Large-Scale Electronic Structure Calculations in solids, P. Giannozzi, in *Computational Approaches to Novel Condensed Matter Systems: Applications to Classical and Quantum Systems*, Proceedings of the 3rd Gordon Godfrey Workshop on Condensed Matter Physics, Editors D. Neilson and M.P. Das (Plenum, New York, 1995), p.67.

LARGE-SCALE ELECTRONIC STRUCTURE CALCULATIONS IN SOLIDS

Paolo Giannozzi

Scuola Normale Superiore, Piazza dei Cavalieri 7

I-56126 Pisa, Italy

1 Introduction

Electronic-structure calculations in solids have considerably evolved from early approaches (*band structure* calculations in periodic model potentials, aimed at reproducing simple crystals) into very sophisticated and powerful techniques. These techniques usually require no or very little experimental input beyond the basic information on atomic composition and some structural data. This is the origin of the (perhaps too ambitious) definitions of *ab-initio*, or *first-principles*, or (perhaps more appropriately) *parameter-free*, which usually label these techniques. In conjunction with the enormous increase in computer power (and the decrease in computer prices), ab-initio methods now allow us to accurately reproduce and even to *predict* electronic and structural properties of *real materials*, and not just the simplest ones. This *predictive power* makes a strong case in favour of ab-initio methods, whenever they are applicable, with respect to *empirical* or *semiempirical* methods. These are far less computationally demanding but also less reliable.

This paper will introduce ab-initio techniques based on Density Functional Theory, in particular those using plane waves and pseudopotentials. Such an approach has proven to be surprising successful and specially well suited to the study of weakly correlated, s-p bonded materials, like for example semiconductors. Relevant aspects for computer implementation will be examined in some detail. In particular, I will examine problems that arise when dealing with complicated systems, that is those described by unit cells containing many atoms. With present algorithms, both computer time and memory requirements scale unfavourably with the size of the unit cell. As a consequence the study of many interesting physical systems such as alloys, disordered or amorphous materials, defects, surfaces and interfaces, is still difficult or impossible. Possible ways to improve the situation will be pointed out.

2 Theoretical Approach

Solving the many-body Schrödinger equation for electrons in a real material is by no means a trivial task even in the presence of simplifying assumptions (such as perfect periodicity for crystals). For completeness, I list here the main possibilities:

i) 'Exact' solutions: Quantum Monte Carlo calculations, Fermionic simulations with Hubbard-Stratonovich transformations. These very recent developments can yield extremely accurate results, but they are also extremely time-consuming so that only very simple problems can be treated.

ii) 'Green's Function approach'. This is based on a perturbative expansion of the self-energy. The most often used approach is the so-called 'GW approximation'. This method is also very accurate (although not undisputed), but it is also computationally cumbersome and it is still limited to simple crystals.

iii) Hartree-Fock-based methods. Hartree-Fock alone is known not to be very accurate, so that some form of correction (Configuration Interaction or perturbative) has to be added to the Hartree-Fock results in order to get better accuracy. This set of techniques, generally used in Quantum Chemistry, has traditionally been applied to molecules rather than to extended system.

iv) Density Functional Theory (DFT), mainly in the Local Density Approximation (LDA). In contrast to the previous approaches, DFT is a *ground-state* theory in which the emphasis is on the *charge density* as the relevant physical quantity. DFT in the LDA has proved to be highly successful in describing structural and electronic properties in a vast class of materials. Furthermore LDA is computationally very simple. For these reasons LDA has become a common tool in first-principles calculations aimed at describing – or even predicting – properties of complex condensed matter systems.

Many good books and papers have been written on DFT and LDA[1], so that I will recall here only the basic facts. DFT looks deceivingly simple, but it hides a number of subtle points. The interested reader should consult the more specialized literature.

We can start from the obvious statement that an external potential $V(\mathbf{r})$ acting on a system of N interacting electrons will determine the charge density $n(\mathbf{r})$ of the ground state. The opposite statement is far less obvious. However this is exactly what has been demonstrated by Hohenberg and Kohn in 1964[2]: there is only one external potential $V(\mathbf{r})$ which yields a given ground-state charge density $n(\mathbf{r})$. The demonstration is elementary and uses a *reductio ad absurdum* argument.

DFT arises from the Hohenberg and Kohn theorem: the ground state energy E is also uniquely determined by the ground-state charge density. In mathematical terms E is a *functional* [1] $E[n(\mathbf{r})]$ of $n(\mathbf{r})$. We can write

$$E[n(\mathbf{r})] = F[n(\mathbf{r})] + \int n(\mathbf{r})V(\mathbf{r})d\mathbf{r} \tag{1}$$

where $F[n(\mathbf{r})]$ is a *universal* functional of the charge density $n(\mathbf{r})$ (and not of V). The variational principle implies that the ground-state energy is *minimised* by the ground-state charge density. In this way, DFT reduces the N-body problem exactly to the determination of a 3-dimensional function $n(\mathbf{r})$ which minimises a functional $E[n(\mathbf{r})]$. Unfortunately this is of little utility as $F[n(\mathbf{r})]$ is not known.

One year later, Kohn and Sham (KS) reformulated the problem[3] and opened the way to practical applications of DFT. First, the system of interacting electrons is mapped on to a fictitious system of non-interacting electrons having the same ground

[1] A functional is a generalisation of the concept of a function: a function associates a value with another value while a functional associates a value with a given function.

state charge density $n(\mathbf{r})$. This is performed by introducing KS orbitals $\psi_i(\mathbf{r})$ for N electrons, such that

$$n(\mathbf{r}) = 2\sum_{i=1}^{N/2} |\psi_i(\mathbf{r})|^2 \tag{2}$$

assuming double occupancy of all states. Charge conservation requires that the KS orbitals obey orthonormality constraints:

$$\int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})d\mathbf{r} = \delta_{ij}. \tag{3}$$

The energy functional is rewritten:

$$E = T_0[n(\mathbf{r})] + \frac{e^2}{2}\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + E_{xc}[n(\mathbf{r})] + \int n(\mathbf{r})V(\mathbf{r})d\mathbf{r}. \tag{4}$$

The first term is the kinetic energy of non-interacting electrons:

$$T_0[n(\mathbf{r})] = -\frac{\hbar^2}{2m}2\sum_{i=1}^{N/2}\int \psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r})d\mathbf{r}, \tag{5}$$

where m is the electron mass. The second term (called the Hartree energy) has the electrostatic interactions between clouds of charge. The third term, called the *exchange-correlation energy*, contains everything also and all our ignorance of the density functional.

Minimization of Eq. (4) under the constraints $\int n(\mathbf{r})d\mathbf{r} = N$ yields the KS equations:

$$(H - \epsilon_i)\,\psi_i(\mathbf{r}) = 0 \tag{6}$$

where

$$\begin{aligned} H &= -\frac{\hbar^2}{2m}\nabla^2 + e^2\int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_{xc}}{\delta n(\mathbf{r})} + V(\mathbf{r}) \\ &\equiv -\frac{\hbar^2}{2m}\nabla^2 + V_{scf}(\mathbf{r}) \end{aligned} \tag{7}$$

is the so-called KS Hamiltonian. [2] In this way the N-electron problem is remapped on to a 1-electron problem with a self-consistent potential V_{scf} which is given by the sum of the external (ionic) potential and a screening potential:

$$V_{scf}(\mathbf{r}) = V(\mathbf{r}) + e^2\int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_{xc}}{\delta n(\mathbf{r})} \tag{8}$$

One still needs a reasonable estimate for the exchange-correlation energy $E_{xc}[n(\mathbf{r})]$. Kohn and Sham[3] introduced the Local Density Approximation, or LDA: they approximated the functional with a *function* of the local density $n(\mathbf{r})$:

$$E_{xc}[n(\mathbf{r})] = \int \epsilon(n(\mathbf{r}))n(\mathbf{r})d\mathbf{r}, \qquad \frac{\delta E_{xc}}{\delta n(\mathbf{r})} \equiv \mu_{xc}(n(\mathbf{r})) = \left(\epsilon(n) - n\frac{d\epsilon(n)}{dn}\right)_{n=n(\mathbf{r})} \tag{9}$$

and for $\epsilon(n(\mathbf{r}))$ used the same dependence on the density as for the homogeneous electron gas (or 'jellium') for which the $n(\mathbf{r})$ is constant.

[2] Functional derivatives $\delta F/\delta n(\mathbf{r})$ are defined implicitly through $\delta F = \int(\delta F/\delta n(\mathbf{r}))\delta n(\mathbf{r})d\mathbf{r}$.

Even in such simple case the exact form of $\epsilon(n)$ is unknown. However, approximate forms have been known for a long time, going back to Wigner[4]. Numerical results from exact Monte-Carlo calculations in jellium by Ceperley and Alder[5] have been parameterized by Perdew and Zunger[6] with a simple analytical form:

$$\begin{aligned} \epsilon_{xc}(n) = & \quad -0.4582/r_s - 0.1423/(1 + 1.0529\sqrt{r}_s + 0.3334 r_s) & , r_s \geq 1 \\ = & \quad -0.4582/r_s - 0.0480 + 0.0311 \ln r_s - 0.0116 r_s + 0.0020 r_s \ln r_s & , r_s \leq 1 \end{aligned} \tag{10}$$

where r_s is the usual parameter in the theory of metals: $r_s = (3/4\pi n)^{1/3}$, and the energy is expressed in Hartree. This form – the 'real' LDA – is often used. More accurate approximations have been recently proposed in Ref.[7]. Usually all forms yield very similar results in condensed-matter calculations (which is not surprising as all the parameterizations are very similar in the range of r_s applicable for solid-state phenomena).

LDA has turned out to be much more successful than expected[8]. Although it is very simple it yields a description of the chemical bond that is superior to that obtained by Hartree-Fock, and it compares well to much weightier Quantum Chemistry methods. For weakly correlated materials such as semiconductors structural and vibrational properties are accurately described: the correct structure is usually found to have the lowest energy, bond lengths, bulk moduli and phonon frequencies are accurate within a few percent or so on[8, 9, 10]. LDA also has some well-known drawbacks. In finite systems the incorrect cancellation of the self-interaction is reflected in an incorrect long-range behaviour[6]. LDA tends to badly overestimate ($\sim 20\%$) cohesive energies and to underestimate to an even worse degree ($\sim 50\%$) the band gaps in insulators. More generally DFT is a ground state theory and KS eigenvalues and eigenvector do not have a clear physical meaning. Moreover LDA is an uncontrolled approximation: it is not clear at all what to do in order to go beyond LDA. There have been several attempts in this direction: recently the *gradient correction*[11] has attracted a lot of interest, but the effectiveness of such approaches is still a controversial issue. Nevertheless LDA is a valuable tool in investigation of material properties.

2.1 Forces in DFT

An important consequence of the variational character of DFT is the possibility of calculating Hellmann-Feynman forces acting on atoms[12]. Forces are the derivative of the total energy with respect to atomic positions $\mathbf{R}_i$:

$$\begin{aligned} \mathbf{F}_i = -\nabla_{\mathbf{R}_i} E \quad = \quad & -\int n(\mathbf{r}) \nabla_{\mathbf{R}_i} V(\mathbf{r}) d\mathbf{r} - \nabla_{\mathbf{R}_i} E_{II}(\mathbf{R}) \\ & -\int \left[\frac{\delta F}{\delta n(\mathbf{r})} + V(\mathbf{r}) \right] \nabla_{\mathbf{R}_i} n(\mathbf{r}) d\mathbf{r}, \end{aligned} \tag{11}$$

where E_{II} is the ion-ion (classical) interaction energy. The electronic part contains an *explicit* dependence on atomic positions through the potential $V(\mathbf{r}) \equiv V_{\{\mathbf{R}\}}(\mathbf{r})$, and an *implicit* dependence through the ground-state charge density $n(\mathbf{r})$. The explicit derivative and the ion-ion term are easy to calculate. The implicit derivative is not. In fact we have to know how the charge density changes when we move the atoms. Fortunately, the variational character of DFT helps us. In fact the last term in Eq. (11) contains the first-order variation of the energy functional around the ground-state energy, which vanishes if the ground state is a minimum. In more mathematical terms, the minimization of the energy under the constraint of constant total charge implies that the

variation of the following function:

$$\min\left\{F[n(\mathbf{r})] + \int n(\mathbf{r})V(\mathbf{r})d\mathbf{r} - \lambda\left(\int n(\mathbf{r})d\mathbf{r} - N\right)\right\}, \tag{12}$$

with respect to an arbitrary variation $\delta n(\mathbf{r})$, must vanish. λ is a Lagrange multiplier and N the total number of electrons. Using elementary variation calculus, this yields

$$\int\left[\frac{\delta F}{\delta n(\mathbf{r})} + V(\mathbf{r}) - \lambda\right]\delta n(\mathbf{r})d\mathbf{r} = 0, \tag{13}$$

that is,

$$\left[\frac{\delta F}{\delta n(\mathbf{r})} + V(\mathbf{r})\right] = \lambda. \tag{14}$$

As a consequence, forces are simply the matrix element of the ground state of the gradient of the external potential plus an ion-ion term:

$$\mathbf{F}_i = -\int n(\mathbf{r})\nabla_{\mathbf{R}_i}V(\mathbf{r})d\mathbf{r} - \nabla_{\mathbf{R}_i}E_{II}(\mathbf{R}). \tag{15}$$

3 Routes to the ground-state

A major goal of electronic structure calculations is to find ground-state ionic positions $\{\mathbf{R}\}$ and the corresponding ground-state electronic states $\{\psi_{gs}\}$ by minimising the total energy $E(\mathbf{R},\psi) = E_{DFT}(\mathbf{R},\psi)+E_{II}(\mathbf{R})$, where the second term is the ion-ion classical Coulomb interaction. We assume the validity of the adiabatic (Born-Oppenheimer) approximation. This goal can be achieved through different strategies.

The most straightforward – and historically the first – approach is to do things one at the time using nested iterations. An *outer* iteration on atomic positions minimises $f(\mathbf{R}) = E(\mathbf{R},\psi_{gs}(\mathbf{R}))$ where the $\psi_{gs}(\mathbf{R})$ are the electronic ground-state KS orbitals for a given atomic configuration $\{\mathbf{R}\}$ (in more formal terms: the $\psi_{gs}(\mathbf{R})$ are 'on the Born-Oppenheimer surface'). The task is much easier if the forces $\mathbf{F}_i = -\nabla_{\mathbf{R}_i}f(\mathbf{R})$ can be calculated.

An *inner* iteration on the potential (or equivalently on the charge density) allows us to find $\psi_{gs}(\mathbf{R})$ and $E(\mathbf{R},\psi_{gs}(\mathbf{R}))$. Starting from some initial guess for the self-consistent potential (for instance, the bare ionic potential) the KS Hamiltonian is diagonalised and solved:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V_{in}(\mathbf{r})\right)\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}). \tag{16}$$

With the resulting charge density:

$$n_{out}(\mathbf{r}) = 2\sum_{i=1}^{N/2}|\psi_i(\mathbf{r})|^2 \tag{17}$$

a new LDA potential is obtained:

$$V_{out}(\mathbf{r}) = V(\mathbf{r}) + e^2\int\frac{n_{out}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \mu_{xc}(n_{out}(\mathbf{r})). \tag{18}$$

This will be equal to the input potential only when self-consistency is attained. The output potential cannot be simply reinserted in the cycle: this will not in general bring convergence. Instead a new V_{in} is generated by some suitable algorithm which takes

into account the V_{in}'s and V_{out}'s of preceding interactions. When self-consistency is achieved one can calculate the total energy $E = E(\mathbf{R})$ as the sum of the electronic and ionic terms. [3] We denote this type of iteration of the potential as 'SCF' ('self-consistent field').

This kind of approach, often referred to as the *wheel in the wheel* algorithm, [4] has been widely and successfully used in a variety of solid-state systems[10]. However this is not the only way to achieve our goal, nor necessarily the best. In particular the relaxation of atoms to their equilibrium positions requires an accurate calculation of forces. It may happen that the relaxation brings into a local minimum and never achieves the minimum-energy structure. The SCF iterations of the potential can be quite slow in some cases (e.g. elongated unit cells, low or no symmetry).

An alternative way to find the electronic ground state for fixed ions consists in minimizing directly the density functional, without solving explicitly the KS equations[13]. In practical calculations the density functional, Eq. (4), is a function of the coefficients of the KS orbitals ψ_i in some basis set. Minimization of such a function under the orthonormality constraints of Eq. (3) can be achieved through the use of well-known conjugate-gradient[14] or similar algorithms. This kind of approach – which we will refer to as 'Direct Minimization' – is becoming increasingly common[15]. Along the same line of thought one can minimize directly the density functional very efficiently using the Direct Inversion in Iterative Subspace (DIIS) method[16] borrowed from Quantum Chemistry.

A more radical point of view has been proposed by Car and Parrinello in 1985[17]: do everything at the same time and on the same footing. This is achieved by merging DFT with classical Molecular Dynamics methods into a new and powerful tool – a Quantum Molecular Dynamics – for the study of real materials. Car and Parrinello introduced a fictitious Lagrangian

$$\mathcal{L} = \frac{\mu}{2} \sum_i \int d\mathbf{r} |\dot{\psi}_i(\mathbf{r}, t)|^2 + \frac{1}{2} \sum_I M_I \dot{\mathbf{R}}_I^2 - E(\mathbf{R}, \psi) \tag{19}$$

which generates the following set of equations of motion:

$$\mu \ddot{\psi}_i = \left(-\frac{\hbar^2}{2m} \nabla^2 + V_{scf}(\psi) \right) \psi_i - \Lambda_{ij} \psi_j, \quad M_I \ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} E. \tag{20}$$

The 'masses' μ associated to the ψ_i's are fictitious: they are used only to generate the dynamics. The Λ_{ij} enforce the orthonormality constraints. The resulting equations of motion can be treated with classical Molecular Dynamics technology. In this way one can simultaneously achieve electronic and ionic relaxation, and even more: one can obtain information on the dynamical behaviour. A good introduction to the Car-Parrinello method and its successes is given in Ref.[18].

The following sections give an overview of technical aspects and of the practical implementation of the SCF approach using plane waves basis sets and pseudopotentials to reproduce the electron-ion interactions. Most of the technical points are relevant for the Car-Parrinello and Direct Minimization approaches as well.

[3] The general properties of the density functional ensure that *for fixed ions* there is only one minimum, corresponding to the ground state.

[4] Another 'wheel' is generally present in the solution of the KS Hamiltonian, see Sec.5.2.

4 Tools

4.1 Supercells

In perfect crystals the natural framework to solve the KS equations takes advantage of periodicity: the crystal is described by a periodically repeated *unit cell* containing one or more atoms (the *basis*), by a crystal lattice and a reciprocal lattice. Using Bloch's theorem, states are classified by a wavevector $\mathbf{k}$ in the Brillouin Zone (BZ) and form energy bands. In some cases, unit cells can be quite large and contain many atoms. This is the case for example with superlattices, artificial layered materials in which the periodicity of the original lattice – the one of the component materials – is replaced by a 'superperiodicity', with a *supercell* which is larger than the original cell.[5] Another example is given by solid C_{60}, both pure and doped with alkali atoms (such as the superconductor K_3C_{60}).

Many interesting physical systems do not exhibit perfect periodicity. Good examples are disordered superlattices (superlattices with intermixing at the interfaces), quantum wells, substitutional alloys, point defects, and surfaces. In all such cases it is convenient to simulate an aperiodic (or 'almost-periodic') system with a periodically repeated fictitious supercell. The form and the size of the supercell depend on the physical system being examinated. For example, the study of point defects requires a defect 'not to see' its periodic replica in order to accurately simulate a truly isolated defect. This approach is actually used also for the extreme case of molecules and clusters. The use of supercells for such demonstrably nonperiodic objects may seem odd but there are important computational advantages in such an approach. [6]

The size of the unit cell – the number of atoms and the volume – is very important. Together with the type of atoms it determines the difficulty of the calculation: large unit cells mean large calculations. Unfortunately many interesting physical systems are described – exactly or approximately – by large unit cells.

4.2 Plane Waves

KS wavefunctions must be expanded in some suitable basis set. For periodic systems, the basis set is formed by Bloch states of wavevector $\mathbf{k}$. A large number of different basis sets have been proposed and used. Most fall into the 'localised basis set' category. Some of the most frequently used are Bloch sums of Linear Combinations of Atomic Orbitals (LCAO), Gaussian-type Orbitals (GTO) and Linearised Muffin-Tin Orbitals (LMTO)[9]. These atomic-like functions are tailored for fast convergence, so that only a few (some tens at most) functions per atom are needed. However they are quite delicate to use. In particular it is difficult to check systematically for convergence (a well-known problem of Quantum Chemistry). Another serious drawback is the difficulty of calculating derivatives of the energy (e.g. forces on atoms)[19]. Forces are very important in determining the structure of a complex system. They are crucial quantities in Car-Parrinello and other Quantum Molecular Dynamics applications. Only recently have important advances been made in this field[20].

An opposite approach is to choose completeley extended, atomic-independent Plane Waves (PW) as basis set. For a given reciprocal lattice $\{\mathbf{G}\}$ and for a given vector $\mathbf{k}$

[5] Strictly speaking, there is nothing 'super' in both the cell and the lattice of a perfect superlattice. One just wants to stress that the unit cell includes several unit cells of the component materials.

[6] The opposite point of view, that is, simulating extended systems with larger and larger clusters, has been traditionally used in Quantum Chemistry. However, it turns out that clusters converge only very slowly in solid state behaviour.

in the BZ, a PW basis set is defined as

$$|\mathbf{k}+\mathbf{G}\rangle = \frac{1}{V} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \quad \frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2 \leq E_{cut}, \tag{21}$$

where V is the crystal volume, and E_{cut} is called the 'kinetic energy cutoff', or simply the cutoff. PW's have many attractive features: they are simple, orthonormal by construction, unbiased and it is very simple to check for convergence (by increasing the cutoff).

A distinct computational advantage of PW's is the existence of very fast algorithms (known as the Fast Fourier-Transform, FFT) to perform the discrete Fourier transforms. This allows simple and fast transformation from reciprocal to real space and vice versa. The basic one-dimensional FFT executes the following transformation:

$$f_i = \sum_{j=0}^{N-1} g_j e^{2\pi i j/N}, \quad i = 0, ..., N-1 \ , \tag{22}$$

and its inverse

$$g_i = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-2\pi i j/N}. \tag{23}$$

The tranformation is usually performed 'in place', that is the result is overwritten on the input vector. This takes $\mathcal{O}(N \log N)$ operations. In three dimensions the discrete Fourier transform maps a function $\widetilde{f}(\mathbf{g}_i)$ in reciprocal space into a function $f(\mathbf{r}_i)$ in the unit cell (and vice versa):

$$\mathbf{g}_i = i_1\mathbf{G}_1 + i_2\mathbf{G}_2 + i_3\mathbf{G}_3, \quad \mathbf{r}_i = \frac{j_1}{N_1}\mathbf{R}_1 + \frac{j_2}{N_2}\mathbf{R}_2 + \frac{j_3}{N_3}\mathbf{R}_3 \tag{24}$$

where $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ ($\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3$) are the three fundamental translations that generate the real-space (reciprocal) lattice, $i_1 = -N_1/2, ..., N_1/2$, and so on. N_1, N_2, N_3 must be sufficiently large to include all available Fourier components; the more Fourier components, the larger the grid in $\mathbf{G}$-space and the finer the grid in $\mathbf{R}$-space. It is easily verified that this 3-d FT can be done in a very fast way by performing 3 inter-nested 1-d FFT. Computers usually have some highly optimised library routines performing FFT.[7]

PW's look like the ideal basis set for solids. Unfortunately their extended character makes it very difficult to accurately reproduce localized functions such as the charge density around a nucleus or even worse, the orthogonalization wiggles of inner ('core') states. In order to describe features which vary on a lengthscale δ, one needs PW's with a cutoff as high as $\sim (2\pi/\delta)^2$, that is one needs $\sim 4\pi(2\pi/\delta)^3/3\Omega$ PW's (where Ω is the dimension of the BZ). A simple estimate for diamond is instructive. The $1s$ wavefunction of the carbon atom has its maximum around 0.3 a.u., so $\delta \simeq 0.1$ a.u. is a reasonable value. Diamond has an fcc lattice ($\Omega = (2\pi)^3/(a_0^3/4)$) with lattice parameter $a_0 = 6.74$ a.u., thus yielding $\sim 250,000$ PW's. This is clearly too much for practical use.

[7]The 'true' original FFT works only when N_1, N_2, N_3 are powers of 2, but algorithms for more general values do exist. While it is a good practise to keep factors in N as small as possible, it is not always true that powers of 2 are the best choice: it depends on the algorithms used and on the computer architecture.

4.3 Pseudopotentials

Core states prevent the use of PW's. However they do not contribute in a significant manner to chemical bonding and to solid-state properties, only outer ('valence') electrons do[8]. Core states remain there 'frozen' in their atomic configuration making calculations more difficult. The idea of replacing the full atom with a much simpler 'pseudoatom' with valence electrons only arises naturally (apparently for the first time, in a 1934 paper by Fermi[21]). 'Pseudopotentials' (PP's) have been widely used in solid state physics starting from the 1960's. In earlier approaches PP's were devised to reproduce some known experimental solid-state or atomic properties such as energy gaps or ionization potentials. Already in this 'empirical' phase they proved to be a valuable tool[22].

A major breakthrough occurred in 1979 with the introduction of Norm-Conserving PP's by Hamann, Schlüter, and Chiang[23]. These are *atomic* potentials which must obey the following conditions. Given a reference atomic configuration
1) all-electron and pseudo-wavefunctions must have the same energy, and
2) they must be the same beyond a given 'core radius' r_c, which is usually located around the outermost maximum of the atomic wavefunction;
3) the pseudo-charge and the true charge contained in the region $r < r_c$ must be the same. [9]

The above conditions ensure *transferability*, the ability to yield correct results for configurations different from the reference one, or in different environments (such as in molecules or in condensed matter). In fact Norm-Conserving PP's are able to accurately mimic core scattering properties in an energy region which is not too far from the energy of the reference state. Norm-Conserving PP do not have singularities. They are relatively smooth functions, whose long-range tail goes like $-Z_v e^2/r$ where Z_v is the number of valence electrons. They are *nonlocal* because it is usually impossible to mimic the effect of orthogonalization to core states on different angular momenta l with a single function. There is a PP for every l:

$$\widehat{V}^{ps} = V_{loc}(r) + \sum_l V_l(r)\widehat{P}_l \tag{25}$$

where $V_{loc}(r) \simeq -Z_v e^2/r$ for large r and $\widehat{P}_l = |l\rangle\langle l|$ is the projection operator on states of angular momentum l. We can recast such a potential in the form

$$\widehat{V}^{ps} = V_{loc}(r) + \sum_{lm} Y_{lm}(\mathbf{r})V_l(r)\delta(r-r')Y^*_{lm}(\mathbf{r}'), \tag{26}$$

for which the nonlocal character is more evident. Although nonlocality is a drawback, because it makes calculations more difficult, it is not a big one [10] in practical applications.

Tables of PP's for all elements have been published by Bachelet, Hamann, and Schlüter in a a much-quoted 1982 paper[24]. Experience has shown that PP's are practically equivalent to the frozen core approximation[25]: PP and all-electron calculations on the same systems yield almost indistinguishable results (except for those cases in which core states are not sufficiently 'frozen').

[8] The border between core and valence electron is often evident on physical grounds but sometimes it is not. For instance, in Cs the $5s$ and $5p$ states can sometimes be considered safely as core states but more often they cannot.

[9] This last condition explains the name 'Norm-Conserving'. There is an historical reason: some older pseudopotentials had an 'orthogonalization hole' problem which caused violation of condition 3

[10] However, strictly speaking, DFT is valid only for *local* external potentials!

4.4 Brillouin-Zone Sampling

In a periodic system, or in a system described by a suitable supercell, the states are classified by a wavevector $\mathbf{k}$ in the BZ and by a band index. For a given $\mathbf{k}$, the KS orbitals are expanded into PW's of appropriate periodicity

$$|\psi_{\mathbf{k},i}\rangle = \sum_{\mathbf{G}} \Psi_i(\mathbf{k}+\mathbf{G})|\mathbf{k}+\mathbf{G}\rangle \tag{27}$$

up to a given energy cutoff E_{cut}. In order to calculate the charge density $n(\mathbf{r})$ in a periodic system one has to sum over an infinite number of $\mathbf{k}$-points:

$$n(\mathbf{r}) = \sum_{\mathbf{k}} \sum_{i} |\psi_{\mathbf{k},i}(\mathbf{r})|^2. \tag{28}$$

In fact, assuming periodic (Born-Von Kàrmàn) boundary conditions

$$\psi(\mathbf{r}+L_1\mathbf{R}_1) = \psi(\mathbf{r}+L_2\mathbf{R}_2) = \psi(\mathbf{r}+L_3\mathbf{R}_3) = \psi(\mathbf{r}), \tag{29}$$

a crystal has $L_1L_2L_3$ allowed $\mathbf{k}$-points. In the limit of an infinite crystal, $L_1L_2L_3 \to \infty$ and the discrete sum becomes an integral over the BZ. It is not obvious at all that this integral can be approximated by a discrete sum over an affordable number of $\mathbf{k}$-points. However experience shows that this is actually possible, at least in crystals with completely filled or completely empty bands.

Symmetry – when present – can be used to reduce the number of calculations to be performed. Only one $\mathbf{k}$-point is left to represent each 'star' – the set of $\mathbf{k}$-points that are equivalent by symmetry – with a 'weight' w_i which is proportional to the number of $\mathbf{k}$-points in the star. The infinite sum over the BZ is replaced by a discrete sum over a set of points $\{\mathbf{k}_i\}$ and 'weights' w_i:

$$\frac{1}{V}\sum_{\mathbf{k}} f_{\mathbf{k}}(\mathbf{r}) \longrightarrow \frac{\Omega}{(2\pi)^3}\sum_{i} w_i f_{\mathbf{k}_i}(\mathbf{r}). \tag{30}$$

The resulting sum is then symmetrized to get the charge density. Other quantities (such as the total energy) which contain sums over the BZ can be dealt with in a similar way.

Suitable sets for BZ sampling in insulators and semiconductors are called 'special points'[26, 27]. Let us consider the case of an fcc crystal having cubic symmetry. The smallest special point grid is formed by the 'mean-value' or Baldereschi point[26]: $\mathbf{k} = (0.6225, 0.2953, 0.0)$ (in units of $2\pi/a_0$). Better accuracy is obtained with the much-used 'two Chadi-Cohen points'[27]: $\mathbf{k}_1 = (1/4, 1/4, 1/4), w_1 = 1$ and $\mathbf{k}_2 = (1/4, 1/4, 3/4), w_2 = 3$. Then there are 6-, 10-point grids and so on, yielding increasing accuracy. [11]

In metals things are more difficult because one needs an accurate sampling of the Fermi surface. The 'Gaussian broadening' and the 'tetrahedron' techniques, or variations of the above[28], are generally used.

When the unit cell gets larger, the BZ become smaller and the need for accurate sampling becomes less stringent. For very large supercells, the sampling obtained with only the Γ point ($\mathbf{k} = 0$) is usually good enough. Of course when supercells are used to simulate aperiodic systems such as clusters and molecules the Γ point is the good choice. There is no point in trying to reproduce the effect of a fictitious periodicity.

[11] Actually the name 'special points' is somewhat misleading in this case. In fact those sets just form uniform grids in the BZ.

5 Algorithmic aspects

5.1 Potential mixing

The calculation of the self-consistent potential requires a way to mix the input and output potentials to yield a new input potential. The simplest approach ('simple mixing') is the following:

$$V_{in}^{(n+1)} = \alpha V_{in}^{(n)} + (1-\alpha) V_{out}^{(n)}, \tag{31}$$

where the superscripts indicate the iteration number and the value of α must be chosen empirically in order to get fast convergence. As a rule relatively big values ($\alpha = 0.3-0.5$) can be chosen for small cells while smaller values are needed for bigger or elongated cells (like for a superlattice). Simple mixing is not very effective especially for big cells, or even worse for surface calculations. Better results are obtained with more sophisticated algorithms, like for example the family of Anderson algorithms[29] and modified Broyden algorithms[30]. The latter seems to be quite effective in situations of difficult convergence.

Eventually, as the size of the cell increases convergence becomes slower and slower. This adverse behaviour can be traced back to charge oscillations ('charge sloshing') that take place in large cells. Preconditioned Conjugate Gradient algorithms are claimed to be less sensitive to this problem[15]. The phenomenon of convergence-slowing-down is well-known in many different fields of computational physics. The ultimate solution could come from application of *multigrid* concepts[31]. Work along such lines is in progress[32].

5.2 Diagonalization of the Hamiltonian

By far the most time-consuming step in an SCF calculation is the solution of the KS equations, Eq. (7). When the eigenfunctions are expanded on a finite basis set the solution takes the form of a secular equation:

$$\sum_{\mathbf{G}'} H(\mathbf{k}+\mathbf{G}, \mathbf{k}+\mathbf{G}') \psi_{\mathbf{k},i}(\mathbf{G}') = \epsilon_{\mathbf{k},i} \psi_{\mathbf{k},i}(\mathbf{G}), \tag{32}$$

where the matrix elements of the Hamiltonian have the form

$$H(\mathbf{k}+\mathbf{G}, \mathbf{k}+\mathbf{G}') = \frac{\hbar^2}{2m}(\mathbf{k}+\mathbf{G})^2 \delta_{\mathbf{G},\mathbf{G}'} + V_{loc}(\mathbf{G}-\mathbf{G}') + V_{NL}(\mathbf{k}+\mathbf{G}, \mathbf{k}+\mathbf{G}'). \tag{33}$$

The local contribution V_{loc} includes both the local term in the PP's and the screening potential. The nonlocal contribution V_{NL} comes from the PP's. The problem is reduced in this way to the well-known problem of finding the lowest M eigenvalues and eigenvectors (only the valence states for insulators, a few more for metals) of an $N \times N$ Hermitian matrix (or a real symmetric matrix when inversion symmetry is present). This task can be performed with well-known bisection-tridiagonalization algorithms, for which very good public-domain computer packages (EISPACK or the more recent LAPACK)[33] are available. Unfortunately this straightforward procedure has serious limitations. In fact:

– the computer time required to diagonalise a $N \times N$ matrix grows as N^3;

– the matrix must be stored in memory, requiring $\mathcal{O}(N^2)$ memory.

As a consequence a calculation requiring more than a few hundred PW's becomes exceedingly time- and memory-consuming. As the number of PW's increases linearly

with the size of the unit cell it is very hard to study systems containing more than a few (say 5-10) atoms.

Both limitations have been overcome with the introduction of *iterative* techniques[34]. These techniques can be used whenever

i) the number of states to calculate M is much smaller than the dimension of the basis set N, and

ii) a reasonable and economical estimate of the inverse operator H^{-1} is available.

Both conditions are satisfied in an SCF calculation in a PW basis set. In fact $M \ll N$ is always true, and the Hamiltonian matrix is dominated by the kinetic energy at large $\mathbf{G}$ (that is, the Hamiltonian is *diagonally dominant*).

Iterative methods are based on a repeated refinement of a trial solution, which is stopped when satisfactory convergence is achieved. The number of iterative steps cannot be predicted in advance. It depends heavily on the structure of the matrix, on the type of refinement used, and on the starting point. A well-known and widely used algorithm is due to Davidson[34]. In this method, a set of correction vectors $|\delta\psi_i\rangle$ to the M trial eigenvectors $|\psi_i\rangle$ are generated as follows:

$$|\delta\psi_i\rangle = \frac{1}{D - \epsilon_i}(H - \epsilon_i)|\psi_i\rangle \tag{34}$$

where the $\epsilon_i = \langle\psi_i|H|\psi_i\rangle$ are the trial eigenvalues. The $|\delta\psi_i\rangle$'s are orthogonalised and the Hamiltonian is diagonalised (with conventional techniques) in the subspace spanned by the trial and correction vectors. A new set of trial eigenvectors is obtained and the procedure is iterated until convergence is achieved. A good set of starting trial vectors is supplied by the eigenvectors at the preceding iteration of the potential.

An important point is the following. The Hamiltonian matrix is never explicitly required excepted for its diagonal part. Only $H\psi_i$ products are required, which can be calculated in a very convenient way by applying the *dual-space technique*[35]. In fact the kinetic term appearing in Eq. (33) is diagonal in $\mathbf{G}$-space, whereas the local potential term is diagonal in real space. Using FFT's one can go quickly back and forth from real to reciprocal space and perform the products where it is more convenient. There is still a nonlocal term which appears to require the storage of the matrix. The trick is to write V_{NL} in a 'separable' form:

$$V_{NL}(\mathbf{k}+\mathbf{G}, \mathbf{k}+\mathbf{G}') = \sum_{\mu=1}^{N_{at}} \sum_{j=1}^{n} f_j^\mu(\mathbf{k}+\mathbf{G}) g_j^\mu(\mathbf{k}+\mathbf{G}'), \tag{35}$$

where n is a small number and N_{at} is the number of atoms in the unit cell. This allows us to perform the products by storing only the f and g vectors. The reduction to separable form is exact and straightforward when the Kleinman-Bylander projection (see Appendix) is used. It is more involved but still possible when conventional PP's are used. The trick can also be used with V_{NL} in real space thus taking advantage of the short-range nature of V_{NL} [36].

5.3 Computational workload

When iterative techniques are used to diagonalize the Hamiltonian the time-consuming step is the calculation of the products $H\psi$ followed by orthogonalization. For a system of N_{at} atoms in the unit cell having M occupied states expanded into N PW's, the calculation of the products $H\psi$ for all occupied states require:

– $\mathcal{O}(MN^2)$ operations if the Hamiltonian is stored as a matrix;

– $\mathcal{O}(MN \log N) + \mathcal{O}(N_{at}MN)$ operations if the dual-space technique is used;
– $\mathcal{O}(MN \log N) + \mathcal{O}(N_{at}M)$ operations if the dual-space technique is used with V_{NL} in real space.

Orthogonalization of each trial eigenfunction to all others requires $\mathcal{O}(M^2N)$ operations. Similar considerations apply to the Car-Parrinello and Direct Minimization approaches.

Let us consider how the computational workload scales with the number of atoms in the unit cell. [12] The number of states M is proportional to N_{at}. The number of PW's is proportional to the volume of the unit cell. For this reason $N \propto N_{at}$ as well. In summary the $H\psi$ products can be performed in $\mathcal{O}(N_{at}^2)$ operations in the best case. Orthogonalization requires $\mathcal{O}(N_{at}^3)$ operations. For very big cells this will be the dominant part whatever we do to speed up the calculation of the $H\psi$ products.

This unfavourable scaling with the size of the cell has a simple physical root. The Hamiltonian eigenstates of an extended system are usually extended. As a first consequence the computational workload needed to calculate them scales at least as the number of orbitals times their size (the dimension of the basis set), i.e. as N_{at}^2. As a second consequence orbital orthogonalization requires additional work proportional to the square of the number of orbitals times their size, i.e. N_{at}^3.

6 Linear-scaling algorithms

General physical considerations indicate that the energy and related properties of a system of non-interacting electrons (such as the one appearing in the Kohn-Sham formulation of DFT) should be calculable with a workload proportional to the size of the system. In fact the local nature of the Schrödinger equation and the condition of local charge neutrality imply that the local density of states at a given point – in contrast with the individual eigenfunctions – does not depend on the details of the system far from that point[37].

The search for better-scaling algorithms has become a very active research field in the last few years and many different proposals have appeared in the literature[38, 39, 40, 41, 42].

One possibility is in some way to force the eigenfunctions to be localized in space and to express the electron density in terms of the latter. This idea is being explored following different methods by several authors. In Yang's approach[38] the system is divided into small *independent* subsystems within which the numbers of electrons are determined by a common chemical potential. The electron density is then determined by conventional diagonalization techniques within each subsystem. Another way to achieve the same goal is to use 'confining potentials'. This idea is currently under active development in a Car-Parrinello framework and looks quite promising[39].

Other authors have focussed on the direct solution of the *density matrix*[40] suitably truncated in real space. Another idea is to use a statistical approach, based on the *maximum entropy* principle[41], to extract the relevant information on the density of states.

In the following, I will briefly describe a different approach[42] in which the use of KS orbitals is completely avoided. The charge and energy densities are obtained directly from the one-electron Green's function without going through any Schrödinger-

[12] We do not address here the problem of convergence slowing-down with increased size of the cell. We only consider the time needed to execute a basic steps, not how many basic steps are needed in a complete calculation.

like equations. This method is in principle exact,[13] it displays the right linear scaling with size, and it is still rather simple to implement. The approach is based on a finite-difference real-space discretization of the Hamiltonian and on the Recursion Method[43] to calculate the electron and energy densities from selected elements of the one-electron Green's function.

The starting point is a well-known identity relating the density matrix of a system to its one-electron Green's function,

$$\rho(\mathbf{r},\mathbf{r}') = 2\frac{1}{2\pi i}\oint_{C_F} G(\mathbf{r},\mathbf{r}';z)dz, \tag{36}$$

where C_F is a contour in the complex energy plane enclosing just the occupied-state eigenvalues. The factor 2 accounts for spin degeneracy and the Green's function $G(\mathbf{r},\mathbf{r}';z)$ is the real-space representation of the resolvent of the Hamiltonian,

$$G(\mathbf{r},\mathbf{r}';\epsilon) \equiv \langle \mathbf{r}|\widehat{G}(\epsilon)|\mathbf{r}'\rangle = \langle \mathbf{r}|\frac{1}{\epsilon - H}|\mathbf{r}'\rangle. \tag{37}$$

According to Eq. (36), the electron density, $n(\mathbf{r}) = \rho(\mathbf{r},\mathbf{r})$, can easily be calculated from knowledge of the *diagonal* elements of the Green's function. The number of points necessary to numerically evaluate the integral (36) to a given accuracy depends on the valence band width and on the minimum distance of the contour from the energy eigenvalues. Both these properties do not depend on the size of the system, at least for insulators. Therefore if the work needed to calculate Eq. (36) is independent of the size, then the work necessary to obtain the full charge density will scale linearly.

We next discretize the problem by using finite differences on a uniform grid $\{\mathbf{r}_i\}$. The finite-difference representation of the Hamiltonian is sparse. If second-order discretization on a cubic uniform grid is used for the Laplacian the only non vanishing matrix elements $H_{ij} \equiv \langle \mathbf{r}_i|H|\mathbf{r}_j\rangle$ are on the diagonal and between neighbouring points,

$$H_{ij} = \begin{cases} \dfrac{1}{12h^2} + V_{scf}(\mathbf{r}_i) & \text{if } i = j \\ -\dfrac{1}{2h^2} & \text{if } |\mathbf{r}_i - \mathbf{r}_j| = h \\ 0 & \text{otherwise,} \end{cases} \tag{38}$$

where h is the spacing between the grid points and a local potential is assumed.[14]

Calculating Eq. (38) by standard factorization techniques would again result in a workload proportional to N_{at}^3. Iterative algorithms to solve elliptic partial differential equations could in principle be used to calculate the inverse of $(\epsilon - H)$ with a cost proportional to N_{at}^2. Of course this price is optimal for the calculation of the *full* inverse matrix. It is too high if only the *diagonal* of the inverse is required.

A convenient way to compute a *single* diagonal element of the Green's function is provided by the Recursion Method of Haydock, Heine, and Kelly[43]. In the Recursion Method, diagonal elements of the Green's function $\langle \phi_0|\widehat{G}|\phi_0\rangle$ are expressed in terms of a continuous fraction whose coefficients are calculated from a chain of orthonormal states recursively generated from $|\phi_0\rangle$,

$$\begin{aligned} b_1|\phi_1\rangle &= H|\phi_0\rangle - a_0|\phi_0\rangle \\ b_2|\phi_2\rangle &= H|\phi_1\rangle - a_1|\phi_1\rangle - b_1|\phi_0\rangle \\ &\cdots \\ b_n|\phi_n\rangle &= H|\phi_{n-1}\rangle - a_{n-1}|\phi_{n-1}\rangle - b_{n-1}|\phi_{n-2}\rangle, \end{aligned} \tag{39}$$

[13] It is also possible to get rid of the KS orbitals by using Thomas-Fermi like approximations for the density functional. Such approximations are unfortunately quite inaccurate.

[14] Nonlocal PP can also be used exploiting the short-range nature of the nonlocal terms.

where

$$a_n = \langle \phi_n | H | \phi_n \rangle, \qquad b_n = \langle \phi_n | H | \phi_{n-1} \rangle. \tag{40}$$

The relevant diagonal element of the Green's function is then given by the continuous-fraction expansion,

$$\langle \phi_0 | \widehat{G}(\epsilon) | \phi_0 \rangle = \cfrac{1}{\epsilon - a_0 - \cfrac{b_1^2}{\epsilon - a_1 - \cfrac{b_2^2}{\epsilon - a_2 - \cdots}}}. \tag{41}$$

The part of the continuous fraction which is not calculated when truncating the chain after n steps (indicated in Eq. (41) by dots '$\cdots$' in the case $n = 2$) is called the *terminator*, $t_n(\epsilon)$. The terminator describes the influence on $\langle \phi_0 | \widehat{G}(\epsilon) | \phi_0 \rangle$ of that portion of the system which is not spanned by the finite chain. In actual calculations, the continuous fraction (41) is either *truncated* after a finite number n of steps ($t_n(\epsilon) = 0$), or is closed by an approximate terminator. In this case the initial state is localized, $|\phi_0\rangle = |\mathbf{r}_i\rangle$, as are the subsequent terms of the chain by virtue of the sparseness of the Hamiltonian (38). Therefore the evaluation of each given step of the chain (39) requires a workload independent of the size of the system. At each step the newly generated state in the chain of states explores further points. The continued fraction can be truncated when the relevant region around a point – the local environment which mostly determines the charge density at that point – is explored. This happens after a number of steps which is independent of the size[37], at least for sufficiently large systems. We conclude that the application of Recursion Method ideas allows us to calculate the electron density – and hence to implement DFT – with a workload which scales linearly with N_{at}. Moreover the procedure is highly suitable for parallel computing as the charge density can be calculated independently at each point of the grid.

The off-diagonal elements of the density matrix necessary to calculate the kinetic energy could be computed along similar lines. The kinetic energy, however, is more conveniently calculated from the relation,

$$E_{kin} = 2 \sum_i \epsilon_i - \int n(\mathbf{r}) V_{scf}(\mathbf{r}) d\mathbf{r}. \tag{42}$$

The sum over occupied-state eigenvalues can again be expressed in terms of diagonal elements of the Green's function:

$$\sum_i \epsilon_i = \frac{1}{2\pi i} \sum_j \oint_{C_F} z G(\mathbf{r}_j, \mathbf{r}_j; z) dz. \tag{43}$$

A first test of this approach has been done in Ref.[42]. It was shown that the number of steps necessary to obtain a satisfactory accuracy was rather large if the chain was simply truncated. This is a consequence of the well-known difficulty of reproducing the properties of an extended system by truncating it to a finite cluster. In fact a truncated chain cannot distinguish between a cluster spanned by the chain functions and the extended system under consideration. Later it has been found [44] that much faster convergence could be achieved when the terminator for free electrons was used. Tests done on large silicon clusters (up to ~ 2000 atoms) have demonstrated the applicability of this approach to realistic cases. A major problem still to be solved concerns accurate force calculation, which still requires an exceedingly high number of chain steps to be performed.

7 Conclusions

Application of iterative techniques, both in SCF and in Car-Parrinello or Direct Minimization approaches, has greatly enhanced the scope of first-principles investigations of real materials[10, 15, 18]. It is now possible to study systems with ~ 100 atoms in the unit cell, including e.g. superlattices, some alloys and disordered materials, simple surfaces, small clusters and fullerene systems.

Many interesting physical systems are still beyond the reach of ab-initio methods based on PW's and PP's. In many cases this is due to the presence of atoms, such as O, F, the rare earths and the transition metals, whose PP's are very hard. A first answer to this limitation is coming from 'ultrasoft' PP's[45]. This approach is computationally quite cumbersome but it is becoming increasingly common[18]. Another idea is to look for new basis sets, hopefully enjoying the advantages of both PW's and localized basis sets but without the respective disadvantages. Two proposals have recently appeared in the literature, 'adaptive grids'[46] and 'wavelets'[47]. Both basis sets have nice properties but their merits in real calculations have to be demonstrated 'on the battlefield'.

Many other interesting systems (e.g. amorphous or disordered materials, alloys, extended defects, carbon microtubules, organic molecules) are described by supercells that are too big ($\sim$ 1000 atoms) for present computers. The development of parallel computers and of parallelized algorithms can bring some of those systems within reach. A recent example of the potential of parallel computing has been the study of the 7×7 reconstruction of the Si (111) surface[48]. To the best of my knowledge this is the largest system (400 atoms) ever studied ab-initio. Eventually the limiting factor is the unfavourable $\mathcal{O}(N_{at}^3)$ scaling of the computer effort with the number N_{at} of atoms. Overcoming such a bottleneck requires new radically different ideas such as those outlined in Sec. 6.

A Appendix

A.1 Generation of Pseudopotentials

The best way to understand what a pseudopotential is is to follow the steps needed in its generation. The basic tool is an atomic LDA program. This usually assumes that the charge density is spherical (a very good approximation even for open-shell atoms) thus allowing separation of variables into the radial and the angular ones. The states are classified in exactly the same way as in an introductory course on quantum mechanics. There is a main quantum number n, an angular momentum $l = 0, .., n-1$, and $m = -l, ..., l$. The atomic configuration is given in terms of occupation numbers (traditionally, $1s^2 2s^2 2p^6 ....$). The radial KS equations are integrated numerically on a logarithmic grid with one of the many well-known methods (this can be done even on a personal computer in a few seconds). A reasonable configuration is chosen (usually the ground-state) and all-electron self-consistent LDA radial wavefunctions $\phi_l(r)$ are obtained.

At this point there are several possible ways to proceed. A very simple and clear approach is due to Kerker[49]. For each l in the valence shell a nodeless [15] pseudo-

[15] This ensures that there are no lower states with the same l. The inner part of pseudo-wavefunctions is unphysical: orthogonalization to the core states yields meaningless objects.

wavefunction is constructed in the following way:

$$\begin{aligned} \phi_l^{ps}(r) = &\quad \phi_l(r) &&, r \geq r_c \\ = &\quad r^{l+1} e^{p(r)} &&, r \leq r_c \end{aligned} \tag{44}$$

where $p(r)$ is a polynomial $p(r) = a + br^2 + cr^3 + dr^4$ whose coefficients are determined by imposing continuity of $\phi_l^{ps}(r)$ and its first and second derivatives at the matching point r_c, plus the norm-conservation requirement 3) (see Sec. 4.3). The r_c can safely be taken at the outermost maximum. A 'dressed' PP (containing the screening LDA potential as well) is now obtained by inverting the radial KS equation at the all-electron eigenvalue ϵ_l. The final PP is obtained by 'unscreening', that is, removal of the LDA potential generated by valence electrons only.

PP's are usually obtained in numerical form on a grid, sometimes fitted to an analytical form like the one used in ref.[24]. However it is also possible to get PP's directly in numerical form[50]. Assuming a simple analytic form for the PP's which depends on a few parameters $\{\lambda_i\}$, one directly minimizes the following function:

$$f(\{\lambda_i\}) = \sum_l |\epsilon_l^{ps} - \epsilon_l|^2 + \sum_l \int_{r>r_c} r^2 \left(\phi_l^{ps}(r) - \phi_l(r)\right)^2 dr, \tag{45}$$

or some equivalent form, using one of the standard minimisation methods[14]. The author has used this procedure for many simple atoms.

A.2 Towards better Pseudopotentials

After the original Hamann, Schlüter, and Chiang paper, many (real or presumed) improvements and extensions have been proposed. They can be classified according to their main aim.

1) The first goal is to improve the reliability and accuracy of the PP's. One should not forget that a PP is an approximation, albeit a good one, to the true atom (in fact even worse: an approximation to the frozen-core approximation). In some cases, e.g. for alkali atoms with one valence electron, the loss of accuracy due to the neglect of non-linearity in exchange-correlation which is implicit in the unscreening procedure can be intolerable. In such cases the simple 'core correction'[51] is very useful. 'Generalized'[52] and 'extended'[53] PP's have also been proposed to improve transferability. The former allow the use of unbound atomic states as reference states; the latter have scattering properties that are correctly reproduced beyond first order in energy differences.

More recent work concerns extension of PP's to gradient-corrected density functionals [54] (PP's should be generated within the same approximation used for subsequent calculations. Using LDA PP's in gradient-corrected calculations is inconsistent), and PP's for GW and Quantum Monte Carlo calculations[55].

2) The second goal is to improve the computational efficiency of PP's. An important step in this direction is due to Kleinman and Bylander[56]. They proposed projecting the PP on to the reference pseudo-wavefunctions ϕ_l^{ps} in the following way,

$$\widehat{V}^{ps} = V_{loc} + V_L + \sum_l \frac{|\delta V_l \phi_l^{ps}\rangle\langle\delta V_l \phi_l^{ps}|}{\langle\phi_l^{ps}|\delta V_l|\phi_l^{ps}\rangle}\ , \quad \delta V_l(r) = V_l(r) - V_L(r) \tag{46}$$

where $V_L(r)$ is an arbitrary function. By construction the original PP and the projected $\widehat{V}^{ps}$ have the same eigenvalues and eigenvectors on the reference states ϕ_l^{ps}. This justifies the hope that the original and projected PP's will yield very similar results on other configurations as well. The Kleinman-Bylander form is much more convenient than the

conventional form (for reasons that will be explained in Sec. 5.2). Unfortunately it can happen that spurious states ('ghosts') appear at energies which are lower than or comparable to the reference energy. In such a case the Kleinman-Bylander projection badly fails. Some recipes have been devised to avoid the 'ghost' problem[57].

Another desirable goal is the reduction of the number of PW's needed for calculations. This depends on the type of atoms involved. In typical semiconductors (e.g. Si, Ge, GaAs, AlAs) 100-150 PW's per atom are sufficient for most applications. However, many atoms – transition metals, first-row elements like F, O, and to a lesser extent N and C – are described by strong PP's, requiring impractically large PW basis sets. One can try to exploit the many 'degrees of freedom' which are present in PP generation to get softer PP's. For instance, a rule of thumb states that the larger the matching point r_c is the smoother and less accurate the resulting PP. One can strike a compromise between accuracy and computer budget by pushing r_c outwards. However this will not give dramatic improvements. If an atom has localised d- or p-states, then in any event it will require a lot of PW's. Several recipes have been proposed to get an 'optimally smooth' PP (for example by acting on the form of pseudowavefunctions in the inner, unphysical region). A very simple and effective recipe is described in Ref.[58].

A more radical solution has been proposed by Vanderbilt[45]. His PP's are quite different from traditional PP's: they are definitely much softer, but also much less straightforward to use. The first interesting application to systems containing Oxygen and Copper have already appeared[59].

References

[1] See e.g. *Theory of the Inhomogeneous Electron Gas*, edited by S. Lundqvist and N. H. March (Plenum, New York, 1983); *Density Functional Theory of Atoms and Molecules*, R.G. Parr and W. Yang (Oxford University Press, New York, 1989); R.M. Dreizler and E.K.U. Gross, *Density Functional Theory*, Springer-Verlag, Berlin (1990).

[2] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[3] W. Kohn and L.J. Sham, Phys. Rev. **140**, A1133 (1965).

[4] E.P. Wigner, Trans. Faraday Soc. **34**, 678 (1938).

[5] D.M. Ceperley and B.J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[6] J. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[7] G. Ortiz and P. Ballone, Europhys. Lett. **23**, 7 (1993); G. Ortiz and P. Ballone, to be published (1994).

[8] R.O. Jones and O. Gunnarson, Rev. Mod. Phys. **61**, 689 (1989).

[9] O.K. Andersen, O. Jepsen, and M. Sob, in: *Electronic Band Structure and Its Applications*, edited by M. Yussouf (Springer, Berlin 1987), p. 1.

[10] W.E. Pickett, Computer Phys. Reports **9**, 115 (1989).

[11] A.D. Becke, Phys. Rev. A **38**, 3098 (1988).

[12] H. Hellmann, *Einführung in die Quantenchemie* (Deuticke, Leipzig, 1937); R.P. Feynman, Phys. Rev. **56**, 340 (1939).

[13] I. Štich, R. Car, M. Parrinello, and S. Baroni, Phys. Rev. B **39**, 4997 (1989).

[14] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes*, 2nd ed., Cambridge University Press (1991)

[15] M.C. Payne, M.P. Teter, D.C. Allen, T.A. Arias, and J.D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[16] J. Hutter, H.P. Lüthi, and M. Parrinello, Comput. Mat. Sci., in press (1994).

[17] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[18] G. Galli and A. Pasquarello, in *Computer Simulation in Chemical Physics*, edited by M.P. Allen and D.J. Tildesley (Kluwer, Amsterdam, 1993), p. 261.

[19] P. Pulay, Mol. Phys. **17**, 197 (1969).

[20] R. Yu, D. Singh, and H. Krakauer, Phys. Rev. B **43**, 6411, (1991); M. Methfessel and M. van Schilfgaarde, Phys. Rev. B **48**, 4937 (1993).

[21] E. Fermi, Nuovo Cimento **11**, 157 (1934).

[22] See the papers in *Solid State Physics*, edited by H.E. Ehrenreich, F. Seitz, and D. Turnbull, vol.24 (Academic Press, New York, 1970).

[23] D.R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).

[24] G.B. Bachelet, D.R. Hamann and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).

[25] A subtle point about the validity of frozen-core approximation is discussed in U. von Barth and C.G. Gelatt, Phys. Rev. B **21**, 2222 (1980).

[26] A. Baldereschi, Phys. Rev. B **7**, 5212 (1973).

[27] D.J. Chadi and M.L. Cohen, Phys. Rev. B **8**, 5747 (1973); H.J. Monkhorst and J.D. Pack, Phys. Rev. B **13**, 5188 (1976).

[28] Two recent references on this subjects: M. Methfessel and A.T. Paxton, Phys. Rev. B **40**, 3616 (1989); J. Hama, M. Watanabe, and T. Kato, J. Phys.: Condens. Matter **2**, 7445 (1990).

[29] D.G. Anderson, J. Assoc. Comput. Mach. **12**, 547 (1965).

[30] D.D. Johnson, Phys. Rev. B **38**, 12807 (1988).

[31] For an introduction: W.L. Briggs, *A Multigrid Tutorial*, SIAM, Philadelphia (1987).

[32] S. Baroni and M. Buongiorno Nardelli, unpublished.

[33] E. Anderson et al, *LAPACK Users' Guide*, SIAM (Philadelphia, 1992).

[34] For a recent review, see E.R. Davidson, Computer Phys. Commun. **53**, 49 (1989).

[35] N. Troullier and J.L. Martins, Phys. Rev. B **43**, 8861 (1991).

[36] R.D. King-Smith, M.C. Payne, and J.S. Lin, Phys. Rev. B **44**, 13063 (1991).

[37] J. Friedel, Adv. Phys. **3**, 446 (1954); F.J. Dyson, unpublished, as quoted in: C. Kittel, *Quantum theory of Solids* (Wiley, New York, 1963), p. 339.

[38] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991); W. Yang, Phys. Rev. A **44**, 7823 (1991).

[39] G.Galli and M. Parrinello, Phys. Rev. Lett **69**, 3547 (1992); F. Mauri, G. Galli, and R. Car, Phys. Rev. B **47**, 9973 (1993); F. Mauri and G. Galli, Phys. Rev. B, in press (1994); P. Ordejón, D.A. Drabold, M.P. Grumbach, and R.M. Martin, Phys. Rev. B **48**, 14646 (1993).

[40] X.-P. Li, R.W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993); S. Goedecker, preprint; M. Daw (unpublished).

[41] D.A. Drabold and O.F. Sankey, Phys. Rev. Lett. **70**, 3631 (1993).

[42] S. Baroni and P. Giannozzi, Europhys. Lett. **17**, 547 (1992).

[43] For a review, see: *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull (Academic, New York, 1980), Vol. 35.

[44] A. Franceschetti and S. Baroni, unpublished; A. Franceschetti, Ph.D. Thesis, SISSA-Trieste, 1993 (unpublished).

[45] D. Vanderbilt, Phys. Rev B **41**, 7892 (1990).

[46] F. Gygi, Europhys. Lett. **19**, 617 (1992).

[47] K. Cho, A. Arias, J.D. Joannopoulos, and P.K. Lam, Phys. Rev. Lett. **71**, 1808 (1993).

[48] I. Štich, M.C. Payne, R.D. King-Smith, J.-S. Lin, and L.J. Clarke, Phys. Rev. Lett. **68**, 1351 (1992); K. Brommer, M. Needels, B. Larson, and J.D. Joannopoulos, Phys. Rev. Lett. **68**, 1355 (1992).

[49] G. Kerker, J. Phys. C **13**, L189 (1980).

[50] U. von Barth and R. Car, unpublished.

[51] S.G. Louie, S. Froyen, and M.L. Cohen, Phys. Rev. B **26**, 1738 (1982).

[52] D.R. Hamann, Phys. Rev. B **40**, 2980 (1989).

[53] E.L. Shirley, D.C. Allan, R.M. Martin, and J.D. Joannopoulos, Phys. Rev. B **40**, 3652 (1989).

[54] G. Ortiz and P. Ballone, Phys. Rev. B **43**, 6376 (1991).

[55] E.L. Shirley and R.M. Martin, Phys. Rev. B **47**, 15413 (1993).

[56] L. Kleinman and D.M. Bylander, Phys. Rev. Lett. 48, 1425 (1982).

[57] X. Gonze, P. Kaeckell, and M. Scheffler, Phys. Rev. B **41**, 12264 (1990); X. Gonze, R. Stumpf, and M. Scheffler, Phys. Rev. B **44**, 8503 (1991).

[58] N. Troullier and J.L. Martins, Phys. Rev. B **43**, 1993 (1991).

[59] K. Laasonen, A. Pasquarello, R. Car, C. Lee, and D. Vanderbilt, Phys. Rev. B **47**, 10142 (1993).