5 / Bandstructure

In the last chapter we have seen how the atomic orbitals can be used as a basis to write down a matrix representation for the Hamiltonian operator, which can then be diagonalized to find the energy eigenvalues. In this chapter we will show how this approach can be used to calculate the energy eigenvalues for an infinite periodic solid. We will first use a few 'toy' examples to show that the bandstructure can be calculated by solving a matrix eigenvalue equation of the form

$$E(\varphi_0) = \left[h(\vec{k})\right](\varphi_0)$$

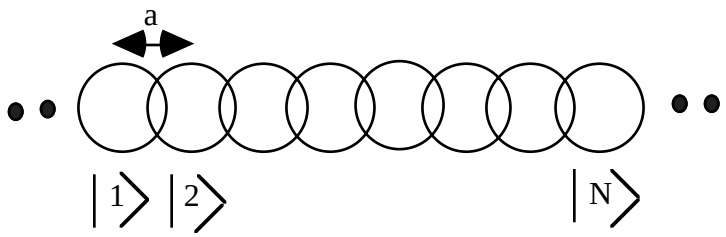
where $\left[h(\vec{k})\right] = \sum_m \left[H_{nm}\right] e^{i\vec{k}.(\vec{d}_m - \vec{d}_n)}$

The matrix $[h(\vec{k})]$ is (bxb) in size, 'b' being the number of basis orbitals per unit cell. The summation over 'm' runs over all neighboring unit cells (including itself) with which cell 'n' has any overlap (that is, for which H_{nm} is non-zero). The sum can be evaluated choosing any unit cell 'n' and the result will be the same because of the periodicity of the lattice. The bandstructure can be plotted out by finding the eigenvalues of the (bxb) matrix $[h(\vec{k})]$ for each value of $\vec{k}$ and it will have 'b' branches, one for each eigenvalue. This is the central result which we will first motivate using toy examples (Section 5.1), then formulate generally for periodic solids (Section 5.2), and then use to discuss the bandstructure of 3-D semiconductors (Section 5.3). We end in Section 5.4 with a discussion of spin-orbit coupling and its effect on the energy levels in semiconductors.

5.1. Toy examples

Let us start with a toy one-dimensional solid composed of 'N' atoms (see Fig.5.1.1).

Fig.5.1.1. A one-dimensional solid



If we use one orbital per atom we can write down a (NxN) Hamiltonian matrix using one orbital per atom:

$$\mathbf{H} = \begin{array}{c|cccccc} & |1\rangle & |2\rangle & \cdots & |N-1\rangle & |N\rangle \\ |1\rangle & E_0 & E_{ss} & & 0 & E_{ss} \\ |2\rangle & E_{ss} & E_0 & & 0 & 0 \\ & \cdots & & \cdots & & \cdots \\ |N-1\rangle & 0 & 0 & & E_0 & E_{ss} \\ |N\rangle & E_{ss} & 0 & & E_{ss} & E_0 \end{array} \qquad (5.1.1)$$

We have used what is called the periodic boundary condition (PBC), namely, that the Nth atom wraps around and overlaps the 1st atom like a ring. This leads to non-zero values for the matrix elements $H_{1,N}$ and $H_{N,1}$ which would normally be zero if the solid were abruptly truncated. The PBC is usually not realistic, but if we are discussing the bulk properties of a large solid then the precise boundary condition at the surface does not matter and we are free to use whatever boundary conditions makes the mathematics the simplest, which happens to be the PBC.

So what are the eigenvalues of the matrix [H] given in Eq.(5.1.1) ? This is essentially the same matrix that we discussed in Chapter 1 in connection with the finite difference method. If we find the eigenvalues numerically we will find that they can all be written in the form (α: integer)

$$E_\alpha = E_0 + 2E_{ss}\cos(k_\alpha a) \quad \text{where} \quad k_\alpha a = \alpha 2\pi/N \qquad (5.1.2)$$

The values of $k_\alpha a$ run from - π to + π and are spaced by $2\pi/N$ as shown in Fig.5.1.2. If N is large the eigenvalues are closely spaced (as on the left); if N is small the eigenvalues are further apart (as on the right).

Why is it that we can write down the eigenvalues of this matrix so simply ? The reason is that because of its periodic nature, the matrix equation $E\ (\psi) = [H]\ (\psi)$ consists of a set of N equations which are all identical in form and can all be written as (n = 1,2,....N)

$$E\,\psi_n = E_0\,\psi_n + E_{ss}\,\psi_{n-1} + E_{ss}\,\psi_{n+1} \qquad (5.1.3)$$

This set of equations can be solved analytically by the ansatz:

$$\psi_n = \psi_0\, e^{ikna} \qquad (5.1.4)$$

Substituting Eq.(5.1.4) into (5.1.3) and canceling the common factor exp[ikna] we obtain

$$E\,\psi_0 = E_0\,\psi_0 + E_{ss}\, e^{-ika}\,\psi_0 + E_{ss}\, e^{ika}\,\psi_0$$

that is, $E = E_0 + 2E_{SS}\cos(ka)$

This shows us that a solution of the form shown in Eq.(5.1.4) will satisfy our set of equations for any value of 'k'. But what restricts the number of eigenvalues to a finite number (as it must be for a finite-sized matrix)?

This is a result of two factors. Firstly, periodic boundary conditions require the wavefunction to be periodic with a period of 'Na' and it is this finite lattice size that restricts the allowed values of 'k' to the discrete set $k_\alpha a = \alpha 2\pi/N$ (see Eq.(5.1.2)). Secondly, values of 'ka' differing by 2π do not represent distinct states on a discrete lattice. The wavefunctions

$$\exp(i\,k_\alpha\,x) \quad \text{and} \quad \exp(i\,[k_\alpha + (2\pi/a)]\,x)$$

represent the same state because at any lattice point $x_n = na$,

$$\exp(i\,k_\alpha\,x_n) = \exp(i\,[k_\alpha + (2\pi/a)]\,x_n)$$

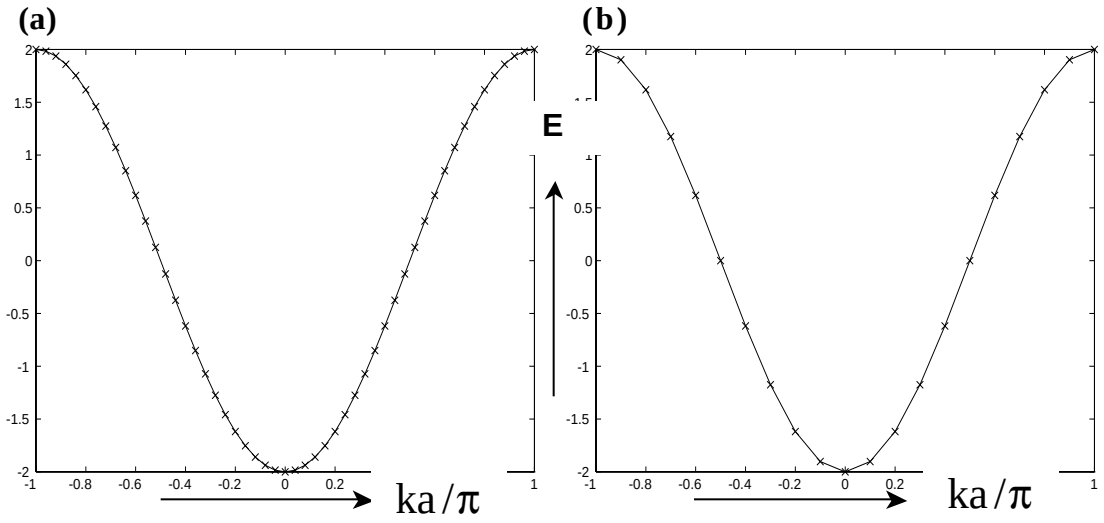


Fig.5.1.2. The solid lines in both (a) and (b) are plots of E vs. ka/π from Eq.(5.1.2) with E_0 = 0, E_{SS} = - 1. The x's denote the eigenvalues of the matrix in Eq.(3.2.1) with (a) N=100 and with (b) N=20.

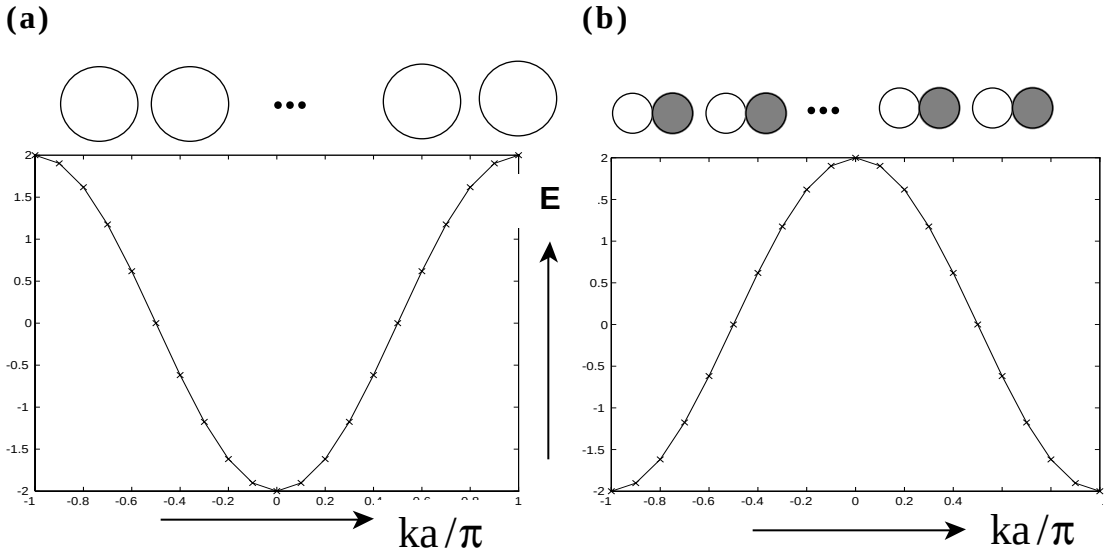


Fig.5.1.3. Same as Fig.5.1.2b with E_0 = 0 and (a) E_{SS} = -1 and with (b) E_{SS} = +1.

They are NOT equal between two lattice points and thus represent distinct states in a continuous lattice. But once we adopt a discrete lattice, values of k_α differing by $2\pi/a$ represent identical states and only the values of $k_\alpha a$ within a range of 2π yield independent solutions. In principle, any range of size 2π is acceptable, but it is common to restrict the values of $k_\alpha a$ to the range (sometimes called the first Brillouin zone)

$$-\pi \leq ka < +\pi \qquad \text{for periodic boundary conditions} \qquad (5.1.5)$$

It is interesting to note that the finite range of the lattice ("Na") leads to a discreteness (in units of "$2\pi/Na$") in the allowed values of 'k' while the discreteness of the lattice ("a") leads to a finite range of allowed 'k' ("$2\pi/a$"). The number of allowed values of 'k'

$$(2\pi/a)/(2\pi/Na) = N$$

is exactly the same as the number of points in the real space lattice. This ensures that the number of eigenvalues (which is equal to the number of allowed 'k' values) is equal to the size of the matrix [H] (determined by the number of lattice points).

When do bands run downwards in k ? In Fig.5.1.1 we have assumed E_{SS} to be negative which is what we would find if we used say Eq.(4.1.11c) to evaluate it (note that the potentials U_L or U_R are negative) and the atomic orbitals were 's' orbitals. But if the atomic orbitals are 'p_X' orbitals as shown in Fig.5.1.3b then the sign of the overlap integral (E_{SS}) would be negative and the plot of E(k) would run downwards in 'k' as shown. Roughly speaking this is what happens in the valence band of common semiconductors which are formed primarily out of atomic 'p' orbitals.

Lattice with a basis - the Peierls' distortion:

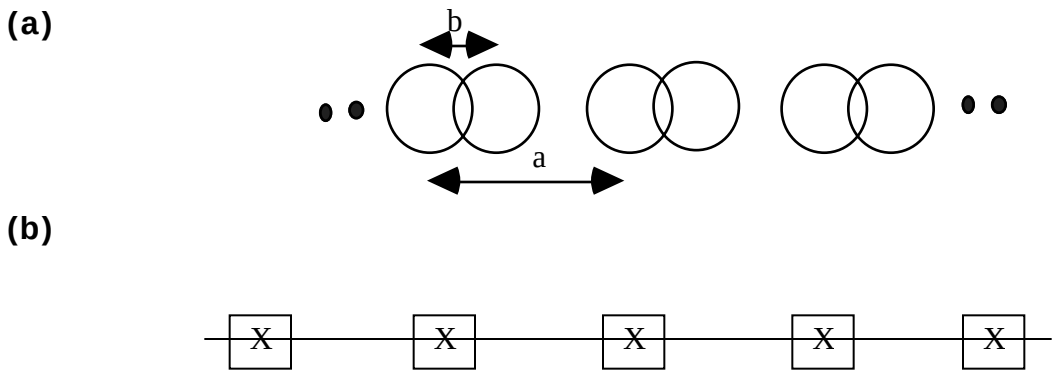


Fig.5.1.4. (a) A one-dimensional solid whose unit cell consists of two atoms (b) Basic lattice defining the periodicity of the solid.

Consider next a one-dimensional solid whose unit cell consists of two atoms as shown in Fig.5.1.4. Actually one-dimensional structures like the one shown in Fig.5.1.1 tend to distort spontaneously into the structure shown in Fig.5.1.4 - a phenomenon that is generally referred to as the Peierls' distortion. We will not go into the energetic considerations that cause this to happen. Our purpose is simply to illustrate how we can find the bandstructure for a solid whose unit cell contains more than one basis orbital. Using one orbital per atom we can write the matrix representation of [H] as

$$
[H] = \begin{array}{c|cccccc|c}
 & |1_A\rangle & |1_B\rangle & |2_A\rangle & |2_B\rangle & |3_A\rangle & |3_B\rangle & \cdots \\
\hline
|1_A\rangle & E_0 & E_{ss} & 0 & 0 & 0 & 0 & .... \\
|1_B\rangle & E_{ss} & E_0 & E_{ss}' & 0 & 0 & 0 & .... \\
|2_A\rangle & 0 & E_{ss}' & E_0 & E_{ss} & 0 & 0 & .... \\
|2_B\rangle & 0 & 0 & E_{ss} & E_0 & E_{ss}' & 0 & .... \\
|3_A\rangle & 0 & 0 & 0 & E_{ss}' & E_0 & E_{ss} & .... \\
|3_B\rangle & 0 & 0 & 0 & 0 & E_{ss} & E_0 & ....
\end{array} \qquad (5.1.6)
$$

Unlike the matrix in Eq.(5.1.1) there are two different overlap integrals E_{ss} and E_{ss}' appearing alternately. As such the ansatz in Eq.(5.1.4) cannot be used directly. But we could combine the elements of the matrix into (2x2) blocks and rewrite it in the form

$$\mathbf{H} = \begin{array}{c|cccc} & |1\rangle & |2\rangle & |3\rangle & \cdots \\ |1\rangle & H_{11} & H_{12} & 0 & .... \\ |2\rangle & H_{21} & H_{22} & H_{23} & .... \\ |3\rangle & 0 & H_{32} & H_{33} & .... \end{array} \qquad (5.1.7)$$

where

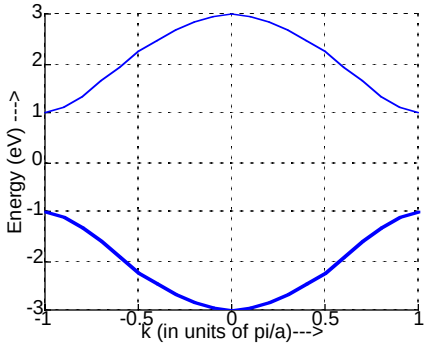
$$H_{nn} = \begin{bmatrix} E_0 & E_{ss} \\ E_{ss} & E_0 \end{bmatrix} \qquad H_{n,n+1} = \begin{bmatrix} 0 & 0 \\ E_{ss}' & 0 \end{bmatrix} \qquad H_{n,n-1} = \begin{bmatrix} 0 & E_{ss}' \\ 0 & 0 \end{bmatrix}$$

The matrix in Eq.(5.1.7) is now periodic and we can write the matrix equation $E\ (\psi) = [H]\ (\psi)$ in the form

$$E\,\varphi_n = H_{nn}\,\varphi_n + H_{n,n-1}\,\varphi_{n-1} + H_{n,n+1}\,\varphi_{n+1} \qquad (5.1.8)$$

where φ_n represents a (2x1) column vector and the element H_{nm} is a (2x2) matrix. We can solve this set of equations using the ansatz : $\varphi_n = \varphi_0\, e^{ikna}$ (5.1.9)

Fig.5.1.5. Bandstructure for the "dimerized" one-dimensional solid shown in Fig.5.1.4 plotted from Eq. (5.1.10) using $E_0 = 0$, $E_{ss} = 2$, $E_{ss}' = 1$.



Substituting Eq.(5.1.9) into (5.1.8) and canceling the common factor exp[ikna] we obtain

$$E\,\varphi_0 = H_{nn}\,\varphi_0 + H_{n,n-1}\,e^{-ika}\,\varphi_0 + H_{n,n+1}\,e^{ika}\,\varphi_0$$

that is,

$$E\,\{\varphi_0\} = \begin{bmatrix} E_0 & E_{ss} + E_{ss}'\,e^{-ika} \\ E_{ss} + E_{ss}'\,e^{ika} & E_0 \end{bmatrix} \{\varphi_0\}$$

We can now find the eigenvalues by setting the determinant to zero:

$$\det \begin{bmatrix} E_0 - E & E_{ss} + E_{ss}' e^{-ika} \\ E_{ss} + E_{ss}' e^{ika} & E_0 - E \end{bmatrix} = 0$$

that is, $$E = E_0 \pm \left(E_{ss}^2 + E_{ss}'^2 + 2E_{ss}E_{ss}' \cos ka \right)^{1/2} \qquad (5.1.10)$$

Eq.(5.1.10) gives us a E(k) diagram with two branches as shown in Fig.5.1.5.

5.2. General result

It is straightforward to generalize this procedure for calculating the bandstructure of any periodic solid with arbitrary number of basis functions per unit cell. Consider any particular unit cell 'n' (Fig.5.2.2) connected to its neighboring unit cells 'm' by a matrix $[H_{nm}]$ of size (bxb), 'b' being the number of basis functions per unit cell. We can write the overall matrix equation in the form

$$\sum_m [H_{nm}]\{\varphi_m\} = E\{\varphi_n\} \qquad (5.2.1)$$

where $\{\varphi_m\}$ is a (bx1) column vector denoting the wavefunction in unit cell 'm'.

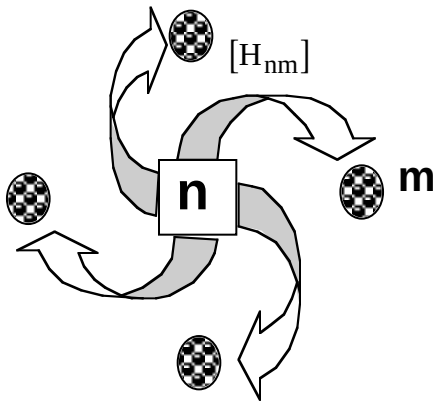


Fig.5.2.1. Schematic picture showing a unit cell 'n' connected to its neighboring unit cells 'm' by a matrix $[H_{nm}]$ of size (bxb), 'b' being the number of basis functions per unit cell. The configuration of neighbors will differ from one solid to another, but in a periodic solid the configuration is identical regardless of which 'n' we choose.

The important insight is the observation that this set of equations can be solved by the ansatz

$$\{\varphi_m\} = \{\varphi_0\} \exp^{i\vec{k}.\vec{d}_m} \tag{5.2.2}$$

provided Eq.(5.2.1) looks the same in every unit cell 'n'. This is a consequence of the periodicity of the lattice and it ensures that when we substitute our ansatz Eq.(5.2.2) into Eq.(5.2.1) we obtain

$$E\{\varphi_0\} = \left[h(\vec{k})\right]\{\varphi_0\} \tag{5.2.3}$$

with $$\left[h(\vec{k})\right] = \sum_m \left[H_{nm}\right] e^{i\vec{k}.(\vec{d}_m - \vec{d}_n)} \tag{5.2.4}$$

independent of which unit cell 'n' we use to evaluate the sum in Eq.(5.2.4). This is the ***central result*** underlying the bandstructure of periodic solids. The summation over 'm' in Eq,.(5.2.4) runs over all neighboring unit cells (including itself) with which cell 'n' has any overlap (that is, for which H_{nm} is non-zero). The size of the matrix $[h(\vec{k})]$ is (bxb), 'b' being the number of basis orbitals per unit cell. The bandstructure can be plotted out by finding the eigenvalues of the (bxb) matrix $[h(\vec{k})]$ for each value of $\vec{k}$ and it will have 'b' branches one for each eigenvalue.

Allowed values of k: In connection with the 1-D example, we explained how 'k' has only a finite number of allowed values equal to the number of unit cells in the solid. To reiterate the basic result, the finite range of the lattice ("Na") leads to a discreteness (in units of " $2\pi/Na$") in the allowed values of 'k' while the discreteness of the lattice ("a") leads to a finite range of allowed 'k' ("$2\pi/a$"). How do we generalize this result beyond one dimension?

This is fairly straightforward if the solid forms a rectangular (or a cubic) lattice as shown in Fig.5.2.2a. In 2-D the allowed values of $\vec{k}$ can be written as

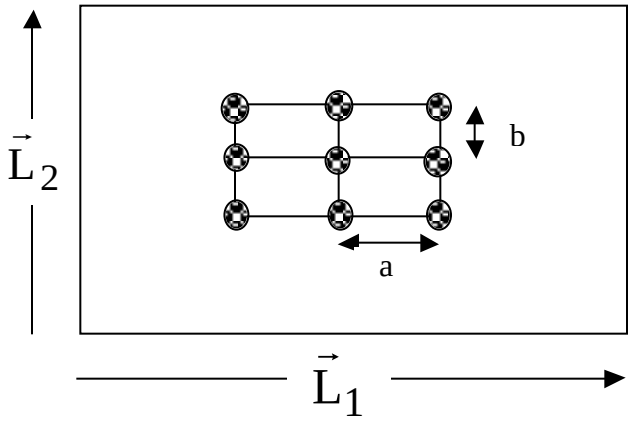
$$\left[\vec{k}\right]_{m,n} = \hat{x}\,(m\,2\pi/Ma) + \hat{y}\,(n\,2\pi/Nb) \tag{5.2.5}$$

where (m,n) are a pair of integers while M, N represent the number of unit cells stacked along the x- and y-direction respectively. This seems like a reasonable extension of the

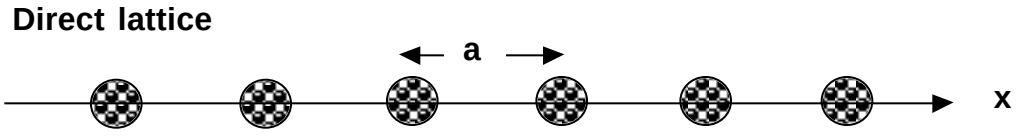
1-D result (cf. Eq.(5.1.2): $k_\alpha = \alpha\,(2\pi/a)$. Formally we could derive Eq.(5.2.5) by writing ($\vec{L}_1 = \hat{x}\,Ma$, $\vec{L}_2 = \hat{y}\,Nb$)

$$\vec{k}.\vec{L}_1 = m\,2\pi \;\rightarrow\; k_x = m\,2\pi/M\,a$$
$$\vec{k}.\vec{L}_2 = n\,2\pi \;\rightarrow\; k_y = n\,2\pi/N\,b$$

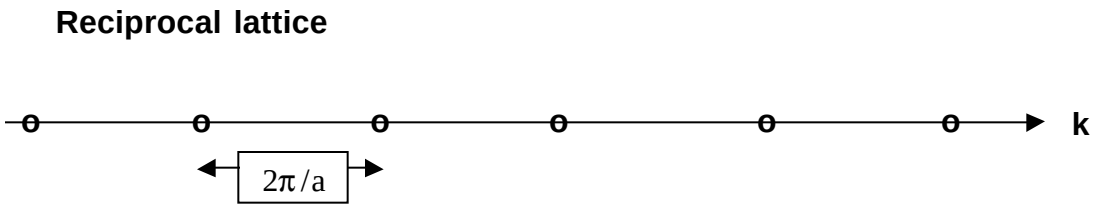
Fig.5.2.2. A finite 2-D rectangular lattice with M unit cells stacked along the x-direction and N unit cells stacked along the y-direction:
($\vec{L}_1 = \hat{x}\,Ma, \vec{L}_2 = \hat{y}\,Nb$)



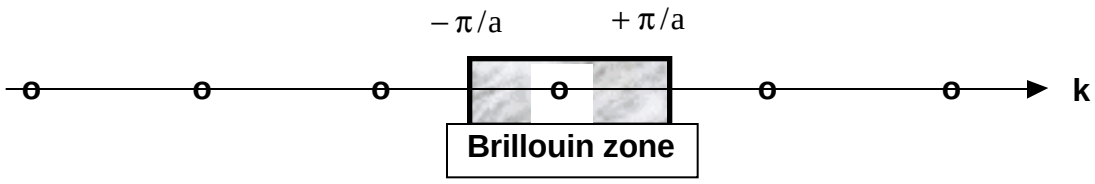
Brillouin zone: Formally, the general procedure for constructing the Brillouin zone starts by constructing the reciprocal lattice (Fig.5.2.2b) in k-space, which can be viewed as the Fourier transform of the direct lattice. In 1-D we know that a set of impulses separated by 'a'



has a Fourier transform consisting of a set of impulses separated by $2\pi/a$



We could then construct the first Brillouin zone centered around k = 0 by connecting it to the neighboring points on the reciprocal lattice and drawing their bisectors:



Similarly for a two dimensional rectangular lattice we can construct a reciprocal lattice and then obtain the first Brillouin zone by drawing perpendicular bisectors of the lines joining $\vec{k} = (0,0)$ to the neighboring points on the reciprocal lattice:

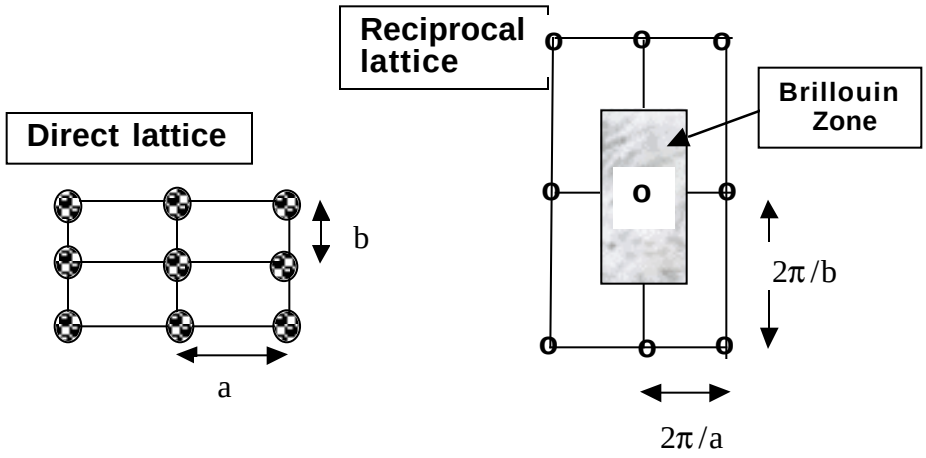


Fig.5.2.3. (a) Rectangular lattice in real space. (b) Corresponding reciprocal lattice.

The Brillouin zone obtained from this procedure defines the allowed range of values of $\vec{k}$

$$-\pi \le k_x a < +\pi \qquad \text{and} \qquad -\pi \le k_y b < +\pi \tag{5.2.6}$$

which agrees with what one might write down from a heuristic extension of Eq.(5.1.5).

Reciprocal lattice: In general, if the direct lattice is not rectangular or cubic, it is not possible to construct the reciprocal lattice quite so simply by inspection. We then need to adopt a more formal procedure as follows. We first note that any point on a direct lattice in 3-D can be described by a set of three integers (m,n,p) such that

$$\vec{R} = m\,\vec{a}_1 + n\,\vec{a}_2 + p\,\vec{a}_3 \qquad (5.2.7)$$

where $\vec{a}_1$, $\vec{a}_2$, $\vec{a}_3$ are called the basis vectors of the lattice. The points on the reciprocal lattice can be written as

$$\vec{K} = M\,\vec{A}_1 + N\,\vec{A}_2 + P\,\vec{A}_3 \qquad (5.2.7)$$

where (M,N,P) are integers and $\vec{A}_1$, $\vec{A}_2$, $\vec{A}_3$ are determined such that

$$\vec{A}_j \cdot \vec{a}_i = 2\pi\,\delta_{ij} \qquad (5.2.8)$$

δ_{ij} being the Kronecker delta (equal to one if i = j, and equal to zero if i ≠ j). Eq.(5.2.8) can be satisfied by writing

$$\vec{A}_1 = \frac{2\pi\,(\vec{a}_2 \times \vec{a}_3)}{\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)}\,,\quad \vec{A}_2 = \frac{2\pi\,(\vec{a}_3 \times \vec{a}_1)}{\vec{a}_2 \cdot (\vec{a}_3 \times \vec{a}_1)}\,,\quad \vec{A}_3 = \frac{2\pi\,(\vec{a}_1 \times \vec{a}_2)}{\vec{a}_3 \cdot (\vec{a}_1 \times \vec{a}_2)} \qquad (5.2.9)$$

It is easy to see that this formal procedure for constructing the reciprocal lattice leads to the lattice shown in Fig.5.2.3b if we assume the real space basis vectors to be $\vec{a}_1 = \hat{x}\,a$, $\vec{a}_2 = \hat{y}\,b$, $\vec{a}_3 = \hat{z}\,c$. Eq, (5.2.9) then yields

$$\vec{A}_1 = \hat{x}\,(2\pi/a)\,,\ \vec{A}_2 = \hat{y}\,(2\pi/b)\,,\ \vec{A}_3 = \hat{z}\,(2\pi/c).$$

Using Eq.(5.2.7) we can now set up the reciprocal lattice shown in Fig.5.2.3b. Of course, in this case we do not really need the formal procedure. The real value of the formal approach lies in handling non-rectangular lattices, as we will now illustrate with a 2-D example.

A 2-D example: The carbon atoms on the surface of a sheet of graphite are arranged in the hexagonal pattern shown in Fig.5.2.4a. It can be seen that the structure is not really periodic. Adjacent carbon atoms do not have identical environments. But if we lump two

atoms together into a unit cell then the lattice of unit cells is periodic: every site has an identical environment (Fig.5.2.4b).

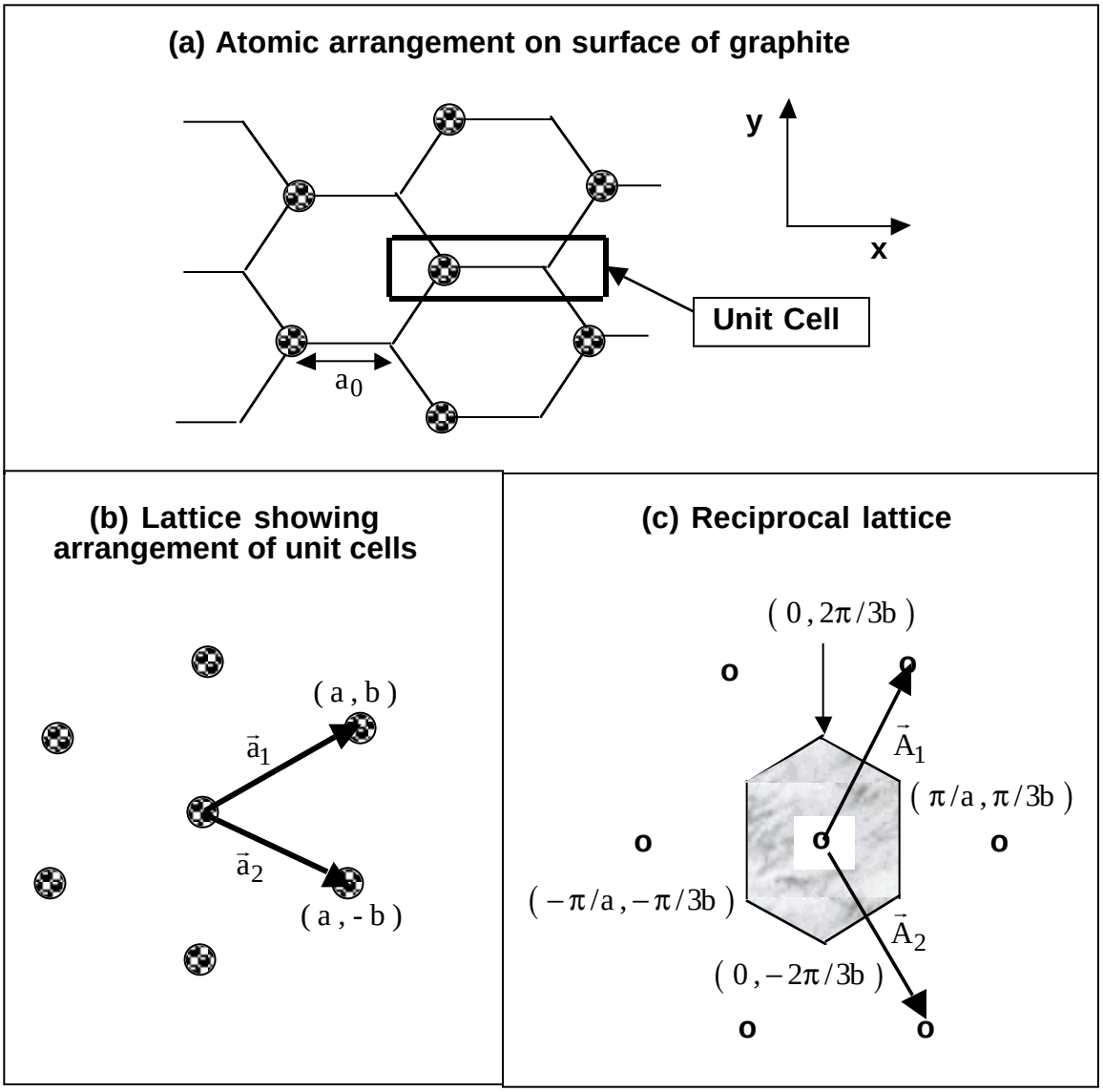


Fig.5.2.4. (a) Arrangement of carbon atoms on the surface of graphite, showing the unit cell of two atoms. (b) Direct lattice showing the periodic arrangement of unit cells with basis vectors $\vec{a}_1$ and $\vec{a}_2$. (c) Reciprocal lattice with basis vectors $\vec{A}_1$ and $\vec{A}_2$ determined such that $\vec{A}_1 \cdot \vec{a}_1 = \vec{A}_2 \cdot \vec{a}_2 = 2\pi$ and $\vec{A}_1 \cdot \vec{a}_2 = \vec{A}_2 \cdot \vec{a}_1 = 0$. Also shown is the Brillouin zone obtained by drawing the perpendicular bisectors of the lines joining the origin (0,0) to the neighboring points on the reciprocal lattice.

Every point on this periodic lattice formed by the unit cells can be described by a set of integers (m,n,p) where

$$\vec{R} = m\,\vec{a}_1 + n\,\vec{a}_2 + p\,\vec{a}_3 \qquad (5.2.10)$$

with $\vec{a}_1 = \hat{x}\,a + \hat{y}\,b \quad , \quad \vec{a}_2 = \hat{x}\,a - \hat{y}\,b \quad , \quad \vec{a}_3 = \hat{z}\,c$

where $a \equiv 3a_0/2$ and $b \equiv \sqrt{3}\,a_0/2$

Here 'c' is the length of the unit cell along the c-axis, which will play no important role in this discussion since we will talk about the electronic states in the x-y plane assuming that different planes along the c-axis are isolated (which is not too far from the truth in real graphite). The points on the reciprocal lattice in the $k_x - k_y$ plane are given by

$$\vec{K} = M\,\vec{A}_1 + N\,\vec{A}_2 \qquad (5.2.11)$$

where (M,N) are integers and $\vec{A}_1$, $\vec{A}_2$ are determined from Eq.(5.2.9):

$$\vec{A}_1 = \frac{2\pi\,(\vec{a}_2 \times \hat{z})}{\vec{a}_1 \cdot (\vec{a}_2 \times \hat{z})} = \hat{x}\left(\frac{\pi}{a}\right) + \hat{y}\left(\frac{\pi}{b}\right)$$

$$\vec{A}_2 = \frac{2\pi\,(\hat{z} \times \vec{a}_1)}{\vec{a}_2 \cdot (\hat{z} \times \vec{a}_1)} = \hat{x}\left(\frac{\pi}{a}\right) - \hat{y}\left(\frac{\pi}{b}\right)$$

Using these basis vectors we can construct the reciprocal lattice shown in Fig.5.2.4c. The Brillouin zone for the allowed k-vectors is then obtained by drawing the perpendicular bisectors of the lines joining the origin (0,0) to the neighboring points on the reciprocal lattice.

The Brillouin zone tells us the range of k-values while the actual discrete values of 'k' have to be obtained from the finite size of the direct lattice, as explained following Eq.(5.2.5). But for a given value of 'k' how do we obtain the corresponding energy eigenvalues? Answer: From Eqs.(5.2.3) and (5.2.4). The size of the matrix $\left[h(\vec{k})\right]$ depends on the number of basis functions per unit cell. If we use the four valence orbitals of Carbon ($2s, 2p_x, 2p_y, 2p_z$) as our basis functions then we will have 4x2=8 basis functions per unit cell (since it contains two carbon atoms) and hence 8 eigenvalues for each value of k.

It is found, however, for graphite that the levels involving $2s, 2p_x, 2p_y$ orbitals are largely decoupled from those involving $2p_z$ orbitals; in other words, there are no matrix elements coupling these two subspaces. Moreover, the levels involving $2s, 2p_x, 2p_y$ orbitals are either far below or far above the Fermi energy, so that the conduction and valence band levels right around the Fermi energy (which are responsible for electrical conduction) are essentially formed out of the $2p_z$ orbitals (Fig.2.4.5).

This means that the conduction and valence band states can be described quite well by a theory that uses only one orbital (the $2p_z$ orbital) per Carbon atom resulting in a (2x2) matrix $\left[h(\vec{k})\right]$ which can be written down by summing over any unit cell and all its four neighboring unit cells (the matrix element is assumed equal to '-t' between neighboring Carbon atoms and zero otherwise):

$$\left[h(\vec{k})\right] = \begin{bmatrix} 0 & -t \\ -t & 0 \end{bmatrix} + \begin{bmatrix} 0 & -t\exp\left(i\vec{k}\cdot\vec{a}_1\right) \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -t\exp\left(i\vec{k}\cdot\vec{a}_2\right) \\ 0 & 0 \end{bmatrix}$$
$$+ \begin{bmatrix} 0 & 0 \\ -t\exp\left(-i\vec{k}\cdot\vec{a}_1\right) & 0 \end{bmatrix} + \begin{bmatrix} 0 & -t \\ -t\exp\left(-i\vec{k}\cdot\vec{a}_2\right) & 0 \end{bmatrix}$$

Defining $h_0 \equiv -t\left(1 + e^{i\vec{k}.\vec{a}_1} + e^{i\vec{k}.\vec{a}_2}\right) = -t\left(1 + 2e^{ik_x a_0}\cos k_y b_0\right)$

we can write

$$h(\vec{k}) = \begin{bmatrix} 0 & h_0 \\ h_0^* & 0 \end{bmatrix}$$

so that the eigenvalues are given by

$$E = \pm\left|h_0\right| = \pm t\sqrt{1 + 4\cos k_y b_0 \cos k_x a_0 + 4\cos^2 k_y b_0}$$

Note that we obtain two eigenvalues (one positive and one negative) for each value of $\vec{k}$ resulting in two branches in the $E(\vec{k})$ plot (cf. Fig.5.1.5) – this is what we expect since we have two basis functions per unit cell. We will discuss the physics of this $E(\vec{k})$ relation (generally called the energy dispersion relation) in the next chapter when we discuss carbon nanotubes, which are basically graphite sheets rolled into cylinders. For the moment our main purpose is to illustrate the procedure for calculating the bandstructure using a 2-D example that involves non-trivial features beyond the 1-D

examples from the last section and yet does not pose serious problems with visualization as the 3-D example in the next section.

5.3. Common semiconductors

All the common semiconductors (like Gallium Arsenide) belong to the diamond structure which has a unit cell consisting of two atoms, a cation (like Gallium) and an anion (like Arsenic). For elemental semiconductors like Silicon, both cationic and anionic sites are occupied by the same atom. For each atom we need to include at least four valence orbitals like 3s, $3p_X$, $3p_Y$ and $3p_Z$ for Silicon. It is common to include the next higher orbital (4s for Silicon) as well giving rise to what is called the sp^3s^* model. In this model we have five orbitals per atom leading to 10 basis orbitals per unit cell. Consequently the matrices $[h(\vec{k})]$ and $[H_{nm}]$ in Eq.(5.2.4) are each (10x10) in size. To perform the summation indicated in Eq.(5.2.4) we need to figure out how the nearest neighbors are located.

The diamond structure consists of two interpenetrating face-centered cubic (FCC) lattices. For example, if we look at GaAs, we find that the Gallium atoms occupy the sites on an FCC lattice. The Arsenic atoms occupy the sites of a different FCC lattice offset from the previous one by a quarter of the distance along the body diagonal - that is, the coordinates of this lattice can be obtained by adding $(\hat{x}+\hat{y}+\hat{z})a/4$ to those of the first one. If a Gallium atom is located at the origin (0 0 0)a/4 then there will be an Arsenic atom located at $(\hat{x}+\hat{y}+\hat{z})a/4$ which will be one of its nearest neighbors. Actually it will have three more Arsenic atoms as nearest neighbors. To see this consider where the nearest Gallium atoms are located. There are four of them on the X-Y face as shown in Fig.5.3.1 whose coordinates can be written as $(\hat{x}+\hat{y})a/2$, $(\hat{x}-\hat{y})a/2$, $(-\hat{x}+\hat{y})a/2$ and $(-\hat{x}-\hat{y})a/2$.

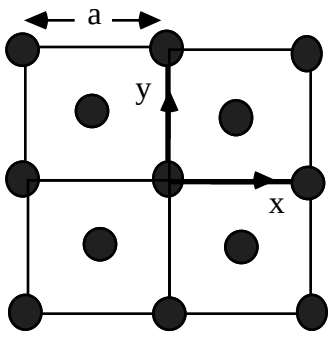


Fig.5.3.1. X-Y face of a face-centered cubic (FCC) lattice, showing the location of atoms.

The coordinates of the corresponding Arsenic atoms are obtained by adding $(\hat{x}+\hat{y}+\hat{z})a/4$:

$(3\hat{x}+3\hat{y}+\hat{z})a/4$, $(3\hat{x}-\hat{y}+\hat{z})a/4$, $(-\hat{x}+3\hat{y}+\hat{z})a/4$ and $(-\hat{x}-\hat{y}+\hat{z})a/4$

Of these the first three are too far away, but the fourth one is a nearest neighbor of the Gallium atom at the origin. Similarly if we consider the neighboring Galium atoms on the Y-Z face and the Z-X face we will find two more nearest neighbors, so that the Gallium atom at the origin (0 0 0) has four nearest neighbor Arsenic atoms located at

$$(\hat{x}+\hat{y}+\hat{z})a/4,\quad (-\hat{x}-\hat{y}+\hat{z})a/4,\qquad (\hat{x}-\hat{y}-\hat{z})a/4\quad \text{and}\quad (-\hat{x}+\hat{y}-\hat{z})a/4.$$

Every atom in a diamond lattice has four nearest neighbors of the opposite type (cation or anion) arranged in a tetrahedron.

To see how we perform the summation in Eq.(5.2.4) let us first consider just the s-orbital for each atom. The matrices $[h(\vec{k})]$ and $[H_{nm}]$ in Eq.(5.2.4) are then each (2x2) in size. We can write $[H_{nn}]$ as

$$\begin{array}{c|cc} & |s_a\rangle & |s_c\rangle \\ |s_a\rangle & E_{sa} & E_{ss} \\ |s_c\rangle & E_{ss} & E_{sc} \end{array} \qquad (5.3.1a)$$

where E_{sa} and E_{sc} are the energies of the 's' orbitals for the anion and cation respectively while E_{ss} represents the overlap integral between an 's' orbital on the anion and an 's' orbital on the cation. The anion in unit cell 'n' overlaps with the cations in three other unit cells 'm' for which

$$\vec{d}_m-\vec{d}_n=(-\hat{x}-\hat{y})a/2,\ (-\hat{y}-\hat{z})a/2\quad \text{and}\quad (-\hat{z}-\hat{x})a/2$$

Each of these contributes a $[H_{nm}]$ of the form

$$\begin{array}{c|cc} & |s_a\rangle & |s_c\rangle \\ |s_a\rangle & 0 & E_{ss} \\ |s_c\rangle & 0 & 0 \end{array} \qquad (5.3.1b)$$

Similarly the cation in unit cell 'n' overlaps with the anions in three other unit cells 'm' for which

$$\vec{d}_m-\vec{d}_n=(\hat{x}+\hat{y})a/2,\quad (\hat{y}+\hat{z})a/2\quad \text{and}\quad (\hat{z}+\hat{x})a/2$$

Each of these contributes a $[H_{nm}]$ of the form

$$\begin{array}{c|cc} & |s_a\rangle & |s_c\rangle \\ |s_a\rangle & 0 & 0 \\ |s_c\rangle & E_{ss} & 0 \end{array} \qquad (5.3.1c)$$

Adding up all these contributions we obtain

$$[h(\vec{k})] = \begin{array}{c|cc} & |s_a\rangle & |s_c\rangle \\ |s_a\rangle & E_{sa} & 4E_{ssg0} \\ |s_c\rangle & 4E_{ssg0}^{*} & E_{sc} \end{array} \qquad (5.3.2)$$

where $4g_0 \equiv 1 + e^{-i\vec{k}.\vec{d}_1} + e^{-i\vec{k}.\vec{d}_2} + e^{-i\vec{k}.\vec{d}_3}$

with $\vec{d}_1 \equiv (\hat{y}+\hat{z})\,a/2$, $\vec{d}_2 \equiv (\hat{z}+\hat{x})\,a/2$ and $\vec{d}_3 \equiv (\hat{x}+\hat{y})\,a/2$

To evaluate the full (10x10) matrix $[h(\vec{k})]$ including sp^3s^* levels we proceed similarly. The final result is

$$\begin{array}{c|cccccccccc} & |s_a\rangle & |s_c\rangle & |X_a\rangle & |Y_a\rangle & |Z_a\rangle & |X_c\rangle & |Y_c\rangle & |Z_c\rangle & |s_a^*\rangle & |s_c^*\rangle \\ |s_a\rangle & E_{sa} & 4E_{ssg0} & 0 & 0 & 0 & 4E_{sapcg1} & 4E_{sapcg2} & 4E_{sapcg3} & 0 & 0 \\ |s_c\rangle & 4E_{ssg0}^{*} & E_{sc} & 4E_{pascg1}^{*} & 4E_{pascg2}^{*} & 4E_{pascg3}^{*} & 0 & 0 & 0 & 0 & 0 \\ |X_a\rangle & 0 & 4E_{pascg1} & E_{pa} & 0 & 0 & 4E_{xxg0} & 4E_{xyg3} & 4E_{xyg2} & 0 & 4E_{pas^*cg1} \\ |Y_a\rangle & 0 & 4E_{pascg2} & 0 & E_{pa} & 0 & 4E_{xyg3} & 4E_{xxg0} & 4E_{xyg1} & 0 & 4E_{pas^*cg2} \\ |Z_a\rangle & 0 & 4E_{pascg3} & 0 & 0 & E_{pa} & 4E_{xyg2} & 4E_{xyg1} & 4E_{xxg0} & 0 & 4E_{pas^*cg3} \\ |X_c\rangle & 4E_{sapcg1}^{*} & 0 & 4E_{xxg0}^{*} & 4E_{xyg3}^{*} & 4E_{xyg2}^{*} & E_{pc} & 0 & 0 & 4E_{s^*apcg1}^{*} & 0 \\ |Y_c\rangle & 4E_{sapcg2}^{*} & 0 & 4E_{xyg3}^{*} & 4E_{xxg0}^{*} & 4E_{xyg1}^{*} & 0 & E_{pc} & 0 & 4E_{s^*apcg2}^{*} & 0 \\ |Z_c\rangle & 4E_{sapcg3}^{*} & 0 & 4E_{xyg2}^{*} & 4E_{xyg1}^{*} & 4E_{xxg0}^{*} & 0 & 0 & E_{pc} & 4E_{s^*apcg3}^{*} & 0 \\ |s_a^*\rangle & 0 & 0 & 0 & 0 & 0 & 4E_{s^*apcg1} & 4E_{s^*apcg2} & 4E_{s^*apcg3} & E_{s^*a} & 0 \\ |s_c^*\rangle & 0 & 0 & 4E_{pas^*cg1}^{*} & 4E_{pas^*cg2}^{*} & 4E_{pas^*cg3}^{*} & 0 & 0 & 0 & 0 & E_{s^*c} \end{array} \qquad (5.3.3)$$

The factors g_1, g_2 and g_3 look much like the factor g_0 obtained above when discussing only the s-orbitals:

$$4g_0 \equiv 1 + e^{-i\vec{k}.\vec{d}_1} + e^{-i\vec{k}.\vec{d}_2} + e^{-i\vec{k}.\vec{d}_3} \qquad (5.3.4a)$$

However, the signs of some of the terms are negative:

$$4g_1 \equiv 1 + e^{-i\vec{k}.\vec{d}_1} - e^{-i\vec{k}.\vec{d}_2} - e^{-i\vec{k}.\vec{d}_3} \quad (5.3.4b)$$

$$4g_2 \equiv 1 - e^{-i\vec{k}.\vec{d}_1} + e^{-i\vec{k}.\vec{d}_2} - e^{-i\vec{k}.\vec{d}_3} \quad (5.3.4c)$$

$$4g_3 \equiv 1 - e^{-i\vec{k}.\vec{d}_1} - e^{-i\vec{k}.\vec{d}_2} + e^{-i\vec{k}.\vec{d}_3} \quad (5.3.4d)$$

The negative signs arise because the wavefunction for p-orbitals changes sign along one axis and so the overlap integral has different signs for different neighbors. This also affects the signs of the overlap integrals appearing in the expression for $[h(\vec{k})]$ in Eq.(5.2.4) : the parameters E_{ss}, $E_{pa,sc}$ and E_{pas^*c} are negative, while the remaining parameters $E_{sa,pc}$, E_{xx} , E_{xy} and E_{s^*apc} are are positive. Note that the vectors

$$\vec{d}_1 \equiv (\hat{y} + \hat{z})\, a/2,\ \vec{d}_2 \equiv (\hat{z} + \hat{x})\, a/2 \text{ and } \vec{d}_3 \equiv (\hat{x} + \hat{y})\, a/2 \quad (5.3.5)$$

connect the cation in one unit cell to a cation in a neighboring cell (or an anion to an anion). Alternatively, we could define these vectors so as to connect the nearest neighbors - this has the effect of multiplying each of the factors g_0, g_1, g_2 and g_3 by a phase factor $\exp[i\vec{k}.\vec{d}]$ where $\vec{d} = (\hat{x} + \hat{y} + \hat{z})a/4$. This is used by most authors (see for example, P.Vogl, H.P.Hjalmarson and J.Dow, "A Semi-Empirical Tight-Binding Theory of the Electronic Structure of Semiconductors", J. Phys. Chem. Solids, vol. 44, p.365-378 (1983)) but it makes no real difference to the result.

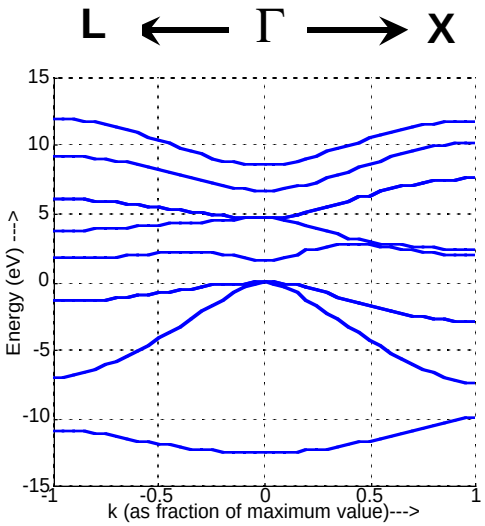


Fig.5.3.2. E ($\vec{k}$) calculated by finding the eigenvalues of the matrix in Eq.(5.3.3) for each value of $\vec{k}$ along the Γ- X (that is, from $\vec{k}$ = 0 to $\vec{k} = \hat{x}\,2\pi/a$) and Γ-L (that is, from $\vec{k}$ = 0 to $\vec{k} = (\hat{x}+\hat{y}+\hat{z})\,\pi/a$) directions. The former is plotted along the positive axis and the latter along the negative axis.

Fig.5.3.2 shows the bandstructure E($\vec{k}$) calculated by finding the eigenvalues of the matrix in Eq.(5.3.3) for each value of $\vec{k}$ along the Γ-X (that is, from $\vec{k}$ = 0 to $\vec{k}$ = $\hat{x}\,2\pi/a$) and Γ-L (that is, from $\vec{k}$ = 0 to $\vec{k}$ = $(\hat{x}+\hat{y}+\hat{z})\pi/a$) directions. We have used the parameters for GaAs given in Vogl et.al.

E_{sa} = -8.3431 eV, E_{pa} = 1.0414 eV, E_{s^*a} = 8.5914 eV

E_{sc} = -2.6569 eV, E_{pc} = 3.6686 eV, E_{s^*c} = 6.7386 eV

$4E_{ss}$ = -6.4513 eV, $4E_{pa,sc}$ = -5.7839 eV, $4E_{pas^*c}$ = -4.8077 eV

$4E_{sa,pc}$ = 4.48 eV, $4E_{s^*apc}$ = 4.8422 eV,
$4E_{xx}$ = 1.9546 eV and $4E_{xy}$ = 5.0779 eV.

5.4. Effect of spin-orbit coupling

The bandstructure we have obtained is reasonably accurate but does not describe the top of the valence band very well. To obtain the correct bandstructure, it is necessary to include spin-orbit coupling as we will describe in this section.

Spinors: Let us first briefly explain how spin can be included explicitly into the Schrodinger equation. Usually we calculate the energy levels from the Schrodinger equation and fill them up with ***two*** electrons per level. More correctly we should view each level as two levels with the same energy and fill them up with ***one*** electron per level as required by the exclusion principle. How could we modify the Schrodinger equation so that each level becomes two levels with identical energies ? The answer is simple. Replace

$$E(\psi) = [H_{op}](\psi) \quad \textbf{\textit{with}} \quad E\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix} = \begin{bmatrix}H_{op} & 0\\ 0 & H_{op}\end{bmatrix}\begin{pmatrix}\psi\\ \bar{\psi}\end{pmatrix} \tag{5.4.1}$$

where $H_{op} = p^2/2m + U(\vec{r})$ $\quad (\vec{p} \equiv -i\hbar\vec{\nabla})$

We interpret ψ as the up-spin component and $\bar{\psi}$ as the down-spin component of the electronic wavefunction. If we now choose a basis set to obtain a matrix representation, the matrix will be twice as big. For example if we were to use just the s-orbital for each atom we would obtain a (4x4) matrix instead of the (2x2) matrix in Eq.(5.3.2):

$$\begin{array}{c|cccc} & |s_a\rangle & |s_c\rangle & |\bar{s}_a\rangle & |\bar{s}_c\rangle \\ |s_a\rangle & E_{sa} & 4E_{ssg0} & 0 & 0 \\ |s_c\rangle & 4E_{ssg0}^* & E_{sc} & 0 & 0 \\ |\bar{s}_a\rangle & 0 & 0 & E_{sa} & 4E_{ssg0} \\ |\bar{s}_c\rangle & 0 & 0 & 4E_{ssg0}^* & E_{sc} \end{array} \tag{5.4.2}$$

Similarly with all 10 orbitals included, the (10x10) matrix becomes a (20x20) matrix :

$$[H_0(\vec{k})] = \begin{bmatrix}h(\vec{k}) & 0\\ 0 & h(\vec{k})\end{bmatrix} \tag{5.4.3}$$

where $[h(\vec{k})]$ is given by Eq.(5.3.3).

Spin-orbit coupling : If we were to calculate the bandstructure using Eq.(5.4.3) instead of Eq.(5.3.3) we would get exactly the same result, except that each line would have a second one right on top of it, which we would probably not even notice if a computer were plotting it out. But the reason we are doing this is that we want to add something called spin-orbit coupling to Eq.(5.4.3).

The Schrodinger equation is a non-relativistic equation. For electrons traveling at high velocities relativistic effects can become significant and we need to use the Dirac equation. Typically in solids the velocities are not high enough to require this, but the electric fields are very high near the nuclei of atoms leading to weak relativistic effects that can be accounted for by adding a spin-orbit correction H_{so} to the Schrodinger equation:

$$E\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix} = [H_0]\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix} + [H_{so}]\begin{pmatrix}\psi\\ \overline{\psi}\end{pmatrix} \tag{5.4.4}$$

$$\text{where} \quad H_0 = \begin{bmatrix} p^2/2m + U(\vec{r}) & 0 \\ 0 & p^2/2m + U(\vec{r}) \end{bmatrix} \tag{5.4.5}$$

$$\text{and} \quad H_{so} = \frac{q\hbar}{4m^2c^2}\begin{bmatrix} E_x p_y - E_y p_x & (E_y p_z - E_z p_y) - i(E_z p_x - E_x p_z) \\ (E_y p_z - E_z p_y) - i(E_z p_x - E_x p_z) & -(E_x p_y - E_y p_x) \end{bmatrix} \tag{5.4.6}$$

c being the velocity of light in vacuum. The spin-orbit Hamiltonian H_{so} is often written as

$$H_{so} = \frac{q\hbar}{4m^2c^2}\vec{\sigma}.\left(\vec{E} \times \vec{p}\right) \tag{5.4.7}$$

where the Pauli spin matrices $\vec{\sigma}$ are defined as

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \qquad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{5.4.8}$$

It is straightforward to show that the two expressions for the spin-orbit Hamiltonian H_{so} in Eqs.(5.4.7) and (5.4.6) are identical. I will not try to justify the origin of the

spin-orbit term for this would take us too far afield into the Dirac equation, but the interested reader may find Problems P.2.2-P.2.5 (at the end of this Chapter) instructive.

Bandstructure with spin-orbit coupling: We already have the matrix representation for the non spin-orbit part of the Hamiltonian, H_0 (Eq.(5.4.5)). It is given by Eq.(5.3.3). We now need to find a matrix representation for H_{SO} and add it to H_0. Let us first see what we would do if we were to just use the s-orbitals for each atom. Usually the spin-orbit matrix elements are significant only if both orbitals are centered on the same atom, so that we expect a matrix of the form

	$\|s_a\rangle$	$\|s_c\rangle$	$\|\bar{s}_a\rangle$	$\|\bar{s}_c\rangle$
$\|s_a\rangle$	a_{11}	0	a_{12}	0
$\|s_c\rangle$	0	c_{11}	0	c_{12}
$\|\bar{s}_a\rangle$	a_{21}	0	a_{22}	0
$\|\bar{s}_c\rangle$	0	c_{21}	0	c_{22}

We would fill up the '11' elements of this matrix by taking the matrix elements of the 11 component of H_{SO} (see Eq.(5.4.6) :

$$a_{11} = \langle s_a | E_x p_y - E_y p_x | s_a \rangle \qquad c_{11} = \langle s_c | E_x p_y - E_y p_x | s_c \rangle$$

To fill up the '12' elements of this matrix we take the matrix elements of the 12 component of H_{SO} (see Eq.(5.4.6) :

$$a_{12} = \langle s_a | (E_y p_z - E_z p_y) - i (E_z p_x - E_x p_z) | s_a \rangle$$
$$c_{12} = \langle s_c | (E_y p_z - E_z p_y) - i (E_z p_x - E_x p_z) | s_c \rangle$$

Similarly we can go on with the '21' and the '22' components. As it turns out, all these matrix elements can be shown to be zero from symmetry arguments if we assume the potential U(r) to be spherically symmetric as is reasonable for atomic potentials. The same is true for the s* orbitals as well. However, some of the matrix elements are non-zero when we consider the X, Y and Z orbitals. These non-zero matrix elements can all be expressed in terms of a single number δ_a for the anionic orbitals:

$$
\begin{array}{c|cccccc}
 & |X_a\rangle & |Y_a\rangle & |Z_a\rangle & |\bar{X}_a\rangle & |\bar{Y}_a\rangle & |\bar{Z}_a\rangle \\
|X_a\rangle & 0 & -i\,\delta_a & 0 & 0 & 0 & \delta_a \\
|Y_a\rangle & i\,\delta_a & 0 & 0 & 0 & 0 & -i\,\delta_a \\
|Z_a\rangle & 0 & 0 & 0 & -\delta_a & i\,\delta_a & 0 \\
|\bar{X}_a\rangle & 0 & 0 & -\delta_a & 0 & i\,\delta_a & 0 \\
|\bar{Y}_a\rangle & 0 & 0 & -i\,\delta_a & -i\,\delta_a & 0 & 0 \\
|\bar{Z}_a\rangle & \delta_a & i\,\delta_a & 0 & 0 & 0 & 0
\end{array}
\qquad (5.4.9a)
$$

and in terms of a single number δ_c for the cationic orbitals:

$$
\begin{array}{c|cccccc}
 & |X_c\rangle & |Y_c\rangle & |Z_c\rangle & |\bar{X}_c\rangle & |\bar{Y}_c\rangle & |\bar{Z}_c\rangle \\
|X_c\rangle & 0 & -i\,\delta_c & 0 & 0 & 0 & \delta_c \\
|Y_c\rangle & i\,\delta_c & 0 & 0 & 0 & 0 & -i\,\delta_c \\
|Z_c\rangle & 0 & 0 & 0 & -\delta_c & i\,\delta_c & 0 \\
|\bar{X}_c\rangle & 0 & 0 & -\delta_c & 0 & i\,\delta_c & 0 \\
|\bar{Y}_c\rangle & 0 & 0 & -i\,\delta_c & -i\,\delta_c & 0 & 0 \\
|\bar{Z}_c\rangle & \delta_c & i\,\delta_c & 0 & 0 & 0 & 0
\end{array}
\qquad (5.4.9b)
$$

If we were to find the eigenvalues of either of these matrices we would obtain four eigenvalues equal to $+\,\delta_c$ (or $+\,\delta_a$) and two eigenvalues equal to $-2\,\delta_c$ (or $-\,2\,\delta_a$). The splitting between these two sets of levels is $3\,\delta_c$ (or $3\,\delta_a$) and is referred to as the spin-orbit splitting Δ_c (or Δ_a):

$$\Delta_c \text{ (or } \Delta_a) = 3\,\delta_c \text{ (or } 3\,\delta_a) \qquad (5.4.10)$$

The spin-orbit splitting is well-known from both theory and experiment for all the atoms. For example, Gallium has a spin-orbit splitting of .013 eV while that for Arsenic is 0.38 eV. It is now straightforward to write down the full matrix representation for H_{SO} making use of Eqs.(5.4.8) and (5.4.9), adding it to Eq.(5.3.3) and then calculating the bandstructure. For GaAs we obtain the result shown in Fig.5.4.1a. For comparison, in Fig5.4.1b we have shown the results obtained directly from Eq.(5.3.3) without adding the spin-orbit part. This is basically the same plot obtained in the last Section (see Fig.5.3.2) except that the energy scale has been expanded to highlight the top of the valence band.

Fig.5.5.1. (a) Bandstructure of GaAs calculated taking spin-orbit interaction into account. The Γ-X direction is plotted along the positive axis while the Γ-L direction is plotted along the negative axis.

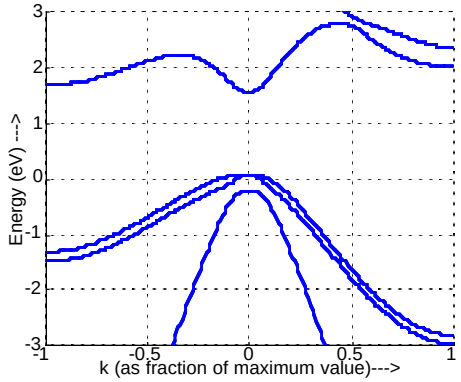
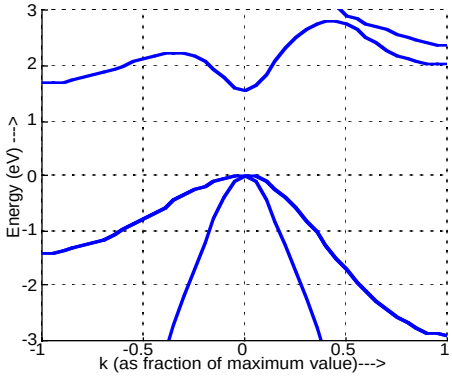


Fig.5.5.1. (b) Bandstructure of GaAs calculated from Eq. (5.2.3) without adding the spin-orbit component.



Heavy hole, light hole and split-off bands: The nature of the valence band wavefunction near the gamma point ($k_X = k_Y = k_Z = 0$) play a very important role in determining the optical properties of semiconductor nanostructures. At the gamma point, the Hamiltonian matrix has a relatively simple form because only g_0 is non-zero, while g_1, g_2 and g_3 are each equal to zero (see Eq.(5.2.3)). Including spin-orbit coupling the Hamiltonian decouples into four separate blocks at the gamma point :

Block I :

	$\|s_a\rangle$	$\|s_c\rangle$	$\|\bar{s}_a\rangle$	$\|\bar{s}_c\rangle$
$\|s_a\rangle$	E_{sa}	$4E_{ss}$	0	0
$\|s_c\rangle$	$4E_{ss}$	E_{sc}	0	0
$\|\bar{s}_a\rangle$	0	0	E_{sa}	$4E_{ss}$
$\|\bar{s}_c\rangle$	0	0	$4E_{ss}$	E_{sc}

Block II :

	$\|s_a^*\rangle$	$\|s_c^*\rangle$	$\|\bar{s}_a^*\rangle$	$\|\bar{s}_c^*\rangle$
$\|s_a^*\rangle$	E_{s^*a}	0	0	0
$\|s_c^*\rangle$	0	E_{s^*c}	0	0
$\|\bar{s}_a^*\rangle$	0	0	E_{s^*a}	0
$\|\bar{s}_c^*\rangle$	0	0	0	E_{s^*c}

Block III :

	$\|X_a\rangle$	$\|Y_a\rangle$	$\|\bar{Z}_a\rangle$	$\|X_c\rangle$	$\|Y_c\rangle$	$\|\bar{Z}_c\rangle$
$\|X_a\rangle$	E_{pa}	$-i\,\delta_a$	δ_a	$4E_{xx}$	0	0
$\|Y_a\rangle$	$i\,\delta_a$	E_{pa}	$-i\,\delta_a$	0	$4E_{xx}$	0
$\|\bar{Z}_a\rangle$	δ_a	$i\,\delta_a$	E_{pa}	0	0	$4E_{xx}$
$\|X_c\rangle$	$4E_{xx}$	0	0	E_{pc}	$-i\,\delta_c$	δ_c
$\|Y_c\rangle$	0	$4E_{xx}$	0	$i\,\delta_c$	E_{pc}	$-i\,\delta_c$
$\|\bar{Z}_c\rangle$	0	0	$4E_{xx}$	δ_c	$i\,\delta_c$	E_{pc}

Block IV :

	$\|\bar{X}_a\rangle$	$\|\bar{Y}_a\rangle$	$\|Z_a\rangle$	$\|\bar{X}_c\rangle$	$\|\bar{Y}_c\rangle$	$\|Z_c\rangle$
$\|\bar{X}_a\rangle$	E_{pa}	$i\,\delta_a$	$-\delta_a$	$4E_{xx}$	0	0
$\|\bar{Y}_a\rangle$	$-i\,\delta_a$	E_{pa}	$-i\,\delta_a$	0	$4E_{xx}$	0
$\|Z_a\rangle$	$-\delta_a$	$i\,\delta_a$	E_{pa}	0	0	$4E_{xx}$
$\|\bar{X}_c\rangle$	$4E_{xx}$	0	0	E_{pc}	$i\,\delta_c$	$-\delta_c$
$\|\bar{Y}_c\rangle$	0	$4E_{xx}$	0	$-i\,\delta_c$	E_{pc}	$-i\,\delta_c$
$\|Z_c\rangle$	0	0	$4E_{xx}$	$-\delta_c$	$i\,\delta_c$	E_{pc}

We can partially diagonalize Blocks III and IV by transforming to the heavy hole (HH), light hole (LH) and split-off (SO) basis using the transformation matrix

$[V] =$

	$\|HH_a\rangle$	$\|LH_a\rangle$	$\|SO_a\rangle$	$\|HH_c\rangle$	$\|LH_c\rangle$	$\|SO_c\rangle$
$\|X_a\rangle$	$1/\sqrt{2}$	$1/\sqrt{6}$	$1/\sqrt{3}$	0	0	0
$\|Y_a\rangle$	$i/\sqrt{2}$	$-i/\sqrt{6}$	$-i/\sqrt{3}$	0	0	0
$\|\bar{Z}_a\rangle$	0	$\sqrt{2/3}$	$-1/\sqrt{3}$	0	0	0
$\|X_c\rangle$	0	0	0	$1/\sqrt{2}$	$1/\sqrt{6}$	$1/\sqrt{3}$
$\|Y_c\rangle$	0	0	0	$i/\sqrt{2}$	$-i/\sqrt{6}$	$-i/\sqrt{3}$
$\|\bar{Z}_c\rangle$	0	0	0	0	$\sqrt{2/3}$	$-1/\sqrt{3}$

and the usual rule for transformation, namely, $[H]_{new} = [V^+]\,[H]_{old}\,[V]$. The transformed Hamiltonian for Block III looks like

	$\lvert HH_a\rangle$	$\lvert LH_a\rangle$	$\lvert SO_a\rangle$	$\lvert HH_c\rangle$	$\lvert LH_c\rangle$	$\lvert SO_c\rangle$
$\lvert HH_a\rangle$	$E_{pa} + \delta_a$	0	0	$4E_{xx}$	0	0
$\lvert LH_a\rangle$	0	$E_{pa} + \delta_a$	0	0	$4E_{xx}$	0
$\lvert SO_a\rangle$	0	0	$E_{pa} - 2\ \delta_a$	0	0	$4E_{xx}$
$\lvert HH_c\rangle$	$4E_{xx}$	0	0	$E_{pc} + \delta_c$	0	0
$\lvert LH_c\rangle$	0	$4E_{xx}$	0	0	$E_{pc} + \delta_c$	0
$\lvert SO_c\rangle$	0	0	$4E_{xx}$	0	0	$E_{pc} - 2\ \delta_c$

Note how the three bands are neatly decoupled so that at the gamma point we can label the energy levels as HH, LH and SO. As we move away from the gamma point, the bands are not decoupled any more and the eigenstates are represented by superpositions of HH, LH and SO.

Similarly Block IV can be transformed using the transformation matrix

$$[\overline{V}] = \begin{array}{c|cccccc} & \lvert\overline{HH}_a\rangle & \lvert\overline{LH}_a\rangle & \lvert\overline{SO}_a\rangle & \lvert\overline{HH}_c\rangle & \lvert\overline{LH}_c\rangle & \lvert\overline{SO}_c\rangle \\ \hline \lvert\overline{X}_a\rangle & 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{3} & 0 & 0 & 0 \\ \lvert\overline{Y}_a\rangle & -i/\sqrt{2} & i/\sqrt{6} & i/\sqrt{3} & 0 & 0 & 0 \\ \lvert Z_a\rangle & 0 & -\sqrt{2/3} & 1/\sqrt{3} & 0 & 0 & 0 \\ \lvert\overline{X}_c\rangle & 0 & 0 & 0 & 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{3} \\ \lvert\overline{Y}_c\rangle & 0 & 0 & 0 & -i/\sqrt{2} & i/\sqrt{6} & i/\sqrt{3} \\ \lvert Z_c\rangle & 0 & 0 & 0 & 0 & -\sqrt{2/3} & 1/\sqrt{3} \end{array}$$

to obtain

	$\lvert\overline{HH}_a\rangle$	$\lvert\overline{LH}_a\rangle$	$\lvert\overline{SO}_a\rangle$	$\lvert\overline{HH}_c\rangle$	$\lvert\overline{LH}_c\rangle$	$\lvert\overline{SO}_c\rangle$
$\lvert\overline{HH}_a\rangle$	$E_{pa} + \delta_a$	0	0	$4E_{xx}$	0	0
$\lvert\overline{LH}_a\rangle$	0	$E_{pa} + \delta_a$	0	0	$4E_{xx}$	0
$\lvert\overline{SO}_a\rangle$	0	0	$E_{pa} - 2\ \delta_a$	0	0	$4E_{xx}$
$\lvert\overline{HH}_c\rangle$	$4E_{xx}$	0	0	$E_{pc} + \delta_c$	0	0
$\lvert\overline{LH}_c\rangle$	0	$4E_{xx}$	0	0	$E_{pc} + \delta_c$	0
$\lvert\overline{SO}_c\rangle$	0	0	$4E_{xx}$	0	0	$E_{pc} - 2\ \delta_c$

It is important to note that the eigenstates (which can be identified by looking at the columns of [V] or $[\overline{V}]$) are not pure upspin or pure downspin states. However, we could view the lower block $[\overline{V}]$ as the spin-reversed counterpart of the upper block [V] since it is straightforward to show that they are orthogonal, as we expect "up" and "down" spin states to be.

5.5. Supplementary notes: Dirac equation

Relativistic electrons are described by the Dirac equation

$$E\begin{pmatrix}\psi\\ \overline{\psi}\\ \phi\\ \overline{\phi}\end{pmatrix} = \begin{bmatrix} mc^2+U & 0 & cp_z & c(p_x-ip_y)\\ 0 & mc^2+U & c(p_x+ip_y) & -cp_z\\ cp_z & c(p_x-ip_y) & -mc^2+U & 0\\ c(p_x+ip_y) & -cp_z & 0 & -mc^2+U\end{bmatrix}\begin{pmatrix}\psi\\ \overline{\psi}\\ \phi\\ \overline{\phi}\end{pmatrix}$$

which can be written compactly as

$$E\begin{Bmatrix}\Psi\\ \Phi\end{Bmatrix} = \begin{bmatrix}\left(mc^2+U\right)I & c\,\vec{\sigma}.\vec{p}\\ c\,\vec{\sigma}.\vec{p} & \left(-mc^2+U\right)I\end{bmatrix}\begin{Bmatrix}\Psi\\ \Phi\end{Bmatrix} \qquad (5.5.1)$$

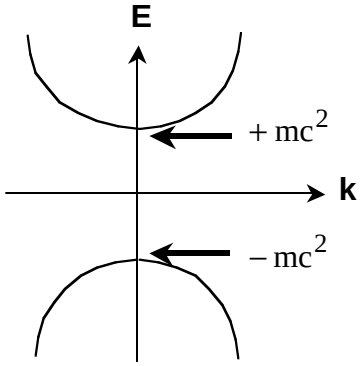
where $\Psi \equiv \begin{Bmatrix}\psi\\ \overline{\psi}\end{Bmatrix}$ and $\Phi \equiv \begin{Bmatrix}\phi\\ \overline{\phi}\end{Bmatrix}$

Assuming U = 0 and substituting a plane wave solution of the form

$$\begin{pmatrix}\Psi\\ \Phi\end{pmatrix} = \begin{pmatrix}\Psi\\ \Phi\end{pmatrix}e^{i\vec{k}.\vec{r}}$$

we can show that the dispersion relation is given by

$$E(\vec{k}) = \pm\sqrt{m^2c^4+c^2\hbar^2k^2}$$



which has two branches as shown.

The negative branch is viewed as being completely filled even in vacuum. The separation between the two branches is $2mc^2$ which is approximately 1 MeV, well outside the range of energies encountered in solid-state experiments. In high energy experiments electrons are excited out of the negative branch into the positive branch resulting in the creation of electron-positron pairs. But in common solid-state experiments energy exchanges are less than 10 eV and the negative branch provides an

inert background. At energies around $E = mc^2$, we can do a binomial expansion of Eq.(5.5.1) to obtain the non-relativistic parabolic relation (apart from an additive constant equal to the relativistic rest energy mc^2) :

$$E(\vec{k}) \approx mc^2 + \left(\hbar^2 k^2/2m\right)$$

Relativistic corrections like the spin-orbit term are obtained by starting from the Dirac equations and eliminating the component Φ using approximate procedures valid at energies sufficiently small compared to mc^2.

Non-relativistic approximation to the Dirac equation: Starting from Eq.(5.5.1) we can show that

$$E\{\Psi\} = \left(mc^2 + U\right)\{\Psi\} + [c\,\vec{\sigma}.\vec{p}]\left[\frac{1}{E + mc^2 - U}\right]^{-1}[c\,\vec{\sigma}.\vec{p}]\{\Psi\}$$

Setting $E \approx mc^2$ on the right hand side, we obtain the lowest order non-relativistic approximation

$$E\{\Psi\} = \left(mc^2 + U\right)\{\Psi\} + \frac{[\vec{\sigma}.\vec{p}]^2}{2m}\{\Psi\} \qquad (5.5.2)$$

which can be simplified to yield Eq.(5.4.1):

$$\left(E - mc^2\right)\{\Psi\} = \begin{bmatrix} U + p^2/2m & 0 \\ 0 & U + p^2/2m \end{bmatrix}\{\Psi\}$$

noting that $[\vec{\sigma}.\vec{p}]^2 = \begin{bmatrix} p^2 & 0 \\ 0 & p^2 \end{bmatrix}$

Effect of magnetic field: One question we will not discuss much in this book is the effect of a magnetic field on the electron energy levels. The effect is incorporated into the Dirac equation by replacing $\vec{p}$ with $\vec{p} + q\vec{A}$:

$$E\begin{Bmatrix}\Psi\\ \Phi\end{Bmatrix} = \begin{bmatrix}\left(mc^2+U\right)I & c\,\vec{\sigma}.\left(\vec{p}+q\vec{A}\right)\\ c\,\vec{\sigma}.\left(\vec{p}+q\vec{A}\right) & \left(-mc^2+U\right)I\end{bmatrix}\begin{Bmatrix}\Psi\\ \Phi\end{Bmatrix}$$

As before (cf. Eq.(5.5.2)) we can obtain the lowest order non-relativistic approximation

$$E\{\Psi\} = \left(mc^2+U\right)\{\Psi\} + \frac{\left[\vec{\sigma}.\left(\vec{p}+q\vec{A}\right)\right]^2}{2m}\{\Psi\}$$

which can be simplified to yield the Pauli equation:

$$\left(E-mc^2\right)\{\Psi\} = \left[U+\left(\vec{p}+q\vec{A}\right)^2/2m\right][I]\{\Psi\} + \mu_B\,\vec{\sigma}.\vec{B}\{\Psi\}$$

where $\mu_B \equiv q\hbar/2m$ (Bohr magneton), and $\vec{B} = \vec{\nabla}\ x\ \vec{A}$

The second term $\mu_B\,\vec{\sigma}.\vec{B}$ is called the Zeeman term. Note that the spin-orbit term in Eq.(5.4.7) can be viewed as the Zeeman term due to an effective magnetic field given by

$$B_{SO} = \left(\vec{E}\ x\ \vec{p}\right)/2mc^2$$

Indeed, one way to rationalize the spin-orbit term is to say that an electron in an electric field sees this effective magnetic field due to "relativistic effects". To obtain the spin-orbit term directly from the Dirac equation it is necessary to go to the next higher order [5.3].

Exercises

E.5.1. Set up the (2x2) matrix given in Eq.(5.1.10) for the one-dimensional dimerized toy solid and plot the E(k) relation, ***cf. Fig.5.1.5.***

E.5.2. Set up the (10x10) matrix given in Eq.(5.3.3) using the parameters for GaAs given in the text and plot the dispersion relation $E(k_X,k_Y,k_Z)$ along Γ-X and Γ-L as shown in ***Fig.5.3.2***.

E.5.3. Set up the (20x20) matrix including the spin-orbit coupling as described in Section 4.4 for GaAs and plot E vs. k along Γ-X and Γ-L for GaAs (Δ_C = .013 eV and Δ_a = .38 eV) and compare with ***Fig.5.4.1a, b***.