess) carrier density in terms of the voltage bias.

- One might think that the density of the holes throughout the p side will be constant because they are the majority carriers and in that case will be simply equal to N_a, assuming full ionisation of the dopants. While this is indeed correct near the contacts, remember that we do not know exactly how the electrochemical potential drops across the device. Almost the entire drop does take place across the junction, but a small part (hard to calculate) will change across the bulk of the p or n side. If it didn't then there would be no current at all, since the current after all depends on the gradient of the electrochemical potential. One might also argue (from observation) that if the number of carriers injected was independent of the potential difference applied from outside then there would be no voltage dependence of the observed characteristics at all.

- The density of holes on the p side (at $x = -x_p$) and of holes on the n side (at $x = x_n$)in equilibrium was related to the barrier voltage, following the method given earlier as

$$\text{zero bias} \qquad \frac{p(x_n, 0)}{p(-x_p, 0)} = e^{-|e|V_0/kT} \tag{6.44}$$

Since the barrier is lowered by an amount very close the forward bias V, the equation becomes

$$\text{with bias} \qquad \frac{p(x_n, V)}{p(-x_p, V)} = e^{-|e|(V_0 - V)/kT} \tag{6.45}$$

Now, with the assumption that the forward bias has very little effect on the majority carrier concentration, implying $p(-x_p, 0) \approx p(-x_p, V)$

$$\begin{aligned} \Delta p(x_n) &= p(-x_p, V) e^{-|e|V_0/kT} . e^{|e|V/kT} - p(-x_p, 0) . e^{-|e|V_0/kT} && (6.46) \\ &\approx p(-x_p, 0) . e^{-|e|V_0/kT} \left(e^{|e|V/kT} - 1 \right) \\ &= \frac{{n_i}^2}{N_d} \left(e^{|e|V/kT} - 1 \right) && (6.47) \end{aligned}$$

- With the "forward bias" the levels at which holes may exist on the n side comes slightly closer to the electrochemical potential (of the holes) on the p side and the injected carrier density increases a little bit. The form of the expression allows an easy generalisation to a situation where the barrier becomes larger (the reverse bias), by replacing V with $-V$

- To complete the expression for the total current, following eqn 6.41, we need an equivalent expression for $\Delta n(-x_p)$, we can write this as

$$\Delta n(-x_p) = \frac{{n_i}^2}{N_a} \left(e^{|e|V/kT} - 1 \right) \tag{6.48}$$

- Using the last two expressions for the excess carriers we complete the calculation as

$$\begin{aligned} J_{total} &= J_{n,diff}(-x_p) + J_{p,diff}(x_n) \\ &= |e| \frac{D_p}{L_p} \Delta p(x_n) + |e| \frac{D_n}{L_n} \Delta n(-x_p) \\ &= \underbrace{|e| {n_i}^2 \left[\frac{D_p}{L_p N_d} + \frac{D_n}{L_n N_a} \right]}_{\text{reverse saturation current}} \left(e^{|e|V/kT} - 1 \right) && (6.49) \\ &= J_0 \left(e^{|e|V/kT} - 1 \right) && (6.50) \end{aligned}$$

The diode equation is surprisingly similar in form to the one derived for a metal-semiconductor junction.

- There are certain modifications which can be done to this expression, like introducing recombination in the barrier region etc, but the basic form of this equation remains intact.

Chapter 7

Band Structure II : The $k.p$ method, spin orbit interaction effects and psuedopotentials

7.1 Using the $k.p$ method

Now let's put together the last two things we did:

1. The tight binding technique tells us that the $\boldsymbol{k} = 0$ wavefunction should closely resemble the atomic wavefunction - which in general are quite well studied.
2. At the same time we saw that the $\boldsymbol{k}.\boldsymbol{p}$ matrix elements involve the same $\boldsymbol{k} = 0$ states, assuming for the moment that there is an extremum at $\boldsymbol{k} = 0$.
3. We should thus be able to calculate the $\boldsymbol{k}.\boldsymbol{p}$ matrix elements using the atomic state wavefucntions.

7.1.1 2×2 $k.p$

Now suppose we have just two (atomic) states $E_1 = 0$ and $E_2 = E_g$ and all the other band bottoms/tops are far away. So it is reasonable that in the vicinity of these two states we should be able to write an arbitrary state as a sum over just two ortho-normal basis states

$$u_n(\boldsymbol{k}) = \sum_m c_m(\boldsymbol{k}) u_m(0) \tag{7.1}$$

The coefficients $c_m(\boldsymbol{k})$ are unknown but the problem can be solved easily. Since

$$\left[\underbrace{\frac{\boldsymbol{p}^2}{2m} + V(\boldsymbol{r})}_{H_0} + \underbrace{\frac{\hbar}{m}\boldsymbol{k}.\boldsymbol{p} + \frac{\hbar^2 k^2}{2m}}_{H_{\boldsymbol{k}}} \right] \sum_m c_m(\boldsymbol{k}) u_m(0) = E(\boldsymbol{k}) \sum_m c_m(\boldsymbol{k}) u_m(0) \tag{7.2}$$

We left multiply the equation with each $u_n(0)$ and form the matrix element. The resulting set of equation gives

$$\begin{vmatrix} E_1 + \dfrac{\hbar^2 k^2}{2m} - E & \dfrac{\hbar}{m}\langle u_1(0)|\boldsymbol{k}.\boldsymbol{p}|u_2(0)\rangle \\ \dfrac{\hbar}{m}\langle u_2(0)|\boldsymbol{k}.\boldsymbol{p}|u_1(0)\rangle & E_2 + \dfrac{\hbar^2 k^2}{2m} - E \end{vmatrix} = 0 \tag{7.3}$$

Now if we call the lower and upper levels valence (v) and conduction(c) bands, and set $E_v = 0$ & $E_c = E_g$ then the solution can be written as

$$E(\boldsymbol{k}) = \frac{1}{2}\left(E_g + \frac{\hbar^2 k^2}{2m}\right) \pm \frac{1}{2}\left(E_g^2 + \frac{4\hbar^2}{m}|\langle u_c|\boldsymbol{k}.\boldsymbol{p}|u_v\rangle|^2\right)^{1/2} \tag{7.4}$$

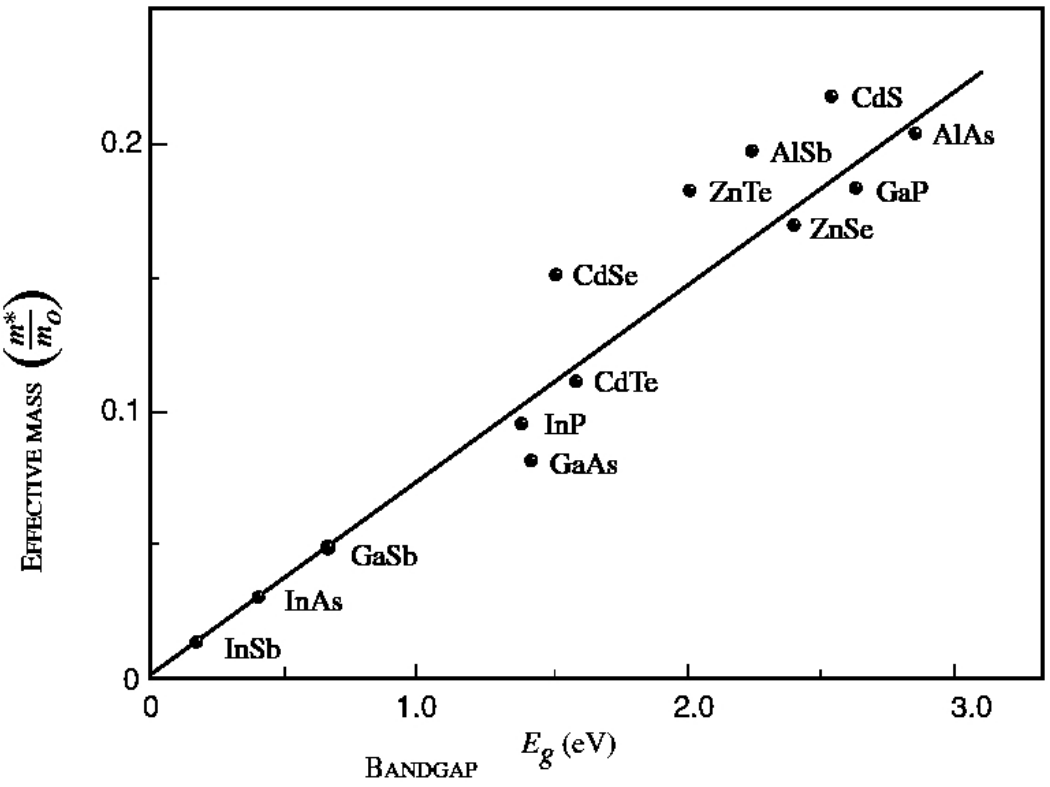


Figure 7.1: Effective mass of the electrons vs bandgap in III-V semiconductors. Compare with the 2×2 $\boldsymbol{k}.\boldsymbol{p}$ prediction. The figure is taken from J. Singh's book *Electronic and Optoelectronic properties of semiconductors*

Expanding for small $\boldsymbol{k}$ we get

$$E_c(\boldsymbol{k}) = \underbrace{E_g}_{\text{gap}} + \frac{\hbar^2}{2m}k^2 + \frac{\hbar^2}{E_g m^2}|\boldsymbol{k}.\boldsymbol{p}_{cv}|^2 \tag{7.5}$$

$$E_v(\boldsymbol{k}) = \underbrace{E_v}_{=0} + \frac{\hbar^2}{2m}k^2 - \frac{\hbar^2}{E_g m^2}|\boldsymbol{k}.\boldsymbol{p}_{cv}|^2 \tag{7.6}$$

where $\boldsymbol{k}.\boldsymbol{p}_{cv} = \langle u_2(0)|\boldsymbol{k}.\boldsymbol{p}|u_1(0)\rangle$ is the off-diagonal element in the matrix 7.3.

- The qualitative prediction about two parabolic bands and large band gap implying large effective electron mass is correct.
- Notice that the valence band effective mass will be larger than the conduction band effective mass. In general this is indeed correct. The "holes" are heavier than the electrons.

7.1.2 4×4, 6×6 and 8×8 $\boldsymbol{k}.\boldsymbol{p}$

It is now natural to ask if we could include more basis states in the framework. Some of the common semiconductors have electronic structures

C	$1s^2 2s^2 2p^2$
Si	$1s^2 2s^2 2p^6 3s^2 3p^2$
Ge	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^2$

and for Gallium Arsenide

Ga $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^1$
As $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^3$

Therefore the outermost electrons (s and p) which form the topmost bands would have atomic wavefunctions, whose spatial parts are (radial part $\times$ spherical harmonic, normalised)

$$\psi_{nlm}(\boldsymbol{r}) = R_{nl}(r) Y_{lm}(\theta,\phi) \tag{7.7}$$

$$Y_{lm}(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}(-1)^{(m+|m|)P_l^{|m|}}(\cos\theta)e^{im\phi} \tag{7.8}$$

$$\int_0^{\pi} d\theta \int_0^{2\pi} d\phi \, |Y_l^m(\theta,\phi)| \sin\theta = 1 \tag{7.9}$$

For the s,p states $l = 0, 1$ and $m = -l, -l+1, ...0, ...l$. The relevant harmonics can be written as

- $l = 0$ for s orbital. So

$$Y_{00} = \frac{1}{\sqrt{4\pi}} \equiv |S\rangle \tag{7.10}$$

- $l = 1$ for p orbitals. So

$$Y_{10}(\theta,\phi) = \sqrt{\frac{3}{4\pi}}\cos\theta = \sqrt{\frac{3}{4\pi}}\frac{z}{r} \equiv |Z\rangle \tag{7.11}$$

$$Y_{1\pm1}(\theta,\phi) = \mp\sqrt{\frac{3}{8\pi}}\sin\theta e^{\pm im\phi} = \mp\sqrt{\frac{3}{8\pi}}\frac{x \pm iy}{r} \equiv \frac{1}{\sqrt{2}}|X \pm iY\rangle \tag{7.12}$$

- We then build eight states out of these as shown including spin, following a standard convention

$$|u_1\rangle = |iS\downarrow\rangle \tag{7.13}$$

$$|u_2\rangle = \left|\frac{X-iY}{\sqrt{2}}\uparrow\right\rangle \tag{7.14}$$

$$|u_3\rangle = |Z\downarrow\rangle \tag{7.15}$$

$$|u_4\rangle = \left|-\frac{X+iY}{\sqrt{2}}\uparrow\right\rangle \tag{7.16}$$

$$|u_5\rangle = |iS\uparrow\rangle \tag{7.17}$$

$$|u_6\rangle = \left|-\frac{X+iY}{\sqrt{2}}\downarrow\right\rangle \tag{7.18}$$

$$|u_7\rangle = |Z\uparrow\rangle \tag{7.19}$$

$$|u_8\rangle = \left|\frac{X-iY}{\sqrt{2}}\downarrow\right\rangle \tag{7.20}$$

- We would then write the state $u_n(\boldsymbol{k})$ as a linear combination of all these 8 states. This is the 8×8 $\boldsymbol{k}.\boldsymbol{p}$ method.

- Notice that the matrix elements will all be matrix elements of the momentum operator (*including spin-orbit effects*)between atomic states. These would be standard integrals involving spherical harmonics. Thus the band structure problem near the extrema is expressed completely in terms of atomic wavefunctions and matrix elements.

- The $\boldsymbol{k}.\boldsymbol{p}$ formulation gives the following generic picture of the hole-bands. See figure 7.2 We will discuss the spin-orbit interaction next.

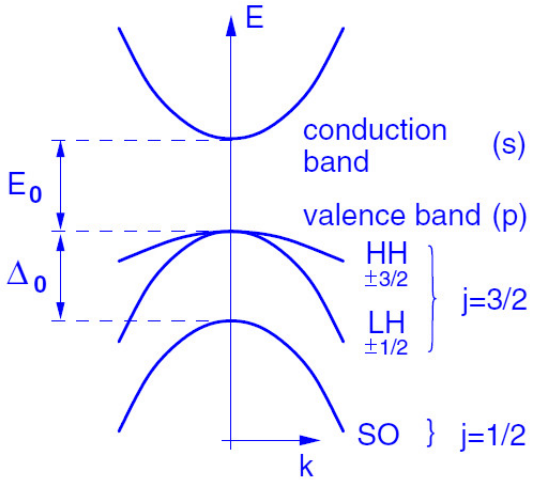


Figure 7.2: The generic structure of the valence and conduction bands near $\boldsymbol{k} = 0$. The figure is taken from Roland Winkler's book *Spin Orbit Coupling effects in 2 dimensional systems*

7.1.3 Spin orbit coupling

The p states of the atoms are not degenerate in reality. The energy difference between these states is caused by the following:

- From the electron's rest frame, the positive charge center is rotating around it.
- The rotating positive charge is equivalent to a current that creates a magnetic field.
- The electron's spin magnetic moment ($\mu_B = \dfrac{e\hbar}{2m}$) senses this field. Hence the up and down states are not equal in energy even in absence of an external magnetic field.
- If one tries to formulate this idea mathematically, the resulting interaction term from this "semi-classical" argument is:

$$V_{LS} = \frac{1}{2}\frac{\mu_0 Z e^2}{4\pi^2 m^2 r^3}\boldsymbol{L}.\boldsymbol{S} \tag{7.21}$$

  Here Ze is the charge in the nucleus and r is the radius of the electron's orbit and m is rest mass of the electron.

- We then use the vector model of the atom to calculate the $\boldsymbol{L}.\boldsymbol{S}$ product in terms of J, L, S of the state.
  - Consider the p states relevant to the semiconductors. These states have $l = 1$ and $s = 1/2$. The j value can go from $|l - s|$ to $|l + s|$. Hence the allowed values are $j = \dfrac{1}{2}, \dfrac{3}{2}$
  - Since $\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S}$, we have $J^2 = L^2 + S^2 + 2\boldsymbol{L}.\boldsymbol{S}$. Hence

$$\boldsymbol{L}.\boldsymbol{S} = \frac{j(j+1) - l(l+1) - s(s+1)}{2} \tag{7.22}$$

  - Write the atomic states in their $|j, m\rangle$ representation and find the value of $\boldsymbol{L}.\boldsymbol{S}$ in each state.
  - For $\left|\dfrac{1}{2}, \pm\dfrac{1}{2}\right\rangle$, $\boldsymbol{L}.\boldsymbol{S} = -2$
  - for $\left|\dfrac{3}{2}, \pm\dfrac{1}{2}\right\rangle$ and $\left|\dfrac{3}{2}, \pm\dfrac{3}{2}\right\rangle$: $\boldsymbol{L}.\boldsymbol{S} = 1$

  - These two sets would then be separated by the spin-orbit interaction at the atomic level and hence the zone center. The $\left|\frac{1}{2},\pm\frac{1}{2}\right\rangle$ states lie lower by an amount usually called Δ_{so}. For Gallium Arsenide $\Delta_{so} = 0.34$eV. It is not an insignificant number and needs to be taken into account.

- *The factor* $1/2$ *in the equation is put in "by hand"* - so that the result agrees with the correct relativistic calculation done by using the Dirac equation. Historically this is called the Thomas factor.

- The resulting effects on atomic spectra are well known.

- However this derivation essentially requires a bound state where the electron "goes round" a proton. In case of an atom that may work fine, but when we talk about nearly free electrons in a band, the correctness of this picture needs to be questioned. In situations where angular momentum is not a good quantum number the formula isn't useful.

- We would see (from the Dirac equation) that the general (*so=spin-orbit*) term in the Hamiltonian which accounts for the interaction of the electrons spin with its momentum and electric field is given by

$$H_{so} = \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma}.\nabla V \times \boldsymbol{p} \tag{7.23}$$

  where V is the electrostatic potential energy. This can arise from the lattice potential as well as external sources like other charges and gate voltages. We need to include this term as a perturbation in our $\boldsymbol{k}.\boldsymbol{p}$ Hamiltonian.

- By putting in the expression for Coulomb potential in eqn.7.23 you should be able to show the semiclassical expression emerging out of it. Notice that the eqn. 7.23 makes no reference to angular momentum.

- So, finally we have the full expression for the perturbing part of the $\boldsymbol{k}.\boldsymbol{p}$ Hamiltonian. Summarising the steps starting from the full wavefunction

$$\psi_{n,\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}.\boldsymbol{r}}u_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{7.24}$$

  satisfies

$$\left[\frac{\boldsymbol{p}^2}{2m} + V(\boldsymbol{r}) + \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma}.\nabla V \times \boldsymbol{p}\right]\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = E_{n,\boldsymbol{k}}\psi_{n,\boldsymbol{k}}(\boldsymbol{r}) \tag{7.25}$$

  Then $u_{n,\boldsymbol{k}}(\boldsymbol{r})$ satisfies

$$\left[\underbrace{\frac{\boldsymbol{p}^2}{2m} + V(\boldsymbol{r}) + \frac{\hbar}{4m^2c^2}\boldsymbol{\sigma}.\nabla V \times \boldsymbol{p}}_{H_0} \;+\; \underbrace{\frac{\hbar}{m}\boldsymbol{k}.p + \frac{\hbar^2}{4m^2c^2}\boldsymbol{\sigma}.\nabla V \times \boldsymbol{k} + \frac{\hbar^2k^2}{2m}}_{H_{\boldsymbol{k}}}\right]u_{n\boldsymbol{k}}(\boldsymbol{r}) = E_{n\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r}) \tag{7.26}$$

  We then write

$$u_{n\boldsymbol{k}}(\boldsymbol{r}) = \sum_m c_m(\boldsymbol{k})u_{m\boldsymbol{0}} \tag{7.27}$$

  where the number of states included in the sum runs over the atomic s and p states relevant to the problem. The eigenvalues are then solved for successive $\boldsymbol{k}$ values and each branch gives us one band. Nothing prevents one from including more states if the effects of other "far away" bands need to be calculated. In this formulation all the matrix elements are matrix elements of the momentum operator between atomic states.

7.1.4 The effective g-factor

The effective mass concept can be carried forward further- we state the final result without proof. The bare g-factor of an electron can be replaced by an effective g-factor to include the effects of a magnetic field. Such that the equation **??** may be written, with γ denoting the state at $\boldsymbol{k} = 0$, and the magnetic field along z-direction, *i.e.*$(0, 0, B)$

$$\begin{aligned} E(\boldsymbol{k}) &= D_{\alpha\alpha}k_\alpha k_\alpha - \frac{e\hbar}{2m}\langle\gamma|L_z|\gamma\rangle\sigma_z B - \mu_B\sigma_z B \quad (7.28)\\ &= D_{\alpha\alpha}k_\alpha k_\alpha - \mu^*\boldsymbol{\sigma}.\boldsymbol{B} \quad (7.29)\end{aligned}$$

Utilising the linear dependence of a part of the dispersion on B, we can club it together with the g-factor, such that the effective moment (μ^*)and the Bohr-magneton are related by

$$\frac{\mu^*}{\mu_B} = 1 + \frac{1}{2m}\,\Im\left(\sum_{\gamma\neq\delta}\frac{\langle\gamma|p_x|\delta\rangle\langle\delta|p_y|\gamma\rangle}{\varepsilon_\gamma - \varepsilon_\delta}\right) \quad (7.30)$$

Here $\Im$ denotes the imaginary part of the expression within the brackets. Notice how the non-diagonal terms ultimately lead to the re-definition of the g-factor. Once again small band gap will lead to a large effect on the final result. Small band gap semiconductors (InSb) are known to have g factors as large as 50-70.

A simplified version of the full expression is often useful for III-V semiconductors.

$$g^* = 2 - \frac{2}{3}\frac{E_p\Delta}{E_g(E_g+\Delta)} \quad (7.31)$$

Here $E_p = \dfrac{|p_{cv}|^2}{2m}$ and Δ is the spin-orbit split at $\boldsymbol{k} = 0$ between the $j = \dfrac{3}{2}$ and $j = \dfrac{1}{2}$ states as shown in figure 7.2.

The relation between the μ^* and g^* is as follows. The magnetic dipole moment is related to the spin of the particle as

$$\boldsymbol{\mu} = g\left(\frac{e}{2m}\right)\boldsymbol{S} \;=\; \frac{g\mu_B}{\hbar}\boldsymbol{S} \quad (7.32)$$

The spin operator for spin-$\frac{1}{2}$ particles is $\boldsymbol{S} = \dfrac{\hbar}{2}\boldsymbol{\sigma}$ where $\boldsymbol{\sigma}$ is a vector formed of the Pauli spin matrices.

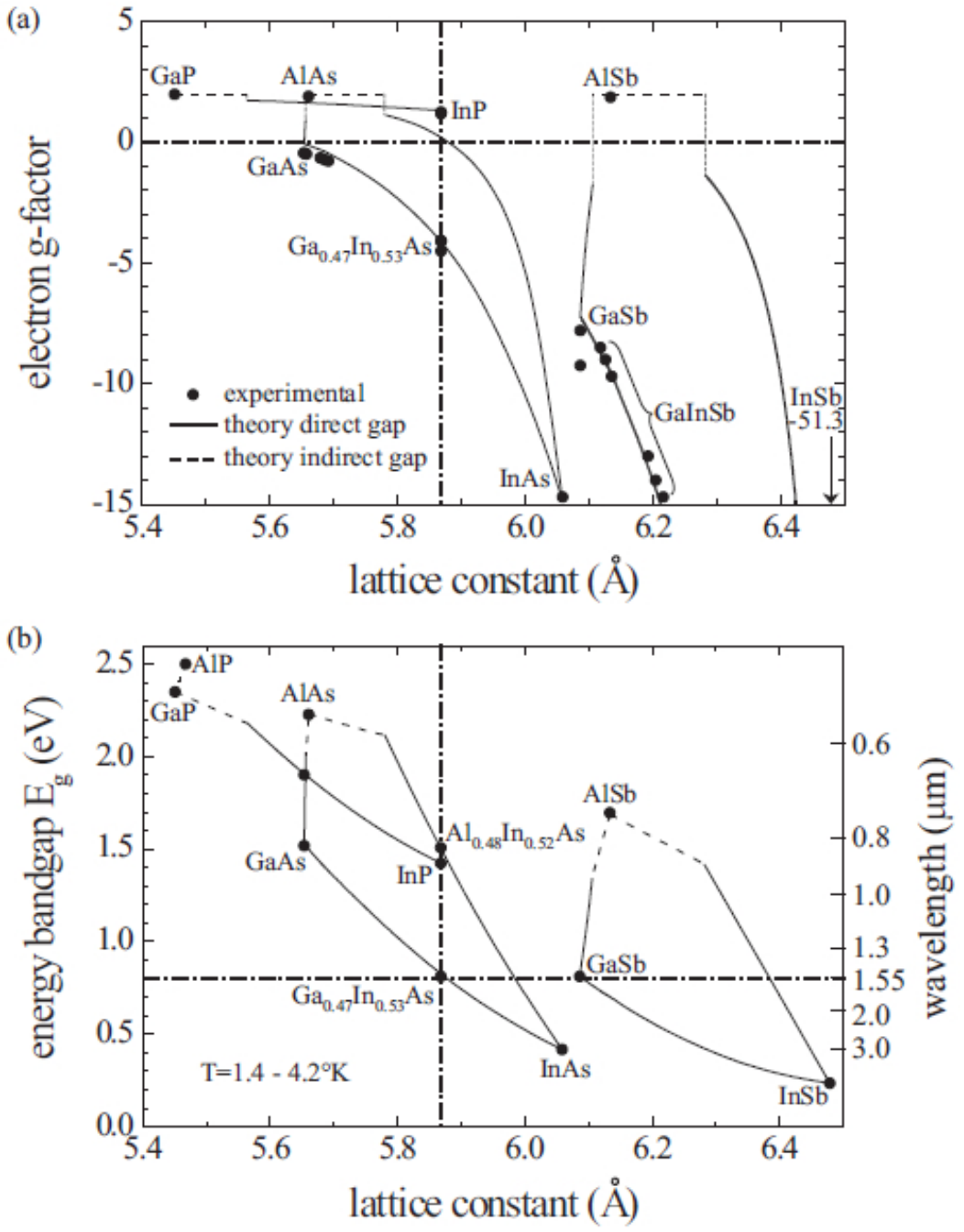


Figure 7.3: (a) A graph of g-factors for conduction electrons in III-V semiconductors as a function of lattice constant. Dots show experimental g^*-factors, solid curves show direct bandgap materials, and dashed curves show indirect bandgap materials. The vertical dash-dotted line indicates the lattice constant of bulk InP, that is normally used for optical communication devices. Bulk ge-factors are plotted for direct bandgap materials (InP, GaAs, GaInAs , InAs, GaSb, GaInSb and InSb), and defect or impurity related g-factors are plotted for indirect bandgap materials (GaP, AlAs, and AlSb). All data were taken at low temperature 1.4 - 4.2.K, except for the InAs data for which $T = 30$K. (b) Energy bandgaps for the same materials at temperatures between 1.4 and 4.2. K. The data is taken from *Electron g-factor engineering of III-V semiconductors for Quantum Communication* by Hideo Kosaka *et al.* arXiv:quant-ph/0102056v2

7.2 Orthogonalised Plane Waves (OPW) and Pseudopotential

First try the following problem:

PROBLEM : Consider a monatomic cubic lattice (again) with lattice constant $a = 5$Å, in which the ion cores sitting in the lattice sites have a Bohr radius of $r = 0.1$Å. Bohr radius of an atom is the radius of the $1s$ state. This state behaves like the tightly bound states (ground state of the simple quantum wells) in the Kronig Penny model that we studied. Estimate by simple arguments, the approximate number of the reciprocal lattice vectors you will have to retain in the plane wave secular determinant (that we derived in the context of Bloch's theorem), to get a good description of this material. The numbers given to you are typical, but not specific to any element. Obviously this will no longer be a "nearly free electron" problem.

The nearly free electron and tight binding approximations represent two opposite limits. The complete solution of the lattice potential should yield the plane wave like delocalised states as well as the localised or tightly bound "core" states. It would require a very large number (upto a million!) of plane wave states to reproduce a deep core state. Also these "plane wave" states and "core" states must be mutually orthogonal, since they belong to different eigenvalues of the same hamiltonian. However if we could somehow write the plane wave states in a way so that they are *by construction* orthogonal to the core states, then the problem would be simplified. Here "simplified" means that a lesser number of components (not a million, but in practice may be about 100) would be needed. This is a crucial point. Let's see how. Some of the intermediate algebra is left for you to complete:

First write down the tight binding state formed out of a core orbital (sum over core states are denoted by c and sum over $\boldsymbol{R}$ denotes sum over all lattice site. Here for simplicity we stick to a monatomic lattice, H denotes the full lattice hamiltonian that contains the potentials created by all the atoms. $\boldsymbol{G}$ denotes reciprocal lattice vectors and $\boldsymbol{k}$ is within BZ1.

As an example: for Si, the "core states" would mean a sum over $1s^2 2s^2 2p^6$

$$\langle \boldsymbol{r}|f_{c\boldsymbol{k}}\rangle = \frac{1}{\sqrt{N}}\sum_{\boldsymbol{R}} e^{i\boldsymbol{k}.\boldsymbol{R}}\phi_c(\boldsymbol{r}-\boldsymbol{R}) \tag{7.33}$$

First we construct the orthogonalised plane wave (OPW) basis states

$$\begin{aligned} |\chi_{\boldsymbol{k}}\rangle &= |\boldsymbol{k}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}\rangle |f_{c\boldsymbol{k}}\rangle \\ |\chi_{\boldsymbol{k}-\boldsymbol{G}}\rangle &= |\boldsymbol{k}-\boldsymbol{G}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle |f_{c\boldsymbol{k}}\rangle \end{aligned} \tag{7.34}$$

We will use these rather than $|\boldsymbol{k}-\boldsymbol{G}\rangle$, to form the eigenvalue equation. The full OPW state is a linear combination

$$|\Psi_{\boldsymbol{k}}\rangle = \sum_{\boldsymbol{G}} C_{\boldsymbol{G}}|\chi_{\boldsymbol{k}-\boldsymbol{G}}\rangle \tag{7.35}$$

1. We will assume that the core state has no dispersion and the energy of the state $\langle \boldsymbol{r}|f_{c\boldsymbol{k}}\rangle$ can be replaced by its atomic value E_c independent of $\boldsymbol{k}$.

2. The main characteristic of the OPW state (eq 7.34) is that it is like a plane wave far from the ion core, but oscillates very fast close to the cores. This allows it to remain orthogonal to the Bloch states (with very little bandwidth) formed out of the core states.

$$\begin{aligned} H|\Psi_{\boldsymbol{k}}\rangle &= \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} \left[H|\boldsymbol{k}-\boldsymbol{G}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle E_c |f_{c\boldsymbol{k}}\rangle \right] \\ &= \lambda|\Psi_{\boldsymbol{k}}\rangle \end{aligned} \tag{7.36}$$

Now left multiply with $\langle \boldsymbol{k}-\boldsymbol{G'}|$ and show that

$$\begin{aligned} \langle \boldsymbol{k}-\boldsymbol{G'}|H|\Psi_{\boldsymbol{k}}\rangle &= \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} \left[\frac{\hbar^2}{2m}(\boldsymbol{k}-\boldsymbol{G})^2 \delta_{\boldsymbol{G}\boldsymbol{G'}} + \langle \boldsymbol{G'}|V|\boldsymbol{G}\rangle + \sum_c E_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle\langle \boldsymbol{k}-\boldsymbol{G'}|f_{c\boldsymbol{k}}\rangle \right] \\ &= \lambda \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} \left[\delta_{\boldsymbol{G}\boldsymbol{G'}} - \sum_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle\langle \boldsymbol{k}-\boldsymbol{G'}|f_{c\boldsymbol{k}}\rangle \right] \end{aligned} \tag{7.37}$$

This means

$$\sum_{\boldsymbol{G}} C_{\boldsymbol{G}} \left[\left(\frac{\hbar^2}{2m}(\boldsymbol{k}-\boldsymbol{G})^2 - \lambda \right) \delta_{\boldsymbol{G}\boldsymbol{G'}} + \langle \boldsymbol{G'}|V|\boldsymbol{G}\rangle + \sum_c (\lambda - E_c)\langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle\langle \boldsymbol{k}-\boldsymbol{G'}|f_{c\boldsymbol{k}}\rangle \right] = 0 \tag{7.38}$$

- The values of λ that will arise as the roots of the determinental equation, are the desired band energies.
- Notice how the core states have contributed an additional term to the potential. The original potential term $\langle \boldsymbol{G'}|V|\boldsymbol{G}\rangle$ is negative in magnitude because it is the attractive Coulomb potential of the atomic nucleii. But the new term is positive because $\lambda > E_c$.
- Numerical computation show that the cancellation is very good, leaving often only 5% of $\langle \boldsymbol{G'}|V|\boldsymbol{G}\rangle$. This effective potential is called the psedopotential.
- This also tells us why the band structure of real materials still has considerable similarity with the nearly free electron result.

7.2.1 The pseudopotential and the pseudo-wavefunction

We can now ask, whether a simplified equation can be found that the state $\Phi_{\boldsymbol{k}} = \sum_{\boldsymbol{G}} C_{\boldsymbol{G}}|\boldsymbol{k}-\boldsymbol{G}\rangle$ satisfy. This is the smoothly varying part of the OPW wavefunction $\Psi_{\boldsymbol{k}}$. We have

$$|\Psi_{\boldsymbol{k}}\rangle = \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} \left[|\boldsymbol{k}-\boldsymbol{G}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle |f_{c\boldsymbol{k}}\rangle \right] \tag{7.39}$$

$$= \sum_{\boldsymbol{G}} C_{\boldsymbol{G}}|\boldsymbol{k}-\boldsymbol{G}\rangle - \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} \sum_c \langle f_{c\boldsymbol{k}}|\boldsymbol{k}-\boldsymbol{G}\rangle |f_{c\boldsymbol{k}}\rangle \tag{7.40}$$

$$= |\Phi_{\boldsymbol{k}}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle |f_{c\boldsymbol{k}}\rangle \tag{7.41}$$

Therefore

$$\begin{aligned} H|\Psi_{\boldsymbol{k}}\rangle &= E|\Psi_{\boldsymbol{k}}\rangle \\ \implies H|\Phi_{\boldsymbol{k}}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle E_c |f_{c\boldsymbol{k}}\rangle &= E|\Phi_{\boldsymbol{k}}\rangle - E\sum_c \langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle |f_{c\boldsymbol{k}}\rangle \end{aligned} \tag{7.42}$$

Therefore

$$H|\Phi_{\boldsymbol{k}}\rangle - \sum_c \langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle E_c|f_{c\boldsymbol{k}}\rangle = E|\Phi_{\boldsymbol{k}}\rangle - E\sum_c \langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle|f_{c\boldsymbol{k}}\rangle \tag{7.43}$$

$$T|\Phi_{\boldsymbol{k}}\rangle + V|\Phi_{\boldsymbol{k}}\rangle + \sum_c (E-E_c)\langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle|f_{c\boldsymbol{k}}\rangle = E|\Phi_{\boldsymbol{k}}\rangle \tag{7.44}$$

$$T|\Phi_{\boldsymbol{k}}\rangle + \underbrace{\left[V + \sum_c (E-E_c)|f_{c\boldsymbol{k}}\rangle\langle f_{c\boldsymbol{k}}|\right]}_{\mathrm{U}}|\Phi_{\boldsymbol{k}}\rangle = E|\Phi_{\boldsymbol{k}}\rangle \tag{7.45}$$

Convince yourself that the operator U is like an integral operator:

$$U|\Phi_{\boldsymbol{k}}\rangle = V|\Phi_{\boldsymbol{k}}\rangle + \sum_c (E-E_c)|f_{c\boldsymbol{k}}\rangle\langle f_{c\boldsymbol{k}}|\Phi_{\boldsymbol{k}}\rangle \tag{7.46}$$

$$\langle \boldsymbol{r}|U|\Phi_{\boldsymbol{k}}\rangle = \langle \boldsymbol{r}|V|\Phi_{\boldsymbol{k}}\rangle + \sum_c (E-E_c)\langle \boldsymbol{r}|f_{c\boldsymbol{k}}\rangle\langle f_{c\boldsymbol{k}}|\boldsymbol{r'}\rangle\langle \boldsymbol{r'}|\Phi_{\boldsymbol{k}}\rangle \tag{7.47}$$

$$\therefore U(\boldsymbol{r})\Phi(\boldsymbol{r}) = V(\boldsymbol{r})\Phi(\boldsymbol{r}) + \sum_c (E-E_c)\int d\boldsymbol{r'}\underbrace{f_{c,\boldsymbol{k}}(\boldsymbol{r})f^*_{c,\boldsymbol{k}}(\boldsymbol{r'})}_{K(\boldsymbol{r},\boldsymbol{r'})}\Phi(\boldsymbol{r'}) \tag{7.48}$$

The pseudo-potential acts on the pseudo-wavefunction and produces the correct eigenvalues! The deep coulomb potential near the nuclei and the sharp rise and fall (nodes) of the wavefunction have been taken away. This OPW+Pseudopotenial method can be used to calculate real band structures of many elements.

7.2.2 What have we still left out?

At the end of the day, no band structure problem is a single electron problem. The coulomb repulsion between the electrons is strong, but it is a remarkable fact that they do not fundamentally alter the band picture that we have calculated so far. Why it can happen was systematically explained by Landau.

Chapter 8

Electrons, lattice vibrations and electromagnetic radiation together

We will address the following questions:

1. How does an electron "see" the lattice vibrations?
2. What conservation laws hold for this interaction?
3. How does an electron interact with electromagnetic waves (or photons)?
4. What kind of transitions are possible due to interaction of electrons with lattice vibrations and e-m radiation?

8.1 Electrons and lattice vibrations

We know that the vibrations of the atoms of the lattice can be modelled as simple harmonic oscillators. The N atoms which make up the lattice move from their equilibrium positions. So the instantaneous potential at a point inside the solid would also change a little. The Bloch states are the eigenstates of a potential generated by all the atoms sitting in their equilibrium positions. Since that potential changes a little due to the oscillatory motion of the atoms, the Bloch electrons sees this as a time varying perturbation. This is what we must model. The complexity in the problem is primarily due to the fact that we must take into account the motion of all the atoms at once and consider the interaction of the (single) electron with them all at once.

First, we consider a hypothetical situation where there is only one oscillator. The co-ordinates of the atoms will be denoted in uppercase X, Y, Z, the co-ordinates of the electron will be in lowercase $\boldsymbol{r} = (x, y, z)$.

- The eigenstates of the electron and the oscillator (assumed to be in its n^{th} state), are respectively

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}.\boldsymbol{r}} u_{\boldsymbol{k}}(\boldsymbol{r}) \tag{8.1}$$

$$U_n(X) = \frac{1}{\sqrt{2^n n!}} \left(\frac{M\omega}{\pi\hbar}\right)^{1/4} e^{-\frac{M\omega}{2\hbar}X^2} H_n\left(\sqrt{\frac{M\omega}{\hbar}}X\right) \tag{8.2}$$

where

$$H_n(y) = (-1)^n\, e^{y^2} \left(\frac{d^n}{dy^n}\right) e^{-y^2} \tag{8.3}$$

- The displacement $\delta\boldsymbol{R} = (X, Y, Z)$ means that the perturbation seen by the electron+oscillator system becomes

$$H_1 = V(x - X, y - Y, z - Z) - V(x, y, z) \approx -\left(X\frac{\partial V}{\partial x} + Y\frac{\partial V}{\partial y} + Z\frac{\partial V}{\partial z}\right) \tag{8.4}$$

- We need to analyse transitions between state that can be denoted by $(n,\boldsymbol{k})$ and $(n',\boldsymbol{k}')$, the notation should be self-explanatory here. The product state $U_n(X)\psi_{\boldsymbol{k}}(\boldsymbol{r})$ will be denoted as $|n\boldsymbol{k}\rangle$. So

$$\begin{aligned}\langle n'\boldsymbol{k}'|H_1|n\boldsymbol{k}\rangle &= \langle n'|X|n\rangle \times \left\langle \boldsymbol{k}' \left| \frac{\partial V}{\partial x} \right| \boldsymbol{k} \right\rangle \\ &= \begin{cases} \sqrt{\dfrac{\hbar n}{M\omega}}\ \delta_{n',n\pm1} \times \left.\dfrac{\partial V}{\partial x}\right|_{kk'} & \text{if } n > n' \\ \sqrt{\dfrac{\hbar n'}{M\omega}}\ \delta_{n',n\pm1} \times \left.\dfrac{\partial V}{\partial x}\right|_{kk'} & \text{if } n' > n \end{cases}\end{aligned} \tag{8.5}$$

- The first term involving Hermite polynomials exist only if $n' = n\pm1$, this means that only one quanta $\hbar\omega$ can be transferred via this interaction.
- The electron-phonon interaction is *NOT* elastic as far as the electron is concerned. But the energy $\hbar\omega$ (where ω is a lattice vibration frequency) is generally extremely small compared to the electron's kinetic energy. Thus an electron *effectively* remains on the same constant energy surface after a scattering, but its crystal momentum can change a lot. in the $k$space

8.1.1 The crystal momentum before and after scattering

The total potential seen by the electron is the sum total (uppercase V) of the contribution of all the unit cells. The contribution of each unit cell is in lowercase v. The displacement of the atoms for a vibration mode with energy $\omega_{\boldsymbol{q}}$ and wavevector $\boldsymbol{q}$ is

$$\delta\boldsymbol{R}_n = \sum_{\boldsymbol{q}} \boldsymbol{A}_{\boldsymbol{q}}\ e^{i(\boldsymbol{q}.\boldsymbol{R}_n - \omega_{\boldsymbol{q}}t)} \tag{8.6}$$

Hence

$$V(\boldsymbol{r},t) = \sum_n v(\boldsymbol{r} - \boldsymbol{R}_n - \delta\boldsymbol{R}_{\boldsymbol{n}}) \tag{8.7}$$

$$\begin{aligned}\delta V(\boldsymbol{r},t) &= -\sum_n\ \delta\boldsymbol{R}_{\boldsymbol{n}}.\nabla_{\boldsymbol{r}} v(\boldsymbol{r} - \boldsymbol{R}_n) \\ &= -\sum_q \boldsymbol{A}_{\boldsymbol{q}}.\nabla_{\boldsymbol{r}} \underbrace{\sum_n e^{i\boldsymbol{q}.\boldsymbol{R}_n} f(\boldsymbol{r} - \boldsymbol{R}_n)}_{\text{tight binding form !}}\ e^{i\omega_{\boldsymbol{q}}t} \\ &= -\sum_q \boldsymbol{A}_{\boldsymbol{q}}.\nabla_{\boldsymbol{r}} \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} e^{i(\boldsymbol{q}+\boldsymbol{G}).\boldsymbol{r}}\ e^{i\omega_{\boldsymbol{q}}t}\end{aligned}$$

The total potential should be the sum total contribution of all the modes

$$\delta V(\boldsymbol{r},t) = -i\sum_{\boldsymbol{q}}\sum_{\boldsymbol{G}} C_{\boldsymbol{G}}\boldsymbol{A}_{\boldsymbol{q}}.(\boldsymbol{q}+\boldsymbol{G}) e^{i(\boldsymbol{q}+\boldsymbol{G}).\boldsymbol{r}}\ e^{i\omega_{\boldsymbol{q}}t} \tag{8.8}$$

The harmonic oscillator states for the entire lattice needs to be written in terms of the occupation number representation and the raising and lowering operators. Recall that

$$a_n = \sqrt{\frac{M\omega}{2\hbar}}\hat{x_n} + i\sqrt{\frac{1}{2\hbar M\omega}}\hat{p_n} \tag{8.9}$$

$${a_n}^\dagger = \sqrt{\frac{M\omega}{2\hbar}}\hat{x_n} - i\sqrt{\frac{1}{2\hbar M\omega}}\hat{p_n} \tag{8.10}$$

$$\therefore \hat{x_n} = \sqrt{\frac{\hbar}{2M\omega}}\ (a_n + {a_n}^\dagger) \tag{8.11}$$

$$\tag{8.12}$$

Hence

$$A_{\boldsymbol{q}} = \frac{1}{\sqrt{N}} \sum_n \sqrt{\frac{\hbar}{2M\omega_q}} \left(a_n e^{i\boldsymbol{q}.\boldsymbol{R}_n} + {a_n}^\dagger e^{i\boldsymbol{q}.\boldsymbol{R}_n}\right) \tag{8.13}$$

$$= \sqrt{\frac{\hbar}{2M\omega_q}} \left(a_{\boldsymbol{q}} + {a_{-\boldsymbol{q}}}^\dagger\right) \tag{8.14}$$

So the potential will have the form

$$\delta V(\boldsymbol{r},t) = \sum_{\boldsymbol{q},\boldsymbol{G}} \sqrt{\frac{\hbar}{2M\omega_{\boldsymbol{q}}}} (a_{\boldsymbol{q}} + a^\dagger_{-\boldsymbol{q}}) C_{\boldsymbol{G}} \varepsilon_{\boldsymbol{q}}.(\boldsymbol{q}+\boldsymbol{G}) e^{i(\boldsymbol{q}+\boldsymbol{G}).\boldsymbol{r}}\ e^{i\omega t} \tag{8.15}$$

Now when this potential is sandwiched between two Bloch states which can be expanded as $\sum_{\boldsymbol{G_m}} C_{\boldsymbol{G_m}} e^{i(\boldsymbol{k}+\boldsymbol{G_m}).\boldsymbol{r}}$ and $\sum_{\boldsymbol{G_l}} C_{\boldsymbol{G_l}} e^{i(\boldsymbol{k'}+\boldsymbol{G_l}).\boldsymbol{r}}$ it is clear that the integral will involve terms like $e^{i(\boldsymbol{k}-\boldsymbol{k'}+\boldsymbol{q}-\boldsymbol{G}_l+\boldsymbol{G}_m)}$. This will lead to a delta function with these arguments and hence a conservation law of the form

$$\boldsymbol{k'} - \boldsymbol{k} = \boldsymbol{q} + \boldsymbol{G} \tag{8.16}$$

since the difference of two reciprocal lattice vector is itself another reciprocal lattice vector. Notice how the conservation law again looks like momentum conservation even though lattice vibrations (phonons) do not carry *any* real momentum at all.

The wavevector of the phonon would often be comparable to the size of the Brillouin zone and hence of the order of the reciprocal lattice vector.

The energy is however be at most $\hbar\omega_D$ (the Debye frequency). This number is about a few meV in most cases.

8.1.2 The generic form of the electron-phonon interaction

If the sum over $\boldsymbol{G}$ in 8.15 is completed we would get an expression that is a function of $\boldsymbol{q}$ and $\boldsymbol{r}$ and one phonon creation/annihlation operator.

$$H_{ep}(\boldsymbol{r},t) = \frac{1}{\sqrt{N}} \sum_{\boldsymbol{q}}\ V(\boldsymbol{q},\boldsymbol{r}) a_{\boldsymbol{q}} e^{-i\omega_{\boldsymbol{q}} t} \quad + \quad h.c. \tag{8.17}$$

The $a_{\boldsymbol{q}}$ term corresponds to one phonon with energy $\omega(\boldsymbol{q})$ and wavevector $\boldsymbol{q}$ being destroyed (the energy goes to the electron) the hermitian conjugate term $a^\dagger_{\boldsymbol{q}}$ term denotes the process in which the electron loses energy and a phonon is created.

8.2 Electrons and electromagnetic radiation

An electron has a charge $q = -|e|$. This interacts with an electromagnetic field via the vector potential $\boldsymbol{A}$ and the scalar potential ϕ

$$H = \frac{1}{2m} \left(\boldsymbol{p} - q\boldsymbol{A}\right)^2 + q\phi \tag{8.18}$$

The electric field is

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t} \tag{8.19}$$

Recall that the divergence of the vector potential can be chosen and we make a choice

$$\nabla \boldsymbol{A} = 0 \tag{8.20}$$

A vector potential corresponding to a propagating wave would be

$$\boldsymbol{A} = \boldsymbol{A}_0 \ \cos \ (\boldsymbol{k}.\boldsymbol{r} - \omega t) \tag{8.21}$$

Correct to first order this coupling leads to a perturbation in the Hamiltonian which has two pieces. These two pieces commute due to the choice 8.20 (prove this).

$$\begin{aligned} H_{em} &= -\frac{q}{2m}(\boldsymbol{A}.\boldsymbol{p} + \boldsymbol{p}.\boldsymbol{A}) \\ &= -\frac{q}{m}\boldsymbol{A}.\boldsymbol{p} \\ &= -\frac{q}{2m}\left[e^{i(\boldsymbol{k}.r-\omega t)} + e^{-i(\boldsymbol{k}.r-\omega t)}\right]\boldsymbol{A}_0.\boldsymbol{p} \\ &= \frac{|e||\boldsymbol{A}_0|}{m}e^{-i\omega t}\boldsymbol{\varepsilon}.\boldsymbol{p} \quad + \quad h.c. \end{aligned}$$

here $\boldsymbol{\varepsilon}$ is the polarisation direction of $\boldsymbol{A}$. The wavevector $\boldsymbol{k}$ of electromagnetic radiation is far smaller than the dimension of the Brillouin zone for optical wavelengths. Since $k = \frac{2\pi}{\lambda}$ and the wavelength of visible light is about half a micron, which is at least a thousand times larger than typical lattice constants.

8.3 Electromagnetic radiation and phonons together

If two perturbations act together, in first order their effects are independent. Here we recall the formula for the transition rates upto second order. If H_1 is the perturbation then the rate of transition from *initial* to *final* ($f \leftarrow i$) state P_{fi} is:

$$P_{fi}^{(1)} = \frac{2\pi}{\hbar}\left|\langle\psi_f|H_1|\psi_i\rangle\right|^2 \delta(E_f - E_i - \hbar\omega) \tag{8.22}$$

$$P_{fi}^{(2)} = \frac{2\pi}{\hbar}\left|\sum_\alpha \frac{\langle\psi_f|H_1|\psi_\alpha\rangle\langle\psi_\alpha|H_1|\psi_i\rangle}{E_i + \hbar\omega - E_\alpha}\right|^2 \delta(E_f - E_i - \hbar\omega_1 - \hbar\omega_2) \tag{8.23}$$

Notice how in the second order process (eqn 8.23) the perturbation "acts" twice. Thus if it has two parts like an electron-phonon and an electron-light interaction, then the combined effect of phonon and light would be captured by the second order process. This is how we understand the indirect transitions.

Appendix A

Boltzmann Transport equation : deviation from equilibrium : drift and diffusion.

References:

1. Chapter 12 & 16, *Solid State Physics*, N. W. Ashcroft and N.D. Mermin
2. page 61-67, *Quantum Heterostructures*, V. V. Mitin, V. A. Kochelap and M. A. Stroscio
3. A.H. Marshak and K. M. van Vliet, Electrical current in solids with position dependent band structure, *Solid State Electronics*, page 417-427, **21**, 1978

Boltzmann Transport equation (BTE) allows us to take a step out of equilibrium thermodynamics. In general the concept of equilibrium implies that there is no net particle flow from one point of a system to another. An equivalent statement is that the electrochemical potential is same throughout the system. Yet, in reality, all electrical devices have currents flowing in them - if it didn't it wouldn't be interesting at all. Often the effects of current flow are also irreversible -like Joule heating. We will see how the presence of electro-magnetic fields, electrochemical potential gradients and thermal gradients drive current. We will do so by trying to calculate the distribution function in a situation slightly away from equilibrium. Our main target is to find expressions for current when "external fields" are present.

A.1 A "handwaving" derivation of the equation

Let's consider the phase space with just two co-ordinates $\mathbf{r}, \mathbf{p}$. The distribution is represented by a point as shown at time t. What happens to the points in the "volume element" after a little while? Unless they are scattered they change their co-ordinates according to the following rule:

$$\begin{aligned} \boldsymbol{r}(t+\delta t) &= \boldsymbol{r}(t) + \frac{\boldsymbol{p}}{m}\delta t \\ \boldsymbol{p}(t+\delta t) &= \boldsymbol{p}(t) + \boldsymbol{F}\delta t \end{aligned} \tag{A.1}$$

If this was all, then it wouldn't have been interesting at all, because it would mean:

$$f(\boldsymbol{r}(t) + \frac{\boldsymbol{p}}{m}\delta t, \boldsymbol{p}(t) + \boldsymbol{F}\delta t, t+\delta t) d^3\boldsymbol{r}' d^3\boldsymbol{p}' = f(\boldsymbol{r}(t), \boldsymbol{p}(t), t) d^3\boldsymbol{r} d^3\boldsymbol{p} \tag{A.2}$$

Now note the following:

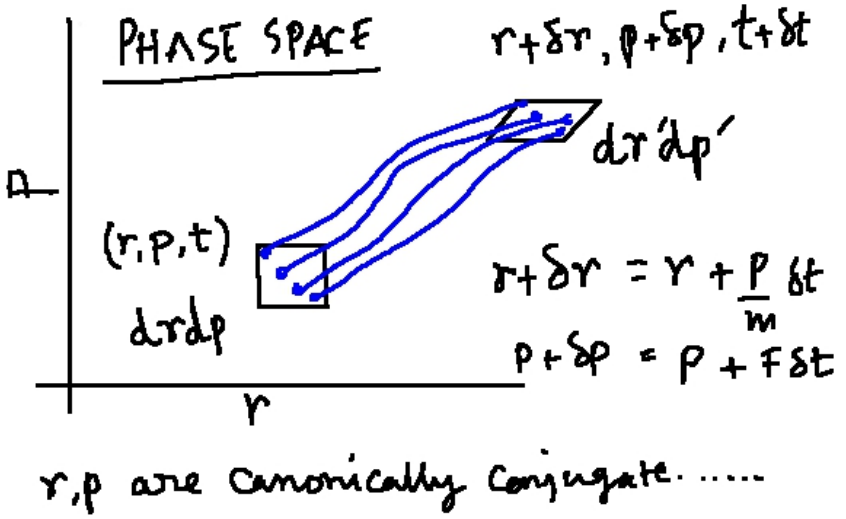


Figure A.1: Flow of points in phase space.

- The volume element around the point distorts, but preserves its volume. Something that was a square at time t, would become a parallelogram at $t+\delta t$. This would happen if the external force is derivable from a potential and the two co-ordinates that we have chosen are canonically conjugate. We will ignore this technicality at this point. Many text books on statistical physics, treat this point carefully...
- The equality fails to hold, because some of the trajectories get scattered by collisions. Thus the amount by which this equality fails to hold, must be attributed to collisions. This leads us to the following

$$\frac{\boldsymbol{p}}{m}.\nabla f + \boldsymbol{F}.\nabla_{\boldsymbol{p}} f + \frac{\partial f}{\partial t} = \frac{\mathrm{d}f}{\mathrm{d}t} \tag{A.3}$$

Now we take another step, to convert this classical equation into a semiclassical one. We change momentum to wavevector. This indeed means that we are using the concept of phase space (clearly defined momentum and position) in quantum mechanical scenario. An analysis of how far this can give correct results, is non-trivial, but we will give an answer later.

A.2 The semiclassical Boltzmann equation

We denote the equilibrium distribution function by $f^0(\mathbf{r},\mathbf{k},t)$ When the distribution function deviates from equilibrium, a "restoring effect" arises in the system, that tries to push the distribution back towards equilibrium. This implies that the collision integral on the right hand side of BTE is assumed to have the form

$$\left.\frac{df}{dt}\right|_{collision} = -\frac{f-f^0}{\tau} \tag{A.4}$$

Later on we will try to determine τ in terms of the scattering mechanisms in some systems. The best justification of the relaxation time approximation is that it works in many cases!
We thus write the BTE as:

$$\frac{\partial f}{\partial t} + \frac{d\mathbf{r}}{dt}.\nabla_r f + \frac{d\mathbf{k}}{dt}.\nabla_k f = -\frac{f-f^0}{\tau} \tag{A.5}$$

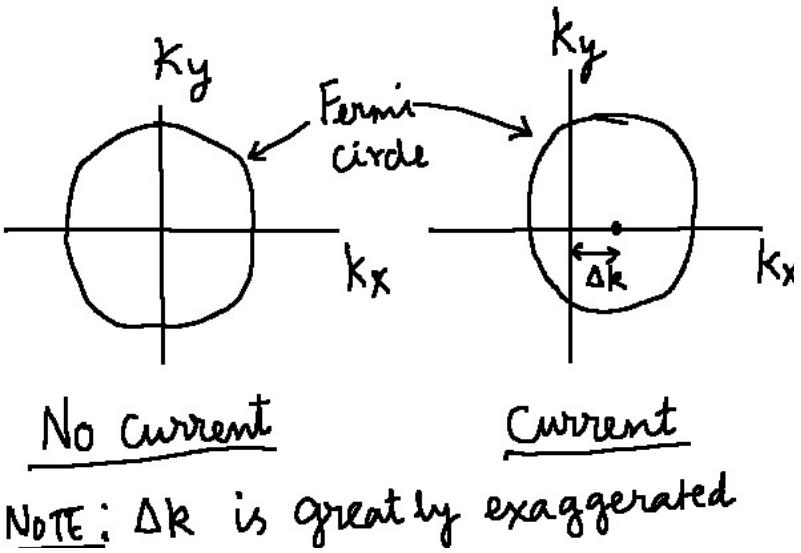


Figure A.2: Displacement of the Fermi circle results in current flow.

If there are electric and magnetic fields in the system, the "semiclassical" equation of motion would be:

$$\hbar \frac{d\mathbf{k}}{dt} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{A.6}$$

We will also assume that f has no explicit time dependence and that $\nabla_r f = 0$ as well - which in general means that there is no density gradient of particles in the system. This assumption is correct if we are dealing with a piece of copper wire at constant temperature, but not necessarily correct for a semiconductor or a even a piece of metal with a thermal gradient. Throughout the lectures we will assume that the charge of each particle is "q". For the most common case of electrons in the conduction band we would need to put $q = -|e|$ to get the correct sign of the terms.

A.3 Electric field only

With these assumptions, equation A.5 in presence of an electric field only reduces to

$$\frac{q}{\hbar}\mathbf{E}.\nabla_k f = -\frac{f - f^0}{\tau} \tag{A.7}$$

Then we make the first order approximation by taking the derivative around the equilibrium value

$$f(\mathbf{k}) = f^0(\mathbf{k}) - \frac{q\tau}{\hbar}\mathbf{E}.\nabla_k f^0 \tag{A.8}$$

$$= f^0(\mathbf{k} - \frac{q\tau}{\hbar}\mathbf{E}) \tag{A.9}$$

This means that the equilibrium distribution function has retained its functional form but just got shifted by a certain amount. Think of how the graph of a function $f(x)$ would be related to $f(x - a)$. In the figure we have drawn it for a Fermi distribution in 2 dimensions. Note that if the relaxation mechanism is strong then τ would be small. On the other hand if the particle suffers very little scattering then τ would be large and the displacement of the Fermi circle (or sphere) would also be large.

PROBLEM: The free electron density in Copper is n=8.5 $\times$ 10^{28}m^{-3} and near room temperature the relaxation time of most metals is of the order of $10^{-15} - 10^{-14}$ sec. From this data estimate the fractional

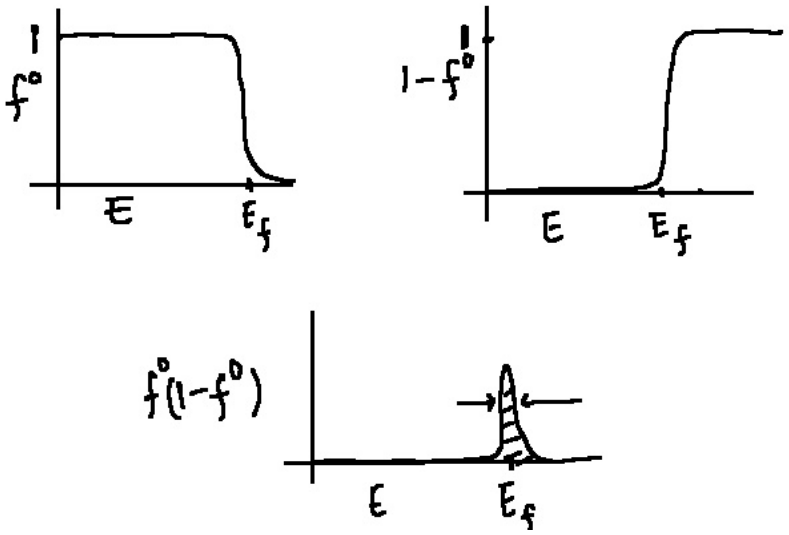


Figure A.3: How does the product $f^0(1-f^0)$ behave?

shift of the distribution on the scale of the Fermi wavevector (k_F) for an electric field of 10V/m,(*i.e.* calculate $\Delta k/k_F$).

Our target is to calculate the current produced by this state:

$$\begin{aligned} \mathbf{j} &= q\sum_{\mathbf{k}} \mathbf{v}\delta f \\ &= \frac{2q}{(2\pi)^3}\int \mathrm{d}^3\mathbf{k}\ \ \mathbf{v}\delta f \end{aligned} \tag{A.10}$$

To proceed we need to evaluate eqn A.9 for the case of Fermi distribution.

$$f^0 = \frac{1}{e^{\beta(E-E_f)}+1} \tag{A.11}$$

$$\begin{aligned} \nabla_k f^0 &= -\left(\frac{1}{e^{\beta(E-E_f)}+1}\right)^2 e^{\beta(E-E_f)}\nabla_k\beta(E-E_f) \\ &= -\beta f^0(1-f^0)\nabla_k E \\ &= -\beta f^0(1-f^0)\hbar\mathbf{v_g} \end{aligned} \tag{A.12}$$

Notice that the Fermi level is not a function of $\mathbf{k}$. The end result of A.12 can also be written as :

$$\nabla_k f^0 = \frac{\partial f^0}{\partial E}\hbar\mathbf{v_g} \tag{A.13}$$

Equations A.12 and A.13 are important results as these derivatives occur frequently in transport related physics. How does the product $f^0(1-f^0)$ behave ? Since f^0 drops sharply around E_f, $(1-f^0)$ must rise sharply around E_f, producing a sharp peak.

PROBLEM : Certain combinations of the Fermi function, occur very frequently in expressions that involve scattering or transitions. It is useful to be familiar with the combination $f^0(1-f^0)$

Make a rough sketch of how $f^0(1-f^0)$ would look as a function of energy. How does the area under the curve of $f^0(1-f^0)$ vary with temperature?

Using eqn A.12 and eqn A.9 we get

$$\delta f = q\tau\beta f^0(1-f^0)\mathbf{E}.\mathbf{v_g} \tag{A.14}$$

Notice that the change occurs only near the Fermi surface. This is the generic reason phenomena like electrical or heat conduction are often referred to as a "Fermi surface property". Now we calculate the current as defined in eqn A.10

$$\begin{aligned}\mathbf{j} &= \frac{q}{4\pi^3}\int \mathrm{d}^3\mathbf{k}\ \ \mathbf{v_g}\left(q\tau\beta f^0(1-f^0)\mathbf{E}.\mathbf{v_g}\right)\\ &= nq\left(\frac{q}{4\pi^3 n}\int \mathrm{d}^3\mathbf{k}\ \ \tau\mathbf{v_g}\otimes\mathbf{v_g}\left(-\frac{\partial f^0}{\partial E}\right)\right).\mathbf{E}\end{aligned} \tag{A.15}$$

Notice that the part within the large brackets is determined by equilibrium properties of the system only. The outer product ($\otimes$) of two vectors is an object with two indices and can be written out like a matrix. For example

$$\mathbf{C} = \mathbf{A}\otimes\mathbf{B} \quad \text{implies} \tag{A.16}$$

$$C_{ij} = A_iB_j \tag{A.17}$$

We will call the quantity inside the bracket as mobility. But it is often not necessary to evaluate this is full generality. We assume that the dispersion relation is spherically symmetric and evaluate the expression for low temperature. Low temperature implies that the Fermi distribution has a sharp drop near E_f and behaves like a step function at that point. The derivative of a step function is a (Dirac) delta function which would pick out the contribution of the integrand around its peak. So we can write

$$\lim_{T\to 0} -\frac{\partial f^0}{\partial E} = \delta(E-E_f) \tag{A.18}$$

Let's go through the steps for evaluating the mobility integral:

$$\overleftrightarrow{\mu} = \frac{q}{4\pi^3 n}\int \mathrm{d}^3\mathbf{k}\ \ \tau\mathbf{v_g}\otimes\mathbf{v_g}\left(-\frac{\partial f^0}{\partial E}\right) \tag{A.19}$$

$$= \frac{q}{n}\int \mathrm{d}E\ \ D(E)\tau\mathbf{v_g}\otimes\mathbf{v_g}\left(-\frac{\partial f^0}{\partial E}\right)$$

$$= \frac{q}{n}\int \mathrm{d}E\ \ D(E)\tau\mathbf{v_g}\otimes\mathbf{v_g}\delta(E-E_f) \quad \text{as T}\to 0 \tag{A.20}$$

$$\tag{A.21}$$

Now since $\mathbf{v_g} = \hbar\mathbf{k}/m$, we can write:

$$\mu_{ij} = \frac{q}{n}\int \mathrm{d}E\ \ D(E)\ \tau\left(\frac{\hbar}{m}\right)^2 k_ik_j\ \ \delta(E-E_f) \tag{A.22}$$

This from works in all dimensions, provided the density n is interpreted correctly. Now μ_{ij} will average to zero if $i\neq j$, due to symmetry. If we fix k_i, we can find corresponding pairs of points at k_j and $-k_j$, which will add up to zero. So we need to calculate only the diagonal terms. Since there is nothing to distinguish the x, y or z directions, all the diagonal components must be equal. This allows us to write:

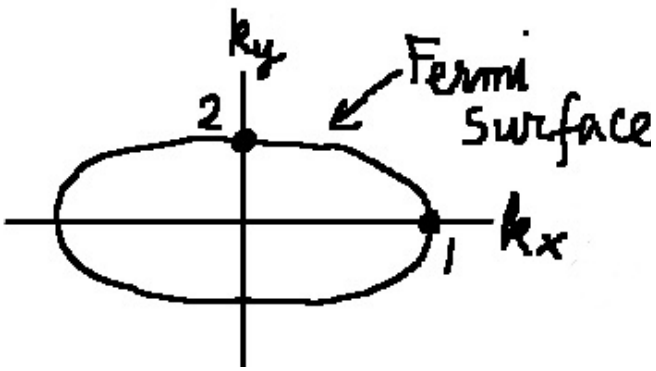


Figure A.4: Along which direction do you expect the effective mass to be higher, at point 1 (along k_x) or point 2 (along k_y)?

$$\begin{aligned}\mu_{ii} &= \frac{q}{n}\int \mathrm{d}E\ \ D(E)\ \tau\left(\frac{\hbar}{m}\right)^2 \frac{k_x^2+k_y^2+k_z^2}{3}\ \delta(E-E_f) \\ &= \frac{q}{3n}\int \mathrm{d}E\ \ D(E)\ \tau\frac{2E}{m}\ \delta(E-E_f) \end{aligned} \tag{A.23}$$

Using the expression for density of states in 3D, eqn A.23 reduces to:

$$\begin{aligned}\mu_{ii} &= \frac{q}{3\pi^2 mn}\int \mathrm{d}E\ k^3\ \tau\ \delta(E-E_f) \\ &= \frac{q\tau}{m} \quad \text{since}\ \ k_F^3 = 3\pi^2 n \end{aligned} \tag{A.24}$$

A.3.1 The temperature dependence of mobility, conductivity

The mobility is a temperature dependent quantity - the T dependence of conductivity for example arises from changes in mobility as well as carrier density of a system. Usually scattering calculations give us the scattering rate ($\tau(E)$) or the collision cross section as a function of E. How do we use this information to calculate $\mu(T)$. Let's consider the diagonal element (say μ_{xx}) from equation A.15 which relates **j** and **E**

$$\begin{aligned}\mu_{ij} &= \frac{q\int_0^\infty dE\ D(E)\tau(E)v_i v_j\left(-\frac{\partial f}{\partial E}\right)}{n} \\ \therefore \mu_{xx} &= \frac{q\int_0^\infty dE\ D(E)\tau(E)v_x^2\left(-\frac{\partial f}{\partial E}\right)}{\int_0^\infty dE\ D(E)f(E)} \end{aligned}$$

Now if we are working in d dimensions, then in general we have

$$\begin{aligned}D(E) &\propto E^{d/2-1} \\ E &\propto mv_x^2\ \frac{d}{2} \end{aligned} \tag{A.25}$$

Using these two results and a partial integration of the denominator we get:

$$\begin{aligned}\mu_{xx} &= \frac{2q}{md}\frac{\int_0^\infty dE\ E^{d/2}\tau(E)\left(-\frac{\partial f}{\partial E}\right)}{\int_0^\infty dE\ \frac{E^{d/2}}{d/2}\left(-\frac{\partial f}{\partial E}\right)} \\ &= \frac{q}{m}\frac{\int_0^\infty dE\ E^{d/2}\tau(E)\frac{\partial f}{\partial E}}{\int_0^\infty dE\ E^{d/2}\frac{\partial f}{\partial E}}\end{aligned}$$

Since $\mu = \frac{q\tau}{m}$, we usually write,

$$\langle\tau(T)\rangle = \frac{\int_0^\infty dE\ E^{d/2}\tau(E)\frac{\partial f}{\partial E}}{\int_0^\infty dE\ E^{d/2}\frac{\partial f}{\partial E}} \tag{A.26}$$

$\tau(E)$ is often available from scattering calculations and the integral gives the energy range over which we need to average it. The presence of the term $\frac{\partial f}{\partial E}$ ensures that the important part is centred at Fermi energy, the spread of the region increases with increasing temperature.

A.4 Conservation of the phase space volume

We will apply the BTE to a situation where the "forces" will have some velocity dependence, like the Lorentz force. So let's prove that the "volume" will still be conserved. Part of the proof is left as an exercise. We will work with two variables only for simplicity. Consider the points (x, p) and a small area element $\delta x\delta p$ around it as before. What happens to the corner points after time δt ? Both $\dot{x}$ and $\dot{p}$ can be functions of x and p, but we do not write all the functional dependances explicitly. See the following table:

point	time= t	time= $t+\delta t$
$1 \to 1'$	$\begin{pmatrix} x \\ p \end{pmatrix}$	$\begin{pmatrix} x+\dot{x}\delta t \\ p+\dot{p}\delta t \end{pmatrix}$
$2 \to 2'$	$\begin{pmatrix} x+\delta x \\ p \end{pmatrix}$	$\begin{pmatrix} x+\delta x+\left(\dot{x}+\frac{\partial\dot{x}}{\partial x}\delta x\right)\delta t \\ p+\left(\dot{p}+\frac{\partial\dot{p}}{\partial x}\delta x\right)\delta t \end{pmatrix}$
$4 \to 4'$	$\begin{pmatrix} x \\ p+\delta p \end{pmatrix}$	$\begin{pmatrix} x+\left(\dot{x}+\frac{\partial\dot{x}}{\partial p}\delta p\right)\delta t \\ p+\delta p+\left(\dot{p}+\frac{\partial\dot{p}}{\partial p}\delta p\right)\delta t \end{pmatrix}$

PROBLEM: Show that the area element $\delta x \delta p$ will become $\delta x' \delta p'$ after time δt where

$$\delta x' \delta p' = \begin{vmatrix} \delta x \left(1 + \frac{\partial \dot{x}}{\partial x} \delta t\right) & \frac{\partial \dot{p}}{\partial x} \delta x \delta t \\ \frac{\partial \dot{x}}{\partial p} \delta p \delta t & \delta p \left(1 + \frac{\partial \dot{p}}{\partial p} \delta t\right) \end{vmatrix} \tag{A.27}$$

Then prove that if x and p are driven by a Hamiltonian such that $\dot{x} = \frac{\partial H}{\partial p}$ and $\dot{p} = -\frac{\partial H}{\partial x}$ then the first order part (in δt) of the expression will be zero. Since equations of motion in a magnetic field can be written in Hamiltonian form as well with canonical momentum defined properly, we can still use the equations.

A.5 Electric and magnetic field

Now let's recall eqn A.6 and allow a magnetic field. Eqn A.8 that described the deviation of the distribution function from equilibrium should now read :

$$f(\mathbf{k}) = f^0(\mathbf{k}) - \frac{q\tau}{\hbar}(\mathbf{E} + \mathbf{v} \times \mathbf{B}).\nabla_k f^0 \tag{A.28}$$

Now because the force term has explicit dependence on $\mathbf{k}$ we can no longer write down the solution by inspection, as we did in eqn A.9. However we now try a solution of the same form, with an unknown vector $\mathbf{Z}$. Our target is to write $\mathbf{Z}$ as a function of $\mathbf{E}$ and $\mathbf{B}$, but free of $\mathbf{k}$ and $\mathbf{v_g}$. Thus we want $\mathbf{Z}$, such that

$$f(\mathbf{k}) \quad = \quad f^0(\mathbf{k} - \frac{q\tau}{\hbar}\mathbf{Z}) \tag{A.29}$$

Hence,

$$\delta f \quad = \quad -\frac{q\tau}{\hbar}\mathbf{Z}.\nabla_k f^0 \tag{A.30}$$

We now use the assumed form (eqn A.30) with eqns A.5 and A.6. This gives:

$$\frac{q}{\hbar}(\mathbf{v} \times \mathbf{B}).\left(\nabla_k f^0 + \nabla_k \delta f\right) + \frac{q}{\hbar}\mathbf{E}.\nabla_k f^0 = -\frac{\delta f}{\tau} \tag{A.31}$$

We already know that $\nabla_k f^0$ points along $\mathbf{v_g}$ and hence the first term in eqn A.31 gives zero. This leaves us with

$$\frac{q}{\hbar}(\mathbf{v} \times \mathbf{B}).\nabla_k \delta f + \frac{q}{\hbar}\mathbf{E}.\nabla_k f^0 = \frac{q}{\hbar}\mathbf{Z}.\nabla_k f^0 \tag{A.32}$$

Now we need to calculate $\nabla_k \delta f$.

$$\begin{aligned} \nabla_k \delta f &= \nabla_k \frac{q\tau}{\hbar}\mathbf{Z}.\nabla_k f^0 \\ &= \frac{q\tau}{\hbar}\nabla_k \left(-\beta f^0(1-f^0)\mathbf{Z}.\hbar\mathbf{v_g}\right) \\ &= -\beta q\tau \left((1-f^0)(\mathbf{Z}.\mathbf{v_g})\nabla_k f^0 + f^0(\mathbf{Z}.\mathbf{v_g})\nabla_k(1-f^0) + f^0(1-f^0)\nabla_k(\mathbf{Z}.\mathbf{v_g})\right) \end{aligned} \tag{A.33}$$

Once again the first two terms in the RHS of A.33 will give zero when dotted with $\mathbf{v} \times \mathbf{B}$ as they are $\propto \mathbf{v_g}$. The only term left is

$$\nabla_k \mathbf{Z}.\mathbf{v_g} = \nabla_k \mathbf{Z}.\frac{\hbar\mathbf{k} - q\mathbf{A}}{m} = \frac{\hbar}{m}\mathbf{Z} \tag{A.34}$$

In eqn A.34, $\mathbf{A}$ denotes the vector potential of the magnetic field, $\mathbf{v_g}$ is related to the canonical momentum in presence of a magnetic field in the usual way. Combining eqns A.33 and A.34 we can write:

$$(\mathbf{v}\times\mathbf{B}).\nabla_k\delta f = -\beta q\tau f^0(1-f^0)(\mathbf{v}\times\mathbf{B}).\frac{\hbar}{m}\mathbf{Z} \tag{A.35}$$

So eqn A.32 now simplifies to:

$$\begin{aligned}
& -\frac{\hbar}{m}\beta q\tau f^0(1-f^0)(\mathbf{v}\times\mathbf{B}).\mathbf{Z} + (\mathbf{E}-\mathbf{Z}).\nabla_k f^0 = 0 \\
\therefore\quad & -\frac{\hbar}{m}\beta q\tau f^0(1-f^0)(\mathbf{v}\times\mathbf{B}).\mathbf{Z} + (\mathbf{E}-\mathbf{Z})\beta f^0(1-f^0)\hbar\mathbf{v_g} = 0 \\
\therefore\quad & \frac{q\tau}{m}(\mathbf{v_g}\times\mathbf{B}).\mathbf{Z} + (\mathbf{E}-\mathbf{Z}).\mathbf{v_g} = 0 \\
\therefore\quad & \frac{q\tau}{m}(\mathbf{B}\times\mathbf{Z}).\mathbf{v_g} + (\mathbf{E}-\mathbf{Z}).\mathbf{v_g} = 0 \\
\therefore\quad & \mathbf{E} = \mathbf{Z} - \frac{q\tau}{m}\mathbf{B}\times\mathbf{Z}
\end{aligned} \tag{A.36}$$

We call $\mathbf{Z}$ as the Hall vector. When both $\mathbf{E}$ and $\mathbf{B}$ fields are present, this quantity in some way, "replaces" the electric field in the transport equation. But we still need to express $\mathbf{Z}$ explicitly in terms of $\mathbf{E}$ and $\mathbf{B}$, with $\mu = q\tau/m$. The proof is left as an exercise.

$$\mathbf{Z} = \frac{\mathbf{E} + \mu\mathbf{B}\times\mathbf{E} + \mu^2(\mathbf{B}.\mathbf{E})\mathbf{B}}{1+\mu^2B^2} \tag{A.37}$$

PROBLEM: If $\mathbf{E} = \mathbf{Z} - \mathbf{A}\times\mathbf{Z}$, then show that

$$\mathbf{Z} = \frac{\mathbf{E} + \mathbf{A}\times\mathbf{E} + (\mathbf{A}.\mathbf{E})\mathbf{A}}{1+A^2}$$

Hint : Try $\mathbf{A}\times\mathbf{E}$ and $\mathbf{A}.\mathbf{E}$

Recall that the relation current with electric field was reduced to a simple (Drude) form for simple parabolic $E(\mathbf{k})$ dispersion. Following this we then write the expression for current in presence of a magnetic field by replacing $\mathbf{E}$ by $\mathbf{Z}$:

$$\mathbf{j} = nq\mu\mathbf{Z} = \sigma_0\mathbf{Z} \tag{A.38}$$

A very general expression with arbitrary $\mathbf{E}$ and $\mathbf{B}$ can be written, but is not very useful. Rather, we consider a situation where the magnetic field points along $\hat{\mathbf{z}}$, and the electric field is in the $x-y$ plane. So we have :

$$\begin{aligned}
\mathbf{E} &= E_x\hat{\mathbf{x}} + E_y\hat{\mathbf{y}} \\
\mathbf{B} &= B_0\hat{\mathbf{z}}
\end{aligned} \tag{A.39}$$

and hence:

$$\begin{aligned}
Z_x &= \frac{E_x - \mu B_0 E_y}{1+\mu^2{B_0}^2} \\
Z_y &= \frac{E_y + \mu B_0 E_x}{1+\mu^2{B_0}^2}
\end{aligned}$$

Eqn A.38 then can be written out in 2×2 matrix form as :

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = \frac{\sigma_0}{1+\mu^2{B_0}^2}\begin{pmatrix} 1 & -\mu B_0 \\ \mu B_0 & 1 \end{pmatrix}\begin{pmatrix} E_x \\ E_y \end{pmatrix} \tag{A.40}$$

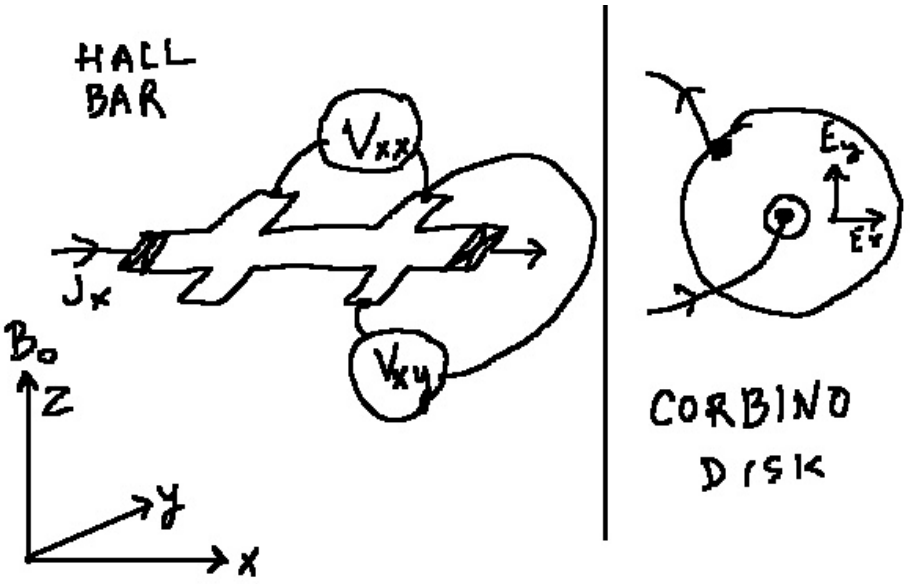


Figure A.5: Two device geometries commonly used in experiments with 2-dimensional systems

Eqn A.40 can be inverted to give the resistivity matrix such that:

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} \rho_0 & \frac{B_0}{nq} \\ -\frac{B_0}{nq} & \rho_0 \end{pmatrix} \begin{pmatrix} j_x \\ j_y \end{pmatrix} \tag{A.41}$$

where we have written ρ_0 for $1/\sigma_0$.

How do we relate A.41 to experimental situations? Consider a rectangular block in the xy plane, with the current injecting contacts placed as shown. Sufficiently away from the contacts, the current component j_y must vanish, because there are no current sourcing/withdrawing contacts on the long sides. This allows us to interpret the ratio E_x/j_x as the longitudinal voltage drop and E_y/j_x as the Hall (transverse) voltage. The off-diagonal terms are linear in B and offers the most common way of measuring the electron density in a 2-dimensional system.

PROBLEM: Consider the "Corbino-disk" geometry shown in figure. Current flow is between the inner (central) contact and the outer (circumferential) contact. Show by symmetry arguments that one of the components of the electric field (E_y in figure) must be zero. Can you roughly sketch the current flow paths from the center to the circumference?

It is important to understand that resistance or conductance can no longer be specified by a single number in presence of a magnetic field. They must be understood in a matrix sense. In fact by inverting the resistivity matrix you can easily show that in a magnetic field both σ_{xx} and ρ_{xx} can be simultaneously zero, which appears counter-intuitive at first glance - but there is no contradiction in it.

PROBLEM: Invert the matrix in the equation **??**. Call this the conductivity matrix whose elements are σ_{ij}. What will be the value of ρ_{xx} if $\sigma_{xx} = 0$, when no magnetic field is present? How would your answer be modified when a finite strong magnetic field is present?

A.6 Moments of the transport equation: Continuity & Drift-diffusion

Taking the moments of a differential equation means multiplying both sides of the equation with some function and integrating over all states/ space. How does that help? The integration "removes" some variable and results in a simpler looking equation. Of course the "simpler" equation is no longer as detailed or informative as the original one - but sometimes we may need focus on a broad feature while removing some details. The BTE refers to the distribution function which is not always possible (or necessary) to know. We show cases where focussing on quantities averaged over the distribution $f(\mathbf{r},\mathbf{k})$ is immensely useful. The simple 1D version reads:

A.6.1 Continuity equation

This is an expected result of course. However it is useful to show the process. The distribution function $f(\mathbf{r},\mathbf{k},t)$ is written as f for simplicity:

$$\frac{\partial f}{\partial t}+\frac{q}{\hbar}\left(\mathbf{E}+\mathbf{v}\times\mathbf{B}\right).\nabla_{\mathbf{k}}f+\mathbf{v}.\nabla_{\mathbf{r}}f=\left.\frac{df}{dt}\right|_{collision} \tag{A.42}$$

We integrate/sum over all $\mathbf{k}$ states. The first term in LHS gives

$$\int\frac{d^3\mathbf{k}}{(2\pi)^3}\frac{\partial f}{\partial t}=\frac{\partial n(\mathbf{r})}{\partial t} \tag{A.43}$$

where $n(\mathbf{r})$ is the conventional particle density at $\mathbf{r}$.
To evaluate the second term in LHS use the following vector identity. f is a scalar and $\mathbf{A}$ is a vector.

$$\mathbf{A}.\nabla f=\nabla.f\mathbf{A}-f\nabla.\mathbf{A} \tag{A.44}$$

This implies:

$$\left(\mathbf{E}+\mathbf{v}\times\mathbf{B}\right).\nabla_{\mathbf{k}}f=\nabla_{\mathbf{k}}.f\left(\mathbf{E}+\mathbf{v}\times\mathbf{B}\right)-f\nabla_{\mathbf{k}}.\left(\mathbf{E}+\mathbf{v}\times\mathbf{B}\right) \tag{A.45}$$

The first term can be converted to a surface integral. Because the surface will grow as k^2, the fields are finite and the Maxwell-Boltzmann and Fermi distributions go to zero as $\sim e^{-k^2}$, this integral will vanish. The next term is also zero, provided we interpret $\mathbf{v}$ as the group velocity. The steps are left as an exercise. Here $\mathcal{E}$ denotes energy.

$$\begin{aligned}\nabla_{\mathbf{k}}.\left(\mathbf{E}+\mathbf{v}\times\mathbf{B}\right) &= \frac{\partial E_i}{\partial k_i}+\varepsilon_{ijk}\frac{\partial}{\partial k_i}\frac{\partial\mathcal{E}}{\partial k_j}B_k\\ &= 0+\varepsilon_{ijk}\frac{\partial^2\mathcal{E}}{\partial k_i k_j}B_k\\ &= 0+0\end{aligned} \tag{A.46}$$

Now, the third term in LHS of A.42 is

$$\begin{aligned}\int\frac{d^3\mathbf{k}}{(2\pi)^3}\mathbf{v}.\nabla_{\mathbf{r}}f &= \nabla_{\mathbf{r}}.\int\frac{d^3\mathbf{k}}{(2\pi)^3}\mathbf{v}f\\ &= \nabla_{\mathbf{r}}.n(\mathbf{r})\langle\mathbf{v}\rangle\end{aligned} \tag{A.47}$$

The RHS of A.42 must give zero when integrated over all *k-space* because the particles which are scattered out of a certain volume must be appearing in some other volume.
Adding A.43, A.46, A.47 gives

$$\frac{\partial n(\mathbf{r})}{\partial t}+\nabla_{\mathbf{r}}.n(\mathbf{r})\langle\mathbf{v}\rangle=0 \tag{A.48}$$

A.6.2 Drift-diffusion equation

We multiply both sides of the BTE by velocity (or momentum) and integrating over all states. Obviously the RHS will give the general expression for current. The LHS will be formed of two or three terms with distinct physical meaning. This is very extensively used in describing electronic transport in metals and semiconductors. There are important assumptions which go into it and one needs to be aware of those! The calculation is somewhat long, so we first do a simplified version in 1D with an electric field only. Then we will show how to generalise it to 3D with both electric and magnetic field.

$$\frac{\partial f}{\partial t} + \frac{q}{\hbar}E\frac{\partial f}{\partial k} + v\frac{\partial f}{\partial x} = -\frac{f - f^0}{\tau} \tag{A.49}$$

multiply by v and integrate over all k. The first term in LHS of A.49 gives:

$$\begin{aligned} \int \frac{dk}{2\pi} v \frac{\partial f}{\partial t} &= \int \frac{dk}{2\pi} \frac{\partial}{\partial t} fv \\ &= \frac{\partial}{\partial t} n\langle v\rangle \end{aligned} \tag{A.50}$$

The second term in LHS of A.49 gives with $v = \dfrac{\hbar k}{m}$:

$$\begin{aligned} \int \frac{dk}{2\pi} v \frac{\partial f}{\partial k} &= \int \frac{dk}{2\pi} \left(\frac{\partial}{\partial k} fv - f\frac{\partial v}{\partial k} \right) \\ &= fv|_{-\infty}^{\infty} - \frac{\hbar}{m} n \end{aligned} \tag{A.51}$$

The third term in LHS of A.49 gives:

$$\begin{aligned} \int \frac{dk}{2\pi} v^2 \frac{\partial f}{\partial x} &= \frac{\partial}{\partial x} \int \frac{dk}{2\pi} v^2 f \\ &= \frac{\partial}{\partial x} n(x) \langle v^2 \rangle \\ &= \frac{\partial}{\partial x} n(x) \left\langle \frac{2E}{m} \right\rangle \\ &= \frac{kT}{m} \frac{\partial n(x)}{\partial x} \end{aligned} \tag{A.52}$$

Notice the use of thermal average kinetic energy in the last step. This will ultimately lead to a relation between mobility and diffusion constant.
The RHS term :

$$\begin{aligned} -\int \frac{dk}{2\pi} v \frac{f - f^0}{\tau} &= \frac{1}{\tau} \int \frac{dk}{2\pi} fv \\ &= \frac{1}{\tau} n\langle v\rangle \end{aligned} \tag{A.53}$$

Now we can put the last four result together, multiply with τ allover and write $J = nq\langle v\rangle$ for the electric current:

$$\tau \frac{\partial}{\partial t} n\langle v\rangle + n\langle v\rangle + \underbrace{\frac{q\tau}{\hbar} E \left(-\frac{\hbar}{m} n \right)}_{\text{drift:}\mu = \frac{q\tau}{m}} + \underbrace{\tau \frac{kT}{m} \frac{\partial n(x)}{\partial x}}_{\text{diffusion:}\propto -D\frac{\partial n(x)}{\partial x}} = 0 \tag{A.54}$$

Notice that the ratio of the drift mobility to diffusion constant is $\dfrac{kT}{q}$, called the Einstein relation. This is correct for a classical distribution only. Notice how $\hbar$ has disappeared, another indication that the result is essentially classical. The relation between drift and diffusion components would be different if full Fermi-Dirac distribution used. However at room temperatures in most devices this holds very well for motion of electrons/holes in a band.

Drift diffusion in 3D

PROBLEM: Now let us remove the simplifying assumptions and take the moment of the BTE after multiplying with $\mathbf{v}$. We need to work with

$$\underbrace{\int \frac{d^3\mathbf{k}}{(2\pi)^3}\mathbf{v}\frac{\partial f}{\partial t}}_{\text{LHS 1}} + \underbrace{\frac{q}{\hbar}\int \frac{d^3\mathbf{k}}{(2\pi)^3}\mathbf{v}\left(\mathbf{E}+\mathbf{v}\times\mathbf{B}\right).\nabla_{\mathbf{k}}f}_{\text{LHS 2}} + \underbrace{\int \frac{d^3\mathbf{k}}{(2\pi)^3}\mathbf{v}\mathbf{v}.\nabla_{\mathbf{r}}f}_{\text{LHS 3}} = -\int \frac{d^3\mathbf{k}}{(2\pi)^3}\mathbf{v}\frac{f-f^0}{\tau} \tag{A.55}$$

Prove the following results. The algebra can be done using the ϵ-δ notation for handling indices of vectors and tensors. The relation between velocity and the wavevector is $m\mathbf{v} = \hbar\mathbf{k} - q\mathbf{A}$

1. LHS 1 : This gives

$$\frac{\partial}{\partial t}n(\mathbf{r})\langle\mathbf{v}\rangle$$

2. LHS 2 : Notice the occurrence of the *averaged velocity* in the Lorentz term. The calculation is somewhat non-trivial. Do it carefully! You should get

$$-\frac{q}{m}n(\mathbf{r})\left(\mathbf{E}+\langle\mathbf{v}\rangle\times\mathbf{B}\right)$$

3. LHS 3 : The diffusion term requires averaging over the distribution. You should get

$$\nabla_{\mathbf{r}}.n(\mathbf{r})\langle v_i v_j\rangle$$

   For Maxwell-Boltzmann distribution $\langle v_i v_j\rangle = \frac{kT}{m}\delta_{ij}$

4. RHS : This gives the current term

$$-n(\mathbf{r})\frac{\langle\mathbf{v}\rangle}{\tau}$$

Adding all the results will give the drift diffusion relation.

Appendix B

Some facts about common semiconductors

B.1 The direct lattice and reciprocal lattice

The common semiconductors like Si, Ge, GaAs, InAs crystallise in the FCC lattice with a two atom basis. As far as we understand this is a happy coincidence! A few facts are very useful to know.

The cubic unit cell has 8 atoms in it, each of the four fcc lattice points contributing two atoms. The second atom is displaced along the diagonal of the cube by $\frac{1}{4}^{\text{th}}$ of the cubes diagonal. So we have for the basis in terms of the cubic lattice vectors $\mathbf{a}_i$

$$\begin{aligned}
\mathbf{d}_1 &= (0,0,0) \\
\mathbf{d}_2 &= \left(\frac{1}{2},\frac{1}{2},0\right) \\
\mathbf{d}_3 &= \left(0,\frac{1}{2},\frac{1}{2}\right) \\
\mathbf{d}_4 &= \left(\frac{1}{2},0,\frac{1}{2}\right) \\
\mathbf{d}_5 &= \left(\frac{1}{4},\frac{1}{4},\frac{1}{4}\right) \\
\mathbf{d}_6 &= \left(\frac{3}{4},\frac{3}{4},\frac{1}{4}\right) \\
\mathbf{d}_7 &= \left(\frac{1}{4},\frac{3}{4},\frac{3}{4}\right) \\
\mathbf{d}_8 &= \left(\frac{3}{4},\frac{1}{4},\frac{3}{4}\right)
\end{aligned} \tag{B.1}$$

Another way to describe the diamond lattice is to use the FCC unit vectors, (still using the cubic cell to define our unit of length).

$$\begin{aligned}
\mathbf{a}_1 &= \frac{a}{2}(\hat{\mathbf{x}}+\hat{\mathbf{y}}) \\
\mathbf{a}_2 &= \frac{a}{2}(\hat{\mathbf{y}}+\hat{\mathbf{z}}) \\
\mathbf{a}_3 &= \frac{a}{2}(\hat{\mathbf{z}}+\hat{\mathbf{x}})
\end{aligned} \tag{B.2}$$

the atoms are at:

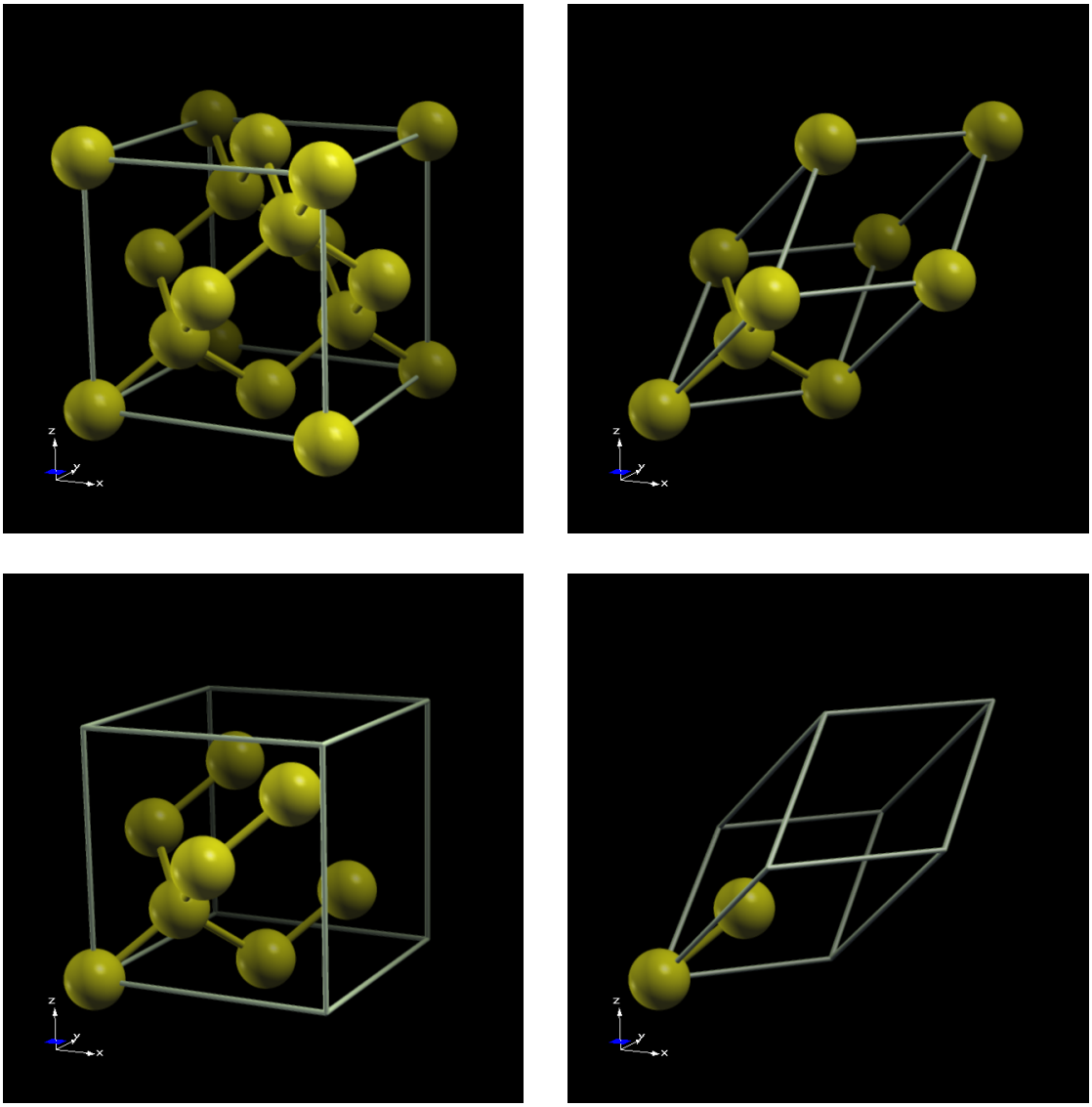


Figure B.1: The diamond lattice. On the left column, we have used the cubic unit cell, on the right we use the FCC primitive vectors. The cubic lattice constant $a = 3.57$Å for diamond, for Si $a = 5.43$Å, for Ge $a = 5.66$Å, for GaAs $a = 5.65$Å

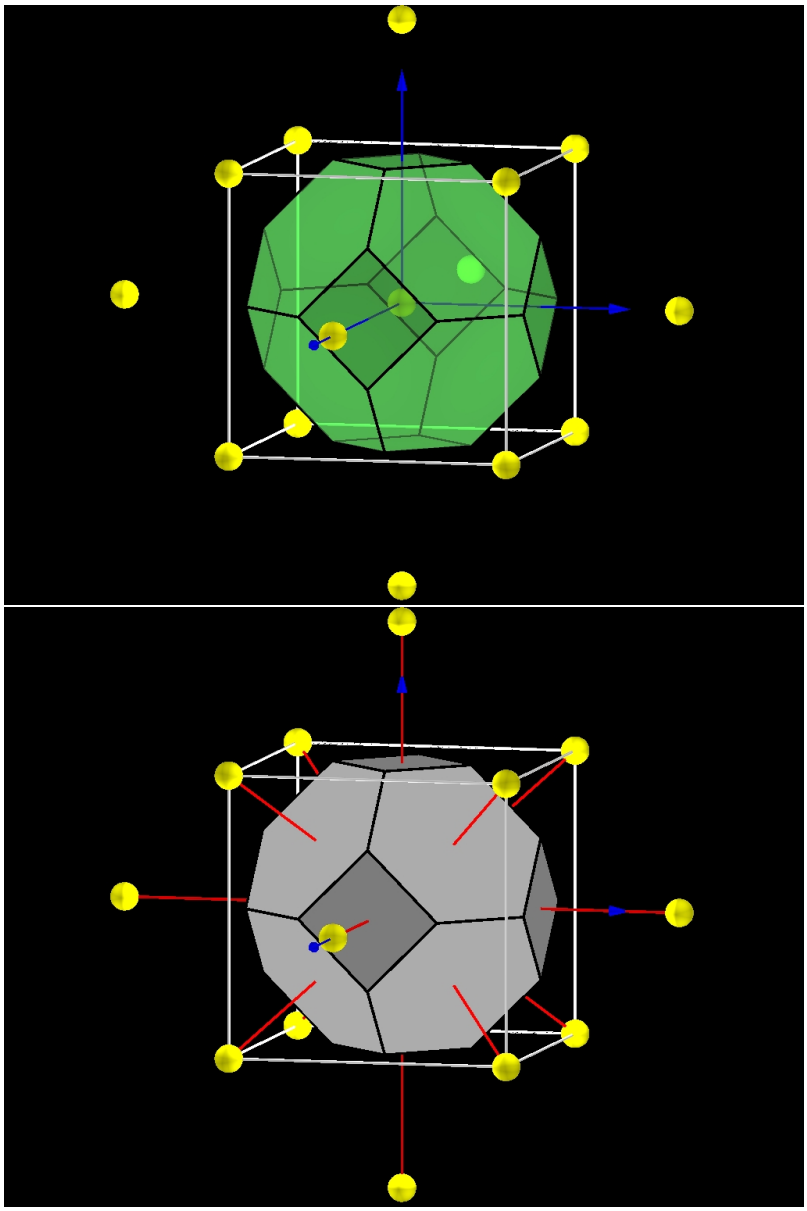


Figure B.2: BCC Wigner-Seitz cell. This is also the Brillouin zone for the FCC lattice.

$$\begin{aligned} \mathbf{d}_1 &= (0,0,0) \\ \mathbf{d}_2 &= a\left(\frac{1}{4},\frac{1}{4},\frac{1}{4}\right) \end{aligned} \tag{B.3}$$

Then the reciprocal lattice vectors form a BCC lattice

$$\begin{aligned} \mathbf{b}_1 &= \frac{2\pi}{a}(-\hat{\mathbf{x}}+\hat{\mathbf{y}}+\hat{\mathbf{z}}) \\ \mathbf{b}_2 &= \frac{2\pi}{a}(\hat{\mathbf{x}}-\hat{\mathbf{y}}+\hat{\mathbf{z}}) \\ \mathbf{b}_3 &= \frac{2\pi}{a}(\hat{\mathbf{x}}+\hat{\mathbf{y}}-\hat{\mathbf{z}}) \end{aligned} \tag{B.4}$$

Now it is very useful to remember the shape of the Wigner Seitz cell of the FCC and BCC lattice. The Wigner Seitz cell of the BCC lattice is essentially the *first Brillouin zone* of the FCC lattice.

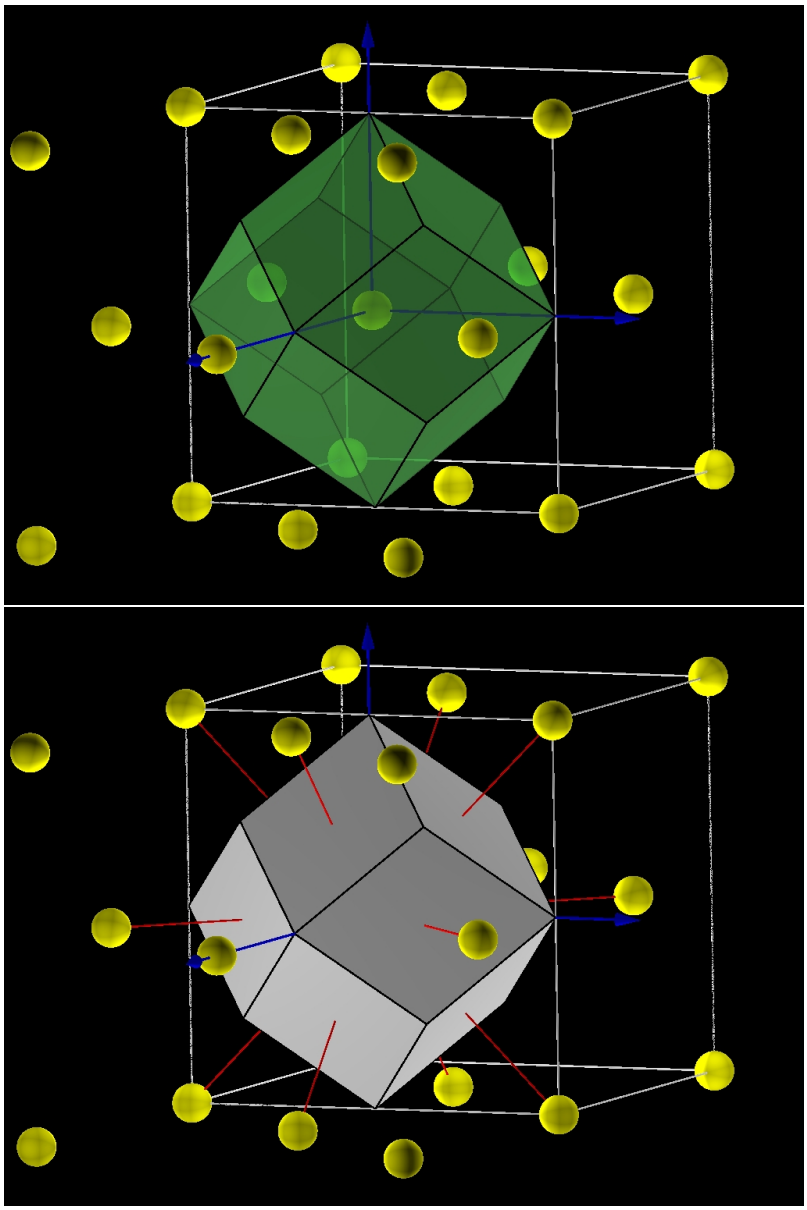


Figure B.3: FCC Wigner-Seitz cell, and hence the Brillouin zone for the BCC lattice.

B.2 The special points

Certain directions in the FBZ are denoted by specific letters. It is useful to know this because the standard band structure diagrams use the notation. Below we give the standard notation for the square, cube, FCC and BCC Brillouin zones.

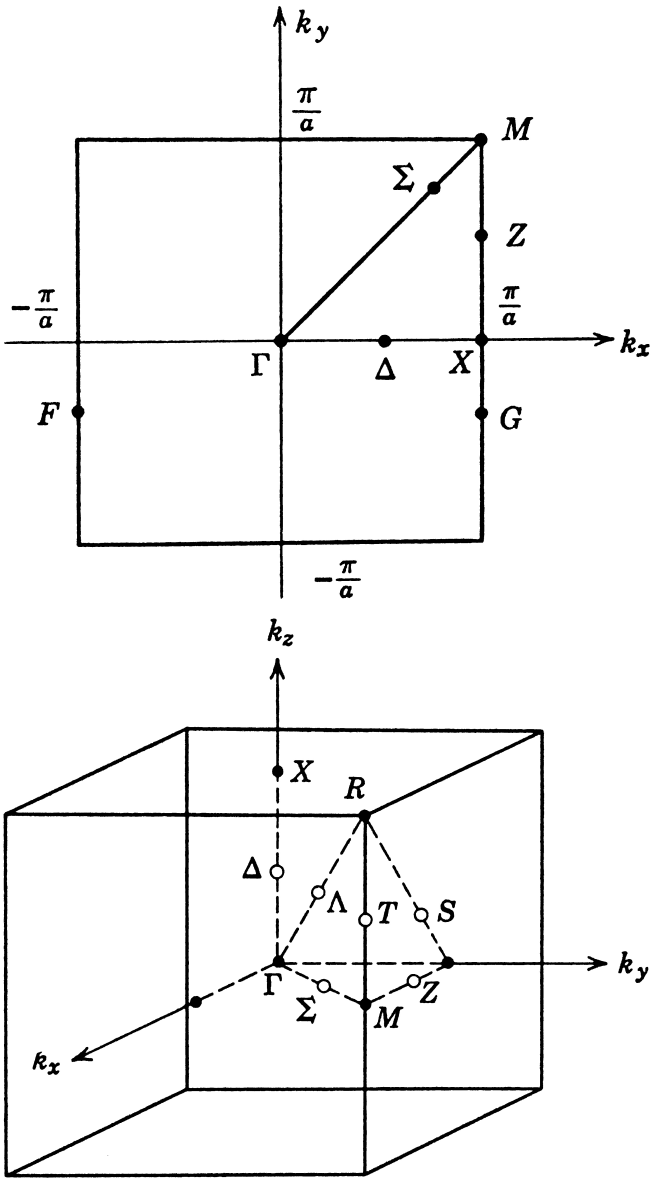


Figure B.4: The notation used to mark directions in the $\boldsymbol{k}$ space of the square and simple cubic lattice . The pictures are taken from C. Kittel, *Quantum Theory of Solids*

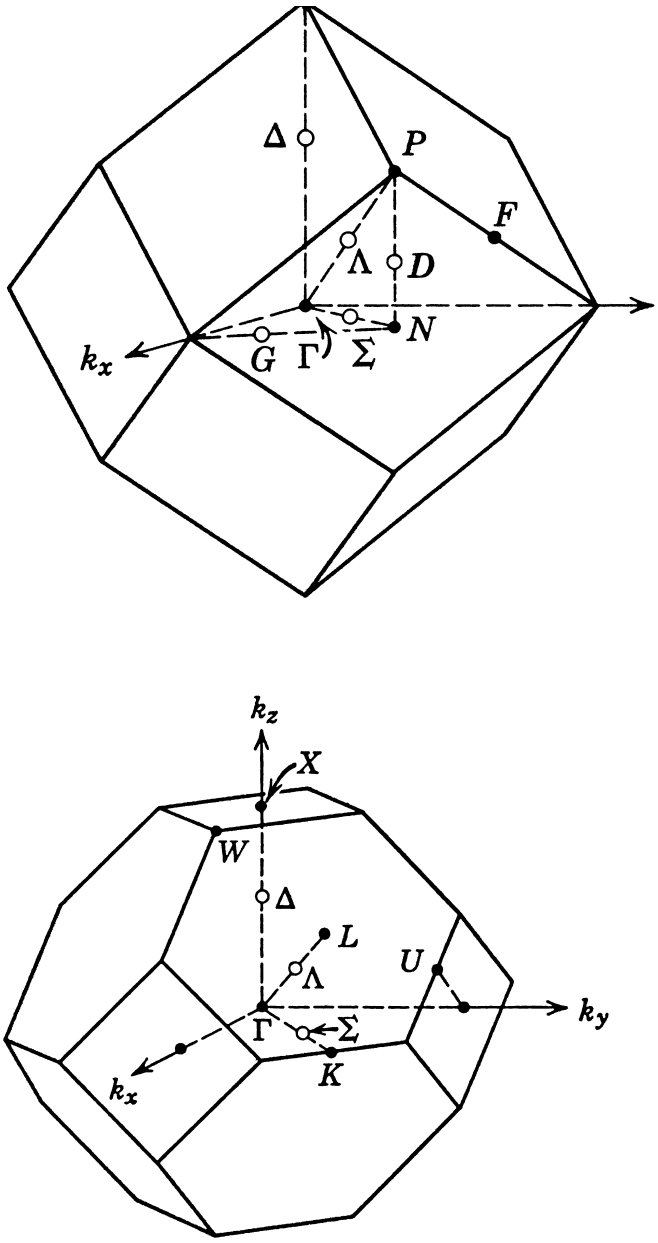


Figure B.5: The notation used to mark directions in the $\boldsymbol{k}$ space of the FCC and BCC lattice . The pictures are taken from C. Kittel, *Quantum Theory of Solids*

B.3 Band Structures

We have discussed the generic nature of band structure near the Γ point before. We now discuss some more features with reference to common semiconductors.

- Compare GaAs (direct bandgap) with Si (indirect gap). Notice where the carriers are for realistic densities. Near $k = 0$ the s character of the wavefunction is dominant and the equal energy contours are spheres. So the dispersion is simply

$$E(k) = \underbrace{\frac{\hbar^2}{2m^*}\left(k_x^2 + k_y^2 + k_z^2\right)}_{\substack{\text{isotropic} \\ \text{effective mass,} \\ m^* = 0.067\ m_0}} \tag{B.5}$$

  But for Si, the minima is at $\frac{2\pi}{a}(0, 0, 0.85)$ and six other equivalent places. This is quite far from the zone center and the p character of the wavefunction would be significant. Near the minima the dispersion is not isotropic. Near $(0, 0, k_{min})$

$$E(k) = \underbrace{\frac{\hbar^2}{2m_t}\left(k_x^2 + k_y^2\right)}_{\substack{\text{transverse} \\ \text{effective mass,} \\ m_t = 0.19\ m_0}} + \underbrace{\frac{\hbar^2}{2m_l}\left(k_z - k_{min}\right)^2}_{\substack{\text{longitudinal} \\ \text{effective mass} \\ m_l = 0.92\ m_0}} \tag{B.6}$$

  Due to symmetry the directions [100], [$\bar{1}$00], [010], [0$\bar{1}$0], [001], [00$\bar{1}$] are equivalent.

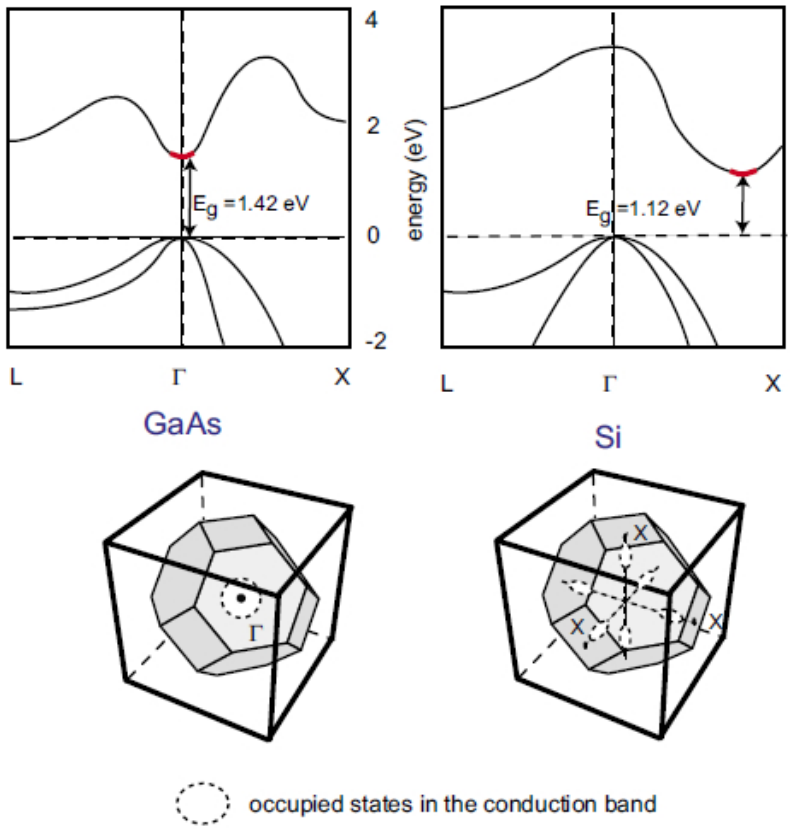


Figure B.6: The location of the band minima The pictures are taken from the book by C. Kittel

Figure B.7 shows simple schematics for some of the most common semiconductors.

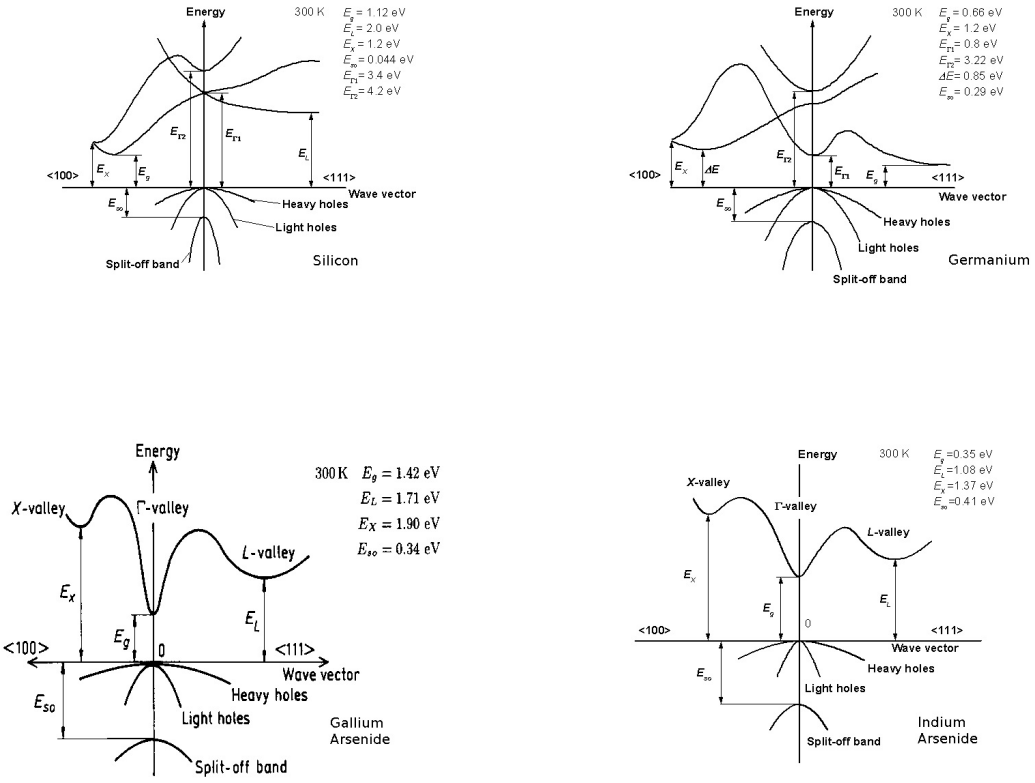


Figure B.7: Some parameters and schematic band structures of common semiconductors

Appendix C

Origin of the Spin-Orbit Coupling term

We need to begin from the Dirac equation that describes the dynamics of an electron. The relativistic energy momentum relation (m=rest mass) is

$$E = \sqrt{p^2c^2 + m^2c^4} \tag{C.1}$$

If we start with a quantum mechanical equation, then at the end the expectation values should satisfy the "classical" equation. Here classical means relativistic but not quantum mechanical. The usual correspondence of

$$E = \langle H \rangle \quad H \rightarrow i\hbar\frac{\partial}{\partial t} \tag{C.2}$$

$$\boldsymbol{p} = \langle \hat{\boldsymbol{p}} \rangle \quad \hat{\boldsymbol{p}} \rightarrow \frac{\hbar}{i}\boldsymbol{\nabla} \tag{C.3}$$

needs more work due to the presence of the square root in eqnC.1.

Also in a relativistic equation we expect x and ct to appear in a symmetrical way. This is not true for Schrödinger equation, since we have ∇^2 but $\dfrac{\partial}{\partial t}$

The wave equation is symmetric in x and ct but this is a second order equation. Can we get a first order equation that would be invariant under Lorentz transformation and would lead to eqnC.1? This is the key question.

C.1 How to take the square root?

Suppose

$$p^2c^2 + m^2c^4 = \left(c\alpha_x p_x + c\alpha_y p_y + c\alpha_z p_z + \beta mc^2\right)^2 \tag{C.4}$$

$$= \left(c\boldsymbol{\alpha}.\boldsymbol{p} + \beta mc^2\right)^2 \tag{C.5}$$

Now what must α and β be like? Whatever they are, they need to satisfy certain properties. We equate like powers of p_x,p_y,p_z to get that.
Continuing after the previous equation

$$\begin{aligned} p^2c^2 + m^2c^4 &= \left(c^2{\alpha_x}^2{p_x}^2 + c^2{\alpha_y}^2{p_y}^2 + c^2{\alpha_z}^2{p_z}^2\right) + \beta^2m^2c^4 \\ &+ c^2p_xp_y\left(\alpha_x\alpha_y + \alpha_y\alpha_x\right) + c^2p_yp_z\left(\alpha_y\alpha_z + \alpha_z\alpha_y\right) \\ &+ \text{similar cross terms.....} \\ &+ mc^3p_x\left(\alpha_x\beta + \beta\alpha_x\right) \\ &+ \text{similar cross terms.....} \end{aligned} \tag{C.6}$$

We need

$$\alpha_x{}^2 = \alpha_y{}^2 = \alpha_z{}^2 = \beta^2 = I \tag{C.7}$$

$$\alpha_x\alpha_y + \alpha_y\alpha_x = 0 \quad \text{and all other permutations} \tag{C.8}$$

$$\beta\alpha_x + \alpha_x\beta = 0 \quad \text{and all other permutations} \tag{C.9}$$

Simple numbers cannot have these properties but matrices can. We need four such matrices. The 2×2 Pauli spin matrices satisfy similar properties, but there are only three of them. We should thus look for 4×4 matrices. Why did we not try 3×3 matrices? Let us make this point clear.

The set od properties lead to two important conclusions about the eigenvalues λ_i of α_x,α_y,α_z,β. These must have

$$\lambda_i = \pm 1 \tag{C.10}$$

$$\sum \lambda_i = 0 \tag{C.11}$$

This is how we prove these two very important assertions. Consider $\alpha_x\alpha_y$

$$\begin{aligned}
\alpha_x\alpha_y &= -\alpha_y\alpha_x \\
\therefore \alpha_x\alpha_x\alpha_y &= -\alpha_x\alpha_y\alpha_x \\
\text{Tr } \alpha_x{}^2\alpha_y &= -\text{Tr } \alpha_x\alpha_y\alpha_x \\
\text{Tr } \alpha_y &= -\text{Tr } \alpha_x\alpha_x\alpha_y \quad \text{cyclic permutation} \\
&= -\text{Tr } \alpha_y
\end{aligned}$$

Same process applies to all the matrix pairs. These must be *traceless*.
We then assume S is a matrix that diagonalises α_x. We know $\alpha_x{}^2 = I$.

$$\begin{aligned}
S^{-1}\alpha_x S &= D \\
S^{-1}\alpha_x S S^{-1}\alpha_x S &= D^2 \\
S^{-1}\alpha_x{}^2 S &= D^2 \\
&= I \\
\therefore D_{ii} &= \pm 1
\end{aligned} \tag{C.12}$$

Clearly this is possible only in even dimension where pairs of ±1 can cancel each other. We would thus need 4×4 matrices. The following is a possible choice (obviously not unique, since any similarity transformation on these will also work)

$$\alpha_i = \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix} \tag{C.13}$$

$$\beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \tag{C.14}$$

When both electric and magnetic fields are present we would have $\boldsymbol{P} = \boldsymbol{p} + q\boldsymbol{A}$, replacing the momentum operator. Also since the matrices still retain some 2×2 characteristic, we would write the column vector for the wave function as $\Psi = \begin{pmatrix} \chi \\ \Phi \end{pmatrix}$ where χ and Φ are both two component objects that Pauli matrices can act on.

$$H\Psi = \left[c\boldsymbol{\alpha}.(\boldsymbol{p} - q\boldsymbol{A}) + \beta m^2c^4\right]\Psi + qV\Psi \tag{C.15}$$

So

$$c\begin{pmatrix} \boldsymbol{\sigma}.\boldsymbol{p} & 0 \\ 0 & \boldsymbol{\sigma}.\boldsymbol{p} \end{pmatrix}\begin{pmatrix} \chi \\ \Phi \end{pmatrix} + qV\begin{pmatrix} \chi \\ \Phi \end{pmatrix} = \begin{pmatrix} E - mc^2 & 0 \\ 0 & E + mc^2 \end{pmatrix}\begin{pmatrix} \chi \\ \Phi \end{pmatrix} \tag{C.16}$$

$$\therefore \left(E - qV - mc^2\right)\chi - c\boldsymbol{\sigma}.\boldsymbol{p}\Phi = 0 \tag{C.17}$$

$$\left(E - qV + mc^2\right)\Phi - c\boldsymbol{\sigma}.\boldsymbol{p}\chi = 0 \tag{C.18}$$

The last two equations can be used to decouple χ and Φ and also used to estimate the relative magnitudes of the "upper" and "lower" components.

$$\begin{aligned}\Phi &= \frac{c\boldsymbol{\sigma}.p}{E-qV+mc^2} \\ &\approx \frac{cmv}{E+mc^2-qV}\chi \\ &= \frac{v}{2c}\chi\end{aligned} \tag{C.19}$$

Electron has a rest mass of $mc^2 = 0.5$Mev and the atomic potentials would not exceed $10-100$eV for the kind of systems we are talking about. For the inner core electrons of heavy atoms this will not hold - but they are not the ones we are concerned with here. So for $\frac{v}{c} << 1$, the upper χ component is dominant and we will try to write the equation for χ with a "small" correction coming from the lower component. This is the way we develop the "non-relativistic" approximation.

- We will call $E - mc^2 = E_S$, reminding us that this is the energy that appears in the Schrödinger equation. The rest energy is subtracted from the "total" energy which the Dirac equation deals with by default.

- We then substitute
$$\Phi = \frac{c\boldsymbol{\sigma}.p}{E-qV+mc^2}\chi \tag{C.20}$$
in eqn. C.17 to do the decoupling. Notice that the decoupling comes at the cost of converting a first order equation to second order, because two momentum operators will now act on the 2-component spinor χ

- The decoupled equation is:
$$\left(E-qV-mc^2\right)\chi - c\boldsymbol{\sigma}.\boldsymbol{P}\left(\frac{1}{E-qV+mc^2}\right)c\boldsymbol{\sigma}.\boldsymbol{P}\chi = 0 \tag{C.21}$$
$$(E_S-qV)\chi - \boldsymbol{\sigma}.\boldsymbol{P}\frac{1}{2m}\underbrace{\left(1+\frac{E_S-qV}{2mc^2}\right)^{-1}}_{\text{expand}}\boldsymbol{\sigma}.\boldsymbol{P}\chi = 0 \tag{C.22}$$

- Retaining the first term only results in the following equation, that is free of c and starts looking like the Schrödinger equation..
$$\left[\frac{1}{2m}\left(\boldsymbol{\sigma}.\boldsymbol{P}\right)\left(\boldsymbol{\sigma}.\boldsymbol{P}\right)+qV\right]\chi = E_S\chi \tag{C.23}$$

- Now recall the following result for Pauli matrices where $\boldsymbol{A}$ and $\boldsymbol{B}$ are vectors
$$\begin{aligned}\left(\boldsymbol{\sigma}.\boldsymbol{A}\right)\left(\boldsymbol{\sigma}.\boldsymbol{B}\right) &= (\sigma_i A_i)(\sigma_j A_j) \\ &= \sum_{i\neq j}\sigma_i\sigma_j A_i B_j + {\sigma_i}^2 A_i B_i \\ &= i\varepsilon_{ijk}\sigma_k A_i B_j + I(\boldsymbol{A}.\boldsymbol{B}) \\ &= i\boldsymbol{\sigma}.\boldsymbol{A}\times\boldsymbol{B} + I(\boldsymbol{A}.\boldsymbol{B})\end{aligned}$$
But the components pf $\boldsymbol{P} = \boldsymbol{p} - q\boldsymbol{A}$ are such operators for which $\boldsymbol{P}\times\boldsymbol{P}\neq 0$, so $\left(\boldsymbol{\sigma}.\boldsymbol{p}\right)\left(\boldsymbol{\sigma}.\boldsymbol{p}\right) = \boldsymbol{p}^2$ but $\left(\boldsymbol{\sigma}.\boldsymbol{P}\right)\left(\boldsymbol{\sigma}.\boldsymbol{P}\right)\neq\boldsymbol{P}^2$

- $$\left(\boldsymbol{\sigma}.\boldsymbol{P}\right)\left(\boldsymbol{\sigma}.\boldsymbol{P}\right) = \boldsymbol{P}^2 + i\boldsymbol{\sigma}.\boldsymbol{P}\times\boldsymbol{P} \tag{C.24}$$

What is $\boldsymbol{P}\times\boldsymbol{P}$? The calculation is a little involved, but the result is quite remarkable. let this operator act on a *scalar* wavefunction ψ

$$\begin{aligned}
[\boldsymbol{P}\times\boldsymbol{P}\psi]_k &= \varepsilon_{ijk}\,(p_i - qA_i)\,(p_j - qA_j)\,\psi \\
&= \varepsilon_{ijk}\,[p_i(p_j - qA_j) - qA_i(p_j - qA_j)]\,\psi \\
&= \varepsilon_{ijk}\left[\frac{\hbar}{i}\frac{\partial}{\partial x_i}\left(\frac{\hbar}{i}\frac{\partial}{\partial x_j} - qA_j\right)\psi - qA_i\left(\frac{\hbar}{i}\frac{\partial}{\partial x_j} - qA_j\right)\psi\right] \\
&= \varepsilon_{ijk}\left[\underbrace{-\hbar^2\frac{\partial^2\psi}{\partial x_i\partial x_j}}_{\substack{\text{symmetric in ij}\\ \text{will sum to zero}}} - q\frac{\hbar}{i}\frac{\partial}{\partial x_i}A_j\psi - qA_i\frac{\hbar}{i}\frac{\partial\psi}{\partial x_j} + \underbrace{qA_iA_j\psi}_{\substack{\text{symmetric in ij}\\ \text{will sum to zero}}}\right] \\
&= \varepsilon_{ijk}\left[-q\frac{\hbar}{i}\left(A_j\frac{\partial\psi}{\partial x_i} + \psi\frac{\partial A_j}{\partial x_i}\right) - q\frac{\hbar}{i}A_i\frac{\partial\psi}{\partial x_j}\right] \\
&= \varepsilon_{ijk}\left[\underbrace{-q\frac{\hbar}{i}\left(A_j\frac{\partial\psi}{\partial x_i} + A_i\frac{\partial\psi}{\partial x_j}\right)}_{\substack{\text{symmetric in ij}\\ \text{will sum to zero}}} - q\frac{\hbar}{i}\frac{\partial A_j}{\partial x_i}\psi\right] \\
\therefore \frac{1}{2m}i\boldsymbol{\sigma}.\boldsymbol{P}\times\boldsymbol{P} &= \frac{q\hbar}{2m}\boldsymbol{\sigma}.\boldsymbol{\nabla}\times\boldsymbol{A}
\end{aligned} \tag{C.25}$$

The eqn C.23 takes the form

$$\left[\frac{1}{2m}(\boldsymbol{p} - q\boldsymbol{A})^2 + qV - \frac{q\hbar}{2m}\boldsymbol{\sigma}.\boldsymbol{B}\right]\chi = E_S\chi \tag{C.26}$$

C.2 The next order of approximation

Now we get back to eqn. C.22 and use the expansion

$$\left(1 + \frac{E_S - qV}{2mc^2}\right)^{-1} = 1 - \frac{E_S - qV}{2mc^2} + ... \tag{C.27}$$

Rewriting the eqn. C.22 slightly we have

$$\underbrace{\frac{1}{2m}\,(\boldsymbol{\sigma}.\boldsymbol{P})\,(\boldsymbol{\sigma}.\boldsymbol{P})\,\chi}_{\text{of order } \frac{v^2}{c^2}} - \underbrace{\frac{1}{2m}\frac{1}{2mc^2}\,(\boldsymbol{\sigma}.\boldsymbol{P})\,(E_S - qV)\,(\boldsymbol{\sigma}.\boldsymbol{P})\,\chi}_{\text{of order } \frac{v^4}{c^4}} = (E_S - qV)\chi \tag{C.28}$$

We can put the value of $(E_S - qV)\chi$ correct to leading order, by looking at the last equation on the left hand side, otherwise the unknown E_S cannot be pulled out. This means approximating the additional term as

$$\begin{aligned}
(E_S - qV)\boldsymbol{\sigma}.\boldsymbol{P}\chi &= \boldsymbol{\sigma}.\boldsymbol{P}(E_S - qV)\chi + \boldsymbol{\sigma}.\underbrace{[E_S - qV, \boldsymbol{P}]}_{\text{commutator}}\chi \\
&\approx (\boldsymbol{\sigma}.\boldsymbol{P})\frac{1}{2m}(\boldsymbol{\sigma}.\boldsymbol{P})(\boldsymbol{\sigma}.\boldsymbol{P}) + \boldsymbol{\sigma}.\underbrace{[\boldsymbol{p}, qV]}_{\boldsymbol{p}\text{ NOT }\boldsymbol{P}}\chi
\end{aligned}$$

Now substitute this in eqn. C.28 and use the rule for $(\boldsymbol{\sigma}.\boldsymbol{A})(\boldsymbol{\sigma}.\boldsymbol{B})$ along with $[\boldsymbol{p}, V(\boldsymbol{r})] = -i\hbar\boldsymbol{\nabla}V(\boldsymbol{r})$ and consider for the moment $\boldsymbol{A} = 0$ so that $\boldsymbol{P} = \boldsymbol{p}$, then we have

$$\left[\underbrace{\frac{\boldsymbol{p}^2}{2m}+qV}_{\text{Schrödinger}}-\underbrace{\frac{p^4}{8m^3c^2}}_{\text{relativistic corr}}-\underbrace{\frac{\hbar}{4m^2c^2}\boldsymbol{\sigma}.\boldsymbol{p}\times\boldsymbol{\nabla}qV}_{\text{spin-orbit}}-\underbrace{\frac{1}{4m^2c^2}\boldsymbol{p}.[\boldsymbol{p},qV]}_{\text{!NOT hermitian!}}\right]\chi=E_S\chi \tag{C.29}$$

The last "non-hermitian" bit requires some extra care. If a hamiltonian is non hermitian, time evolution of χ will not be unitary and integrated probability density would not be conserved. The apparent "error" has crept in because we have unwittingly dropped one term that should be considered in the calculation of total probability - this we will not go into. However the "corrected" version of the last term is called the Darwin term and it happens to affect only s atomic states. Many textbooks discuss this in detail - see R. Shankar (*Principles of Quantum Mechanics, Chapter 20*) for example.