"Tight Binding" Method: Linear Combination of Atomic Orbitals (LCAO)

W. E. Pickett
(June 6, 1999)

This write-up is a homemade introduction to the tight binding representation of the electronic structure of crystalline solids. This information is important in the parametrization of the band structures of real solids and for the underlying character of model Hamiltonians for correlated electron studies, as well as for other uses. Those interested in the parametrization of band structures of real materials should consult the book by

D. A. Papaconstantopoulos,
Handbook of the Band Structure of Elemental Solids, (Plenum, New York,1986).

For the general TB method, one should consult the original classic work by

J. C. Slater and G. F. Koster, Phys. Rev. **94**, 844 (1954).

I. INTRODUCTION TO TIGHT BINDING THEORY

Since a crystal is made up of a periodic array of atoms, it may seem peculiar that when we think of Bloch electron wavefunctions in solids it is often in terms of wavy modulations that don't pay much attention to just where the atoms sit. Indeed, in simple metals and covalent semiconductors that is a good picture: the crystal potential that enters into the Hamiltonian is a smooth function and atomic sites *per se* are not critical in the understanding (although they are in the underlying description of covalent semiconductors).

There is in fact a common picture – the tight binding model – that is based on the "collection of atoms" viewpoint. It is most appropriate when electrons move through the crystal slowly (or not at all, as in insulators) and therefore 'belong' to an atom for an appreciable time before they move on. The electrons are in some sense *tightly bound* to the atom and only hop because staying put on a simple atom costs a bit too much energy. The TB model is not readily applicable to simple (free or nearly free electron) metals, but it is quite good for a wide variety of other solids. It is interesting that covalent semiconductors can be described well from either viewpoint.

A. Crystal as a collection of atoms

A good approximation for the electron's potential $V(\vec{r})$ in a crystal is the sum of atomic potentials:

$$V(\vec{r}) = \sum_{\vec{R}} V_{at}(\vec{r} - \vec{R}), \tag{1.1}$$

where the sum runs over lattice vectors. We will not worry about the considerable non-uniqueness of this decomposition. For parametrization purposes, this non-uniqueness is irrelevant (only the *form* of the resulting parameters is relevant), while if the method is intended for real electronic structure calculations, the non-uniqueness often is used to make numerical procedures as convenient as possible.

This potential is periodic by construction:

$$\begin{aligned} V(\vec{r}+\vec{R}_o) &= \sum_{\vec{R}} V_{at}(\vec{r}+\vec{R}_o-\vec{R}) \\ &= \sum_{\vec{R}} V_{at}(\vec{r}-(\vec{R}-\vec{R}_o)) \\ &= \sum_{\vec{R}'} V_{at}(\vec{r}-\vec{R}') \\ &\equiv V(\vec{r}), \end{aligned} \tag{1.2}$$

where the change of summation index $\vec{R} \to \vec{R}' = \vec{R} - \vec{R}_o$ was made. ($\vec{R}_o$ is a lattice vector.)

The crystal Hamiltonian is ($\hbar^2/2m \equiv 1$)

$$H = -\nabla^2 + V(\vec{r}). \tag{1.3}$$

Later we will generalize the situation to cover several atoms in the unit cell.

B. Periodic array of atomic orbitals

Shouldn't the electron wavefunction in the crystal be related to the *atomic* orbitals, which satisfy

$$H_{at}\phi_n \equiv (-\nabla^2 + V_{at})\phi_n = \varepsilon_n \phi_n. \tag{1.4}$$

We might try a linear combination

$$\Phi_n(\vec{r}) = \sum_{\vec{R}} \phi_n(\vec{r} - \vec{R}); \tag{1.5}$$

is this a Block function that can be put in the form

$$\Phi(\vec{r}) = e^{i\vec{k}\cdot\vec{r}} u_{\vec{k}}(\vec{r})? \tag{1.6}$$

Indeed it is (**prove** it), *but* only for $\vec{k} = 0$. We want, and need, Bloch-like functions for arbitrary $\vec{k}$ within the first Brillouin zone.

C. Bloch Sums

A much better choice of candidate for a crystal wavefunction is to form the "Bloch sums" of atomic orbitals, given by

$$B_{n,\vec{k}}(\vec{r}) = N^{-\frac{1}{2}} \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}} \phi_n(\vec{r} - \vec{R}). \tag{1.7}$$

This is called a Bloch sum because it produces a function that satisfies the Bloch condition for wavevector $\vec{k}$:

$$\begin{aligned} N^{\frac{1}{2}} B_{n,\vec{k}}(\vec{r} + \vec{R}_o) &= \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}} \phi_n(\vec{r} + \vec{R}_o - \vec{R}) \\ &= \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}} \phi_n(\vec{r} - (\vec{R} - \vec{R}_o)) \\ &= \sum_{\vec{R}'} e^{i\vec{k}\cdot(\vec{R}' + \vec{R}_o)} \phi_n(\vec{r} - \vec{R}') \\ &= e^{i\vec{k}\cdot\vec{R}_o} \sum_{\vec{R}'} e^{i\vec{k}\cdot\vec{R}'} \phi_n(\vec{r} - \vec{R}'), \end{aligned} \tag{1.8}$$

so

$$B_{n,\vec{k}}(\vec{r} + \vec{R}_o) = e^{i\vec{k}\cdot\vec{R}_o} B_{n,\vec{k}}(\vec{r}). \tag{1.9}$$

This result is equivalent to the Bloch form. (**Prove** it by manipulating it into the block form, and find out what $u_{\vec{k}}(\vec{r})$ is.)

It suffices to confine $\vec{k}$ in Eq. (1.7) to the 1st Brillouin zone. If the $\vec{k}$ point were of the form $\vec{k}_r + \vec{K}$, where $\vec{k}_r$ (the *reduced wavevector*) is in the 1st Brillouin zone and $\vec{K}$ is a reciprocal lattice vector, note that

$$e^{i\vec{k}\cdot\vec{R}} = e^{i\vec{k}_r\cdot\vec{R}} \tag{1.10}$$

for any and all lattice vectors $\vec{R}$. (**Why?**)

D. Proceeding toward the eigenfunctions

For many solids there will be several types (s, p or d) of atomic states in the valence region, and that is why we have kept the index n. In the solid these atomic states will mix with each other due to the overlap of atomic orbitals on neighboring atoms (as we will see). A Bloch sum of atomic orbitals itself is not an eigenfunction for the crystal. It is important to allow the valence wavefunction in the solid to be some of *each* of the atomic functions, with the actual amounts to be determined by solving Schrödinger's equation. Thus we try expressing the electron wavefunction in the crystal as a bit $b_{n,\vec{k}}$ of each of the Bloch sums,

$$\psi_{\vec{k}}(\vec{r}) = \sum_n b_n(\vec{k}) B_{n,\vec{k}}(\vec{r}), \tag{1.11}$$

where the coefficients b gives the amount of Bloch sum B_n in the crystal wavefunction. The Bloch sums become the *basis functions* that we express the wavefunction in terms of.

Now we want to learn how to find the wavefunctions. The condition is that they be solutions of the Schrödinger equation

$$H\psi_{\vec{k}} = \varepsilon_{\vec{k}} \psi_{\vec{k}}. \tag{1.12}$$

The eigenvalues $\varepsilon_{\vec{k}}$ will be the *energy bands* of the crystal. But how do we solve for $\varepsilon_{\vec{k}}$ and the coefficients $b_n(\vec{k})$?

E. The matrix equation

In quantum mechanics generally, a good try is to take *matrix elements* (integrals between basis functions) and reduce the problem to a matrix equation. In this case, the thing to do is to multiply the Schrödinger equation on the left by another Bloch sum $B^*_{m,\vec{k}}$ and integrate over the crystal. (Note that we have not chosen a Bloch function corresponding to another wavevector $\vec{k}' \neq \vec{k}$. **Why** didn't we do so?)

The result is

$$\sum_n H_{m,n}(\vec{k}) b_n(\vec{k}) = \varepsilon_{\vec{k}} \sum_n S_{m,n}(\vec{k}) b_n(\vec{k}), \tag{1.13}$$

where

$$H_{m,n}(\vec{k}) \equiv \int B^*_{m,\vec{k}}(\vec{r}) H B_{n,\vec{k}}(\vec{r}), \tag{1.14}$$

and

$$S_{m,n}(\vec{k}) \equiv \int B^*_{m,\vec{k}}(\vec{r}) B_{n,\vec{k}}(\vec{r}). \tag{1.15}$$

These matrices are called the Hamiltonian matrix and the overlap matrix, respectively, where m and n are the matrix indices. Writing the matrices implicitly (without displaying the indices), the equation becomes

$$H(\vec{k}) b(\vec{k}) = \varepsilon_{\vec{k}} S(\vec{k}) b(\vec{k}) \tag{1.16}$$

or

$$\{H(\vec{k}) - \varepsilon_{\vec{k}} S(\vec{k})\} b(\vec{k}) = 0. \quad (1.17)$$

This is a linear algebra problem (generalized eigenvalue problem), but we have to know what the matrices H and S are.

F. The H and S matrices

Well, we never said that this wouldn't get a little messy. However, the messiness does clear up soon. Substituting in the Bloch sum forms for B^*_m and B_n within the integral, we have the expression

$$\begin{aligned} H_{m,n}(\vec{k}) &= (N^{-\frac{1}{2}})^2 \sum_{\vec{R}_1,\vec{R}_2} e^{i\vec{k}\cdot(\vec{R}_2-\vec{R}_1)} \\ &\quad \times \int \phi^*_m(\vec{r}-\vec{R}_1) H \phi_n(\vec{r}-\vec{R}_2) \\ &\equiv \frac{1}{N} \sum_{\vec{R}_1,\vec{R}_2} e^{i\vec{k}\cdot(\vec{R}_2-\vec{R}_1)} H_{m,n}(\vec{R}_2-\vec{R}_1). \quad (1.18) \end{aligned}$$

Because the Hamiltonian is cell periodic, the matrix element on the right hand side depends only on the difference between $\vec{R}_2$ and $\vec{R}_1$. We can then change the summation index $\vec{R}_2$ to $\vec{R}_2 - \vec{R}_1 = \vec{R}$, and the index $\vec{R}_1$ no longer appears on the right hand side. Then the sum over $\vec{R}_1$ just gives the factor N, the number of unit cells in our "crystal." Then we obtain

$$H_{m,n}(\vec{k}) = \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}} H_{m,n}(\vec{R}). \quad (1.19)$$

There should be no confusion between $H_{m,n}(\vec{k})$ and $H_{m,n}(\vec{R})$ in practice; in fact, they are *lattice Fourier transforms* of each other.

G. The Real Space TB Matrix Elements

The expression for the real space integral is

$$H_{m,n}(\vec{R}) = \int \phi^*_m(\vec{r}) H \phi_n(\vec{r}-\vec{R}), \quad (1.20)$$

i.e. it indicates the amount by which the Hamiltonian H *couples* atomic orbital ϕ_m on the site at the origin to the atomic orbital ϕ_n that is located at site $\vec{R}$. Physically, $H_{m,n}(\vec{R})$ is the amplitude that an electron in orbital ϕ_n at site $\vec{R}$ will hop to the orbital ϕ_m at the origin under the action of the Hamiltonian. One limit is easy to see: if $|\vec{R}|$ is so large that either one or the other of the orbitals is vanishingly small everywhere (no overlap), then the integral is negligible. Thus we can confine ourselves to values of $\vec{R}$ that connect an atom to only a few near neighbors.

All of this discussion applies as well to S, by just removing the Hamiltonian H from inside the integral. $S_{m,n}(\vec{R})$ in fact is called the *overlap* of $\phi_m(\vec{r})$ and $\phi_n(\vec{r}-\vec{R})$. Note that, if the orbitals are normalized (as is always possible, and is always assumed), the $S_{m,m}(\vec{0})$=1 for all m, *i.e.* the diagonal elements of S are unity.

H. Several atoms in the unit cell

So far, the notation has been limited to a elemental crystal. More generally, one encounters compounds where there are various atoms in the cell, which can be labelled by their position $\vec{\tau}_i$ with respect to the origin of the cell ($\vec{R}$). Then the basis orbitals are

$$\phi_{m,i,R} \equiv \phi_{m,i}(\vec{r}-\vec{R}-\vec{\tau}_i). \quad (1.21)$$

Then, generalizing from the Bloch sum defined in Eq. (1.7), one has the basis Bloch sums

$$B_{m,i,\vec{k}}(\vec{r}) = N^{-\frac{1}{2}} \sum_{\vec{R}} e^{i\vec{k}\cdot(\vec{R}+\vec{\tau}_i)} \phi_n(\vec{r}-\vec{R}-\vec{\tau}_i). \quad (1.22)$$

Then, instead of Eq. (1.18) we obtain

$$\begin{aligned} &H_{m,i;n,j}(\vec{k}) \\ &\quad = (N^{-\frac{1}{2}})^2 \sum_{\vec{R}_1,\vec{R}_2} e^{i\vec{k}\cdot(\vec{R}_2+\vec{\tau}_j-\vec{R}_1-\vec{\tau}_i)} \\ &\quad \times \int \phi^*_m(\vec{r}-\vec{R}_1-\vec{\tau}_i) H \phi_n(\vec{r}-\vec{R}_2-\vec{\tau}_j) \\ &\quad = \frac{1}{N} e^{i\vec{k}\cdot(\vec{\tau}_j-\vec{\tau}_i)} \sum_{\vec{R}_1,\vec{R}_2} e^{i\vec{k}\cdot(\vec{R}_2-\vec{R}_1)} \\ &\quad \times H_{m,i;n,j}(\vec{R}_2+\vec{\tau}_j-\vec{R}_1-\vec{\tau}_i) \\ &\quad = e^{-i\vec{k}\cdot\vec{\tau}_i} \Big(\sum_{\vec{R}} H_{m,i;n,j}(\vec{R}) e^{i\vec{k}\cdot\vec{R}}\Big) e^{i\vec{k}\cdot\vec{\tau}_j} \\ &\quad = e^{-i\vec{k}\cdot\vec{\tau}_i} H^o_{m,i;n,j}(\vec{k}) e^{i\vec{k}\cdot\vec{\tau}_j}. \quad (1.23) \end{aligned}$$

Here the notation in the next-to-last line has been shortened using

$$H_{m,i;n,j}(\vec{R}) \equiv H_{m,i;n,j}(\vec{R}+\vec{\tau}_j-\vec{\tau}_i). \quad (1.24)$$

The overlap matrix behaves in an exactly analogous way.

Eq. (1.23) can be viewed as the matrix $\hat{H}^0(\vec{k})$ transformed by the unitary transformation

$$U_{m,i;n,j}(\vec{k}) = e^{i\vec{k}\cdot\vec{\tau}_j}\delta_{n,m}\delta_{i,j}, \tag{1.25}$$

which can easily be shown to obey the unitarity conditions

$$\hat{U}^\dagger\hat{U} = \hat{1} = \hat{U}^{-1}\hat{U}. \tag{1.26}$$

However, a unitary transformation of a Hermitian matrix does not affect its eigenvalues, but merely transforms the eigenvectors. So, unless there is some specific reason for doing so, the phase factors in the last line of Eq. 1.23) (*i.e.* $\hat{U}$) can be disregarded. Looked at the other way, the one may *include* any additional unitary transformation that one desires. There are occasionally good reasons for making the transformation – for checking various aspects of the code, for example, or more importantly for reducing the secular equation to a real equation in cases where there is a center of inversion in the crystal but the original choice of orbitals produced a complex secular equation.

II. APPLICATIONS OF THE TIGHT BINDING MODEL

A. Single Site Terms

It may not have been obvious, but the mathematics is finished; the problem has been solved. Obtaining the energy bands $\varepsilon_{\vec{k}}$ (in general there will be several bands of them, labelled $\varepsilon_{\vec{k},n}$) and the expansion coefficients $b_n(\vec{k})$ of the wavefunctions amounts to solving the equation (1.16) or (1.17), with the matrix H given by Eq. (1.19) and S given by an analogous expression:

$$\begin{aligned} S_{m,n}(\vec{k}) &= N^{-1}\sum_{\vec{R}_1,\vec{R}_2} e^{i\vec{k}\cdot(\vec{R}_2-\vec{R}_1)}S_{m,n}(\vec{R}_2-\vec{R}_1) \\ &= \sum_{\vec{R}} e^{i\vec{k}\cdot\vec{R}}S_{m,n}(\vec{R}). \end{aligned} \tag{2.27}$$

First we will look at the $\vec{R} = \vec{0}$ terms, where both orbitals are on the same site (at our origin). Now it is helpful to write the crystal Hamiltonian as the atomic Hamiltonian for the atom at the origin, plus the potential from all of the other atoms:

$$\begin{aligned} H &= -\nabla^2 + V_{at}(\vec{r}) + \sum_{\vec{R}\neq 0} V_{at}(\vec{r}-\vec{R}) \\ &= -\nabla^2 + V_{at}^{sph}(|\vec{r}|) + V_{at}^{non-sph}(\vec{r}) + \sum_{\vec{R}\neq 0} V_{at}(\vec{r}-\vec{R}) \\ &= H_{ata^{sph}}(\vec{r}) + \Delta V(\vec{r}). \end{aligned} \tag{2.28}$$

The integral results primarily from the first part, which gives the "atomic" eigenvalues $\{\varepsilon_m\}$ for a spherical atomic Hamiltonian H_{at}^{sph}. Then because atomic orbitals on the same atom are orthogonal to each other, the on-site matrix element is

$$\int \phi_m^*(\vec{r})H_{at}(\vec{r})\phi_n(\vec{r}) = \varepsilon_n\delta_{m,n}, \tag{2.29}$$

i.e. just the atomic eigenvalue.

B. Crystal Field Splitting

The quantity $\Delta V(\vec{r})$ in Eq. (2.22) is not well specified, but we do know that it has the symmetry of the atom in questions (defining the origin in this equation). This symmetry is never spherical in a solid, but discrete, such as a mirror plane (m), a 3-fold rotation in an axis containing inversion together with a mirror plane ($\bar{3}m$), etc. This *crystal field*, that is, the non-spherical potential that arises because the atom is in a crystal, will split some eigenvalues that would be degenerate in a spherical potential. The most common such situation is for the five d orbitals in a cubic crystal field, which are split into the threefold representation called t_{2g} (xy, yz, zx) and the twofold representation known as e_g $(x^2-y^2, 3z^2-1)$. Crystal fields are also very important for $4f$ and $5f$ ions, where the designation may become more complex because intra-atomic coupling (correlation) is important. Although not widely recognized, the oxygen ion in an axial site (such as in the perovskite structure) has crystal field split p levels: twofold (x, y) and a singlet (z). What this means is that instead of a single on-site ("atomic") energy ε_d for a transition metal ion in a cubic site, one has two energies $\varepsilon_{t_{2g}}$ and ε_{e_g}, or even more disticnt ones if the symmetry is lower.

C. Three Center vs. Two Center

A general Hamiltonian matrix element in Eq. (1.20) contains atomic potentials on a third atom, besides the sites upon which the orbitals are centered. The general form of the matrix elements then contains *three center integrals*. Slater and Koster introduced, and advocated, the use of the *two-center approximation* (2CA), in which three center contributions to the parameters are neglected. There had already been suggestions that such three center terms were negligible and, while pointing out that they are not negligible in a serious calculation (borne out by many subsequent studies), SK suggested that for the purposes of parametrizing information (from experiment or from band calculation) a 2CA is likely to be reasonable. Table I in their paper provides the expression for s, p, and d matrix elements in terms of

two center integrals, denoted $(ss\sigma), (pd\pi), (dd\delta)$, etc. The 2CA is very widely, almost universally, used.

As an example of the validity, or lack thereof, of the 2CA, we consider the case of Nb studied by Pickett and Allen. [1] They performed a full three center fit to the augmented plane wave bands of L. F. Mattheiss for Nb (also Mo) at 55 k points in the irreducible zone, using an $s+p+d$ basis set (9×9) matrix, and including parameters out to third neighbors. (Several of the 31 parameters are found to be negligibly small.) As a test of the 2CA in Nb, we can look at the pd parameters for nearest neighbors in the bcc lattice, which lie along (111) directions. The values of the independent paramters (in mRy), together with the expression in the 2CA, are

$$-42 = H_{x,xy} \to \frac{1}{3}(pd\sigma) + \frac{1}{3\sqrt{3}}(pd\pi) \tag{2.30}$$

$$+16 = H_{x,x^2-y^2} \to 0(pd\sigma) + \frac{1}{3\sqrt{3}}(pd\pi) \tag{2.31}$$

$$-51 = H_{x,yz} \to \frac{1}{3}(pd\sigma) - \frac{2}{3\sqrt{3}}(pd\pi). \tag{2.32}$$

Since the 2nd equation determines $(pd\pi) = 28$ mRy (= 0.38 eV), the other two equations can be used separately to "determine" $(pd\sigma)$. The two values thus determined are -142 mRy and -131 mRy. Since these do not differ greatly, one can use the 2CA with $(pd\sigma)$ = -136 mRy (the mean) with acceptable loss of accuracy.

The dd parameters are the crucial ones in Nb. For nearest neighbors there are four three center parameters. Denoting $(dd\sigma), (dd\pi), (dd\delta)$ by S, P, D, respectively, the corresponding values and 2CA expressions are given by

$$-22 = H_{xy,xy} \to \frac{1}{3}S + 0P + \frac{4}{9}D \tag{2.33}$$

$$-35 = H_{xy,zx} \to \frac{1}{3}S - \frac{1}{9}P - \frac{2}{9}D \tag{2.34}$$

$$-31 = \sqrt{3}H_{xy,3z^2-r^2} \to 0S - \frac{2}{3}P + \frac{2}{3}D \tag{2.35}$$

$$+31 = H_{3z^2-r^2,3z^2-r^2} \to \frac{2}{9}S + \frac{2}{3}P + \frac{1}{3}D \tag{2.36}$$

The 3rd of these equations has been multiplied through by $\sqrt{3}$ to facilitate solution by elimination using various combinations of the equations. Subtracting the 2nd from the 1st gives

$$+13 = 0S + \frac{1}{9}P + \frac{2}{3}D, \tag{2.37}$$

which can be combined with the 3rd and 4th equations (separately) to produce the solutions: {P,D}={xx,yy} and {P,D}={xx,yy}.

D. Introducing the usual terminology

Now we consider an intersite term. However, we do not intend here to do a serious calculation; rather, we want to identify the important quantities (often called *parameters*) and choose likely values and learn some things about the general behavior of the electronic system. In fact, we'll just define a TB parameter

$$t_{m,n}(\vec{R}) \equiv H_{m,n}(\vec{R}) \tag{2.38}$$

because t is the usual notation for these integrals. This integral (see above) indicates how easy it is for an electron – whose behavior after all is determined by the Hamiltonian – to "hop" from orbital n on site $\vec{R}$ to orbital m at the origin. The $\{t\}$ constants are called *hopping parameters* or *hopping amplitudes*. The on-site ($\vec{R}=0$) term was given in Eq. (2.23), which we repeat here to jog the memory:

$$t_{m,n}(\vec{0}) = \varepsilon_n \delta_{m,n}, \tag{2.39}$$

which is the atomic energy level corresponding to orbital ϕ_n.

For the overlap matrix we have similarly the notation

$$s_{m,n}(\vec{R}) = S_{m,n}(\vec{R}); \quad s_{m,n}(0) = \delta_{m,n}. \tag{2.40}$$

The latter result expresses the orthonormality of atomic orbitals.

E. Example: s functions on a simple lattice

The simplest case to consider is when there is only one, s-like function on each atom at sites in a Bravais lattice (*i.e.* one atom per cell). Then the TB matrix equation is a 1×1 equation, which gives a direct expression for $\varepsilon_{\vec{k}}$. Since the atomic orbital is a spherically symmetric function, there is no aggravating angular dependence to worry about in Eq. (1.20) for the hopping parameters, and in fact the value of the hopping parameter is the same for all equivalent neighbors – all 1st neighbors have the same value, all 2nd neighbors have the same (different, usually smaller) value, etc. Let's look at some cases.

F. 1D linear chain of atoms

Taking some atom as our origin, we have two 1st neighbors, at $\pm a$, and both have the same hopping amplitude t_1. Then the sum in Eq. (1.19) contains the on-site term and those from neighbors:

$$\begin{aligned} H_{s,s}(k) &= \varepsilon_s + t_1 \sum_{\vec{R}} e^{ikR} \\ &= \varepsilon_s + t_1(e^{ikR} + e^{-ikR}) \\ &= \varepsilon_s + 2t_1 cos(ka). \end{aligned} \tag{2.41}$$

Likewise,

$$S_{s,s}(k) = 1 + s_1 \sum_{\vec{R}} e^{ikR} = 1 + 2s_1 cos(ka). \tag{2.42}$$

The solution to Eq. (1.16) is immediate:

$$\varepsilon_k = \frac{\varepsilon_s + 2t_1 cos(ka)}{1 + 2s_1 cos(ka)} \tag{2.43}$$

It is very simple to add the effects of interaction with 2nd neighbors, which lie at $\pm 2a$ from the reference atom ("at the origin"). The sum over the complex exponential factors leads to the result $2cos(2ka)$. Denoting the corresponding integrals by t_2 and s_2, the energy band *dispersion relation* becomes

$$\varepsilon_k = \frac{\varepsilon_s + 2t_1 cos(ka) + 2t_2 cos(2ka)}{1 + 2s_1 cos(ka) + 2s_2 cos(2ka)} \tag{2.44}$$

G. General Features

It is time to reflect. Eq. (2.29) is the simple *1D tight binding band* that arises commonly in modelling the behavior of 1D materials. Actually, it is almost always even simpler, because the s overlap parameters are usually neglected for further simplicity. This is discussed in the next subsection. However, at the cost of a little messiness in deriving the expression, we have obtained a very simple form of ε vs. k dispersion relation ε_k. In the simplest case it has only a single parameter, $t_1 \equiv t$, and by construction – because the wavefunctions were built to satisfy the Bloch condition – it has precisely the correct periodicity. These are two general features of the TB representation: (i) simplicity, and (ii) expressions involving trig functions that automatically have the correct periodicity.

Another physical requirement is that the parameter $s_{n,m}(\vec{R})$ in Eq. (2.26) cannot be greater than unity in absolute value, so there can be no problem with the denominator vanishing in Eq. (2.29) or (2.30). In fact, to be reasonable, the various overlap parameters $s_{m,n}(\vec{R})$ should be small in magnitude compared with unity (they may be of either sign). The "self overlap" term is unity by normalization, and putting one of the orbitals on a different site must reduce the magnitude of the overlap. Hence the denominator in Eq. (2.29) should always take the form of a correction, an *adjustment*, and not give very large alteration of the denominator, or else the description loses its realism.

If the overlap matrix is approximated by the unit matrix, so that Eq. (2.29) becomes

$$\varepsilon_k = \varepsilon_s + 2t_1 cos(ka), \tag{2.45}$$

the interpretation of the parameters occurring in the dispersion relation is simple. The band minimum and maximum are $\varepsilon_s - |2t_1|$ and $\varepsilon_s + |2t_1|$ respectively. This means that the center of the band lies at ε_s and the bandwidth W is given by

$$W = 4|t_1|. \tag{2.46}$$

This generalizes to square/cubic lattices in 2D and 3D, where the bandwidth is given by $2z|t_1|$, where z is the usual notation for the number of nearest neighbors. This can be changed somewhat by 2nd neighbor terms, but the rule-of-thumb is this simple relationship between bandwidth, coordination number, and nearest neighbor hopping amplitude:

$$W = 2z|t_1|. \tag{2.47}$$

H. Character of the Overlap Correction

We can obtain insight into the effect of including the overlap matrix S by expanding Eq. (2.30), assuming s_1, s_2 are small. We obtain, shortening $cos(ka) \to C_1(k), cos(2ka) \to C_2(k)$,

$$\begin{aligned} \varepsilon_k \approx\ & (\varepsilon_s + 2t_1 C_1(k) + 2t_2 C_2(k)) \times \\ & (1 - 2s_1 C_1(k) - 2s_2 C_2(k)) \\ =\ & \varepsilon_s + 2(t_1 - s_1)C_1(k) + 2(t_2 - s_2)C_2(k) \\ & -4t_1 s_1 C_1(k)^2 - 4t_2 s_2 C_2(k)^2 \\ & -4(t_1 s_2 + t_2 s_1)C_1(k)C_2(k). \end{aligned} \tag{2.48}$$

The effect is to add more wiggles (Fourier components) into the dispersion relation. Also, it doesn so in a somewhat different way than simply adding more neighbors (more t's) into the expansion. Probably it is more efficient, when using the TB parametrization to fit given complicated band structures, to use an overlap matrix rather than simply add more hopping parameters, but this question has not really been studied systemmatically.

III. THE OVERLAP MATRIX

The overlap matrix deserves comment before we continue. In a large majority of cases where the TB method is used for simplification or for pedagogical reasons, the overlap matrix is "set to unity":

$$S_{m,n}(\vec{R}) \to \delta_{m,n}\delta_{\vec{R},\vec{0}}. \tag{3.49}$$

It is important to understand why, first, this is in principle a reasonable thing to do, and second, it leads not only to simplification but to additional approximation.

A. Löwdin Orthogonalization

Let us return to Eq (1.16),

$$\hat{H}b = \varepsilon\hat{S}b, \tag{3.50}$$

where the notation has been simplified even further by dropping the $\vec{k}$ argument on each quantity, but adding a "hat" on the matrices. We can make a transformation that effectively eliminates S.

This is done by introducing the square root, denote it $\hat{S}^{\frac{1}{2}}$, of the overlap matrix S:

$$\hat{S} = \hat{S}^{\frac{1}{2}}\hat{S}^{\frac{1}{2}} = (\hat{S}^{\frac{1}{2}})^2. \tag{3.51}$$

We will also need the *inverse* of the square root of $\hat{S}$ (equal to the square root of the inverse):

$$\hat{S}^{-\frac{1}{2}}\hat{S}^{\frac{1}{2}} = \hat{1} = \hat{S}^{\frac{1}{2}}\hat{S}^{-\frac{1}{2}}. \tag{3.52}$$

Aha! I hear you say. Square roots and inverses of matrices do not always exist. That is true; however, $\hat{S}$ is a *positive matrix* from its physical and mathematical definition. This means that all of its eigenvalues are positive, in which case its square root and its inverse do exist. These matrices can be obtained by (1) performing the similarity transformation that transforms $\hat{S}$ to a diagonal matrix, (2) taking the square root or the inverse of all diagonal elements as desired, and (3) performing the inverse similarity transformation. It is readily shown that the corresponding matrices obey the properties given above for the square root or the inverse. We don't say more about this because for now it is only necessary to know that it is possible.

With these matrices we can now carry out the following steps:

$$\begin{aligned}
\hat{H}b &= \varepsilon\hat{S}b \\
\hat{H}\hat{S}^{-\frac{1}{2}}\hat{S}^{\frac{1}{2}}b &= \varepsilon\hat{S}^{\frac{1}{2}}\hat{S}^{\frac{1}{2}}b \\
(\hat{S}^{-\frac{1}{2}}\hat{H}\hat{S}^{-\frac{1}{2}})(\hat{S}^{\frac{1}{2}}b) &= \varepsilon(S^{\frac{1}{2}}b) \\
\hat{\mathcal{H}}\overline{b} &= \varepsilon\overline{b},
\end{aligned} \tag{3.53}$$

where

$$\hat{\mathcal{H}} = \hat{S}^{-\frac{1}{2}}\hat{H}\hat{S}^{-\frac{1}{2}}. \tag{3.54}$$

The generalized eigenvalue problem $\hat{H}b = \varepsilon\hat{S}b$ has been transformed into a conventional eigenvalue problem $\hat{\mathcal{H}}\overline{b} = \varepsilon\overline{b}$ *with the same eigenvalues.* The eigenvectors, which contain the information about how much each atomic orbital contributes to the eigenfunction, have been changed, and the Hamiltonian matrix that needs to be diagonalized has been transformed. This observation, together with the lack of any overlap matrix in the new equation, indicates that the transformation

$$\overline{b} = S^{\frac{1}{2}}b \tag{3.55}$$

reflects a transformation of the underlying atomic orbitals into a set of mutually orthogonal orbitals, *even if they were originally residing on different atoms.* This orthogonalization procedure is called Löwdin orthogonalization, after the eminent quantum chemist Per-Olov Löwdin (pronounced *lovedean*).

What this orthogonalization procedure does is to add in (either sign is possible) to any given orbital $\phi_j(\vec{r})$ a fraction of every other orbital that it is not orthogonal to. The amount it must mix in is related to the corresponding element of $S_{m,n}(\vec{R})$. The effect is to make each of the orbitals more spread out, and thereby to increase the number of matrix elements that are needed in $\hat{\mathcal{H}}$ compared to what were needed in $\hat{H}$. This fact is bothersome and tends to get "forgotten," because the objective of the TB formalism is simplicity, and therefore few parameters. The conventional applications of TB theory for (semi)empirical studies therefore usually presumes ("we assume, for simplicity") that the original local orbitals $\{\phi_m(\vec{r}-\vec{R})\}$ have been orthogonalized.

If the off-diagonal elements of S (*i.e.* the overlaps) are small, an approximate square-root matrix can be obtained by expansion. Writing

$$\delta\hat{S} \equiv \hat{S} - 1, \tag{3.56}$$

we can write

$$\hat{S}^{\frac{1}{2}} \approx 1 + \frac{1}{2}\delta\hat{S} + \frac{1}{8}(\delta\hat{S})^2 + ... \tag{3.57}$$

This expansion is readily evaluated without diagonalizing the overlap matrix.

IV. RETURN TO APPLICATIONS

A. 2D Square Lattice of s Orbitals

Henceforward we neglect the $\hat{S}$ matrix, since most of the rest of the world does so. For the square

lattice, the sum over nearest neighbors runs over the sites $(a,0),(0,a),(-a,0),(0,-a)$. The sum becomes

$$\sum_{\vec{R}} \rightarrow \sum_{j=\pm 1} e^{ik_x aj} + \sum_{p=\pm 1} e^{ik_x ap} = 2(cos(k_x a) + cos(k_y a)), \quad (4.58)$$

and the dispersion relation is

$$\varepsilon_{\vec{k}} = \varepsilon_s + 2t_1(cos(k_x a) + cos(k_y a)). \quad (4.59)$$

Now, suppose we want to account for "hopping" to the second neighbors at the points $(\pm a, \pm a)$ with amplitude t_2, we take advantage of $e^{v+w} = e^v e^w$ to get the additional term, and the band looks like

$$\varepsilon_{\vec{k}} = \varepsilon_s + 2t_1(cos(k_x a) + cos(k_y a)) + 4cos(k_x a)cos(k_y a). \quad (4.60)$$

Once you do a few of these, the pattern becomes simple. When you need to deal with atomic p or d orbitals, the procedure becomes more intricate: the $t(\vec{R})$ factor depends on $\vec{R}$ and cannot simply be pulled out from under the summation. But that is another course.

V. COMMENTS OF FITTING OF PARAMETERS

The tight binding hopping parameters (and overlap parameters, if used) are commonly fit to an independently calculated band structure. Denote the parameters by a vector $\vec{t}$ and the set of eigenvalues to be fit by $\{\varepsilon_K\}$, where K will include both $\vec{k}$ and band indices. The mathematical problem is to minimize the "residual" $\mathcal{R}$

$$\delta\mathcal{R} = \delta\sum_K g_K[E_K(\vec{t}) - \varepsilon_K]^2 = 0. \quad (5.61)$$

Here g_K is a weight to be chosen as desired (for example, to fit eigenvalues around the Fermi energy more closely than elsewhere), and $\{E_K\}$ are the eigenvalues resulting from the TB secular equation, which depend on the parameters.

Carrying out the minimization with respect to $\vec{t}$ gives

$$\sum_K g_K(E_K(\vec{t}) - \varepsilon_K)\nabla_t E_K(\vec{t}) = 0. \quad (5.62)$$

Linearizing around $\vec{t} \approx \vec{t}_o$:

$$\nabla_t E_K(\vec{t}) \approx \nabla_t E_K(\vec{t}_o),$$
$$E_K(\vec{t}) \approx E_K(\vec{t}_o) + (\vec{t} - \vec{t}_o)\cdot\nabla_t(E_K(\vec{t}_o), \quad (5.63)$$

leads to

$$\sum_K g_k[E_K(\vec{t}_o) - \varepsilon_K + (\vec{t} - \vec{t}_o)\cdot\nabla_t(E_K(\vec{t}_o)]$$
$$\nabla_t(E_K(\vec{t}_o) = 0, \quad (5.64)$$

which can be written in a matrix form

$$\vec{b} + (\vec{t} - \vec{t}_o)\cdot A = 0. \quad (5.65)$$

This equation can be transposed to

$$\vec{t} = \vec{t}_o - \vec{b}\cdot A^{-1}, \quad (5.66)$$

where b and the matrix A are defined by

$$\vec{b} = \sum_K g_k E_K(\vec{t}_o) - \varepsilon_K)\nabla_t E_K(\vec{t}_o),$$
$$A_{p,q} = \sum_K g_k[\nabla_t E_K(\vec{t}_o)]_p[\nabla_t E_K(\vec{t}_o)]_q. \quad (5.67)$$

The Hellman-Feynman theorem gives a simple method of calculating the derivatives. Given the matrix eigenvalue equation

$$\hat{H}(\vec{k})|k\alpha,\vec{t}> = E_{\vec{k},\alpha}|\vec{k}\alpha,\vec{t}> \quad (5.68)$$

where α is the band index, the derivative is

$$\nabla_t E_{\vec{k}\alpha}(\vec{t}) = < k\alpha,\vec{t}|\left(\nabla_t\hat{H}(\vec{k})\right)|\vec{k}\alpha,\vec{t}> . \quad (5.69)$$

Since the Hamiltonian matrix elements depend linearly on the components of $\vec{t}$, the derivative in this last equation is trivial and can be evaluated exactly by finite difference (numerically; of course, if the matrix is in analytic form it can be obtained trivially). A note to programmers: for Nb, where a 9×9 matrix was used, with 31 parameters and 55 $\vec{k}$ points, there are

$$\frac{9\times 10}{2}\times 55\times 31 = 76,725 \quad (5.70)$$

such derivatives!

Eq. (5.58) provides an iterative procedure for fitting the parameters. Like all such schemes, it will converge if the starting point is sufficiently close to a good minimum. There will in general be many local minima, and the criterion for a best fit may be somewhat subjective as well. However, some study of these problems (unpublished) seems to indicate that non-uniqueness of the fit is not a big problem. Although a converged result is certainly not unique, restarting the iteration procedure in practice leads to sets of parameters whose differences have little physical significance, *i.e.* the large parameters are fairly well determined.

[1] W. E. Pickett and P. B. Allen, Solid State Commun. (1973).