Chapter 2

DENSITY FUNCTIONAL THEORY AND METHODS

Condensed matter physics and materials science are concerned fundamentally with understanding and exploiting the properties of interacting electrons and atomic nuclei. This has been well known since the development of quantum mechanics. With this comes the recognition that, at least in principal, almost all properties of materials can be addressed given suitable computational tools for solving this particular problem in quantum mechanics. Unfortunately, the electrons and nuclei that compose materials comprise a strongly interacting many body system, and this makes the direct solution of Schrodinger's equation an extremely impractical proposition. Rather, as was stated concisely by Dirac in 1929, progress depends on the development of sufficiently accurate, but tractable, approximate techniques [41].

Thus the development of density functional theory (DFT) and the demonstration of the tractability and accuracy of the local density approximation (LDA) to it defined an important milestone in condensed matter physics. First principles quantum mechanical calculations based on the LDA and extensions, like generalized gradient approximations, have emerged as one of the most important components of the theorist's toolbox. These methods are also starting to have significant impact in many areas of materials science, though there remains much to be done. A real challenge is posed by the highly complex nature of most real materials. Related to this, there has been considerable progress in developing DFT based methods suitable for large systems containing many hundreds of atoms in a unit cell. In addition, widely available user friendly DFT codes, with implementation of many property calculations are available. It seems very reasonable to expect these trends to continue and for DFT calculations to become ubiquitous tools in materials science.

It is worth noting that the DFT of Hohenberg and Kohn [73] was predated by the LDA, which was developed and applied by Slater [177] and his co-workers

(see Ref. [179]). Nonetheless, the impact of local density approximation (LDA) calculations in solid state physics remained limited until the late 1970's, when several calculations demonstrating the feasibility and accuracy of the approach in determining properties of solids appeared [215, 216, 217, 120]. The contribution of these and other pioneers in this field should not be underestimated. There has been a great deal written about why the LDA should or should not be adequate for calculating properties of this or that material. There is, however, no doubt that the most convincing arguments derive from the direct comparison of detailed calculations with experiment. The utility of the LDA was demonstrated in the early calculations of these and other groups, and it is this that has led to the widespread application of these tools.

As mentioned, DFT based calculations have become one of the most frequently used theoretical tools in condensed matter physics, and there are now several excellent reviews of the subject including those by Lundqvist and March, [112] Callaway and March, [29], Dreizler and da Provincia, [42] Ernzerhof, Perdew and Burke [44] and Parr and Yang [137]. The reader is warned that this chapter is not along those lines, *i.e.* it is not a comprehensive review of DFT. Rather its purpose is much more limited – to present a very limited sketch, emphasizing those aspects that are necessary groundwork for the material to follow. For a general exposition of DFT, the reader is referred to the excellent reviews mentioned above.

2.1 Density Functional Theory

The theorem upon which DFT and the LDA are based is that of Hohenberg and Kohn. It states that the total energy, E, of a non-spin-polarized system of interacting electrons in an external potential (for our purposes the Coulomb potential due to the nuclei in a solid) is given exactly as a functional of the ground state electronic density, ρ.

$$E = E[\rho]. \tag{2.1}$$

They further showed that the true ground state density is the density that minimizes $E[\rho]$, and that the other ground state properties are also functionals of the ground state density. The extension to spin-polarized systems is straightforward; E and the other ground state properties become functionals of the spin density, which in the general case is given as a four component spinor [198, 155]. In the collinear case, where the spin-up and spin-down densities suffice,

$$E = E[\rho_\uparrow, \rho_\downarrow]. \tag{2.2}$$

Unfortunately, the Hohenberg-Kohn theorem provides no guidance as to the form of $E[\rho]$, and therefore the utility of DFT depends on the discovery of sufficiently accurate approximations. In order to do this, the unknown functional, $E[\rho]$, is rewritten as the Hartree total energy plus another, but presumably smaller, unknown functional, called the exchange-correlation (xc) functional, $E_{xc}[\rho]$.

$$E[\rho] = T_s[\rho] + E_{ei}[\rho] + E_H[\rho] + E_{ii}[\rho] + E_{xc}[\rho]. \tag{2.3}$$

Here $T_s[\rho]$ denotes the single particle kinetic energy, $E_{ei}[\rho]$ is the Coulomb interaction energy between the electrons and the nuclei, $E_{ii}[\rho]$ arises from the interaction of the nuclei with each other, and $E_H[\rho]$ is Hartree component of the electron-electron energy,

$$E_H[\rho] = \frac{e^2}{2} \int \mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' \ \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.4}$$

As mentioned, $E_{xc}[\rho]$ is an unknown functional. However, several useful approximations to it are known. The simplest is the local density approximation (LDA). In the LDA, $E_{xc}[\rho]$ is written as

$$E_{xc}[\rho] = \int \mathrm{d}^3\mathbf{r} \ \rho(\mathbf{r})\epsilon_{xc}(\rho(\mathbf{r})), \tag{2.5}$$

where $\epsilon_{xc}(\rho)$ is approximated by a local function of the density, usually that which reproduces the known energy of the uniform electron gas. The other commonly used approximations are the generalized gradient approximations (GGAs), where the local gradient as well as the density is used in order to incorporate more information about the electron gas in question, *i.e.* $\epsilon_{xc}(\rho)$ is replaced by $\epsilon_{xc}(\rho, |\nabla\rho|)$. The weighted density approximation (WDA) is an approximation that incorporates more non-local information about the electron gas via a model pair correlation function. This is exact in important limits: the uniform electron gas and arbitrary single electron systems. It greatly improves the energies of atoms, and often yields bulk properties that are much improved as well. Nonetheless, the WDA is more computationally demanding than the LDA or GGA, and as such relatively few WDA studies have been reported for solids.

The relationship between the various approximations can be understood using the exact expression for the exchange correlation energy in terms of the pair correlation function [100, 143],

$$E_{xc}[n] = \int\int \mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' \ \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\bar{g}[n, \mathbf{r}, \mathbf{r}'] \tag{2.6}$$

$$= \int\int \mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' \quad \frac{n(\mathbf{r})\bar{n}_{xc}[n,\mathbf{r},\mathbf{r}']}{|\mathbf{r}-\mathbf{r}'|}, \tag{2.7}$$

where $\bar{g}$ is the coupling constant average (from e^2=0 to e^2=1 in atomic units) of the pair correlation function of the electron gas in question, and $\bar{n}_{xc}$ is the coupling constant averaged exchange correlation hole. Since the exchange correlation hole must be a depletion containing exactly one electron charge, E_{xc} is invariably negative. The physical meaning of this expression is that the exchange correlation energy is given by the Coulomb interaction of each electron with its exchange correlation hole, reduced in magnitude by a kinetic energy contribution, which corresponds to the energy required to dig out the hole. This reduction is accounted for by using the coupling constant average instead of the full strength pair correlation function, and includes contributions to the kinetic energy beyond the single particle level. The spherically symmetric Coulomb interaction in Eqn. 2.7 means that only the spherical average of the exchange correlation hole needs to be correct to obtain the correct energy, a fact that is important in the success of simple approximations like the LDA [112, 44].

The local density approximation consists of the replacement in Eqn. 2.7,

$$n(\mathbf{r}')\bar{g}[n,\mathbf{r},\mathbf{r}'] \rightarrow n(\mathbf{r})\bar{g}^h(n(\mathbf{r}),|\mathbf{r}-\mathbf{r}'|), \tag{2.8}$$

where $\bar{g}^h$ is the function $\bar{g}$ for the uniform electron gas. This reproduces the exact energy for the uniform electron gas. The weighted density approximation retains the non-locality using integration with a model function $\bar{g}^w$.

$$n(\mathbf{r}')\bar{g}[n,\mathbf{r},\mathbf{r}'] \rightarrow n(\mathbf{r})'\bar{g}^w(\bar{n}(\mathbf{r}),|\mathbf{r}-\mathbf{r}'|), \tag{2.9}$$

where $\bar{n}(\mathbf{r})$ is the weighted density, determined using the sum rule,

$$\int \mathrm{d}^3\mathbf{r}' n(\mathbf{r}')[\bar{g}^w(\bar{n}(\mathbf{r}),|\mathbf{r}-\mathbf{r}'|)-1] = -1. \tag{2.10}$$

This approximation violates the exact symmetry $\bar{g}[n,\mathbf{r},\mathbf{r}'] = \bar{g}[n,\mathbf{r}',\mathbf{r}]$, but nonetheless has been quite successful in describing structural properties of materials in the admittedly few calculations reported to date.

Modern GGA functionals appear formally like the LDA, but with the local function including not just the density (or spin densities in the LSDA), but also the local gradient (or spin density gradients). However, these are not gradient expansions, but rather sophisticated methods to obtain as good an energy as possible using known exact sum rules and scaling relationships for the electron gas based on Eqn. 2.7 and/or fits to data bases [102, 141, 142, 44, 11, 12,

13, 14]. In contrast to the WDA, GGA calculations have been performed for a wide variety of materials, and the GGA is in fact the method of choice for many first principles studies of materials. The behavior of the GGA relative to the LDA is well understood on the basis of the many comparative studies that have been done. From the results of these, the following conclusions may be drawn: (1) GGAs significantly improve the ground state properties of light atoms and molecules, clusters and solids composed of them; (2) many properties of $3d$ transition metals are greatly improved; for example, unlike the LSDA the correct *bcc* ground state of Fe is obtained; (3) the description of Mott-Hubbard insulators, like the undoped phases of high-T_c cuprates, is not significantly improved over the LSDA; (4) GGA functionals usually favor magnetism more than the LSDA, and as a result the magnetic energies for some $3d$ transition metals may be overestimated; and (5) structural properties are generally improved, although GGAs sometimes lead to overcorrection of the LDA errors in lattice parameters. In some materials containing heavy elements (*e.g.* in $5d$ compounds) these degrade agreement with experiment relative to the LDA. It should be emphasized, however, that there is ongoing work aimed at developing even better GGA functionals, and it is quite possible that an improved form that alleviates the above deficiencies will be found.

Kohn and Sham [90] wrote the electron density as a sum of single particle densities, and used the variational property to obtain a prescription for determining the ground state energy and density, given the functional E_{xc}. In particular, they showed that the correct density is given by the self-consistent solution of a set of single particle Schrodinger-like equations, known as the Kohn-Sham (KS) equations, with a density dependent potential,

$$\{T + V_{ei}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r})\}\,\varphi_i(\mathbf{r}) = \epsilon_i \varphi_i(\mathbf{r}), \tag{2.11}$$

with the density given by a Fermi sum over the occupied orbitals,

$$\rho(\mathbf{r}) = \sum_{occ} \varphi_i^*(\mathbf{r})\varphi_i(\mathbf{r}). \tag{2.12}$$

Here the highest occupied orbital is determined by the electron count, the φ_i are the single particle Kohn-Sham orbitals, the ϵ_i are the corresponding Kohn-Sham eigenvalues, T is the single particle kinetic energy operator, V_{ei} is the Coulomb potential due to the nuclei, V_H is the Hartree potential and V_{xc} is the exchange correlation potential. Both V_H and V_{xc} depend on ρ.

$$V_H(\mathbf{r}) = e^2 \int \mathrm{d}^3\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \tag{2.13}$$

and

$$V_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho]}{\delta \rho(\mathbf{r})}. \tag{2.14}$$

In this framework, a calculation entails the self-consistent solution of Eqns. 2.11 and 2.12. That is, a density must be found such that it yields an effective potential that when inserted into the Schrodinger-like equations yields orbitals that reproduce it. Thus, instead of having to solve a many-body Schrodinger equation, using DFT we have the far easier problem of determining the solution to a series of single particle equations, along with a self-consistency requirement.

In solids, Bloch's theorem provides a further simplification that facilitates DFT based calculations: Because the charge density has the periodicity of the lattice, so does the single particle KS Hamiltonian. Thus KS orbitals with different Bloch momenta are coupled only indirectly through the density dependent potential. Accordingly, in DFT based (but not, for example, in Hartree-Fock) calculations, the single particle KS equations may be solved separately on a grid of sampling points in the symmetry irreducible wedge of the Brillouin zone, and the resulting orbitals used to construct the charge density.

2.2 Solution of the Single Particle Kohn-Sham Equations

DFT based electronic structure methods are classified according to the representations that are used for the density, potential and, most importantly, the KS orbitals. The choice of representation is made to minimize the computational and human (*e.g.* programming) costs of calculations, while maintaining sufficient accuracy. These competing material and application dependent goals have led to the development and use of a wide variety of techniques.

This book is concerned with two particular approaches, planewave pseudopotential methods and the LAPW method. It is certainly possible to avoid the explicit use of a basis in constructing the KS orbitals, for example, by numerically solving the differential equations on grids. However, nearly all approaches that have been proposed for solids, including the planewave pseudopotential and LAPW methods, do rely on a basis set expansion for the KS orbitals. Here the discussion is confined to methods that do use a basis, in which case the KS orbitals are:

$$\varphi_i(\mathbf{r}) = \sum_\alpha c_{i\alpha}\phi_\alpha(\mathbf{r}), \tag{2.15}$$

where the $\phi_\alpha(\mathbf{r})$ are the basis functions and the $c_{i\alpha}$ are expansion coefficients. Since, given a choice of basis, these coefficients are the only variables in the problem (note that the density depends only on the KS orbitals), and since the total energy in DFT is variational, solution of the self-consistent KS equations

amounts to determining the $c_{i\alpha}$ for the occupied orbitals that minimize the total energy.

To proceed, note that the energy can be rewritten using the single particle eigenvalues to eliminate the unknown functional, $T_s[\rho]$.

$$E[\rho] = E_{ii}[\rho] + \sum_{occ} \epsilon_i + E_{xc}[\rho] - \int \mathrm{d}^3\mathbf{r}\rho(\mathbf{r})(V_{xc}(\mathbf{r}) + \frac{1}{2}V_H(\mathbf{r})), \quad (2.16)$$

where the sum is over the occupied orbitals, and ρ, V_H and V_{xc} are given by Eqns. 2.12, 2.13 and 2.14, respectively.

It is very common to separate the determination of the $c_{i\alpha}$ and the determination of the self-consistent charge density in density functional calculations. The solutions for the density and the $c_{i\alpha}$ are done hierarchically in this case, as shown schematically in Fig. 2.1. In this scheme, it is necessary to repeatedly determine the $c_{i\alpha}$ that solve the single particle equations 2.11 for fixed charge density. This may be done using standard matrix techniques. Specifically, given the basis, the Kohn-Sham Hamiltonian and overlap matrices, $\mathbf{H}$ and $\mathbf{S}$ are constructed and the matrix eigenvalue equation,

$$(\mathbf{H} - \epsilon_i\mathbf{S})\mathbf{c}_i = 0, \quad (2.17)$$

is solved at each $\mathbf{k}$-point in the irreducible wedge of the Brillouin zone. This can be done efficiently using standard linear algebra routines, such as EISPACK. Here the square matrices $\mathbf{H}$ and $\mathbf{S}$ are of rank equal to the number of basis functions, n_b and the $\mathbf{c_i}$ are vectors containing the n_b coefficients, $c_{i\alpha}$ for each KS orbital i.

If the true occupied KS orbitals can be expressed as linear combinations of the basis functions, then optimizing the $c_{i\alpha}$ will yield the exact self-consistent solution. On the other hand, if the exact KS orbitals cannot be expressed exactly in terms of the chosen basis, this procedure will yield an approximate solution that is optimal in the sense that it gives the lowest possible total energy for this basis. The quality of a basis set can, therefore, be measured by the extent to which the total energy evaluated using the orbitals of Eqn. 2.15 differs from the true KS energy.

Efficiency, bias, simplicity and completeness are common terms that are used in discussing the relative merits of electronic structure techniques. These refer to the number of the basis functions needed to achieve a given level of convergence, whether or not the basis favors certain regions of space over others (*e.g.* by being more flexible near atomic nuclei than elsewhere), the difficulty in calculating matrix elements and whether the basis can be improved arbitrarily by adding additional functions of the same type.

Planewave basis sets are notoriously inefficient in the above sense for most solids. This, however, is not necessarily a defect, since it just reflects the fact that they are unbiased. Further, planewaves form a complete set and they are a simple basis. The completeness means that, at least in principle, arbitrary accuracy can be obtained by increasing the number of planewaves in the basis, and more importantly that the convergence of a calculation can be monitored by varying the planewave cutoff. Further, because of the simplicity of this basis, implementation of planewave codes is relatively straightforward, and matrix elements of many operators can be calculated quickly. The fact that wavefunctions expanded in planewaves can be transformed efficiently from reciprocal space (coefficients of the planewave expansion) to real space (values on a real space grid) using fast Fourier transforms (FFTs) means that many operators can be made diagonal. In particular, the kinetic energy and momentum operators are diagonal in reciprocal space, and the operation of local potentials is diagonal in real space.

It is apparent from Eqn. 2.15 that the most efficient basis set consists of the KS orbitals themselves (or equivalently, linear combinations of the KS orbitals). In this case, an exact calculation is achieved using a basis set size equal to the number of occupied orbitals. Even though, in general, the KS orbitals are unknown at the beginning of a calculation, this property can be exploited in constructing basis sets. In particular, if the KS orbitals for a similar Hamiltonian are known, inclusion of these in the basis will often result in a great improvement. A simple example is the case where a small perturbation, ΔH (*e.g.* spin-orbit) is added to a Hamiltonian, H_0 for which a solution has already been generated. Using the KS orbitals of H_0 as a basis, the matrix elements with the perturbed Hamiltonian can be readily constructed as those of ΔH with the addition of the eigenvalues of H_0 on the diagonal. The construction and diagonalization of the Hamiltonian in this space can often be done quite rapidly, even for complicated ΔH because of its small dimension.

More common examples are the use of atomic and muffin-tin orbitals in electronic structure calculations. In the former, an atomic H_0 is assumed in constructing basis functions for each site. Despite the fact that crystal potentials are often significantly different from atomic potentials, even in the vicinity of an ion, linear combination of atomic orbitals (LCAO) methods have been quite successful, particularly for large systems, where the efficiency of this basis is an important advantage. However, although this basis is clearly complete, problems often arise when attempts are made to add large numbers of basis functions to obtain highly converged calculations. This is because atomic orbitals centered at a single site are already complete. Thus LCAO's which have orbitals centered at each site are over-complete and, because of this, the overlap matrix, $\mathbf{S}$, in Eqn. 2.17 becomes ill-conditioned for large basis sets.

Muffin-tin orbital derived basis sets will be discussed in the chapters on the LAPW method. Here it suffices to state that they are based on solutions of radial Schrodinger's equation with a better approximation to the crystal potential in the vicinity of the site in question than that used in constructing LCAOs and that methods using them can be constructed to largely avoid the over-completeness problem.

2.3 Self-Consistency in Density Functional Calculations

As mentioned, the Hohenberg-Kohn theorem shows that the total energy is variational, and this is the key to its usefulness. The true ground state density is that density which minimizes the energy. When approximations are made to $E_{xc}[\rho]$, such as the LDA, there is no longer a true variational principle, and there is no guarantee that the energy obtained by minimizing the now approximate energy functional will be higher than the exact ground state energy. Clearly then, the relative quality of different approximations cannot be determined by determining which of them yields the lower energy. Furthermore, the true ground state density is not in general the density that minimizes the total energy as determined using approximate functionals. There is, in fact, no prescription for determining what the exact ground state density is from approximate functionals. Accordingly, calculations proceed by minimizing the approximate energy functional, recognizing that, although the resulting energy may be lower (or higher) than the true ground state energy, a good approximation to the energy functional should give a good energy and density and that the procedure is exact for the true energy functional.

Since the single particle kinetic energy, $T_s[\rho]$ appearing in Eqn. 2.3 is unknown in this form, the minimization proceeds via the KS equations. Then the variation is with respect to the orbitals, or in a basis set expansion, the coefficients, $c_{i\alpha}$. With a fixed basis, these are the only parameters that can be varied; otherwise there are additional parameters that determine the basis functions. In any case, the problem may be stated as follows: Find the coefficients (and other parameters, if any) that minimize the energy functional, Eqn. 2.16, subject to the constraint that the orbitals remain orthonormal.

The direct minimization of the total energy with respect to the $c_{i\alpha}$ was proposed quite early on by Bendt and Zunger [15] (see also Payne *et al.* [139]) and is at the heart of the Car-Parrinello (CP) and related methods [30]. Nonetheless, and in spite of potential computational advantages, this type of approach has not yet become popular for methods that use non-planewave basis sets. This is a result of the complexity of the optimization problem; there are typically hundreds or thousands of parameters even for small problems and the objective function is highly nonlinear with many local extrema (corresponding to missed KS orbitals, with occupation of higher lying ones).

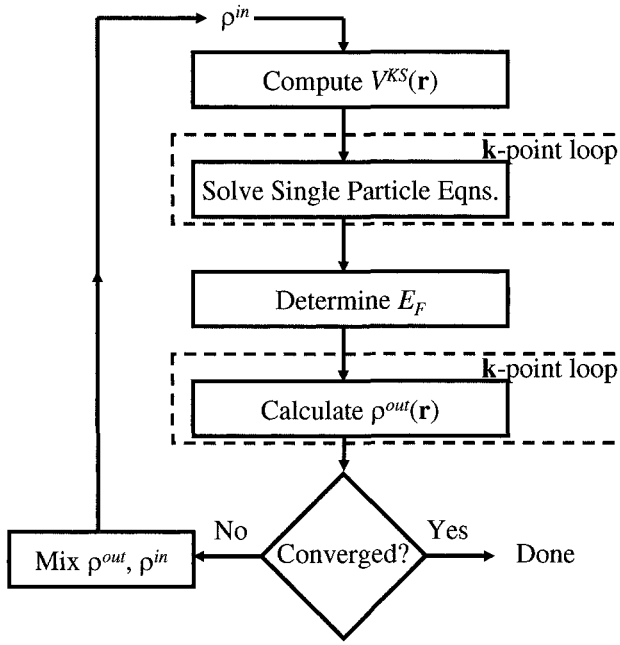

Figure 2.1. Schematic flow-chart for self consistent density functional calculations.

Because of these complications, the historically dominant approach has been to refine the density iteratively by solving Eqns. 2.11 and 2.12 alternately. This is the basis of the standard self-consistency cycle illustrated in Fig. 2.1. Given a charge density, Eqn. 2.17 is diagonalized, ensuring that the orbitals are orthonormal and that no orbitals are missed. This eliminates almost all local extrema. An output charge density is constructed from the eigenvectors using Eqn. 2.12, and then mixed with the input to yield a refined input for the next iteration. The simplest mixing scheme is straight mixing:

$$\rho_{in}^{i+1} = (1-\alpha)\rho_{in}^{i} + \alpha\rho_{out}^{i}, \tag{2.18}$$

where the superscript refers to the iteration number and α is the mixing parameter. For sufficiently small α, the iterations converge. However, the radius of convergence can be small, and rapidly becomes smaller as the size of the unit cell increases, particularly for metallic and/or magnetic systems. As a result, considerable effort has been devoted to devising more sophisticated mixing procedures, using information from previous iterations to accelerate the convergence. The most common of these is Broyden's method and variants thereof [28, 186]. These will be discussed in detail in Chapter 5.

As mentioned, this is a hierarchical approach to the optimization. The diagonalization may viewed as an optimization (minimization of the residuals); this is the lowest level of the hierarchy. The next level, which can also be regarded as an optimization (minimization of the difference between the input and output densities) is the search for a self-consistent charge density. Experience has shown that this approach is very robust. It is, however, inefficient in the following sense.

Exact eigenvectors are calculated (Eqn. 2.17) for the current single particle Hamiltonian at each step of the iteration to self-consistency, including early iterations for which the charge density is poor. However, these are of little interest; the only eigenvectors that are relevant are those for the self-consistent charge density. In the early iterations, approximate eigenvectors would serve as well. This observation provides useful a insight into the CP method, and suggests avenues for speeding up non-planewave approaches.

The eigenvectors and the input charge density may be viewed as independent quantities to be optimized (recall that at the minimum the input and output densities are equal removing this independence). The iteration to self-consistency is nothing more than a sequence of moves towards the minimum. In the hierarchical approach, the eigenvector moves are to the exact solution for the current density, while the moves of the charge density are determined by the mixing. Given the true charge density, a single move of the eigenvectors yields the true minimum. Meanwhile, the complex nonlinear dependence of the single particle Hamiltonian on the density makes the charge density moves less effective.

In the CP method, the eigenvector moves are based on an iterative refinement rather than exact diagonalizations. One or more steps of an iterative diagonalization are used to generate the refined eigenvectors, which are then used to construct a charge density move. For planewave basis sets, FFT dependent algorithms (discussed in Chapter 3) can be used to perform these refinements in a small fraction of the time needed for exact diagonalizations. Thus, even though there may be an increase in the total number of iterations needed, their cost is much lower, particularly for large systems. This underlies, at least in part, the efficiency of the CP method.

The question often arises as to why CP like algorithms have not yet been widely applied in non-planewave methods. Certainly, the basic idea of iteratively refining the eigenvectors along with the charge density is applicable to any method using a basis set. The first complication is that, in order for CP based algorithms to be worthwhile, an eigenvector move based on iterative refinement should be much faster than one using exact diagonalization. The second is that these refinements need to be effective in the sense of rapidly converging to the true eigenvectors. This can be difficult in techniques with non-orthogonal basis sets and poorly conditioned matrices, including the LAPW method.

As mentioned, the main raison d'etre of non-planewave basis sets is to reduce the size of the secular equation (Eqn. 2.17) for materials with hard pseudopotentials. In general, the price to be paid for this efficiency is an increase in the cost of computing matrix elements. In planewave based methods, the cost of synthesizing the Hamiltonian matrix is often negligible compared to the cost of diagonalizing it. However, it is often the case with non-planewave methods that the cost of synthesizing the Hamiltonian matrix rivals or even exceeds the cost of diagonalizing it.

The key step in an iterative refinement is the operation of the Hamiltonian on wavefunctions. This can be done either by synthesizing the Hamiltonian matrix and then performing matrix-vector multiplies, or directly as in the CP method.

In the LAPW method, as it is normally implemented, the cost of computing the Hamiltonian and overlap matrices is smaller than the diagonalization time, but only by a factor of two to five, depending on the details of the system (it scales with system size in the same way as the diagonalization). This limits the gains that can be obtained by adopting an iterative refinement of the eigenvectors if the Hamiltonian matrix is synthesized. On the other hand, the cost of operating the Hamiltonian directly on a wavefunction is at least equal to the cost of computing a single row of the Hamiltonian matrix. Further, in the LAPW method the dimension of the Hamiltonian is typically only an order of magnitude larger than the number of occupied states. Since the number of operations of the Hamiltonian in iterative approaches is several times the number of orbitals (depending on the exact scheme used), the potential gains from this approach are limited as well. What is needed then are new algorithms for operating the Hamiltonian on trial wavefunctions, *i.e.* linear combinations of the basis functions. Progress for the LAPW method in this direction is discussed in the final chapter of this book.

2.4 Spin-Polarized Systems

In the generalization of DFT to spin-polarized systems [198], the charge density is augmented by a magnetization density, $\mathbf{m}(\mathbf{r})$. This is in general a continuous three dimensional vector field; both the magnitude and direction of $\mathbf{m}(\mathbf{r})$ vary from place to place. In nature, magnetism is often non-collinear, *i.e.* the magnetization direction does in fact vary from place to place. This non-collinearity arises for many reasons [156, 157], *e.g.* Fermi surface effects leading to spin spirals, frustration of exchange interactions as in triangular lattice systems, or between spin-orbit and exchange, like in U_3P_4, and other relativistic effects like the Dzyaloshinskii-Moriya interaction, which leads to the helical magnetic order of MnSi. However, many interesting magnetic systems either are collinear or are well approximated as collinear. In this case, which we discuss first, the direction dependence of $\mathbf{m}(\mathbf{r})$ reduces to a sign and therefore

the theory may be formulated in terms of two scalar fields, a spin-up density, $\rho_\uparrow(\mathbf{r})$ and a spin-down density, $\rho_\downarrow(\mathbf{r})$. Then

$$\rho(\mathbf{r}) = \rho_\uparrow(\mathbf{r}) + \rho_\downarrow(\mathbf{r}), \tag{2.19}$$

and

$$m(\mathbf{r}) = \rho_\uparrow(\mathbf{r}) - \rho_\downarrow(\mathbf{r}). \tag{2.20}$$

In this case, the Hohenberg-Kohn theorem is generalized to state that the true ground state total energy is a variational functional of the spin densities [198, 45].

$$E = E[\rho, \mathbf{m}] = E[\rho_\uparrow, \rho_\downarrow], \tag{2.21}$$

where the first part of the equation applies in the general non-collinear case as well. The energy may then be decomposed as in Eqn. 2.3. The Coulomb terms remain functionals of the total density, but T_s and E_{xc} become functionals of the two spin densities. The variational principle is invoked to generate the spin-polarized KS equations of spin density functional theory.

$$(T + V_{ei}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc,\sigma}(\mathbf{r}))\varphi_{i\sigma}(\mathbf{r}) = \epsilon_{i\sigma}\varphi_{i\sigma}(\mathbf{r}), \tag{2.22}$$

where σ is the spin index and

$$\rho_\sigma(\mathbf{r}) = \sum_{occ} \varphi^*_{i\sigma}(\mathbf{r})\varphi_{i\sigma}(\mathbf{r}), \tag{2.23}$$

with the highest occupied orbital again determined by the electron count and

$$V_{xc,\sigma} = \frac{\delta E_{xc}[\rho_\uparrow, \rho_\downarrow]}{\delta \rho_\sigma(\mathbf{r})}. \tag{2.24}$$

The total energy expression then becomes

$$\begin{aligned} E &= E_{ii} + \sum_{occ} \epsilon + E_{xc}[\rho_\uparrow, \rho_\downarrow] - \frac{1}{2}\int \mathrm{d}^3\mathbf{r} V_H(\mathbf{r})\rho(\mathbf{r}) - \\ &\int \mathrm{d}^3\mathbf{r} \left\{\rho_\uparrow(\mathbf{r})V_{xc,\uparrow}(\mathbf{r}) + \rho_\downarrow(\mathbf{r})V_{xc,\downarrow}(\mathbf{r})\right\}, \end{aligned} \tag{2.25}$$

where we are implicitly using the fact that the Hartree potential of a Coulomb system is twice the Hartree energy.

These equations are to be solved self-consistently, as in the non-spin-polarized case. The differences are:

1 The density is replaced by two spin densities.

2 There are separate sets of KS orbitals for the two spin components, and two sets of single particle equations need to be solved to obtain them.

3 V_{xc} is spin dependent; this is the only term in the single particle Hamiltonian that is explicitly spin-dependent.

4 In the total energy expression E_{xc} is a functional of the two spin densities. E_{xc} favors spin polarized solutions, T_s opposes them. Whether or not a material is magnetic depends on the balance between these terms.

Finally, because of the additional degrees of freedom contained in the spin density, spin-polarized KS equations often have multiple self-consistent solutions, corresponding to different stable spin configurations. Determining which of these is the ground (lowest energy) state and if there are any solutions that have been missed may require an exhaustive search. However, a constrained density functional technique, known as the fixed spin-moment method [205, 158], greatly simplifies the search in ferromagnetic systems. This procedure is discussed in chapter 5.

2.5 Non-Collinear Magnetism

In the case of general non-collinear magnetizations, the spin-dependent KS equations will no longer decouple. Instead it is meaningful to introduce an exchange-correlation field, which plays the role of an internal magnetic field, $\mathbf{b}_{\mathrm{xc}}$ [156, 175]. This is given by the functional derivative of the exchange-correlation energy with respect to the magnetization

$$\mathbf{b}_{\mathrm{xc}}(\mathbf{r}) = -\frac{\delta E_{xc}[\rho, \mathbf{m}]}{\delta \mathbf{m}(\mathbf{r})} . \tag{2.26}$$

With this auxiliary field, we can write the KS equation

$$\left(T + V_{ei}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r}) - \mathbf{b}_{xc} \cdot \boldsymbol{\sigma}\right) \varphi_i(\mathbf{r}) = \epsilon_i \varphi_i(\mathbf{r}) , \tag{2.27}$$

which is now in a spinor form, *i.e.* $\varphi_i(\mathbf{r})$ is a general two-component spinor

$$\varphi_i(\mathbf{r}) = \begin{pmatrix} \alpha_i(\mathbf{r}) \\ \beta_i(\mathbf{r}) \end{pmatrix} . \tag{2.28}$$

In Eqn. 2.27, the first four operators are spin-independent, *i.e.* they act in the same way on both spin components, while the last operator is spin dependent

through the action of the Pauli spin matrix vector, with its Cartesian components

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.29}$$

The spin-independent part of the xc potential, $V_{xc}(\mathbf{r})$, is here the functional derivative of E_{xc} with respect to the charge density. It is now easy to see, by comparing Eqns. 2.22 and 2.27, that whenever the x and y components of the xc magnetic field vanish we return to the collinear case with decoupled equations for the spin components. Then $V_{xc,\sigma} = V_{xc} - \sigma b_{xc,z}$, where σ here is $\pm$ depending on spin character.

In order to get an iterative cycle, we need to obtain the charge and magnetization densities from our eigen-spinors in 2.28. This is done via

$$\rho(\mathbf{r}) = \sum_{\text{occ}} \varphi_i^\dagger(\mathbf{r})\varphi_i(\mathbf{r}) \tag{2.30}$$

and

$$\mathbf{m}(\mathbf{r}) = \sum_{\text{occ}} \varphi_i^\dagger(\mathbf{r})\boldsymbol{\sigma}\varphi_i(\mathbf{r}). \tag{2.31}$$

This scheme [132], which does not build in any geometrical approximations for the charge, magnetization, potential, nor xc magnetic field, *i.e.* a full potential formulation, has now been implemented in various electronic structure codes.

In the LSDA [198], the exchange-correlation energy is given by Eq. 2.5, but with a spin polarized energy density $\epsilon_{xc}(\rho_\uparrow(\mathbf{r}), \rho_\downarrow(\mathbf{r}))$. This gives a xc magnetic field from Eq. 2.26 that is locally parallel to the magnetization at each point in space

$$\mathbf{b}_{\text{xc}}(\mathbf{r}) = -\widehat{\mathbf{m}}(\mathbf{r})\,\rho(\mathbf{r}) \left[\frac{\partial \varepsilon_{xc}(\rho_\uparrow, \rho_\downarrow)}{\partial m}\right]_{\rho=\rho(\mathbf{r}), m=|\mathbf{m}(\mathbf{r})|}, \tag{2.32}$$

where $\widehat{\mathbf{m}}(\mathbf{r})$ is the unit vector along the direction of the magnetization density at point $\mathbf{r}$. For other approximations to the exchange correlation energy, the magnetic field is not as simple, and does not necessarily point in the same direction as the magnetization. Unfortunately, a complete GGA formulation has not yet been developed. In particular, all present formulations neglect gradients of the transverse component of the magnetization density. However, these gradients might not be essential, and there have been applications with these incomplete GGA formulae.

The secular matrix problem, Eq. 2.17, that enters a non-collinear calculation has double the size of that in a non-magnetic calculation or equivalently double the size of the secular equation for each of the individual spin components in a collinear calculation. That said, it is clear that these calculations are more time consuming. In addition, the relevant magnetic cell is usually a multiple of the

chemical cell, and not least the extra degree of freedom with the magnetization as a vector instead of a scalar slows the iterative convergence, and allows for many more metastable states that must in general be sorted out.

In addition to treating commensurate magnetic cells, there exists a beautiful method [72, 155] to deal with non-commensurate spin density waves. It uses the fact that in case of so-called spin spirals, the magnetization is rotated in between different unit cells in the crystal, but is otherwise unchanged, at least in the absence of spin-orbit coupling. Since the magnitude of the magnetization is translationally invariant, one can introduce general symmetry operations that combine translations with spin rotations. This method has been used to calculate the observed helical or cycloidal spin density waves in, for instance, some transition metal systems [94] and in rare earth metals [133].

2.6 The LDA+U Method

The current approximations to the exchange-correlation functionals $E_{\rm xc}$ have clear limitations when it comes to systems with so-called correlated electrons, *e.g.* some transition metal oxides or rare earth compounds. In these d or f metal systems, the electronic states are close to localization and the Coulomb repulsion between the electrons within an open shell is of a completely different nature than in the homogeneous electron gas, upon which LSDA and GGA are based. This should, in principle, be remedied in a more exact version of the DFT. However, it is not known how to write the appropriate functionals in a standard orbital independent way. In the mean time, a completely different approach has been developed, which is to add a Hubbard like on-site repulsion on top of the usual Kohn-Sham Hamiltonian, [3, 4], *i.e.* to add,

$$E_U = U/2 \sum_{i \neq j} n_i n_j \,, \tag{2.33}$$

to the ordinary DFT xc energy, while subtracting a double counting term. Here n_i is the occupation number of orbital $i = \{m_\ell, \sigma\}$ in the relevant atomic shell ℓ. This method, known as LDA+U, was first developed to be able to cope with so-called Mott insulators, *i.e.* systems where LDA and GGA incorrectly predict a metallic state. By construction, the resulting potential is orbital dependent. This now allows for localization of occupied orbitals. However, since DFT also incorporates exchange and correlation in some sense, care has to be taken in order to correct for double counting. Unfortunately, there is not a unique way to make this correction. For instance, one can assume integer occupation numbers, which is relevant in the atomic limit, or equal non-integer occupation numbers for all orbitals, the so-called around mean field approach [145]. It has been observed that the different treatments of the double counting term can lead to qualitatively different physics, especially at intermediate values of U, and so it is important to note which scheme is being used. However, in general the effect of

the Hubbard U in the LDA+U method is to drive the orbital occupations towards integers, and to favor insulating states over metallic ones. This may or may not be the correct physics in a given system. Especially in the case of metals, fluctuations not included in the LDA+U scheme can work against the tendency of the LDA+U method towards integer orbital occupations. As a result, the the LDA can provide a better description of the electronic structure than the LDA+U method, even in some moderately to strongly correlated metals. On the other hand, for correlated Mott-Hubbard insulators, the LDA description is unphysical, while the LDA+U approach provides a very reasonable description of the electronic structure.

The LDA+U method has evolved since its first suggestion. The most general versions use a parametrized screened Hartree-Fock interaction for electrons within one atomic shell [108, 185]. The renormalization of the bare exchange parameters is due to screening or correlations and depends strongly on the specific system. These parameters, *e.g.* U and J (in general there are $\ell + 1$ independent parameters for a shell of angular momentum ℓ), can be estimated by constrained DFT calculations, but often they are used as free parameters. In this Hartree-Fock like scheme, local density matrices, *i.e.* n_{ij}, are used instead of the occupation numbers in Eq. 2.33, which leads to a rotationally invariant formulation. This is what is generally implemented in codes.

The LDA+U method involves the identification of local atomic-like orbitals to which the non-LDA, orbital dependent, interaction is to be applied. Schick *et al.* [162] describe the implementation of the LDA+U approach in the context of the LAPW method. In their implementation, the Hubbard term is applied by projecting the bands onto the LAPW radial functions of selected angular momentum character (see Chapters 4 and 5, for an explanation of the radial functions), and using these projections to define the density matrix that then defines the U and J dependent parts of the Hamiltonian for the next self-consistent iteration.