The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2137-3

Joint ICTP-IAEA Advanced Workshop on Multi-Scale Modelling for Characterization and Basic Understanding of Radiation Damage Mechanisms in Materials

12 - 23 April 2010

Elements of first-principles electronic structure calculations

J.J. Kohanoff
The Queen's University of Belfast
U.K.

Elements of first-principles electronic structure calculations

Jorge Kohanoff

Atomistic Simulation Centre,
School of Mathematics and Physics,
Queen's University Belfast,
Belfast BT7 1NN,
Northern Ireland

November 13, 2008

General Bibliography

- J. Kohanoff, *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods* (Cambridge University Press, 2006).
- R. M. Martin, *Electronic Structure: Basic Theory and Prectical Methods* (Cambridge University Press, 2003).

Contents

Chapter 1

The problem of the structure of matter

The microscopic description of the physical and chemical properties of a material system is a complex problem. In general, we have a collection of atoms interacting with forces that derive from some potential field. This ensamble of particles may be isolated (molecules and clusters), extended (solids, surfaces, wires, and liquids), or a combination of both (molecules in solution). However, in all cases we can unambiguously describe the system by a number of nuclei and electrons interacting through Coulombic (electrostatic) forces. Formally, we can write the hamiltonian of such a system in the following general form:

$$\begin{aligned} \hat{H} &= -\sum_{I=1}^{P} \frac{\hbar^2}{2M_I}\nabla_I^2 - \sum_{i=1}^{N} \frac{\hbar^2}{2m}\nabla_i^2 + \frac{e^2}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P} \frac{Z_I Z_J}{\mid R_I - R_J \mid} + \qquad (1.1) \\ &+ \frac{e^2}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{1}{\mid r_i - r_j \mid} - e^2 \sum_{I=1}^{P}\sum_{i=1}^{N} \frac{Z_I}{\mid R_I - r_i \mid} \qquad (1.2) \end{aligned}$$

where $R = \{R_I\}$, $I = 1...P$, is a set of P nuclear coordinates, and $r = \{r_i\}$, $i = 1...N$, is a set of N electronic coordinates. Z_I and M_I are the P nuclear charges and masses, respectively. Electrons are fermions, so that the total electronic wave function must be antisymmetric with respect to exchange of two electrons. Nuclei can be fermions, bosons or distinguishable particles, according to the particular problem under examination. All the ingredients are perfectly known and, in principle, all the properties can be derived by solving the following Schrödinger equation:

$$\hat{H}\ \Psi_i(r,R) = E_i\ \Psi_i(r,R) \qquad (1.3)$$

In practice, this problem is almost impossible to treat in a full quantum mechanical framework. Only in a few cases a complete analytic solution is available, and numerical solutions are also limited to a very small number of particles. There are several features that contribute to this difficulty. First, this is a multicomponent many-body system, where each component (each nuclear species and the electrons) obbey a particular statistics. Moreover, the complete wave function cannot be easily factorized because of Coulombic correlations. In other words, the full Schrödinger equation cannot be easily decoupled into a set of equations so that, in general, we have to deal with $(3P + 3N)$ coupled degrees of freedom. The dynamics is an even more difficult problem, and very few and limited numerical techniques have been devised to solve it. The usual choice is to resort to some sensible approximations.

1.1 Adiabatic approximation (Born-Oppenheimer)

The first observation is that the time scale associated to the motion of the nuclei is usually much slower than that associated to electrons. In fact, the small mass of the electrons as compared to that of the protons (the most unfavoralbe case) is about 1 in 2000, meaning that their velocity is much larger. In this spirit, it was proposed in the early times of quantum mechanics that the electrons can be adequately described as following instantaneously the motion of the nuclei, staying always in the same stationary state of the electronic hamiltonian [1]. This stationary state will vary in time because of the Coulombic coupling of the two sets of degrees of freedom but, if the electrons were, *e.g.* in the (many electron) ground state, they will remain there forever. In other words, as the nuclei follow their dynamics, the electrons instantaneously adjust their wave function according to the nuclear wave function. This approximation ignores the possibility of having non-radiative transitions between different electronic eigenstates. Transitions can only arise through the coupling with an electromagnetic field, and these will not be addressed in the following.

All this can be cast in a formal mathematical framework by proposing a solution to Eq. (1.1) of the following form:

$$\Psi(R,r,t) = \sum_n \Theta_n(R,t)\Phi_n(R,r) \tag{1.4}$$

where $\Phi_n(R,r)$ are the eigenstates of the electronic hamiltonian

$$\hat{h}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{ne} = \hat{H} - \hat{T}_n - \hat{V}_{nn} \tag{1.5}$$

i.e.

$$\hat{h}_e \Phi_n(R,r) = \varepsilon_n(R)\Phi_n(R,r) \quad . \tag{1.6}$$

In this partial differential equation on the r variables, R enters as a parameter. This expansion, which is always mathematically possible, is called the expansion in the *adiabatic basis*, because $\Phi_n(R,r)$ are solutions of the time-independent electronic Schrödinger equation, corresponding to a particular nuclear configuration. Eq. (1.6) has to be solved for all nuclear configurations R where the nuclear wave function is non-zero.

By replacing the above ansatz into the full Schrödinger equation we obtain:

$$\begin{aligned}\left[i\hbar\frac{\partial}{\partial t} + \sum_{I=1}^{P}\frac{\hbar^2}{2M_I}\nabla_I^2 - \varepsilon_q(R)\right]\Theta_q(R,t) = \sum_n\sum_{I=1}^{P}\frac{\hbar^2}{2M_I}\ \left\langle \Phi_q \left| \nabla_I^2 \right| \Phi_n \right\rangle\ \Theta_n(R,t) + \\ +2\sum_n\sum_{I=1}^{P}\frac{\hbar^2}{2M_I}\ \vec{\nabla}_I\Theta_n(R,t)\cdot\left\langle \Phi_q \left| \vec{\nabla}_I \right| \Phi_n \right\rangle\end{aligned} \tag{1.7}$$

which constitutes a set (infinite, in principle) of coupled partial differential equations containing off-diagonal terms.

Therefore, the reduction of the full wave function to an expression of the type

$$\Psi(R,r,t) = \Theta_n(R,t)\ \Phi_n(R,r) \tag{1.8}$$

is not completely correct because, even if the system was initially *prepared* in a *pure* state like the above one, the off-diagonal terms will mix (excite) the different electronic eigenstates along the temporal evolution. These are precisely the non-radiative transitions alluded above. If this is the case, then the dynamics is said to be *non-adiabatic*. However, if the off-diagonal terms can be neglected, then an expression like (1.8) is valid because the nuclear dynamics has no means to cause electronic transitions, and the electrons remain always in the same (n) adiabatic state (ground or excited). In this case, the dynamics is said to be *adiabatic*.

When is it possible to neglect the non-adiabatic couplings ? The condition is that

$$\left| \sum_{I=1}^{P} \frac{\hbar^2}{M_I} \left\langle \Theta_q \left| \vec{\nabla}_I \right| \Theta_n \right\rangle \cdot \left\langle \Phi_q \left| \vec{\nabla}_I \right| \Phi_n \right\rangle \right| \ll |\varepsilon_q(R) - \varepsilon_n(R)| \tag{1.9}$$

or, equivalentely,

$$\frac{m}{M} \left| \frac{\hbar\ \Omega_v}{\varepsilon_q(R) - \varepsilon_n(R)} \right| \ll 1 \tag{1.10}$$

where Ω_v is the maximum frequency of rotation of the electronic wave function due to the nuclear motion, and the energies in the denominator correspond to the electronic adiabatic eigenstates (the energy gap if $q = 1$ and $n = 0$). Notice that the mass ratio m/M is always smaller than 5×10^{-4}, thus justifying the adiabatic approximation unless a very small gap occurs [2]. Typical electronic excitations are of the order of 1 eV, while typical nuclear excitations (phonons) are of the order of 0.01 eV. This indicates that there is a clear separation of energy (and consequently time) scales.

There are situations in which this approximation is not adequate, but they are more the exception than the rule. In metallic systems, for instance, it can be argued that the adiabatic approximation breaks down because the energy gap is zero, and electronic excitations of vanishing energy are possible. However, since typical temperatures (at most a few thousand degrees) are usually much smaller than the electronic Fermi temperature, then the width of the region where the electronic Fermi-Dirac distribution varies (*i.e.* region associated with transport porperties) is very small. This means that the excitations are confined to a narrow region around the Fermi surface, and that most properties are not affected by neglecting non-adiabatic contributions due to these few electrons. The usual treatment of transport phenomena in metals is to start from the adiabatic description, and to introduce non-adiabatic terms (in the form of electron-phonon interactions) afterwards, perturbatively. In terms of the ratio of energy scales, it can also be realized that the relevant excitations in metals at small wave numbers are not electron-hole pairs which, besides being very few, carry a small oscillator strength. The relevant energy scale is dictated by the plasmon (collective charge excitation), which is again typically of the order of a few eV.

1.2 Classical nuclei approximation

Therefore, according to the adiabatic approximation, the total wave function can be written in the form of expression (1.8), where $\Theta_n(R,t)$ is the nuclear wave function. At room temperature the thermal wavelength is $\lambda_T = (e^2/k_B T) \approx 0.1$ Å, so that regions of space separated by more than λ_T do not present quantum phase coherence. Interatomic distances are normally of the order of 1 Å and, then, the total nuclear wave function can be considered as an incoherent superposition of individual nuclear wave packets:

$$\Theta_n(R,t) = \Pi_I \, \Theta_n^{(I)}(R_I, R_I^{(c)}(t), t) \tag{1.11}$$

where $R_I^{(c)}(t)$ are the centers of the individual wave packets. The details of the decoherence process, i.e. the quantum-to-classical transition, are a field on its own, and are still matter of debate [3]. This approximation receives the name of time-dependent Hartree approximation, and implies that there are no quantum correlations between the wave functions of the different nuclei. Exchange effects are also absent in this expression, although they could be recovered by proposing a total wave function in the form of a Slater determinant (for odd-spin nuclei), thus leading to the so-called time-dependent Hartree-Fock approximation. However, atomic nuclei exhibit exchange effects only at very low temperatures, e.g. below 5 K in the case of H. On the other hand, nuclear masses are typically large enough that the individual nuclear wave packets are quite localized, provided that the curvature of the potential felt is sufficiently strong. For instance, a proton in a typical molecular bonding environment has a width of about 0.25 Å. The combination of these two observations allows us to propose that, in most cases, atomic nuclei can be treated as *classical* particles.

The time-dependent (adiabatic) Schrödinger equation

$$i\hbar \frac{\partial \Theta_n(R,t)}{\partial t} = \left(-\sum_{I=1}^{P} \frac{\hbar^2}{2M_I} \nabla_I^2 + \varepsilon_n(R) \right) \Theta_n(R,t) \tag{1.12}$$

implies, through Ehrenfest theorem, the following evolution equation for the mean values of the position and momentum operators:

$$\begin{aligned} i\hbar \frac{d\langle R \rangle}{dt} &= \langle [H,R] \rangle = i\hbar \frac{\langle P \rangle}{M} \Longrightarrow M \frac{d\langle R \rangle}{dt} = \langle P \rangle \qquad (1.13) \\ i\hbar \frac{d\langle P \rangle}{dt} &= \langle [H,P] \rangle = -i\hbar \langle \nabla \varepsilon_n(R) \rangle \end{aligned}$$

which combined give rise to the following Newtonian equation of motion:

$$M \frac{d^2 \langle R \rangle}{dt^2} = -\langle \nabla \varepsilon_n(R) \rangle \tag{1.14}$$

In principle, this expression is valid only for the mean value of the position operator. The *classical nuclei approximation* consists of identifying this mean value with the cartesian coordinates of the classical particle. This can be easily understood if the nuclear wave function is represented as a product of Dirac's delta functions whose centers are located at the classical positions. In that case, $\langle R \rangle = R_{cl}(t)$ and $\langle \nabla \varepsilon_n(R) \rangle = \nabla \varepsilon_n(R_{cl})$. This latter is strictly valid only for δ-functions or for harmonic potentials. In the general case, the leading error of this approximation is proportional to the anharmonicity of the potential and to the spatial extension of the wave packet. The expression for the classical equation of motion becomes, then:

$$M \frac{d^2 R_{cl}(t)}{dt^2} = -\nabla \varepsilon_n(R_{cl}) \tag{1.15}$$

where $\varepsilon_n(R_{cl})$ is the n^{th} adiabatic *potential energy surface* (PES), and there is one equation of motion for each different PES.

The final expression for the equation of motion is achieved by using Hellmann-Feynman theorem [4]:

$$\frac{\partial \varepsilon_n(\lambda)}{\partial \lambda} = \left\langle \Phi_n(R) \left| \frac{\partial \hat{h}_e(\lambda)}{\partial \lambda} \right| \Phi_n(R) \right\rangle \tag{1.16}$$

and therefore

$$M_I \frac{d^2 R_I^{(n)}(t)}{dt^2} = -\left\langle \Phi_n(R^{(n)}) \left| \frac{\partial \hat{h}_e(R^{(n)})}{\partial R_I^{(n)}} \right| \Phi_n(R^{(n)}) \right\rangle - \frac{\partial V_{nn}(R^{(n)})}{\partial R_I^{(n)}} \tag{1.17}$$

where

$$\hat{h}_e(R, r) = -\sum_{i=1}^{N} \frac{\hbar^2}{2m} \nabla_i^2 + \frac{e^2}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{1}{\mid r_i - r_j \mid} - e^2 \sum_{I=1}^{P} \sum_{i=1}^{N} \frac{Z_I}{\mid R_I - r_i \mid} \tag{1.18}$$

and

$$V_{nn}(R) = \frac{e^2}{2} \sum_{I=1}^{P} \sum_{J \neq I}^{P} \frac{Z_I Z_J}{\mid R_I - R_J \mid} \,. \tag{1.19}$$

The numerical integration of the above Newtonian equation of motion receives the name of *first-principles Molecular Dynamics,* and $\varepsilon_n(R)$ is the first-principles potential. The solution of the stationary problem, i.e. $\nabla \varepsilon_n(R) = 0$ is also an important issue, usually called *geometry optimization.* In either case, in order to obtain $\varepsilon_n(R)$ and its gradient it is necessary to solve the time-independent electronic Schrödinger equation (1.6). This is a field on its own, and has traditionally received the name of *electronic structure.*

Chapter 2

The electronic problem

This is, then, the key problem in the structure of matter: to solve the Schrödinger equation for a system of N interacting electrons in the external Coulombic field created by a collection of atomic nuclei (and may be some other external field, *e.g.* electromagnetic). This is a very difficult problem in *many-body* theory and, in fact, the exact solution is known only in the case of the homogeneous electron gas, for atoms with a small number of electrons, and for a few small molecules. These exact solutions are always numerical. At the analytical level, one always has to resort to approximations. However, the effort of devising schemes to solve this problem is really worth, because the knowledge of the electronic ground state of a system gives access to many of its properties, *e.g.* relative stability of different structures, equilibrium structural information, mechanical stability and elastic properties, pressure-temperature (P-T) phase diagrams, dielectric properties, lattice vibrations and spectral functions, (non-electronic) transport properties like diffusivity, viscosity, ionic conductivity, etc. Excited electronic states give also access to another wealth of measurable phenomena like electronic transport and optical properties.

2.1 The physical origin of many-body effects

As early as in 1903-1913, Gouy and Chapman considered the problem of the variation in the electronic charge distribution at the electrodes of a battery, upon varying the potential difference between the electrodes (see Fig. 2.1). This was one the first many-body problems addressed in the literature. Two main concepts were identified in this context. One is the *screening length*, which is a measure of the distance at which the charge at the electrode, is counteracted – or *screened* – by charges of the opposite sign in the electrolytic solution that are attracted towards the electrode. The second is the *plasma frequency*, which measures the frequency of colective charge oscillations appearing due to the restoring force generated by the displaced charge density. The theoretical solution to this problem, which is essentially the electronic many-body problem, *i.e.* an electron interacting with the electric field of other electrons, has been discussed by Debye and Hückel in 1923.

Consider a system of interacting electrons and write Poisson's equation for the potential generated by the charge at the origin plus the charge distribution $n(\mathbf{r})$, in a uniform

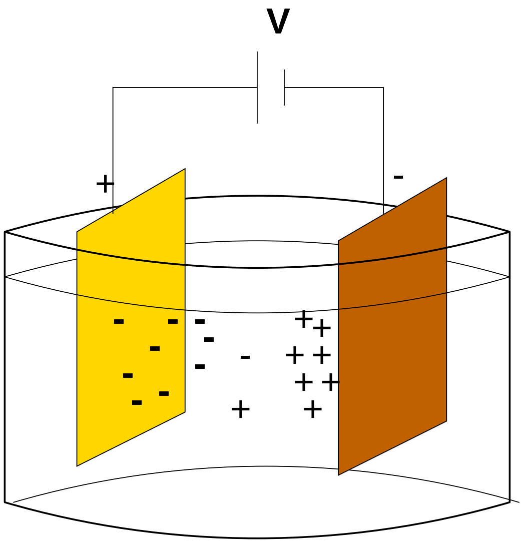

Figure 2.1: Schematics of an electrochemical cell.

positive background:

$$\nabla^2 v_H(\mathbf{r}) = -4\pi\{-e\delta(\mathbf{r}) - en(\mathbf{r}) + en\} \ , \tag{2.1}$$

where n is the average charge density.

Defining the electron-electron pair distribution (or correlation) function $g(\mathbf{r}) = n(\mathbf{r})/n$, it can be seen that $g(\mathbf{r})$ represents the probability of finding an electron at $\mathbf{r}$ given that there is an electron at the origin. Clearly, the presence of this electron *discourages* the other electrons to approach the origin, because of the Coulomb repulsion. Typically, the pair distribution function will interpolate from zero at the origin to 1 at infinity, as schematically shown in Fig. 2.3. This is the physical meaning of what is usually called *correlation*.

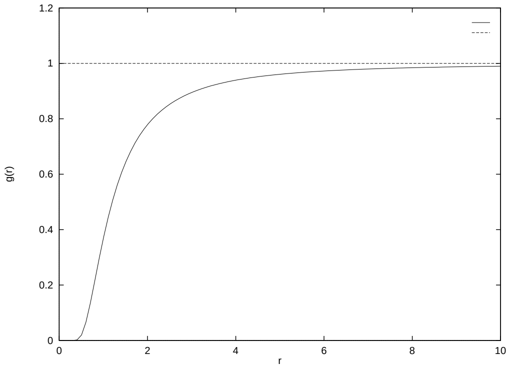

Figure 2.2: Schematic pair correlation function.

In terms of $g(\mathbf{r})$ Poisson's equation (for the potential energy V_H) reads:

$$\nabla^2 V_H(\mathbf{r}) = -4\pi e^2 \{\delta(\mathbf{r}) + n[g(\mathbf{r}) - 1]\} \ . \tag{2.2}$$

The essence of the many-body problem consists of finding an appropriate expression for the pair correlation function. Notice that this is a general physical concept, which is not confined to quantum systems. For instance, in systems of classical charged particles (imagine classical macroions), the probability of finding two particles too close is much smaller than if they are far away. The quantum character will introduce additional elements into the pair distribution function.

Coming back to classical ions, $g(\mathbf{r})$ can be calculated at different levels of approximation. The simplest possibility is to consider that the probability of finding an electron at $\mathbf{r}$ given that there is another electron at the origin, is given by Boltzmann's thermal distribution

$$g(\mathbf{r}) \approx \exp\left(\frac{-V_H(\mathbf{r})}{k_B T}\right) \ . \tag{2.3}$$

This is the so-called Poisson-Boltzmann approximation, which is widely used for describing the electrostatics of classical fluids. Linearization of the above expression is possible as long as the exponent is small, what is not valid at small r. However, at reasonably large distances the linearized Poisson-Boltzmann equation reads:

$$\nabla^2 V_H(\mathbf{r}) = -4\pi e^2 \delta(\mathbf{r}) + \left(\frac{4\pi n e^2}{k_B T}\right) V_H(\mathbf{r}) \ , \tag{2.4}$$

whose analytic solution is a potential of the Yukawa form

$$V_H(r) = \frac{e^2}{r} \exp(-r/l_{DH}) \ , \tag{2.5}$$

where $l_{DH} = \sqrt{k_B T / 4\pi n e^2}$ is the Debye-Hückel screening length. This approximation is sometimes also called Debye-Hückel, or linear screening.

The expression for the screening length arises as a consequence of Boltzmann's approximation. Let us now assume that linearization is still possible, but that the pair correlation function is not necessarily given by the above expression. In that case we can use the equilibrium condition that the electrostatic potential is compensated by the chemical potential, to obtain the following expression for the screening length:

$$l_s = \sqrt{\frac{(\partial\mu/\partial n)_T}{4\pi e^2}} \ , \tag{2.6}$$

which for an ideal Fermi gas ($\mu = E_F = \alpha n^{2/3}$) becomes $l_{TF} = \sqrt{E_F / 6\pi n e^2}$, and receives the name of Thomas-Fermi screening length. This is the ideal Fermi gas version of linear screening, and it is a reasonably good approximation for describing simple metals.

However, linear screening is not all, and we would like to extend this theory to more general situations. We consider then a system of interacting electrons that verify the Schrödinger equation

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + v_{ext}(\mathbf{r}) + v_H(\mathbf{r}) + v_{scr}(\mathbf{r})\right\}\psi_k(\mathbf{r}) = \varepsilon_k\ \psi_k(\mathbf{r}) \tag{2.7}$$

where v_H is the Hartree potential satisfying Poisson's equation (2.1), and v_{scr} is a screening potential which takes into account the fact that our test electron is just one more amongst the other electrons instead of being an infinitesimal perturbation. In other words, what happens is that the presence of this electron modifies (or displaces) the charge density in a non trivial way. The screening potential is then due to the displaced electronic charge, and this is intimately related to the pair correlation function.

By transforming the above into a scattering-like integral equation

$$\psi_k(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} + \frac{2m}{\hbar^2}\int\left[-\frac{1}{4\pi}\frac{e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')}}{|\mathbf{r}-\mathbf{r}'|}\right] v_H(\mathbf{r}')\ \psi_k(\mathbf{r}')\ d\mathbf{r}'\ , \tag{2.8}$$

and constructing the charge density for the Fermi system as

$$n(\mathbf{r}) = \sum_{k\leq k_F} |\psi_k(\mathbf{r})|^2\ , \tag{2.9}$$

where k_F is the Fermi momentum, it is possible to find the expression for the displaced charge density. It is interesting to remark that in the above integral equation the wavefunction appears in the integrand. This can be viewed in the following way: the test electron (the impurity) is screened by the other electrons. This screening affects the scattering between electrons and impurity, which in turn affects again the screening, and so on until scattering and screening are consistent with each other in the sense that *input* and *output* wavefunctions are the same. This property is called *self-consistency.* Solving this self-consistent problem is the main goal of many-body theory, and is equivalent to finding the pair correlation function $g(\mathbf{r})$.

The solution of this problem is extremely difficult, as it involves the calculation of an infinite series of corrections of higher and higher order to the scattering problem. However, approximations are aboundant. For instance, by replacing the wave function in the integrand with $\psi_k(\mathbf{r}') = \exp(i\mathbf{k}\cdot\mathbf{r}')$ the displaced charge density becomes:

$$\delta n(\mathbf{r}) = -\frac{mk_F^2}{2\pi^3\hbar^2}\int v_H(\mathbf{r}')\ \frac{j_1(2k_F|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|^2}\ d\mathbf{r}'\ , \tag{2.10}$$

where $j_1(x)$ is the spherical Bessel function of order 1.

This equation can be solved in conjunction with Poisson's equation by transforming both to reciprocal space. The resulting Hartree potential can be written in the following way: $v_H(k) = 4\pi e^2/k^2\varepsilon(k)$, where $\varepsilon(k)$ is the dielectric function which relates the screened and the bare potential (or the electric field and the polarization). This is called the Random Phase Approximation (RPA), or sometimes also Lindhard approximation. The corresponding dielectric function is:

$$\varepsilon_{RPA}(k) = 1 + \frac{2mk_Fe^2}{\pi\hbar^2k^2}\left\{1 + \frac{k_F}{k}\left(\frac{k^2}{4k_F^2} - 1\right)\ln\left|\frac{k-2k_F}{k+2k_F}\right|\right\}\ . \tag{2.11}$$

There are two interesting limits of equation (2.11). One is for small values of k, where $\varepsilon_{RPA}(k) \rightarrow 1+k_F^2/k^2$, and therefore v_H, when transformed back to real space, corresponds to the result of Thomas-Fermi or linear screening theory. The other is for $k \rightarrow 2k_F$, where Lindhard's function exhibits a logarithmic singularity. While the small k limit gives a short-range contribution of the Yukawa type (exponentially decreasing), this singularity gives rise to a long-range (large r) contribution of the type $v_H(r) \rightarrow -2Ae^2\cos(2k_F r)/r^3$. This damped oscillatory behavior of spatial frequency $2k_F$ receives the name of Friedel oscillations, and is very important in metals.

In the above we have not taken into account the exchange interaction between electrons. The particles were, however, treated as fermions for constructing the electronic density. In addition, dynamical fluctuations were completely ignored, and we have treated only the static limit. Lindhard's theory can be extended to time-dependent phenomena in a fairly simple way. Now the dielectric function will have in addition a frequency dependency, and thus will provide information about electronic excitations. The main features of $\varepsilon_{RPA}(k,\omega)$ are the existence of a continuum of single-particle excitations (electron-hole pairs), and a collective excitation of frequency $\omega = \omega_p + \alpha k^2$, where $\omega_p = \sqrt{4\pi e^2 n/m}$ is the *plasma frequency* mentioned at the beginning of this section. Therefore, plasma oscillations are a dynamical many-body effect.

This is the very basic quantum many-body theory for electronic systems. In the following we are going to discuss different approaches to the problem of many electrons in the presence of an external electrostatic field. We first introduce quantum chemical approaches like Hartree-Fock, and then focus on density functional theory, describing the different approximations to exchange and correlation that have been proposed, including some recent developments.

2.2 Quantum many-body theory: chemical approaches

The simplest approximation may be considered the one proposed by Hartree (as early as in 1928, in the beginnings of the age of quantum mechanics) [5]. It consists of postulating that the *many-electron* wave function can be written as a product of *one-electron* wave functions. Each of these latter verifies a one-particle Schrödinger equation where the potential is actually an effective potential that takes into account the interaction with the other electrons in a mean field way:

$$\Phi(\mathbf{R},\mathbf{r}) = \Pi_i \; \varphi_i(\mathbf{R},\mathbf{r}_i) \tag{2.12}$$

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V_{eff}^{(i)}(\mathbf{R},\mathbf{r})\right)\varphi_i(\mathbf{R},\mathbf{r}) = \epsilon_i \; \varphi_i(\mathbf{R},\mathbf{r}) \tag{2.13}$$

with

$$V_{eff}^{(i)}(\mathbf{R},\mathbf{r}) = V_{ext}(\mathbf{R},\mathbf{r}) + \int \frac{\sum_{j\neq i}^{N} \rho_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}' \tag{2.14}$$

where

$$\rho_j(r) = |\varphi_j(\mathbf{R},\mathbf{r})|^2 \tag{2.15}$$

is the electronic density associated with particle j. The second term in the RHS of (2.14) is the *classical* electrostatic potential generated by the charge distribution $\sum_{j\neq i}^{N}\rho_j(r)$. Notice that this charge density does not include the charge associated with particle i, so that the Hartree approximation is (correctly) self-interaction free. In this approximation, the energy of the many-body system is not just the sum of the eigenvalues of Eq. (2.13) because the formulation in terms of an effective potential makes the electron-electron interaction to be counted twice. The correct expression for the energy is:

$$E_H = \sum_{n=1}^{N} \epsilon_n - \frac{1}{2}\int\int \frac{\rho(\mathbf{r})\ \rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\ d\mathbf{r}' \quad . \tag{2.16}$$

The set of N coupled partial differential equations (2.13) can be solved by minimizing the energy with respect to some variational parameters in a trial wave function or, alternatively, by re-calculating the electronic densities with the solutions of (2.13) as in (2.15), casting them back into the expression for the effective potential (2.14), and solving again the Schrödinger equation. This procedure is repeated several times, until self-consistency in the input and output wave function (or potential) is achieved. This procedure is called *self-consistent Hartree* (or self-consistent field, or simply SCF) approximation.

The Hartree approximation treats the electrons as distinguishable particles. A step forward is to introduce Pauli exclusion principle (Fermi statistics for electrons) by proposing a many-electron wave function in the form of a Slater determinant:

$$\Phi(\mathbf{R},\mathbf{r}) = SD\left\{\varphi_j(\mathbf{R},\mathbf{r}_i)\right\} \ = \frac{1}{\sqrt{N\ !}}\begin{vmatrix} \varphi_1(\mathbf{R},\mathbf{r}_1) & \varphi_1(\mathbf{R},\mathbf{r}_2) & \cdots & \varphi_1(\mathbf{R},\mathbf{r}_N) \\ \varphi_2(\mathbf{R},\mathbf{r}_1) & \varphi_2(\mathbf{R},\mathbf{r}_2) & \cdots & \varphi_2(\mathbf{R},\mathbf{r}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_N(\mathbf{R},\mathbf{r}_1) & \varphi_j(\mathbf{R},\mathbf{r}_2) & \cdots & \varphi_N(\mathbf{R},\mathbf{r}_N) \end{vmatrix} \tag{2.17}$$

This wave function introduces particle exchange in an exact manner [6, 7]. The approximation is called *Hartree-Fock* (HF), and has been for a long time the way of choice of chemists for calculating the electronic structure of molecules. In fact, it provides a very reasonable picture for atomic systems and, although many-body correlations (arising from the fact that, due to the two-body Coulomb interactions, the total wave function cannot necessarily be separated as a sum of products of single-particle wave functions) are completely absent, it also provides a reasonably good description of interatomic bonding. Hartree-Fock equations look similar to Hartree equations — notice that also in HF the self-interaction cancels exactly because $\rho_j(r) = \varphi_j^*(r)\varphi_j(r)$ —, except for the fact that the exchange integrals introduce additional coupling terms in the differential equations:

$$\begin{aligned} &\left(-\frac{\hbar^2}{2m}\nabla^2 + V_{ext}(\mathbf{R},\mathbf{r}) + \int \frac{\sum_{j=1}^{N}\rho_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'\right)\varphi_i(\mathbf{R},\mathbf{r}) - \\ - &\sum_{j=1}^{N}\left(\int \frac{\varphi_j^*(\mathbf{r}')\varphi_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathbf{d}\mathbf{r}'\right)\varphi_j(\mathbf{R},\mathbf{r}) = \sum_{j=1}^{N}\lambda_{ij}\,\varphi_j(\mathbf{R},\mathbf{r}) \end{aligned} \tag{2.18}$$

which is typically solved by a canonical transformation of the orbitals such that the matrix of Lagrange multipliers λ_{ij} becomes diagonal, i.e. $\mathcal{H}_{HF}(\mathbf{R})\tilde{\varphi}_i(\mathbf{R},\mathbf{r}) = \varepsilon_i\tilde{\varphi}_i(\mathbf{R},\mathbf{r})$

Nowadays, the HF approximation is routinely used as a starting point for more elaborated calculations like Møller-Plesset perturbation theory of second (MP2) or fourth (MP4) order, or by configuration interaction methods (CI) that use a many-body wave function made of a linear combination of Slater determinants, as a means for introducing electronic correlations. Several CI schemes have been devised during the past 40 years, and this is still an active area of research. Coupled clusters (CC) and complete active space (CAS) methods are currently two of the most popular ones [8].

Parallel to the development of this line in electronic structure theory, Thomas and Fermi proposed, at about the same time as Hartree (1927-1928), that the full electronic density was the fundamental variable of the many-body problem, and derived a differential equation for the density without resorting to one-electron orbitals [9, 10]. The *Thomas-Fermi* approximation was actually too crude because it did not include exchange and correlation effects, and was also unable to sustain bound states because of the approximation used for the kinetic energy of the electrons. However, it set up the basis for the later development of *Density Functional Theory* (DFT), which has been the way of choice in electronic structure calculations in condensed matter physics during the past 20 years and, recently, it also became accepted by the quantum chemistry community because of its computational advantages compared to wave function methods [11].

2.3 Density Functional Theory

The total energy of an inhomogeneous system composed by N interacting electrons is given by the following expectation value:

$$E = \left\langle \Phi \left| \hat{T} + \hat{V}_{ext} + \hat{V}_{ee} \right| \Phi \right\rangle = \left\langle \Phi \left| \hat{T} \right| \Phi \right\rangle + \left\langle \Phi \left| \hat{V}_{ext} \right| \Phi \right\rangle + \left\langle \Phi \left| \hat{V}_{ee} \right| \Phi \right\rangle$$

where $\mid \Phi\rangle$ is the N-electron wave function, which has neither the form given by the Hartree approximation (2.12) nor the Hartree-Fock form (2.17). In fact, this wave function has to include correlations amongst electrons, and its general form is basically unknown. $\hat{T}$ is the kinetic energy operator, $\hat{V}_{ext}$ is the interaction with an external field, and $\hat{V}_{ee}$ is the electron-electron interaction. We are going to concentrate now on this latter, which is the one that introduces many-body effects.

$$V_{ee} = \left\langle \Phi \left| \hat{V}_{ee} \right| \Phi \right\rangle = \left\langle \Phi \left| \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{1}{|r_i - r_j|} \right| \Phi \right\rangle = \int \frac{\rho_2(r,r')}{|r - r'|}\, dr\, dr' \tag{2.19}$$

with

$$\rho_2(r,r') = \frac{1}{2} \left\langle \Phi \left| \Psi^\dagger(r)\, \Psi^\dagger(r')\, \Psi(r')\, \Psi(r) \right| \Phi \right\rangle \tag{2.20}$$

the two-body density matrix expressed in real space, being Ψ and $\Psi^\dagger$ the destruction and creation operators for electrons, which obbey the anticonmutation relations

$\left\{\Psi(r), \Psi^\dagger(r')\right\} = \delta(r - r')$. We define now the two-body direct correlation function $g(r, r')$ in the following way:

$$\rho_2(r, r') = \frac{1}{2}\rho_1(r, r)\,\rho_1(r', r')\,g(r, r') \tag{2.21}$$

where $\rho_1(r, r') = \left\langle \Phi \left| \Psi^\dagger(r)\ \Psi(r') \right| \Phi \right\rangle$ is the one-body density matrix (in real space), whose diagonal elements $\rho(r) = \rho_1(r, r)$ are simply the electronic density. With this definition, the electron-electron interaction is written:

$$V_{ee} = \frac{1}{2}\int \frac{\rho(r)\,\rho(r')}{|r - r'|}\,dr\,dr' + \frac{1}{2}\int \frac{\rho(r)\,\rho(r')}{|r - r'|}\,\left[g(r, r') - 1\right]dr\,dr' \ . \tag{2.22}$$

The first term is the classical electrostatic interaction energy corresponding to a charge distribution $\rho(r)$. The second term includes correlation effects of both, classical and quantum origin. Basically, $g(r, r')$ takes into account the fact that the presence of an electron at r excludes the possibility that a second electron comes to a position r' very close to r, because of the Coulomb repulsion. In other words, it says that the probability of finding two electrons (two particles with charges of the same sign, in the general case) is reduced with respect to the probability of finding them at infinite distance. This is so already at the classical level, and it is further modified at the quantum level. Purely quantum exchange further decreases this probability in the case of electrons having the same spin projection.

To understand the effect of exchange, let us imagine that we stand on an electron with spin ↑, and we look at the density of the other $(N - 1)$ electrons. Pauli principle forbids the presence of electrons with spin ↑ at the origin, but it says nothing about electrons with spin ↓, which can be at the origin without any inconvenient. Therefore:

$$g_X(r, r') \longrightarrow \frac{1}{2} \qquad \text{for} \qquad r \to r' \tag{2.23}$$

If we postulate a total wave function of the form of a Slater determinant, as in Hartree-Fock theory (2.17), we can rewrite the electron-electron interaction as

$$V_{ee} = \frac{1}{2}\int \frac{\rho(r)\,\rho(r')}{|r - r'|}\,dr\,dr' - \frac{1}{2}\int \frac{\rho(r)\,\rho(r')}{|r - r'|}\left[-\frac{\rho_1^2(r, r')}{2\rho(r)\,\rho(r')}\right]dr\,dr' \tag{2.24}$$

meaning that the exact expression for the exchange depletion (also called exchange hole) in the HF limit is:

$$g_X(r, r') = 1 - \frac{1}{2}\frac{\rho_1^2(r, r')}{\rho(r)\,\rho(r')} \tag{2.25}$$

The calculation of the correlation hole — $g_C(r, r')$ — is a major problem in many-body theory and, up to the present, it is an open problem in the general case of an inhomogeneous electron gas. The exact solution is know numerically, and also in some analytic derivations, for the homogeneous electron gas. There are several approximations that go beyond the homogeneous limit by including slowly varying densities through its spatial gradients (gradient corrections), and recently also expressions for the exchange-correlation energy that aim at taking into account very weak, nonlocal interactions of the van der Waals type [12].

We can say, then, that the energy of a many-body electronic system can be written in the following way:

$$E = T + V_{ext} + \frac{1}{2}\int \frac{\rho(r)\,\rho(r')}{|r-r'|}\,dr\,dr' + E_X + E_C \tag{2.26}$$

where

$$V_{ext} = \sum_{I=1}^{P} \left\langle \Phi \left| \sum_{i=1}^{N} v_{ext}(r_i - R_I) \right| \Phi \right\rangle = \sum_{I=1}^{P} \int \rho(r)\, v_{ext}(r - R_I)\, dr \quad , \tag{2.27}$$

$$T = \left\langle \Phi \left| -\frac{\hbar^2}{2m} \sum_{i=1}^{N} \nabla_i^2 \right| \Phi \right\rangle = -\frac{\hbar^2}{2m} \int \left[\nabla_r^2\, \rho_1(r, r') \right]_{r'=r} dr \quad . \tag{2.28}$$

and E_X and E_C are the exchange and correlation energies, respectively.

2.3.1 Thomas-Fermi theory

Thomas and Fermi (1927) gave a prescription for constructing the total energy in terms only of the electronic density [13]. They used locally the expression for the kinetic, exchange and correlation energies of the homogeneous electron gas to construct the same quantities for the inhomogeneous system in the following way $E_\alpha = \int \varepsilon_\alpha[\rho(r)]\,dr$, where $\varepsilon_\alpha[\rho(r)]$ is the energy density (corresponding to the piece α), calculated locally for the value of the density at that point in space. For the homogeneous electron gas we have

$$\rho = \frac{1}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \epsilon_F^{3/2} \tag{2.29}$$

where ϵ_F is the Fermi energy, or the kinetic energy of the electron that occupies the highest occupied eigenstate. The kinetic energy of the homogenous gas is $T = 3/5\,\rho\,\epsilon_F$, meaning that the kinetic energy density is:

$$t[\rho] = \frac{3}{5}\frac{\hbar^2}{2m}(3\pi^2)^{2/3}\rho^{5/3} \tag{2.30}$$

Then, the kinetic energy is written $T_{TF} = C_k \int \rho(r)^{5/3} dr$, with $C_k = 3(3\pi^2)^{2/3}/10 = 2.871$ atomic units. The inhomogeneous system is thought of as if it was locally homogeneous. This is the same approximation that has been used later in the framework of modern DFT, with the name of *local density approximation* (LDA), but here applied to the kinetic energy. Neglecting exchange and correlation in expression (2.26) we arrive to Thomas-Fermi theory:

$$E_{TF}[\rho] = C_k \int \rho(r)^{5/3} dr + \int v_{ext}(r)\,\rho(r)\,dr + \frac{1}{2}\int\int \frac{\rho(r)\,\rho(r')}{|r-r'|}\,dr\,dr' \tag{2.31}$$

It can be seen that E_{TF} depends only on the electronic density. This equation has to be solved by minimizing the energy *functional* with respect to $\rho(r)$, subjected to the constraint that the total integrated charge be equal to the number of electrons: $\int \rho(r)\, dr = N$. This can be put in terms of functional derivatives:

$$\frac{\delta}{\delta\rho(r)} \left(E_{TF}[\rho] - \mu \int \rho(r)\ dr \right) = 0 \tag{2.32}$$

or, equivalently,

$$\mu = \frac{5}{3} C_k\, \rho(r)^{2/3} + v_{ext}(r)\ + \int \frac{\rho(r')}{|r-r'|}\, dr \tag{2.33}$$

with μ the chemical potential. This equation can be inverted to obtain the density as a *unique* function of the external potential. This integral form is somewhat inconvenient, but it can be easily done by Fourier transforming the equation to obtain $\rho(g)$.

Exchange can be easily added to the expression above by considering Slater's expression for the homogeneous electron gas: $\varepsilon_X[\rho] = -C_X\, \rho^{4/3} dr$, with $C_X = 3(3/\pi)^{1/3}/4$, such that (2.33) is modified by adding the term $-(4/3) C_X\, \rho(r)^{1/3}$. This level of approximation is called *Thomas-Fermi-Dirac* theory.

Correlation can also be easily added by using any approximation to the homogeneous electron gas, for instance the one proposed by Wigner: $\varepsilon_C[\rho] = -0.056\, \rho^{4/3}/[0.079 + \rho^{1/3}]$.

This is the best that can be done at the *local* level. Additional corrections to the kinetic, exchange, and correlation energies due to nonlocality have been postulated in the form of gradient corrections, e.g. as given by the von Weiszäcker functional [14]:

$$T_{vW} = \frac{1}{8} \int \frac{|\nabla\rho|^2}{\rho}\, d\mathbf{r} \qquad . \tag{2.34}$$

Also terms that correct the linear response properties of the functional have been proposed [15, 16], and even the second order response functions have been incorporated to this approach [17]. This has been developed in the optics that an ultimate explicit expression for the energy in terms of the electronic density does really exist. But ... what guarantees us that the energy can be written as a functional purely dependent on the density ?

2.3.2 Hohenberg-Kohn theorem

In 1964, P. Hohenberg and W. Kohn [18] formulated and proved a theorem which put on a solid mathematical basis the former ideas, which were first proposed by Thomas and Fermi. The theorem is divided into two parts:

Theorem: The external potential is univocally determined by the electronic density, except for a trivial additive constant.

Proof: Let $\hat{H} = \hat{T} + \hat{U} + \hat{V}$, and $E_0 = \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle$ the minimum possible energy for this hamiltonian. Due to the variational principle, for any $\Psi' \neq \Psi$:

$$E_0 < \left\langle \Psi' \left| \hat{H} \right| \Psi' \right\rangle = \left\langle \Psi' \left| \hat{H}' \right| \Psi' \right\rangle + \left\langle \Psi' \left| \hat{H} - H' \right| \Psi' \right\rangle = E'_0 + \int \rho(r)\,(v(r) - v'(r))\,dr\ ,$$

where $\hat{T} + \hat{U}$ depends only on the electronic subsystem, and is the same for both H and H'. Now we can simply reverse the situation of Ψ and Ψ' (H and H'), and we readily obtain:

$$E'_0 < \left\langle \Psi \left| \hat{H}' \right| \Psi \right\rangle = \left\langle \Psi \left| \hat{H} \right| \Psi \right\rangle + \left\langle \Psi \left| \hat{H}' - H \right| \Psi \right\rangle = E_0 - \int \rho(r)\,(v(r) - v'(r))\,dr\ .$$

Therefore: $E_0 < E'_0 + A$ and, simultaneously, $E'_0 < E_0 - A$. Adding these two inequalities, it turns out that $E_0 + E'_0 < E'_0 + E_0$, which is absurd.
$\Longrightarrow$ There are no $v(r) \neq v'(r)$ that correspond to the same electronic density for the ground state.

Corollary: Since $\rho(r)$ univocally determines $v(r)$, then it also determines the ground state wave function Ψ_{GS}.

Theorem: Let $\tilde{\rho}(r)$ be a non-negative density normalized to N. Then: $E_0 < E_v[\tilde{\rho}]$, for

$$E_v[\rho] = T[\rho] + U[\rho] + \int \rho(r)\,v(r)\,dr$$

with

$$U[\rho] = \frac{1}{2} \int\int \frac{\rho(r)\,\rho(r')}{|r - r'|}\,dr\,dr' + E_X[\rho] + E_C[\rho]$$

Proof: Let $\tilde{\Psi}$ be the wave function corresponding to $\tilde{\rho}$. Then,

$$\left\langle \tilde{\Psi} \left| \hat{H} \right| \tilde{\Psi} \right\rangle = T[\tilde{\rho}] + U[\tilde{\rho}] + \int \tilde{\rho}(r)\,v(r)\,dr = E_v[\tilde{\rho}] \geq E_v[\rho] = E_0 = \left\langle \Psi \left| \hat{H} \right| \Psi \right\rangle \quad .$$

The inequality follows from Rayleigh-Ritz's variational principle for the wave function, but applied to the electronic density.

Therefore, the variational principle says

$$\delta \left\{ E_v[\rho] - \mu \left(\int \rho(r)\,dr - N \right) \right\} = 0$$

so that a general Thomas-Fermi-like equation is obtained:

$$\mu = \frac{\delta E_v[\rho]}{\delta \rho} = v(r) + \frac{\delta F[\rho]}{\delta \rho}$$

where $F[\rho] = T[\rho] + U[\rho]$. The knowledge of $F[\rho]$ implies the solution of the ground state density. It has to be remarked that $F[\rho]$ is a *universal* functional which does not depend explicitly on the external potential. It depends only on the electronic density. In the Hohenberg-Kohn formulation, $F[\rho] = \left\langle \Psi \left| \hat{T} + \hat{U} \right| \Psi \right\rangle$, where Ψ is the ground state wave function. This two theorems are the basis of *density functional theory.*

In Hohenberg-Kohn theorem, the electronic density determines the external potential. But ... it is also needed that the density corresponded to some antisymmetric wave function deriving from a potential, and this is not always the case. However, density functional theory can be reformulated in such a way that this is not necessary [19]. We define

$$F[\rho] = \min_{\{\Psi\}\to\rho} \left\{ \left\langle \Psi \left| \hat{T} + \hat{U} \right| \Psi \right\rangle \right\}$$

for non-negative ρ such that $\int \rho(r)\, dr = N$ and $\int \left| \nabla \rho^{1/2}(r) \right|^2 dr < \infty$, arising from an antisymmetric wave function. In other words, the search is performed in the subspace of all the antisymmetric Ψ that give rise to the same density ρ.

DFT is exact for the electronic ground state provided that $F[\rho]$ is known. However, it does not say anything about (many-body) excited states. A similar theory can be formulated for excitations of symmetry different to that of the ground state, by resorting to an orthogonal subspace variational principle [20], but in general this is a very hard problem that only now is beggining to be approached with some degree of success [21].

2.3.3 Kohn-Sham equations

We have already briefly discussed about the electron-electron interaction potential U, and we have seen that we can have a reasonably good description by separating the electrostatic (classical Coulomb energy), exchange and correlation contributions. The biggest difficulty is to deal with the correlation. This is, in fact, an active field of research which has produced significant improvements in the past decade. We shall discuss this later on but, for the moment being, let us mention that this issue is quite under control for most systems of interest. On the contrary, there is a problem with the expression of the kinetic energy $\left\langle \Psi \left| \hat{T} \right| \Psi \right\rangle$ in terms of the electronic density. The only expression we have seen up to now is the one proposed by Thomas and Fermi, which is local in the density. This is a severe shortcoming because this model does not hold bound states, and also the electronic shell structure is absent. The main problem with it is that the kinetic operator is inherently non-local, though short-ranged.

In 1965, W. Kohn and L. Sham [22] proposed the idea of replacing the kinetic energy of the interacting electrons with that of an equivalent non-interacting system, because this latter can be easily calculated. Any density $\rho(r)$ that derives from an antisymmetric wave function can be written:

$$\rho(r) = \sum_{i=1}^{\infty} \sum_{s=1}^{2} n_{i,s} \left| \varphi_{i,s}(r) \right|^2 \tag{2.35}$$

where $\{\varphi_{i,s}(r)\}$ are natural spin orbitals, and $\{n_{i,s}\}$ are the occupation numbers of these orbitals. In that case, the kinetic energy can be written as

$$T = \sum_{s=1}^{2} \sum_{i=1}^{\infty} n_{i,s} \left\langle \varphi_{i,s} \left| -\frac{\nabla^2}{2} \right| \varphi_{i,s} \right\rangle \quad . \tag{2.36}$$

These occupation numbers are actually an artifact arising from the fact that we write the density in terms of a set of single-particle orbitals associated with non-interacting fermions. The interacting many-body wave function has to be identified with an occupation N, and not with a set of occupation numbers. However, bearing in mind this conceptual difference, we can always think of $n_{i,s}$ as the occupation of orbital i and spin s. For the moment being we shall suppose that the equivalent non-interacting system, *i.e.* a system of non-interacting fermions whose density coincides with that of the interacting system, does exist. We shall call this the *non-interacting reference system* of density $\rho(r)$, which is described by the hamiltonian

$$\hat{H}_R = \sum_{i=1}^{N_s} \sum_{s=1}^{2} \left(-\frac{\nabla_i^2}{2} + v_R(r) \right) \tag{2.37}$$

This hamiltonian has no electron-electron interactions and, thus, its eigenstates can be expressed in the form of Slater determinants

$$\Psi_s(r) = \frac{1}{\sqrt{N!}} SD \left[\varphi_{1,s}(r_1)\ \varphi_{2,s}(r_2)\ \cdots\ \varphi_{N_s,s}(r_{N_s}) \right]$$

where we have choosen, at $T = 0$, the occupation numbers to be 1 for $i \leq N_s (s = 1, 2)$, and 0 for $i > N_s (s = 1, 2)$. This means that the density is written as

$$\rho(r) = \sum_{i=1}^{N_s} \sum_{s=1}^{2} |\varphi_{i,s}(r)|^2 \tag{2.38}$$

while the kinetic term is

$$T_R[\rho] = \sum_{s=1}^{2} \sum_{i=1}^{N_s} \left\langle \varphi_{i,s} \left| -\frac{\nabla^2}{2} \right| \varphi_{i,s} \right\rangle \quad . \tag{2.39}$$

The single-particle orbitals $\{\varphi_{i,s}(r)\}$ are the N_s lowest eigenfunctions of $\hat{h}_R = -\frac{\nabla^2}{2} + v_R(r)$, *i.e.* $\hat{h}_R\, \varphi_{i,s}(r) = \varepsilon_{i,s}\, \varphi_{i,s}(r)$.

With this definition, we can rewrite the density functional in the following form:

$$E_{KS}[\rho] = T_R[\rho] + \int \rho(r)\, v(r)\, dr + \frac{1}{2} \int \int \frac{\rho(r)\, \rho(r')}{|r - r'|}\, dr\, dr' + E_X[\rho] + \tilde{E}_C[\rho] \tag{2.40}$$

where the fact that $T_R[\rho]$ is the kinetic energy of the non-interacting reference system implies that the correlation piece of the true kinetic energy has been ignored, and has to be taken into account somewhere else. In practice this is done by redefining the correlation energy functional in such a way as to include kinetic correlations.

In this way we have expressed the density functional in terms of $N = N_\uparrow + N_\downarrow$ orbitals. If we now vary these orbitals over the space of functions in 3-dimensions that give reasonable densities (density and gradient of integrable square), we are sure that we are covering also the densities which are in the domain of definition of $E[\rho]$. In other words, we have parametrized the electronic density with a set of N orbitals, and now we will try to minimize the energy functional by applying the variational principle on the orbitals instead of the density. In principle these orbitals are a mathematical object constructed in order to render the problem more tractable, and do not have a sense by themselves, but only in terms of the density. In practice, however, it is customary to think them as single-particle physical eigenstates, but in general they are not. Only in the case that correlations are weak, this can have a real sense.

Now we have to minimize expression (2.40) with respect to $\{\varphi_{i,s}(r)\}$, but taking into account that the orbitals have to be orthogonal, *i.e.* $\int \varphi^*_{i,s}(r)\, \varphi_{j,u}(r)\, dr = \delta_{ij}\delta_{su}$, because they are the N lowest eigenfunctions of a unique potential. This constraints enter into the minimization problem as Lagrange multipliers:

$$\Omega_{KS}\left[\{\varphi_{i,s}(r)\}\right] = E_{KS}\left[\{\varphi_{i,s}(r)\}\right] - \sum_{s=1}^{2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s} \varepsilon_{ij,s} \int \varphi^*_{i,s}(r)\, \varphi_{j,s}(r)\, dr \tag{2.41}$$

Minimizing this functional with respect to each $\varphi_{i,s}(r)$ gives the following set of coupled differential equations:

$$\frac{\delta\Omega_{KS}\left[\{\varphi_{i,s}(r)\}\right]}{\delta\varphi^*_{i,s}(r)} = \left\{-\frac{\nabla^2}{2} + v(r) + \int \frac{\rho(r')}{|r-r'|}\, dr' + \frac{\delta E_{XC}[\rho]}{\delta\rho}\right\} \varphi_{i,s}(r) - \sum_{j=1}^{N_s} \varepsilon_{ij,s}\varphi_{j,s}(r) \; = 0 \tag{2.42}$$

The effective potential (see below) is hermitian and, therefore, the matrix $\varepsilon_{ij,s}$ is symmetric and can be diagonalized by a unitary transformation that keeps invariant the total wave function (the Slater determinant), and thus the density. Such a procedure brings us to the final result, which are the well-known, self-consistent *Kohn-Sham equations:*

$$\left\{-\frac{\nabla^2}{2} + v_{eff}(r)\right\} \varphi_{i,s}(r) = \varepsilon_{i,s}\varphi_{i,s}(r) \tag{2.43}$$

where the effective potential $v_{eff}(r)$ is defined as:

$$v_{eff}(r) = v(r) + \int \frac{\rho(r')}{|r-r'|}\, dr' + \mu_{XC}[\rho] \tag{2.44}$$

and the electronic density is constructed with the solutions of Kohn-Sham equations

$$\rho(r) = \sum_{i=1}^{N_s}\sum_{s=1}^{2} |\varphi_{i,s}(r)|^2 \tag{2.45}$$

The exchange correlation potential $\mu_{XC}[\rho]$ defined above is simply the functional derivative of the exchange-correlation energy $\delta E_{XC}[\rho]/\delta\rho$.

The solution of Kohn-Sham equations has to be obtained by an iterative procedure, in the same way as we have seen for Hartree and Hartree-Fock equations. As in these methods, the total energy cannot be written simply as the sum of the eigenvalues $\varepsilon_{i,s}$, but double counting terms have to be substracted:

$$E_{KS}[\rho] = \sum_{i=1}^{N_s}\sum_{s=1}^{2} \varepsilon_{i,s} - \frac{1}{2}\int\int \frac{\rho(r)\,\rho(r')}{|r-r'|}\, dr\, dr' + \left\{E_{XC}[\rho] - \int \rho(r)\; \mu_{XC}[\rho]\, dr\right\} \tag{2.46}$$

Interpretation

By introducing the non-interacting reference system we were able to take into account the most important part of the kinetic energy. The missing part (correlations) is due to the fact that the full many-body wave function is not a single Slater determinant (otherwise Hartree-Fock theory would be exact). If we think of a true non-interacting system, then DFT is exact while Thomas-Fermi theory is quite a poor approximation, which is reasonably good only when the electronic density is very smooth, like in alkali metals.

The price we have to pay for having a good description of the kinetic energy is that, instead of solving a single equation for the density in terms of the potential, we have to solve a system of N Euler equations. It can be easily seen that Kohn-Sham equations are very similar to Hartree equations, with the difference that the effective potential includes exchange and correlation, which are absent in Hartree theory. Therefore, the computational cost is of the same order of Hartree, but much less than Hartree-Fock which includes the exact non-local exchange. Now let us make some observations:

- The true wave function is not the Slater determinant of Kohn-Sham orbitals, although it is determined by the density.

- The correlation functional has to be modified to account for the missing part in the kinetic energy $T_R[\rho]$, which corresponds to a non-interacting system. The exchange functional remains unchanged.

- Nothing ensures that the non-interacting reference system will always exist. In fact, there are examples like the carbon dimer C_2, which do not satisfy this requirement. However, this is not an unsurmountable shortcoming. In that case we can always consider a linear combination of Slater determinants that include single-particle eigenstates $\varphi_{i,s}(r)$ with $i > N_s$. This is equivalent to extend the domain of definition of the occupation numbers $n_{i,s}$ from the integer values 0 and 1, to a continuum between 0 and 1. In such a way we are including excited single-particle states in the density. At this point, minimization of the energy functional has to be carried out not only with respect to Kohn-Sham orbitals, but also with respect to the occupation numbers [23]. The introduction of excited single-particle states does not mean that the system is in a true excited state. This is only an artifact of the representation. The true wave function is the correlated ground state.

- Janak's theorem is valid [24]. The ionization energy is given by: $I = -\mu = -\varepsilon_{\max}$ (if the effective potential vanishes at long distances), while the eigenvalues are defined as the derivatives of the total energy with respect to the occupation numbers: $\varepsilon_{i,s} = \partial E/\partial n_{i,s}$.

- In DFT there is no Koopman's theorem which would allow us to calculate excitation energies as the difference between the ground state energy of an $(N+1)$-electron system and that of an N-electron system. Excitations in DFT are still an open issue because DFT is a ground state theory, not valid for excited states. Nevertheless, it has been possible to devise some extensions which made possible to deal with excited states within a DFT-like framework, in addition to the traditional many-body scenarios.

Summary

We have described a theory that is able to solve the complicated many-body electronic ground state problem by mapping the many-body Schrödinger equation into a set of N coupled single-particle equations. Therefore, given an external potential, we are in a position to find the electronic density, the energy, and any ground state property we want to (e.g. stress, phonons, etc.). The density of the non-interacting reference system is equal to that of the true interacting system. Up to now the theory is exact. We have not introduced any approximation into the electronic problem. All our ignorance about the many-fermion problem has been displaced to the $\tilde{E}_C[\rho]$ term, while the remaining terms in the energy are well-known.

In the next section we are going to discuss the exchange and correlation functionals. But now, we would like to know how far is $T_R[\rho]$ from $T[\rho]$. Both are the expectation values of the kinetic operator, but in different states. The non-interacting one corresponds to the expectation value in the ground state of the kinetic operator, while the interacting one corresponds to the ground state of the full hamiltonian. This mean that $T_R[\rho] \leq T[\rho]$, implying that $\tilde{E}_C[\rho]$ contains a positive contribution arising from the kinetic correlations.

2.4 Exchange and correlation

We have displaced the ignorance about the quantum many-body problem towards the exchange and correlation functional $E_{XC}[\rho]$. If we knew the exact expression for the kinetic energy including correlation effects, *i.e.* $T[\rho]$, then

$$E_{XC}[\rho] = \frac{1}{2} \int \int \frac{\rho(r)\rho(r')}{|r-r'|} \, [g(r,r') - 1] \; dr \, dr' \tag{2.47}$$

Since we are using the uncorrelated expression for the kinetic energy, *i.e.* the one for non-interacting fermions $T_R[\rho]$, we have to use a slightly different expression: $\tilde{E}_{XC}[\rho] = E_{XC}[\rho] + T[\rho] - T_R[\rho]$. It can be shown that the kinetic contribution to the correlation energy (the kinetic contribution to exchange is just Pauli's principle, which is already contained in $T_R[\rho]$ and in the density when adding up the contributions of the N lowest eigenstates) can be taken into account by averaging the pair correlation function $g(r,r')$ over the strength of the electron-electron interaction, *i.e.*

$$\tilde{E}_{XC}[\rho] = \frac{1}{2} \int \int \frac{\rho(r)\rho(r')}{|r-r'|} \, [\tilde{g}(r,r') - 1] \, dr \, dr' \tag{2.48}$$

where

$$\tilde{g}(r,r') = \int_0^1 g_\lambda(r,r') \, d\lambda \tag{2.49}$$

and $g_\lambda(r,r')$ is the pair correlation function corresponding to the hamiltonian $\hat{H} = \hat{T} + \hat{U} + \lambda \hat{V}_{ee}$ [25]. If we separate the exchange and correlation contributions, then we have:

$$\tilde{g}(r,r') = 1 - \frac{1}{2}\frac{\rho_1^2(r,r')}{\rho(r)\rho(r')} + \tilde{g}_C(r,r') \tag{2.50}$$

with $\rho_1(r,r')$ the one-body density matrix, which in general is a non-diagonal operator. The diagonal elements of it constitute the electronic density. For the homogeneous electron gas the expression for ρ_1 is well-known, so that the exchange pair correlation assumes the analytic closed form

$$g_X(r,r') = g_X(|r-r'|) = 1 - \frac{9}{2}\left(\frac{j_1(k_F\,|r-r'|)}{k_F\,|r-r'|}\right)^2 \tag{2.51}$$

where $j_1(x) = [\sin(x) - x\cos(x)]/x^2$ is the spherical Bessel function of order 1. We are now going to define the exchange-correlation hole $\tilde{\rho}_{XC}(r,r')$ in the following form:

$$\tilde{E}_{XC}[\rho] = \frac{1}{2}\int\int \frac{\rho(r)\tilde{\rho}_{XC}(r,r')}{|r-r'|}\,dr\,dr' \tag{2.52}$$

or $\tilde{\rho}_{XC}(r,r') = \rho(r')\,[\tilde{g}(r,r')-1]$. This means that $\tilde{E}_{XC}[\rho]$ can be written as the interaction between the electronic charge distribution and the charge distribution that has been displaced by exchange and correlation effects, *i.e.* by the fact that the presence of an electron at r reduces the probability for a second electron to be at r', in the vicinity of r. Actually, $\tilde{\rho}_{XC}(r,r')$ is the exchange-correlation hole averaged over the strength of the interaction, which takes into account kinetic correlations. The properties of $\tilde{g}(r,r')$ and $\tilde{\rho}_{XC}(r,r')$ are very interesting and instructive:

1. $\tilde{g}(r,r') = \tilde{g}(r',r)$ (symmetry)
2. $\int \tilde{g}(r,r')\,\rho(r')\,dr' = \int \tilde{g}(r,r')\,\rho(r)\,dr = N-1$ (normalization)
3. $\int \tilde{\rho}_{XC}(r,r')\,dr' = \int \tilde{\rho}_{XC}(r,r')\,dr = -1$

This means that the exchange-correlation hole contains exactly *one* displaced electron. This sum rule is very important, and it has to be verified by any approximation used for $\tilde{\rho}_{XC}(r,r')$. If we separate the exchange and correlation contributions, it is easy to see that the displaced electron comes exclusively from the exchange part, and it is a consequence of the form in which the electron-electron interaction has been separated. In the Hartree term we have included the interaction of the electron with itself. This unphysical contribution is exactly cancelled by the exchange interaction of the full charge density with the displaced density. However, exchange is more than that. It is a nonlocal operator whose local component is minus the self-interaction. On the other hand, the correlation hole integrates to zero $\int \tilde{\rho}_C(r,r')\,dr' = 0$ so that the correlation energy correponds to the interaction of the charge density with a neutral charge distribution.

A general discussion on DFT and applications can be found in Ref. [26].

2.4.1 The Local Density Approximation (LDA)

This has been for a long time the most widely used approximation to the exchange-correlation energy. It has been proposed in the seminal paper by Kohn and Sham, but the philosophy was already present in Thomas-Fermi theory. The main idea is to consider the general inhomogeneous electronic systems as locally homogeneous, and then to use the exchange-correlation hole corresponding to the homogeneous electron gas for which there are very good approximations and also exact numerical (quantum Monte Carlo) results. This means:

$$\tilde{\rho}_{XC}^{LDA}(r,r') = \rho(r)\,\left\{\tilde{g}^h\left[|r-r'|\,,\rho(r)\right]-1\right\} \tag{2.53}$$

with $\tilde{g}^h\left[|r-r'|\,,\rho(r)\right]$ the pair correlation function of the homogeneous gas, which depends only on the distance between r and r', evaluated at the density ρ^h which locally equals $\rho(r)$. Within this approximation, the exchange-correlation energy density is defined as:

$$\epsilon_{XC}^{LDA}[\rho] = \frac{1}{2}\int\frac{\tilde{\rho}_{XC}^{LDA}(r,r')}{|r-r'|}\,dr' \tag{2.54}$$

and the exchange-correlation energy becomes

$$E_{XC}^{LDA}[\rho] = \int\rho(r)\,\epsilon_{XC}^{LDA}[\rho]\,dr\ . \tag{2.55}$$

In general, the exchange-correlation energy density is not a functional of ρ. From its very definition it is clear that it has to be a non-local object, because it reflects the fact that the probability of finding an electron at r depends on the presence of other electrons in the surroundings, through the exchange-correlation hole.

Looking at expression (2.53), it may seem that there is an inconsistency in the definition. The exact expression would indicate to take $\rho(r')$ instead of $\rho(r)$. However, this would make of $\epsilon_{XC}^{LDA}[\rho]$ a non-local object which would depend on the densities at r and r', and we want to parametrize it with the homogeneous gas, which is characterized by only one density, and not two. This is the essence of the LDA, and it is equivalent to postulate:

$$\tilde{g}(r,r') = \tilde{g}^h[|r-r'|\,,\rho(r)]\left(\frac{\rho(r)}{\rho(r')}\right) \tag{2.56}$$

Therefore, there are in fact two approximations embodied in the LDA:

1. The exchange-correlation hole is centered at r, and interacts with the electronic density at r. The true XC hole is actually centered at r' instead of r.

2. The pair correlation function (g) is approximated by that of the homogeneous electron gas of density $\rho(r)$ corrected by the density ratio $\rho(r)/\rho(r')$ to compensate the fact that the LDA XC hole is centered at r instead of r'.

2.4.2 The Local Spin Density Approximation

In magnetic systems or, in general, in systems where open electronic shells are involved, better approximations to the exchange-correlation functional can be obtained by introducing the two spin densities , $\rho_\uparrow(r)$ and $\rho_\downarrow(r)$, such that $\rho(r) = \rho_\uparrow(r) + \rho_\downarrow(r)$, and $\zeta(r) = (\rho_\uparrow(r) - \rho_\downarrow(r)) \, / \, \rho(r)$ is the magnetization density. The non-interacting kinetic energy (2.39) splits trivially into *spin-up* and a *spin-down* contributions, and the external and Hartree potential depend on the full density $\rho(r)$, but the approximate XC functional — even if the exact functional should depend only on $\rho(r)$ — will depend on both spin densities independently, $E_{XC} = E_{XC}[\rho_\uparrow(r), \rho_\downarrow(r)]$. Kohn-Sham equations then read exactly as in (2.43), but the effective potential $v_{eff}(r)$ now acquires a spin index:

$$\begin{aligned} v^\uparrow_{eff}(r) &= v(r) + \int \frac{\rho(r')}{|r-r'|} \, dr' + \frac{\delta E_{XC}[\rho_\uparrow(r), \rho_\downarrow(r)]}{\delta \rho_\uparrow(r)} \\ v^\downarrow_{eff}(r) &= v(r) + \int \frac{\rho(r')}{|r-r'|} \, dr' + \frac{\delta E_{XC}[\rho_\uparrow(r), \rho_\downarrow(r)]}{\delta \rho_\downarrow(r)} \end{aligned} \qquad (2.57)$$

The density given by expression (2.45) contains a double summation, over the spin states and over the number of electrons in each spin state, N_s. These later have to be determined according to the single-particle eigenvalues, by asking for the lowest $N = N_\uparrow + N_\downarrow$ to be occupied. This defines a Fermi energy ε_F, such that the occupied eigenstates have $\varepsilon_{i,s} < \varepsilon_F$.

In the case of non-magnetic systems $\rho_\uparrow(r) = \rho_\downarrow(r)$, and everything reduces to the simple case of double occupancy of the single-particle orbitals, and then the calculations spare half of the computer time.

The equivalent of the LDA in spin-polarized systems is the *local spin density approximation* (LSDA), and it basically consists of replacing the XC energy density with a spin-polarized expression:

$$E^{LSDA}_{XC}[\rho_\uparrow(r), \rho_\downarrow(r)] = \int [\rho_\uparrow(r) + \rho_\downarrow(r)] \; \varepsilon^h_{XC}[\rho_\uparrow(r), \rho_\downarrow(r)] \, dr \quad , \qquad (2.58)$$

obtained, for instance, by interpolating between the fully-polarized and fully-unpolarized XC energy densities using an appropriate expression that depends on $\zeta(r)$. The standard practice is to use von Barth and Hedin's interpolation formula [27]:

$$f(\zeta) = \frac{(1+\zeta)^{4/3} + (1-\zeta)^{4/3} - 2}{2^{4/3} - 2} \qquad , \qquad (2.59)$$

or a more realistic formula based on the RPA, given by Vosko, Wilk and Nussair [28].

A thorough discussion of the LDA and the LSDA can be found in Ref. [29]. In the following we reproduce the main aspects related to these approximations.

Why does the LDA work so well in many cases ?

1. It satisfies the sum rule that the XC hole contains exactly one displaced electron:

$$\int \tilde{\rho}^{LDA}_{XC}(r,r')\,dr' = \int \rho(r)\ \tilde{g}^h[|r-r'|\,,\rho(r)]\,dr' = -1 \tag{2.60}$$

because for each r, $\tilde{g}^h[|r-r'|\,,\rho(r)]$ is the pair correlation function of the homogeneous gas at density $\rho(r)$, and then the second integral above is nothing but the integral of the XC hole of the homogeneous gas, for which the approximations and numerical results available, carefully take into account that the integral has to be -1.

2. Even if the exact $\tilde{\rho}_{XC}$ has no spherical symmetry, in the expression for the XC energy what really counts is the spherical average of the hole:

$$E_{XC}[\rho] = -\frac{1}{2}\int \rho(r)\left(\frac{1}{R(r)}\right) dr$$

with

$$\frac{1}{R(r)} = \int \frac{\tilde{\rho}_{XC}(r,r')}{|r-r'|}\,dr' = 4\pi \int_0^\infty s\ \tilde{\rho}^{SA}_{XC}(r,s)\ ds$$

and

$$\tilde{\rho}^{SA}_{XC}(r,s)\ = \frac{1}{4\pi}\int_\Omega \tilde{\rho}_{XC}(r,r')\,d\Omega\ .$$

This spherical average $\tilde{\rho}^{SA}_{XC}(r,s)$ is reproduced to a good extent by the LDA, whose $\tilde{\rho}_{XC}$ is already spherical.

The LDA exhibits the following general trends:

- It favours more homogeneous systems.
- It overbinds molecules and solids.
- Chemical trends are usually correct.
- For “good” systems (covalent, ionic, and metallic bonds): geometries are good, bond lengths, bond angles and phonon frequencies are within a few %, while dielectric and piezoelectric constants are about 10% too large.
- For ”bad” systems (weakly bound) bond lengths are too short (overbinding).
- In atoms, the XC potential does not decay as $-e^2/r$, thus affecting the dissociation limit and ionization energies.

What do we normally use for the LDA ?

For the exchange energy density it is adopted the form deduced by Dirac [30]:

$$\epsilon_X[\rho] = -\frac{3}{4}(\frac{3}{\pi})^{1/3}\rho^{1/3} = -\frac{3}{4}(\frac{9}{4\pi^2})^{1/3}\frac{1}{r_s} = -\frac{0.458}{r_s}a.u. \tag{2.61}$$

where $\rho^{-1} = 4\pi r_s^3/3$, and r_s is the radius of the sphere that, on average, contains one electron.

For the correlation, a widely used approximation is Perdew and Zunger's parametrization [23] of Ceperley and Alder quantum Monte Carlo, essentially exact results [31]:

$$\epsilon_C[\rho] = \begin{cases} A \ln r_s + B + C\, r_s \ln r_s + D\, r_s, & r_s \leq 1 \\ \gamma \;/\; (1 + \beta_1 \sqrt{r_s} + \beta_2 r_s), & r_s > 1 \end{cases} \tag{2.62}$$

For $r_s \leq 1$ the expression arises from the the Random Phase Approximation (RPA) — calculated by Gell-Mann and Brückner [32] — which is valid in the limit of very dense electronic systems. For very weak densities, Perdew and Zunger have fitted a Padé approximant to the Monte Carlo results. Another popular parametrization is that proposed by Vosko, Wilk and Nusair [33].

When does the LDA fail ?

- In atomic systems, where the density has large variations.
- In weak molecular bonds, e.g. hydrogen-bonds, because in the bonding region the density is very small and the binding is rather dominated by the inhomogeneities.
- In van der Waals — closed shell — systems, because the nonlocality of the exchange interaction becomes important when the binding is due to dynamical fluctuations of the charge density of neutral objects.
- In metallic surfaces, because the XC potential decays exponentially while it should follow a power law.
- In negatively charged ions, because the LDA fails to cancel exactly the electronic self-interaction, due to the approximative character of the exchange. Self-interaction corrected functionals have been proposed [23], although they are not satisfactory from the theoretical point of view because the potential depends on the electronic state, while it should be the same for all states.
- The energy band gap in semiconductors turns out to be very small. The reason is that when one electron is removed from the ground state, the exchange hole becomes screened, and this is absent in the LDA. On the other hand, also Hartree-Fock has the same shortcoming, and the band gap turns out to be too large.

How can the LDA be improved ?

Once the extent of the approximations involved in the LDA has been understood, then one can start constructing better approximations. The amount of work done into that direction is really overwhelming. Sometimes is difficult even to cope with the new developments, simply because there is not a unique and obvious way of improving the LDA.

One of the key observations is that the *true* pair correlation function – $g(r, r')$ – actually depends on the electronic density at two different points, r and r'. The homogeneous

$g(r, r')$ is quite well-known (see Eq. (2.51) for the exchange part, and [34] for correlation), but it corresponds to a density which is the same everywhere. Therefore, the question is which of the two densities to use in an inhomogeneous environment. Early efforts went into the direction of calculating the pair correlation function at an *average* density $\bar{\rho}(r)$, which in general is different from $\rho(r)$, and incorporates information about the density at neighboring points. Clearly there is no unique recipe for the averaging procedure, but there is at least a crucial condition that this averaging has to verify, namely the normalization condition [35, 36]:

$$\int \tilde{\rho}_{XC}^{WDA}(r, r')\, dr' = \int \rho(r')\ \tilde{g}^h[|r - r'|\,, \bar{\rho}(r)]\, dr' = -1 \quad . \tag{2.63}$$

Approach of this type receives the name of weighted density approximations (WDA). There is still a lot of freedom in choosing the averaging procedure, provided that normalization is verified and, indeed, several different approximations have been proposed [35, 36, 37, 38]. One problem with this approach is that the $r \rightarrow r'$ symmetry of $g(r, r')$ is now broken. These efforts went along the direction of improving the location of the center of the XC hole. An exploration in the context of realistic electronic structure calculations was carried out by D. Singh but the results reported were not significantly better the LDA [39].

Actually, it can be done better than this by attacking the problem with the correct many-body tools. For instance, one could try to solve Dyson's equation for the electronic Green's function, starting from the LDA solution for the bare Green's function (see section XXXX below). Another possibility proposed in the context of strongly correlated systems, *e.g.* exhibiting narrow d or f bands, where the limitation of the LDA is at describing strong on-site correlations of the Hubbard type, is to introduce these features *a posteriori*. The LDA+U approach of Anisimov *et al.* [40] considers the mean-field solution of the Hubbard model on top of the LDA solution, where the Hubbard on-site interaction U are computed for the d or f orbitals by differentiating the LDA eigenvalues with respect to the occupation numbers.

Undoubtedly, and probably because of its computational efficiency and its similarity to the LDA, the most popular approach has been to introduce semilocally the inhomogeneities of the density, by expanding $E_{XC}[\rho]$ as a series in terms of the density and its gradients. This approach, known as generalized gradient approximation (GGA), is easier (and cheaper) to implement in practice than full many-body approaches, and has been quite successful in improving some features over the LDA.

2.4.3 Generalized Gradient Approximations

The exchange-correlation energy has a gradient expansion

$$E_{XC}[\rho] = \int A_{xc}[\rho]\, \rho(r)^{4/3}\, dr + \int C_{xc}[\rho]\ \mid \nabla\rho(r) \mid^2 / \rho(r)^{4/3}\, dr + \cdots \tag{2.64}$$

which is asymptotically valid for densities that vary slowly in space. The LDA retains only the leading term of Eq. (2.64). It is well-known that a straigthforward evaluation of this expansion is ill-behaved, in the sense that it is not monotonically convergent, and it exhibits singularities that cancel out only when an infinite number of terms is resummed, like in the random phase approximation (RPA). In fact, the first-order correction worsens the results, and the second order correction is is plagued of divergencies [41]. The largest errors of this approximation actually arise from the gradient contribution to the correlation term. Provided that the problem of the correlation term can be cured in some way, as the real space cutoff method proposed by Langreth and Mehl [42], the biggest problem remains with the exchange energy.

Many papers have been devoted to the improvement of the exchange term within DFT. The early work of Gross and Dreizler [43] provided a derivation of the second-order expansion of the exchange density matrix, which was later re-analyzed and extended by Perdew [44]. This expansion contains not only the gradient, but also laplacian of the density. The same type of expansion was obtained, using Wigner distribution – phase space – techniques, by Ghosh and Parr [45].

One of the main lessons learnt from these works is that the gradient expansion has to be carried out very carefully in order to retain all the relevant contributions to the desired order. The other important lesson is that these expansions easily violate one or some of the *exact* conditions required for the exchange and the correlation holes. For instance the normalization condition, the negativity of the exchange density, and the self-interaction cancellation (the diagonal of the exchange density matrix has to be minus a half of the density). Perdew has shown that imposing these conditions to functionals that originally do not verify them, results in a remarkable improvement of the quality of exchange energies [44]. On the basis of this type of reasoning, a number of modified gradient expansions have been proposed along the years, mainly between 1986 and 1996. These have received the name of generalized gradient approximations (GGA).

GGA are either based on theoretical developments that reproduce the exact results in some known limits – 0 and ∞ density, or the correlation potential in the He atom –, or that are generated by fitting a number of parameters to a molecular database (training set). Normally, these improve some of the drawbacks of the LDA. The basic idea of GGAs is to express the exchange-correlation energy in the following form:

$$E_{XC}[\rho] = \int \rho(r)\,\varepsilon_{XC}[\rho(r)]\,dr + \int F_{XC}[\rho(r), \nabla\rho(r)]\,dr \tag{2.65}$$

where the function F_{XC} is asked to satisfy a number of formal conditions for the exchange-correlation hole, like sum rules, long-range decay, etc. This cannot be done by considering directly the bare gradient expansion (2.64). What is needed from the functional is a form that mimicks a resumation to infinite order, and this is the main idea of the GGA, for which there is not a unique recipe. Naturally, not all the formal properties can be enforced at the same time, and this differentiates one functional from another. A thorough comparison of different GGA can be found in Ref. [46]. In the following we quote a number of them:

1. Langreth-Mehl exchange-correlation functional [42].

$$\varepsilon_X = \varepsilon_X^{LDA} - a\frac{\mid \nabla\rho(r) \mid^2}{\rho(r)^{4/3}}\left(\frac{7}{9} + 18\,f^2\right)$$

$$\varepsilon_C = \varepsilon_C^{RPA} + a\frac{\mid \nabla\rho(r) \mid^2}{\rho(r)^{4/3}}\left(2e^{-F} + 18\,f^2\right)$$

where $F = b \mid \nabla\rho(r) \mid / \rho(r)^{7/6}$, $b = (9\pi)^{1/6} f$, $a = \pi/(16(3\pi^2)^{4/3})$, and $f = 0.15$.

2. Perdew-Wang '91 exchange functional [47].

$$\varepsilon_X = \varepsilon_X^{LDA} \left(\frac{1 + a_1 s \sinh^{-1}(a_2 s) + (a_3 + a_4 e^{-100 s^2}) s^2}{1 + a_1 s \sinh^{-1}(a_2 s) + a_5 s^4} \right)$$

where $a_1 = 0.19645$, $a_2 = 7.7956$, $a_3 = 0.2743$, $a_4 = -0.1508$, and $a_5 = 0.004$.

3. Perdew-Wang '91 correlation functional [47].

$$\varepsilon_C = \varepsilon_C^{LDA} + \rho\, H[\rho, s, t]$$

with

$$H[\rho, s, t] = \frac{\beta}{2\alpha} \ln\left(1 + \frac{2\alpha}{\beta} \frac{t^2 + At^4}{1 + At^2 + A^2 t^4}\right) + C_{c0}\left[C_c(\rho) - C_{c1}\right] t^2 e^{-100 s^2}$$

and

$$A = \frac{2\alpha}{\beta}\left[e^{-2\alpha\varepsilon_C[\rho]\,/\,\beta^2} - 1\right]^{-1}$$

where $\alpha = 0.09$, $\beta = 0.0667263212$, $C_{c0} = 15.7559$, $C_{c1} = 0.003521$, $t = |\nabla\rho(r)|\,/\,(2k_s\rho)$ for $k_s = (4k_F/\pi)^{1/2}$, and $\rho\,\varepsilon_C[\rho] = \varepsilon_C^{LDA}[\rho]$.

4. Becke '88 exchange functional [48].

$$\varepsilon_X = \varepsilon_X^{LDA} \left(1 - \frac{\beta}{2^{1/3} A_x} \frac{x^2}{1 + 6\beta x \sinh^{-1}(x)}\right)$$

for $x = 2(6\pi^2)^{1/3} s = 2^{1/3} \mid \nabla\rho(r) \mid / \rho(r)^{4/3}$, $A_x = (3/4)(3/\pi)^{1/3}$, and $\beta = 0.0042$.

5. Closed-shell Lee-Yang-Parr correlation functional [49].

$$\varepsilon_C = -a \frac{1}{1 + d\rho^{-1/3}} \left\{\rho + b\rho^{-2/3}\left[C_F \rho^{5/3} - 2t_W + \frac{1}{9}\left(t_W + \frac{1}{2}\nabla^2\rho\right)\right] e^{-c\rho^{-1/3}}\right\}$$

where

$$t_W = \frac{1}{8}\left(\frac{|\nabla\rho|^2}{\rho} - \nabla^2\rho\right)$$

and $C_F = 3/10(3\pi^2)^{2/3}$, $a = 0.04918$, $b = 0.132$, $c = 0.2533$, and $d = 0.349$. This correlation functional is not based on the LDA as the others, but it has been derived as an extension of the Colle-Salvetti expression for the electronic correlation in Helium, to other closed-shell systems.

6. Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional [50].

We define first the enhancement factor F_{XC} over the local exchange:

$$E_{XC}[\rho] = \int \rho(r)\, \varepsilon_X^{LDA}[\rho(r)]\, F_{XC}(\rho, \zeta, s)\, dr$$

where ρ is the local density, ζ is the relative spin polarization, and $s = |\nabla\rho(r)| \,/\, (2k_F\rho)$ is the dimensionless density gradient, as in Perdew-Wang '86. We first concentrate on the exchange enhancement factor

$$F_X(s) = 1 + \kappa - \frac{\kappa}{1 + \mu s^2/\kappa} \quad ,$$

where $\mu = \beta(\pi^2/3) = 0.21951$, where $\beta = 0.066725$ is related to the second order gradient expansion [47]. This form: a) satisfies the uniform scaling condition, b) recovers the correct uniform electron gas limit because $F_x(0) = 1$, c) obeys the spin-scaling relationship, d) recovers the LSDA linear response limit for $s \longrightarrow 0$ ($F_X(s) \longrightarrow 1 + \mu s^2$), and e) satisfies the *local* Lieb-Oxford bound [51], $\varepsilon_X(r) \geq -1.679\rho(r)^{4/3}$, *i.e.* $F_X(s) \leq 1.804$, for all r, provided that $\kappa \leq 0.804$. PBE choose the largest allowed value $\kappa = 0.804$. Other authors have proposed the same form, but with values of κ and μ fitted empirically to a database of atomization energies [52, 53]. The proposed values of κ violate Lieb-Oxford inequality.

The correlation energy is written in a form similar to Perdew-Wang '91 [47], *i.e.*

$$E_C^{GGA} = \int \rho(r) \;\left[\varepsilon_C^{LDA}(\rho,\zeta) + H[\rho,\zeta,t]\right] \, dr$$

with

$$H[\rho,\zeta,t] = (e^2/a_0)\gamma\phi^3 \ln\left\{1 + \frac{\beta\gamma}{t}^2 \left[\frac{1 + A\,t^2}{1 + At^2 + A^2t^4}\right]\right\} \quad .$$

Here, $t = |\nabla\rho(r)| \,/\, (2\phi k_s\rho)$ is a dimensionless density gradient, $k_s = (4k_F/\pi a_0)^{1/2}$ is the Thomas-Fermi screening wave number, and $\phi(\zeta) = [(1+\zeta)^{2/3} + (1-\zeta)^{2/3}]/2$ is a spin-scaling factor. The quantity β is the same as for the exchange term $\beta = 0.066725$, and $\gamma = (1 - \ln 2)/\pi^2 = 0.031091$. The function A has the following form:

$$A = \frac{\beta}{\gamma}\left[e^{-\varepsilon_C^{LDA}[\rho] \,/\, (\gamma\phi^3 e^2/a_0)} - 1\right]^{-1} \quad .$$

So defined, the correlation correction term H satisfies the following properties: a) it tends to the correct second-order gradient expansion in the slowly-varying (high density) limit ($t \longrightarrow 0$), b) it approaches minus the uniform electron gas correlation $-\varepsilon_C^{LDA}$ for rapidly varying densities ($t \rightarrow \infty$), thus making the correlation energy vanish (this results from the correlation hole sum rule), c) it cancels the logarithmic singularity of ε_C^{LDA} in the high density limit, thus forcing the correlation energy to scale to a constant under uniform scaling of the density.

This GGA retains the correct features of LDA (LSDA), and combines them with the nonlocality features which are supposed to be the most energetically important. It sacrifices a few correct, but less important features, like the correct second-order gradient coefficients in the slowly-varying limit, and the correct nonuniform scaling of the exchange energy in the rapidly varying density region.

Becke '88 exchange functional is usually combined with Lee-Yang-Parr correlation to form the popular BLYP approach. Being some of the coefficients obtained by optimizing the agreement of certain structural and energetic quantities, the BLYP functional is not well-justified from a theoretical point of view. On the contrary, the PBE functional [50] is very satisfactory because it verifies many of the exact conditions for the XC hole, and it does not contain any fitting parameters. In addition, its quality is better than that of BLYP [54].

The different recipies for GGAs verify only some of the mathematical properties known for the exact exchange-correlation hole. In the following table we show which properties are verified by some popular functionals.

	Property	E_{XC}^{LDA}	E_{XC}^{LM}	E_{XC}^{PW91}	E_X^{B88}	E_C^{LYP}
1	$\rho_X(r,r') \leq 0$	y	-	y	-	-
2	$\int \rho_X(r,r')\,dr' = -1$	y	-	y	-	-
3	$\int \rho_C(r,r')\,dr' = 0$	y	-	y	-	-
4	$E_X[\rho] < 0$	y	y	y	y	-
5	$E_C[\rho] \leq 0$	y	n	n	-	n
6	$E_X[\rho], E_{XC}[\rho] \geq -c\int \rho^{4/3}dr$ [a]	y	n	y	y	-
7	$E_X[\rho_\lambda] = \lambda E_X[\rho]$ [b]	y	y	y	y	-
8	$E_C[\rho_\lambda] < \lambda E_C[\rho], \quad \lambda < 1$ [e]	y	n	y	-	n
9	$\lim_{\lambda\to\infty} E_C[\rho_\lambda] > -\infty$	n	y[f]	y[f]	-	y
10	$\lim_{\lambda\to 0} \frac{1}{\lambda} E_C[\rho_\lambda] > -\infty$	y	n	y	-	y
11	$\lim_{\lambda\to\infty} E_X[\rho_\lambda^x] > -\infty$ [c]	n	n	y	n	-
12	$\lim_{\lambda\to 0} E_X[\rho_\lambda^x] > -\infty$	y	n	y	y	-
13	$\lim_{\lambda\to\infty} \frac{1}{\lambda} E_X[\rho_{\lambda\lambda}^{xy}] > -\infty$ [d]	y	n	y	y	-
14	$\lim_{\lambda\to 0} \frac{1}{\lambda} E_X[\rho_{\lambda\lambda}^{xy}] > -\infty$	n	n	y	n	-
15	$\lim_{\lambda\to\infty} \lambda E_C[\rho_\lambda^x] > -\infty$	n	y[f]	y	-	n
16	$\lim_{\lambda\to 0} \frac{1}{\lambda} E_C[\rho_\lambda^x] = 0$	n	n	y	-	n
17	$\lim_{\lambda\to\infty} E_C[\rho_{\lambda\lambda}^{xy}] = 0$	n	n	y	-	n
18	$\lim_{\lambda\to 0} \frac{1}{\lambda^2} E_C[\rho_{\lambda\lambda}^{xy}] > -\infty$	n	y[f]	y	-	n
19	$\epsilon_X(r) \to -\frac{1}{2r}, \quad r \to \infty$	n	n	n	yn[g]	-
20	$v_X(r) \to -\frac{1}{r}, \quad r \to \infty$	n	n	n	n	-
21	$v_X(r), v_C(r) \to$ finite, $\quad r \to 0$	y	n	n	n	n
22	LDA limit for constant $\rho(r)$	y	n	y	y	n

[a] $1.44 < c < 1.68$

[b] $\rho_\lambda(r) = \lambda^3\rho(\lambda r)$; [c] $\rho_\lambda^x(r) = \lambda\rho(\lambda x, y, z)$; [d] $\rho_{\lambda\lambda}^{xy}(r) = \lambda^2\rho(\lambda x, \lambda y, z)$

[e] Note that $E_C[\rho_\lambda] < \lambda E_C[\rho], \quad \lambda < 1$ is equivalent to $E_C[\rho_\lambda] > \lambda E_C[\rho], \quad \lambda > 1$.

[f] But it diverges to $+\infty$

[g] "y" for exponential $\rho(r)$, but "n" in general, *e.g.* $\epsilon_X^{B88}(r) \to -1/r$ for a gaussian.

The general trends of GGAs concerning improvements over the LDA are the following:

1. Improves binding energies (they give better atomic energies)

2. Improves bond lengths of IIA and IIB homonuclear dimers

3. Improves energetics, geometries and dynamical properties of water, ice and water clusters. BLYP seems to have the best agreement with experiment. In general, they improve the description of hydrogen-bonded systems, although this is not very clear for the case of the $F \cdots H$ bond.

4. Si, Ge, GaAs are better described in the LDA, except for the binding energies.

5. For 4d-5d transition metals it is not clear whether GGA improves over LDA or not.

6. There is no improvement for the gap problem (and consequently for the dielectric constant), because this feature has to do with the description of the screening of the exchange hole when one electron is removed, and this is usually not taken into account by GGA. An exception is PBE.

7. They do not satisfy the known asymptotic behaviour, *e.g.* for isolated atoms:

   - $v_{XC}(r) \sim -e^2/r$ for $r \to \infty$, while $v_{XC}^{LDA,GGA}(r)$ vanish exponentially.
   - $v_{XC}(r) \to const$ for $r \to 0$, while $v_{XC}^{LDA}(r) \to const$, but $v_{XC}^{GGA}(r) \to -\infty$.

There seems, then, to exist a limit in the accuracy that GGAs can reach. The main responsibility for this is of the exchange term, whose non-locality is not fully taken into account. A particularly problematic issue is that GGA functionals still do not compensate completely the self-interaction.

This has motivated the development of approximations which combine gradient corrected functionals with exact, Hartree-Fock-type exchange. An example is the approximation known as B3LYP, which reproduces very well the geometries and binding energies of molecular systems, at the same level of correlated quantum chemistry approches like second order Møller-Plesset perturbation theory (MP2) or even at the level of coupled cluster and CI methods, although at a significantly lower computational cost [55]. Even if the idea is appealing and physically sensible, the derivation is not rigorous and the functional also involves a number of fitting parameters. An alternative approach are screened exchange functionals, where the exact exchange interaction is reduced by using an empirical screening function, e.g. an error function.

In the past few years there have been serious attempts to go beyond the GGA. Some, like the meta-GGA described in the following Section, are simple and more or less successful, although not completely satisfactory from the theoretical point of view. Another, better-founded approach is the so-called *exact exchange* (EXX), where the Kohn-Sham potential contains a *local* exchange term obtained from the exact Hartree-Fock exchange. This tends to be rather expensive computationally.

2.4.4 Meta-GGA

The second order gradient expansion of the exchange energy introduces a term proportional to the squared gradient of the density. If this expansion is further carried out to fourth order, as originally done by Gross and Dreizler [43] and later resumed by Perdew

[44], it also appears a term proportional to the square of the Laplacian of the density. The Laplacian term was also derived using a different route by Ghosh and Parr [45], although it was then dropped out when considering the gradient expansion only up to second order. More recently, a general derivation of the exchange gradient expansion up to sixth order, using second order density response theory, was given by Svendsen and von Barth [56]. The fourth order expansion of that paper was then used by Perdew *et al.* [57] to construct a practical meta-generalized gradient approximation (meta-GGA) that incorporates additional semilocal information in terms of the Laplacian of the density. The philosophy for constructing the functional is the same as that of PBE, namely to retain the good formal properties of the lower level approximation (the PBE GGA in this case), while adding others.

The gradient expansion of the exchange enhancement factor F_X is

$$F_X(p,q) = 1 + \frac{10}{81}p + \frac{146}{2025}q^2 - \frac{73}{405}qp + Dp^2 + 0(\nabla^6) \quad , \tag{2.66}$$

where

$$p = |\nabla\rho| \,/\, [4(3\pi^2)^{2/3}\rho^{8/3}]$$

is the square of the reduced density gradient, and

$$q = \nabla^2\rho \,/\, [4(3\pi^2)^{2/3}\rho^{5/3}]$$

is the reduced Laplacian of the density.

The first two coefficients of the expansion are exactly known. The third one is the result of a difficult many-body calculation, and has only been estimated numerically by Svendsen and von Barth, to an accuracy better than 20%. The fourth coefficient D has not been explicitly calculated to the date.

In the same spirit of PBE, Perdew, Kurth, Zupan and Blaha (PKZB) proposed an exchange enhancement factor which verifies some of the formal relations, and reduces to the gradient expansion (2.66) in the slowly-varying limit of the density. The expression is formally identical to that of PBE:

$$F_X^{MGGA}(p,\bar{q}) = 1 + \kappa - \frac{\kappa}{1 + x/\kappa} \quad , \tag{2.67}$$

where

$$x = \frac{10}{81}p + \frac{146}{2025}\bar{q}^2 - \frac{73}{405}\bar{q}p + \left[D + \frac{1}{\kappa}(\frac{10}{81})^2\right]p^2$$

is a new inhomogeneity parameter that replaces $0.21951p$ in PBE. The variable q in the gradient expansion (the reduced Laplacian) is also replaced by a new variable $\bar{q}$ defined as

$$\bar{q} = 3\tau \,/\, [2(3\pi^2)^{2/3}\rho^{5/3}] - 9/20 - p/12 \quad ,$$

which reduces to q in the slowly-varying limit, but remains finite at a nucleus while q diverges. In the above expression $\tau[\rho] = \tau_\uparrow + \tau_\downarrow$ is the kinetic energy density for the noninteracting system, with

$$\tau_\sigma = \frac{1}{2}\sum_\alpha^{occup} |\nabla\psi_{\alpha\sigma}(r)|^2$$

$\sigma =\uparrow, \downarrow$. The connection between τ and the density is give n by the second-order gradient expansion

$$\tau^{GEA} = \frac{3}{10}(3\pi^2)^{2/3}\rho^{5/3} + \frac{1}{72}\frac{|\nabla\rho|^2}{\rho} + \frac{1}{6}\nabla^2\rho \ \ .$$

The formal conditions requested for this functional are: a) the spin-scaling relation, b) the uniform density-scaling relation [58], and the Lieb-Oxford inequality [51]. Actually, a value of $\kappa = 0.804$, corresponding to the largest value ensuring that the inequality is verified for all possible densities, is chosen in [57] (exactly as in [50]). The coefficient D still remains undetermined. In the absence of theoretical estimations, PKZB proposed to fix D by minimizing the absolute error in the atomization energies for a molecular data set. The value so obtained is $D = 0.113$. The meta-GGA recovers the exact linear response function up to fourth order in $k/2k_F$. This is not the case of PBE-GGA (and other GGA's), which recovers only the LSDA linear response, and at the expenses of sacrificing the correct second-order gradient expansion (the coefficient of p is larger than 10/81 by a factor of 1.778).

The correlation part of the meta-GGA retains the correct formal properties of PBE GGA correlation, such as the slowly-varying limit and the finite limit under uniform scaling. In addition, it is required that the correlation energy be self-interaction free, *i.e.* to vanish for a one-electron system. PKZB proposed the following form:

$$\begin{aligned} E_C^{MGGA}[\rho_\uparrow, rho_\downarrow] &= \int\{\rho\varepsilon_C^{GGA}(\rho_\uparrow,\rho_\downarrow,\nabla\rho_\uparrow,\nabla\rho_\downarrow)\left[1 + C\left(\frac{\sum_\sigma \tau_\sigma^W}{\sum_\sigma \tau_\sigma}\right)^2\right] - \\ &- (1+C)\sum_\sigma\left(\frac{\tau_\sigma^W}{\tau_\sigma}\right)^2 \rho_\sigma\varepsilon_C^{GGA}(\rho_\sigma, 0, \nabla\rho_\sigma, 0)\} \ \ , \end{aligned} \tag{2.68}$$

where ε_C^{GGA} is the PBE-GGA correlation energy density, and τ_σ^W is the von Weiszäcker kinetic energy density given by expression (2.34) above, which is *exact* for a one-electron density. Therefore, the correlation energy vanishes for any one-electron density, irrespectively of the value of the parameter C. For many-electron systems the self-interaction cancellation is not complete, but the error is shifted to fourth order in the gradient, thus having no effect on systems with slowly-varying density. As for the exchange term, there is no theoretical estimate available for the parameter C. Perdew *et al.* obtained a value of $C = 0.53$ by fitting it to PBE-GGA surface correlation energies for jellium. Atomic correlation energies also agree, but are less accurate. Just as an example, the correlation energy for He is -0.84 Hartree in LSDA, -0.68 H in PBE-GGA, and -0.48 H in MGGA, which basically coincides with the exact value [59].

Unlike the PBE-GGA, the meta-GGA has two fitted parameters, C and D. The reason for them is actually the unavailability of first-principles theoretical estimates for them. Other authors have proposed similar expansions containing the kinetic energy density in addition to the density gradients. These, however, took the road of constructing the functional in a semiempirical way, by fitting a large number of parameters (of the order of 10 or 20) to chemical data, instead of using well-founded theoretical values [60, 61]. The quality of the results of different meta-GGA functionals is quite similar. An assesment of the general quality of the PKZB meta-GGA in comparison to GGA and hybrid exact exchange - GGA models of the B3LYP type, has been published very recently [62]. The conclusion is that the kinetic energy density is a useful additional ingredient. Atomization energies are quite improved n PKZB meta-GGA with respect to PBE-GGA, but unfortunately geometries and frequencies are worsened. In particular, bond lengths are far too long. Adamo *et al.* [62] argued that a possible reason could be that in this functional the long-range part of the exchange hole, which would help to localize the exchange hole, thus favoring shorter bond lengths, is missing. Intriguingly enough, one of the semiempirical meta-GGA functionals [61] gives very good geometries and frequencies. According to the preceeding discussion, this effect on geometries is due to the non-local properties of the exchange functional, a factor that the kinetic energy density, being still a semilocal object, cannot account for. Therefore, this agreement has to be originated in error cancellations between exchange and correlation.

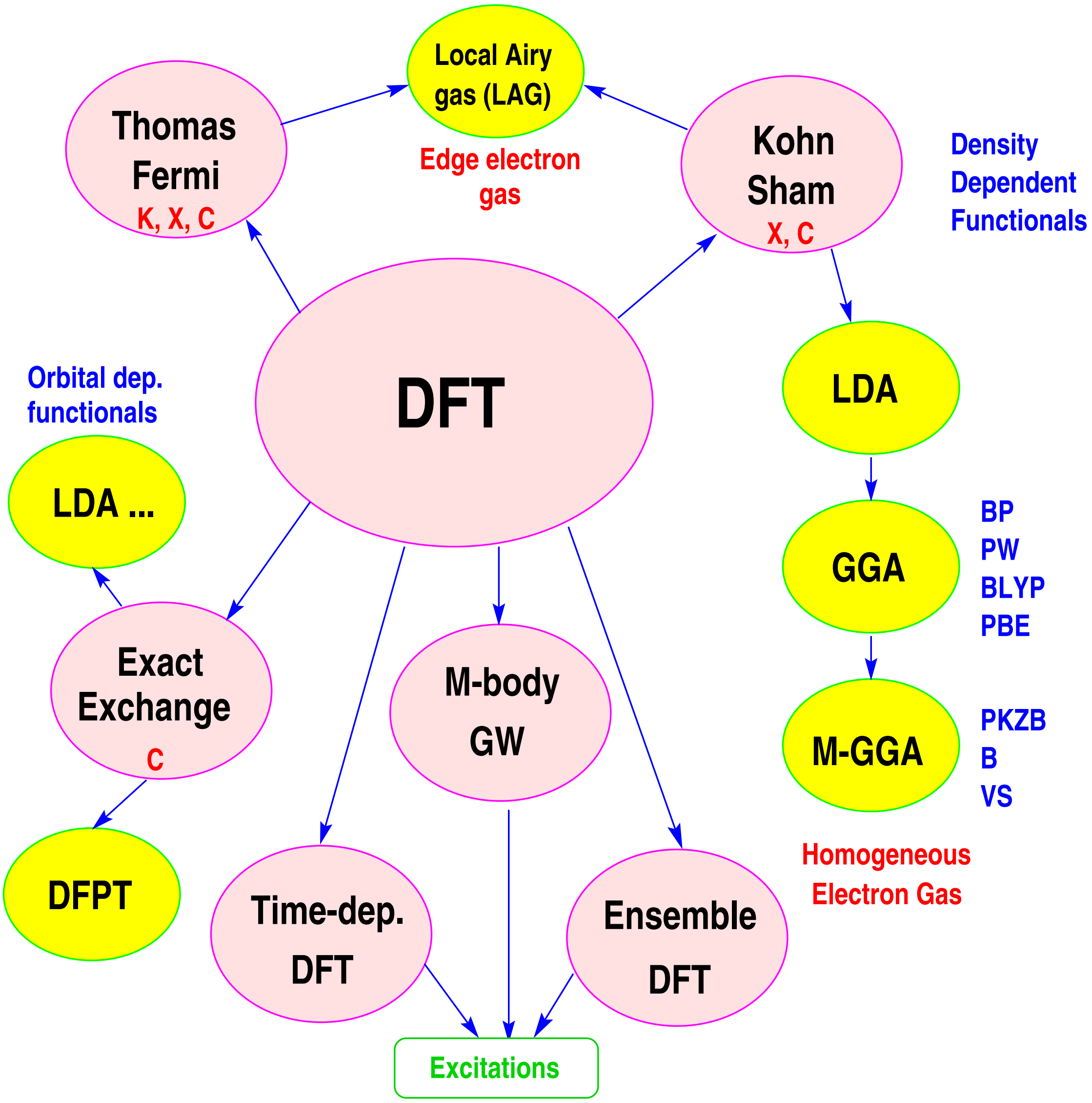

Figure 2.3: A global view of density functional theory.

Chapter 3

A brief review of solid state theory

Atoms can arrange in many different ways: molecules are constituted by a finite (small) number of atoms, typically combining more than one element, and have a specific composition and ordering that defines univocally its properties. Clusters are pretty much like molecules, but the composition and ordering is not well defined as in molecules. They are typically made of one or two (binary clusters) elements. Solids, on the other hand, are macroscopic objects constituted by a a huge number of atoms, of the order of the Avogadro number (6×10^{23}). This justifies the standard approach of describing solids as an infinite collection of atoms. Crystalline solids are those in which a small number of atoms (a *basis*) is infinitely replicated along d different directions in space, where d is the dimensionality of space. Bulk solids are replicated in three dimensions, surfaces in two, and wires in one. These directions are defined by d linearly independent vectors.

There are infinitely many ways of characterizing a crystalline solid, depending on the choice of the set of atoms that are replicated. However, there is only one choice with the minimal number of atoms that contains the whole symmetry of the system. This is called the *unit (or Wigner-Seitz) cell*, and it contains all the information about the point group symmetry underlying the crystalline structure. The vectors that serve to reconstruct the infinite solid from the unit cell are also unique, and are called *unit (or primitive) vectors*. The set of points in space defined as integer combinations of the primitive vectors receives the name of *Bravais lattice*, of which there are only 32 (in 3 dimensions). The combination of the translational symmetry emboddied in the Bravais lattice plus the point group symmetry of the basis, gives rise to 132 space groups, which are sufficient to classify all the known crystalline solids. Sometimes it is convenient to describe the solid in terms of a cell containing more atoms than the unit cell (*conventional cell*) in order to simplify the description of the symmetry properties, e.g. to have orthogonal lattice vectors. For instance a body centered cubic (bcc) unit cell can be also described as a simple cubic cell containing two atoms in the basis, and a face centered cubic cell (fcc) is equivalent to a simple cubic cell with a 4-atom basis. We shall call $\{\mathbf{a_i}\}_{i=1,2,3}$ the unit vectors, and the volume of the unit cell is going to be $\Omega = \mathbf{a_1} \cdot (\mathbf{a_2} \times \mathbf{a_3})$. The Wigner-Seitz cell can be constructed by drawing a line perpendicular to each unit vector exactly at its mid point.

The properties of the infinite system are connected to those of the unit cell by means of Bloch's theorem:

Theorem(Bloch): the wave function of an electron in an external periodic potential $V(\mathbf{r}) = V(\mathbf{r} + \mathbf{a}_i)$ can be written as the product of a function with the same periodicity of the potential, and a purely imaginary phase factor arising from the translational symmetry, *i.e.*

$$\Psi_{\mathbf{k}}(\mathbf{r}) = e^{i\ \mathbf{k}\cdot\mathbf{r}}\ u_{\mathbf{k}}(\mathbf{r}) \tag{3.1}$$

with $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_i)$. This implies that:

$$\Psi_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_i) = e^{i\ \mathbf{k}\cdot\mathbf{a}_i}\ \Psi_{\mathbf{k}}(\mathbf{r}) \quad . \tag{3.2}$$

The reciprocal lattice vectors are defined by the relation $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$, such that $e^{i\ \mathbf{a}_i\cdot\mathbf{b}_j} = 1$. This implies:

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2 \times \mathbf{a}_3}{\Omega}; \quad \mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3 \times \mathbf{a}_1}{\Omega}; \quad \mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1 \times \mathbf{a}_2}{\Omega} \tag{3.3}$$

and the volume of the reciprocal cell is: $\mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3) = (2\pi)^3/\Omega$. The cell defined by the reciprocal vectors corresponding to the primitive vectors is called the first *Brillouin zone*, or Brillouin zone for short (BZ). The idea is that any vector outside the BZ can be written as $\mathbf{k} = \mathbf{k}' + \mathbf{G}$ with $\mathbf{k}'$ inside the BZ and $\mathbf{G} = n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3$, with n_i integer numbers. In other words, the whole reciprocal space can be covered by translating the BZ with vectors of the reciprocal lattice. It is clear that $e^{i\ \mathbf{G}\cdot\mathbf{a}_i} = 1 \Rightarrow e^{i\ \mathbf{k}\cdot\mathbf{a}_i} = e^{i\ \mathbf{k}'\cdot\mathbf{a}_i}$.

Such a periodic wave function obbeys the Schrödinger equation

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right)\Psi_{\mathbf{k}}(\mathbf{r}) = \varepsilon_{\mathbf{k}}\ \Psi_{\mathbf{k}}(\mathbf{r}) \tag{3.4}$$

where $\varepsilon_{\mathbf{k}}$ is the energy of the wave. It is easy to see that there is a family of solutions $\Psi_{\mathbf{k}+\mathbf{G}}(\mathbf{r})$ and $\Psi_{\mathbf{k}'+\mathbf{G}'}(\mathbf{r})$ with the same energy, provided that $|\mathbf{k} + \mathbf{G}| = |\mathbf{k}'{+}\mathbf{G}'|$, for instance $\mathbf{k}' = \mathbf{k}{+}\,(\mathbf{G} - \mathbf{G}')$. It is then clear that for every vector $\mathbf{k}'$ we can always find another vector $\mathbf{k}$ in the first BZ such that $\varepsilon_{\mathbf{k}} = \varepsilon_{\mathbf{k}'}$. Since wave functions of identical energy mix together, then the solutions of the eigenvalue problem have to be searched in the degenerate subspaces $\{\Psi_{\mathbf{k}+\mathbf{G}}(\mathbf{r})\}$ with $\mathbf{G}$ all the reciprocal lattice vectors. Therefore, we can focus on the solution of the eigenvalue problem for $\mathbf{k}$ vectors in the first BZ, and then obtain trivially the solution for any vector outside the first BZ that is connected with $\mathbf{k}$ through a lattice vector $\mathbf{G}$.

In conclusion, the calculation of the wave function for each of the infinite number of electrons in the infinite solid, is mapped — via Bloch theorem — onto the calculation of the wave function for a finite number of electrons in the unit cell, at an infinite number of $\mathbf{k}$ vectors in the first BZ.

A more detailed treatment of solid state theory can be found in any specific book on the subject [63].

3.1 Brillouin Zone sampling

Let us consider a supercell of length $\mathbf{L}$, periodically replicated, which contains N unit cells ($\mathbf{a} = \mathbf{L}/N$). The BZ of the supercell will have dimensions $\frac{2\pi}{L_i} = \frac{2\pi}{N}\frac{N}{L_i} = \frac{2\pi}{N a_i} = \frac{1}{N}\left(\frac{2\pi}{a_i}\right)$, so that it will be contained N times inside the BZ of the unit cell. If the supercell is infinite — as it is the case of perfect crystalline structures —, then its BZ will be contained infinite times in $2\pi/\mathbf{a}$. Notice that a wave vector $\mathbf{k} = \mathbf{j}2\pi/\mathbf{L}$ will correspond to the periodicity of a supercell of length $\mathbf{L}/\mathbf{j}$. In this way we can take into account periodicities that go beyond the size of the supercell just by considering the appropriate $\mathbf{k}$ vectors in the BZ.

Let us consider two electrons in the unit cell. Periodic boundary conditions (PBC) imply that only the lowest state $\Psi_0^{(1)}(\mathbf{r})$ in the box will be occupied by the two electrons with opposite spins. The electronic density will be $\rho(\mathbf{r}) = 2\left|\Psi_0^{(1)}(\mathbf{r})\right|^2$, and will be periodic in the box $\rho(\mathbf{r}+\mathbf{a}) = \rho(\mathbf{r})$. The wave function must also verify $\Psi_0^{(1)}(\mathbf{r}+\mathbf{a}) = \Psi_0^{(1)}(\mathbf{r})$, but including the exponential phase factor. Let us now construct a supercell by doubling the unit cell along one of the primitive vectors. There are 4 electrons, so that two states will be occupied, $\Psi_0^{(2)}(\mathbf{r})$ and $\Psi_1^{(2)}(\mathbf{r})$, which verify the periodic boundary conditions $\Psi_0^{(1)}(\mathbf{r}+2\mathbf{a}) = \Psi_0^{(1)}(\mathbf{r})$ and $\Psi_0^{(1)}(\mathbf{r}+2\mathbf{a}) = \Psi_0^{(1)}(\mathbf{r})$. Since the supercell consists of the replication of two identical units, then one of the two states should correspond to the replication of state $\Psi_0^{(1)}(\mathbf{r})$, which in addition verifies the periodicity condition at $\mathbf{r}+\mathbf{a}$. The other state does not need to verify the PBC on the wave function, but it must do it on the density, which now is written $\rho(\mathbf{r}) = 2\left(\left|\Psi_0^{(2)}(\mathbf{r})\right|^2 + \left|\Psi_1^{(2)}(\mathbf{r})\right|^2\right)$.

We are going to define now $\tilde{\rho}^{(1)}(\mathbf{r})$ as the part of the charge density of the doubled system that belongs to the original unit cell. Since in that part there are only 2 electrons, it has to be normalized dividing by the number of replications (2 in this case). But a wave function with period $2\mathbf{a}$ and a density of period $\mathbf{a}$ is equivalent to a wave function with wave vector $\mathbf{k} = \pi/\mathbf{a}$ (just replace $N = 2$ in the first expression of this section). Therefore, $\tilde{\Psi}_0^{(2)}(\mathbf{r}) = \Psi_{\mathbf{k}=0}^{(1)}(\mathbf{r})$ and $\tilde{\Psi}_1^{(2)}(\mathbf{r}) = \Psi_{\mathbf{k}=\pi/\mathbf{a}}^{(1)}(\mathbf{r})$, where the tilde indicates the restriction to the original unit cell. The density then reads:

$$\tilde{\rho}^{(1)}(\mathbf{r}) = 2\left(\frac{1}{2}\left|\Psi_{\mathbf{k}=0}^{(1)}(\mathbf{r})\right|^2 + \frac{1}{2}\left|\Psi_{\mathbf{k}=\pi/\mathbf{a}}^{(1)}(\mathbf{r})\right|^2\right) \tag{3.5}$$

If we now replicate the doubled box, so that we have 4 replicas of the unit cell, then it is easy to realize that we are introducing states with $\mathbf{k} = \pi/2\mathbf{a}$ and $\mathbf{k} = -\pi/2\mathbf{a}$. The wave functions associated to these two wave vectors have a period $4\mathbf{a}$, while the wave function of period $2\mathbf{a}$ obtained for the doubled supercell (corresponding to $\mathbf{k} = \pi/\mathbf{a}$) is also periodic in $4\mathbf{a}$. This means that the density restricted to the original unit cell is:

$$\tilde{\rho}^{(1)}(\mathbf{r}) = 2\left(\frac{1}{4}\left|\Psi_{\mathbf{k}=0}^{(1)}(\mathbf{r})\right|^2 + \frac{1}{4}\left|\Psi_{\mathbf{k}=\pi/\mathbf{a}}^{(1)}(\mathbf{r})\right|^2 + \frac{1}{2}\left|\Psi_{\mathbf{k}=\pi/2\mathbf{a}}^{(1)}(\mathbf{r})\right|^2\right) \tag{3.6}$$

where we have used the trivial fact that $\Psi_{\mathbf{k}=\pi/2\mathbf{a}}(\mathbf{r}) = \Psi_{\mathbf{k}=-\pi/2\mathbf{a}}^*(\mathbf{r})$.

The generalization to an arbitrary number of replications and to 3 dimensions is straightforward, and leads to the following well-known expression for the electronic density:

$$\rho(\mathbf{r}) = \sum_{\mathbf{k}} \omega_{\mathbf{k}} \left|\Psi_{\mathbf{k}}(\mathbf{r})\right|^2 \quad . \tag{3.7}$$

In 2 and 3 dimensions, the symmetry of the unit cell can be exploited to reduce the portion of the BZ that has to be sampled in the summation above. This introduces the concept of the irreducible wedge of the BZ as the minimal portion that contains all the necessary information to describe the whole BZ. In a simple cubic unit cell, it will be an octant. The **k**-points at the boundaries of the irreducible wedge count less, because they are shared with other wedges. The multiplicity factor is $1/M$, with M the number of different irreducible wedges that share this point. Points inside the wedge count 1, and the Γ-point ($\mathbf{k} = 0$) counts $1/M_{\mathrm{max}}$, with M_{max} the total number of wedges needed to fill the whole BZ. This is the way to calculate the weights $\omega_{\mathbf{k}}$. In practice, a finite number of **k**-points is used to represent the full BZ integration (integration because the summation becomes an integral for the infinite system). The number needed will depend on the size of the supercell and on the specific features of the system. For instance metals need a very fine sampling, while semiconductors can be reasonably represented with a few, carefully selected **k**-points. Sets of special **k**-points for the different symmetries, whose use accelerates the convergence of the BZ summation with the number of points, have been worked out by Baldereschi [64], and Chadi and Cohen [65]. A more general, unbiased recipe for all symmetries, and specially for metallic systems, has been proposed by Monkhorst and Pack [66].

It is interesting to remark that the individual wave functions of wave vector $\mathbf{k} \neq \mathbf{0}$ do not fullfil the PBC because of the phase factor. They do verify PBC, but in a larger supercell of size $\pi/\ \mathbf{k}$. By varying the wave vector from 0 to $\pi/\mathbf{a}$, we scan different boundary conditions in the unit cell from periodic to antiperiodic. The electronic density, however, is always periodic because the phase factor is irrelevant.

The magnitude of the errors introduced by sampling the BZ integral with a finite number of **k**-points can always be reduced by using a denser set of points.

Let us consider now a supercell contaning just a few unit cells. The larger BZ of the unit cell can be reproduced by transporting the smaller BZ of the supercell with reciprocal lattice vectors **G**. Suppose that the supercell was sampled with the Γ-point only. When transported with the BZ of the supercell, the Γ-point will *refold* onto a set of **k**-points other from Γ, but inside the BZ of the unit cell. This means that the choice of a larger supercell is equivalent to consider the unit cell, but with a finer sampling of its BZ. The correspondence is, however, not perfect because there refolded **k**-points cannot be choosen at will, as in the case of true **k**-points. They are univocally determined by the shape of the supercell. In the limit of an infinitely large supercell, its BZ becomes a point (the Γ-point, in fact), and its transportation is equivalent to a uniform and infinitely fine sampling of the BZ of the unit cell.

3.1.1 BZ sampling for aperiodic systems

At finite temperatures the point group symmetry of a bulk solid is broken, and all the discussion of the Brillouin zone becomes less evident. If the temperature is such that the system remains in a well-defined crystal structure, and thermal vibrations of the atoms are circumscribed to the vicinity of their equilibrium positions, then the concepts of discrete translational invariance, unit cell, Brillouin zone, and **k**-points, hold.

If the system becomes diffusive (solid or fluid), it has point defects, presents a surface, or it is a molecule — amongst other possibilities — then, strictly speaking, the replication of a relatively small supercell is not the correct description of the infinite system, which is intrinsically aperiodic. For instance in liquids, PBC break the homogeneity property.

This description, however, is much better than to consider isolated clusters as models for such aperiodic systems, and it represents a very valid alternative to costly, large supercells. This for what concerns the PBC on the atomic nuclei, but ... what about the electronic wave functions ? It is clear that if there is no translational invariance at all, the whole machinery derived from Bloch's theorem breaks down. However, the spurious periodicity introduced by the use of PBC, implies a fictitious translational invariance and, given that this invariance exists anyway, then the solid state machinery is restored. The difference with bulk solids is that the periodic replication of the supercell is not a physical fact, but an artifact which is useful to accelerate convergence with respect to the size of the system. In this perspective, the use of a BZ sampling in the case of aperiodic systems is correct because it takes into account the electronic periodicities at the same level as the nuclear periodicities, which were introduced through PBC.

The most important point of the supercell approach is that, in order to be meaningful, physical properties have to be converged with the respect to the size of the supercell. This means, *e.g.* in the case of molecules or point defects, that the images corresponding to adjacent replicas of the supercell should not interact significantly. In some cases, this might be difficult to achieve, particularly in polar (or charged) systems with large electrostatic (long-range) fields.

Chapter 4

Solving the electronic problem in practice

The central problem in electronic structure at the single-particle approximation level is, then, to self-consistently solve a set of N coupled, 3-dimensional, partial differential equations. In the Kohn-Sham formulation of DFT for infinite systems, this set of equations reads:

$$\hat{H}_{KS}\,\Psi_{\mathbf{k},i}(\mathbf{r}) = \left(-\frac{\nabla^2}{2} + v_{ext}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \mu_{XC}[\rho]\right)\Psi_{\mathbf{k},i}(\mathbf{r}) = \varepsilon_{\mathbf{k},i}\,\Psi_{\mathbf{k},i}(\mathbf{r}) \tag{4.1}$$

where the electronic density is expressed as a BZ average,

$$\rho(\mathbf{r}) = \sum_{\mathbf{k}\in BZ}\omega_{\mathbf{k}}\sum_{i=1}^{N_{\mathbf{k}}}|\Psi_{\mathbf{k},i}(\mathbf{r})|^2 \tag{4.2}$$

and $N_{\mathbf{k}}$ is the number of electronic states that are occupied at that particular $\mathbf{k}$-point. If the system is insulating, then this number is independent of $\mathbf{k}$ and equal to the number of electrons N (if there is spin degeneracy the number of independent eigenfunctions is $N/2$, so that the sum is performed up to $N/2$, and the result multiplied by 2. For metallic systems the occupation numbers $N_{\mathbf{k}}$ are determined by asking that the associated eigenvalues $\{\varepsilon_{\mathbf{k},i};\quad i = 1\cdots N_{\mathbf{k}}\}$ be smaller than a certain value ϵ_F (the Fermi level). This latter is self-consistently adjusted to fulfill the normalization condition: $\sum_{\mathbf{k}\in BZ} N_{\mathbf{k}}\omega_{\mathbf{k}} = N$.

The external potential $v_{ext}(\mathbf{r})$ represents the interaction between the electrons and the nuclei, and can is expressed in the following way:

$$v_{ext}(\mathbf{r}) = -e^2\sum_{I=1}^{P}\frac{Z_I}{|\mathbf{r}-\mathbf{R}_I|} \tag{4.3}$$

At this stage, the solution of Kohn-Sham equations requires two important choices:

1. **How to represent the single-particle wave functions**
2. **How to treat the electron-nuclear interactions**

The representation of the wave functions implies the choice of a basis set. Many possibilities have been explored since the early times of solid state theory and quantum chemistry, which can be divided into four main groups:

1. Extended basis sets: basis functions are delocalized, floating or centered at the nuclear positions.
2. Localized basis sets: basis functions are localized, mainly centered at the nuclear positions, but not uniquely.
3. Mixed basis sets: a combination of extended and localized basis functions.
4. Augmented basis sets: an extended or localized basis set is augmented with atomic-like wave functions in some region around the nuclei.

When dealing with extended systems (solids or liquids), it has to be ensured that Bloch's theorem is verified, in the sense that the combination of basis orbitals representing a solution to the Schrödinger equation must have the periodicity of the supercell.

Expanding the wave functions on some generic basis set $\mid \phi_{\mathbf{k}}^{(m)} >$:

$$\Psi_{\mathbf{k},j}(\mathbf{r}) = e^{i\ \mathbf{k}\cdot\mathbf{r}} \sum_{n=1}^{M} C_{\mathbf{k},j}^{(n)}\, \phi_{\mathbf{k}}^{(n)}(\mathbf{r}) \quad , \tag{4.4}$$

the Schrödinger equation becomes a matrix equation (*secular equation*):

$$\sum_{m=1}^{M} \left(\mathcal{H}_{nm}^{\mathbf{k}} - \varepsilon_{\mathbf{k},j} S_{nm}^{\mathbf{k}} \right) C_{\mathbf{k},j}^{(m)} = \mathbf{0} \tag{4.5}$$

where $\mathcal{H}_{nm}^{\mathbf{k}} =< \phi_{\mathbf{k}}^{(n)} \left| \hat{\mathcal{H}} \right| \phi_{\mathbf{k}}^{(m)} >$ and $S_{nm}^{\mathbf{k}} =< \phi_{\mathbf{k}}^{(n)} \mid \phi_{\mathbf{k}}^{(m)} >$. In the above expressions, M is the size of the basis set and j is a band index which labels the eigenvalues at fixed $\mathbf{k}$ according to their energy. The number of occupied bands is $N/2$ (in the following we focus on the spin unpolarized case) with N the number of electrons in the unit cell. The overlap matrix $S_{nm}^{\mathbf{k}}$ appears in the secular equation because the basis functions do not need to be mutually orthogonal. In fact, in many electronic structure methods the basis set is non-orthogonal.

The electron-nuclear interaction is given by the bare Coulomb interaction. A first class of methods deals with all the electrons in the system, both those participating in the chemical bonding (*valence electrons*) and those tightly bound to the nuclei, which are almost unchanged with respect to the atomic case (*core electrons*). These are generically named *all-electron methods.* They can be constructed in a straightforward way by using finely tuned localized basis sets, like in quantum chemistry methods, or by separating the space in atomic spheres (as closely packed as possible) plus an interstitial region. In this latter, the wave functions of the valence electrons are expanded in some basis set in the interstitial region, and are augmented with atomic-like solutions inside the spheres (Muffin Tins – MT) while the wave functions for the core electrons are obtained as solutions of the atomic problem but taking into account the perturbation produced by the presence of the other atoms. The augmentation is done in such a way that the logarithmic derivatives of the radial part of the wave functions at the MT radius, $d\ln[R_l(E, r_c)]/dE$, are continuous. The matching conditions depend on the eigenvalue associated to that wave function. In principle, these can be fulfilled by recalculating the logarithmic derivatives at the correct eigenvalue every self-consistent iteration. Another possibility which is faster and more stable is to linearize the matching conditions around a reference energy (or a few reference energies) which is (are) representative of the eigenvalue that the wave function assumes in that particular environment. These are called *linear methods* [67].

Since core electrons usually do not participate in chemical bonding, it is possible to integrate out the corresponding degrees of freedom by considering a screened interaction between the valence electrons and the *ionic cores*, *i.e.* nuclei plus core electrons. Because of orthogonalization to the core wave functions the valence wave functions typically have several nodes, and when there are no core electrons of the same symmetry, then the valence wave function peaks very strongly close to the nucleus. These two features are quite inconvenient from the point of view of the representation of these wave functions. In principle, a matching procedure as in all-electron methods can be adopted. Another possibility is to realize that a good description of the valence wave functions inside the ionic cores is, in most cases, unnecessary, because one is usually concerned with bonding properties. In that case, there is no lack of crucial information if the inner solution (inside the core radius) is replaced with a smooth, nodeless pseudo-wave function, which behaves much better from the numerical point of view. This pseudo-wave function is not the solution of the original atomic problem, but the solution of a pseudo-atomic problem where the true potential has been replaced by a pseudopotential. This type of approximations receive the name of *pseudopotential* methods. Pseudopotential theory will be discussed in detail below.

4.1 All-electron methods

There are three main all-electron methodologies according to the basis set used in the interstitial region:

- **Localized basis sets**: This methodology is the most widely used in the quantum chemistry community, which basically aims at describing molecular systems instead of solids. Bloch's theorem and periodicity of the potential are irrelevant issues in those cases. The most popular, because it makes it possible to calculate matrix elements analytically, is the Gaussian-type orbitals basis set (**GTO** – Boys 1950, McWeeny 1953). Two other basis sets which are widely used are Slater-type orbitals (**STO**) and atomic orbitals (**LCAO**). These methods deal with all the electrons, core and valence, at the same level.

- **Muffin Tin Orbitals (MTO)**: the wave functions in the interstitial region are expanded in spherical Hankel functions centered at the nuclear positions. Hankels are solutions of the spherically symmetric Schrödinger equation in the absence of a potential (as it is the case in the interstitial region) which are regular at the origin, rapidly varying inside the Muffin Tin (MT), and exponentially decaying outside.

The method using nonlinear matching conditions [68], although originally formulated in a different form in terms of Green functions, is due to Korringa, Kohn and Rostocker (**KKR**) [69]. The linear method, or **LMTO**, was originally proposed by O. K. Andersen [67]. The LMTO is one the most popular all-electron methods because it is very fast. The fastness arises from the fact that the basis functions can be finely tuned so that a small number of them is enough to have a reasonable description of the system. The treatment of the potential in the interstitial region is an expensive part of the calculation with MTO's. Methods that inlcude self-consistently this contribution receive the name of *full-potential* (e.g. **FP-LMTO** [70]). A much faster method, although not very accurate, consists of approximating the potential in the interstitial region with a constant value, and increasing the size of the MT spheres until they touch each other (optimal close packing). This method is known as the atomic sphere approximation (**LMTO-ASA**), and has been very widely used in the past. Nowadays, FP methods have superseeded it. MTO have a drawback when they are used to study open structures. The interstitial vaccum is poorly described unless empty spheres (MT spheres wth zero charge) to fill the emty space are included in the basis set. However, this renders more difficult the comparison of different structures at the level of the energetics, and forces on the nuclei cannot be computed. Modern developments along the LMTO line have very recently overcome these difficulties.

- **Augmented Plane Waves (APW)**: the wave functions in the interstitial region are expanded in (floating) plane waves, which are the solutions of the Schrödinger equation for free electrons in a box (see below), and matched to atomic-like solutions inside the spheres. Only the lowest angular momenta ($l = 0, 1, 2, ..., l_{\max}$) are present inside the spheres, so that only these projections of the plane waves (PW) are matched [71]. The components of the PW with angular momentum $l > l_{\max}$ are allowed to penetrate inside the spheres without forcing any matching condition. This is the full-potential version of the APW (**FP-APW**) method. The wave functions are written as:

$$\mathcal{A}(\mathbf{p},\mathbf{r}) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} a_{lm}\, Y_{lm}(\theta,\varphi)\, R_l(E,r)\, \eta(r_c - r) + e^{i\,\mathbf{p}\cdot\mathbf{r}}\, \eta(r - r_c) \quad , \tag{4.6}$$

so that the secular equation becomes

$$\begin{aligned} & \langle \mathcal{A}_{\mathbf{k},i} \,| \mathcal{H} - E |\, \mathcal{A}_{\mathbf{k},j} \rangle = & (4.7) \\ = & \left(\frac{\hbar^2}{2m}\mathbf{k}_i\cdot\mathbf{k}_j - E\right)\delta_{ij} + \frac{4\pi r_c^2}{\Omega}\left(\frac{\hbar^2}{2m}\mathbf{k}_i\cdot\mathbf{k}_j - E\right)\frac{j_1\left(|\mathbf{k}_i - \mathbf{k}_j|\, r_c\right)}{|\mathbf{k}_i - \mathbf{k}_j|} + & (4.8) \\ + & \frac{4\pi r_c^2}{\Omega}\sum_{l=0}^{\infty}(2l+1)\, P_l(\cos\theta_{ij})\, j_l(k_i r_c)\, j_l(k_j r_c)\left[\frac{R_l'(E,r_c)}{R_l(E,r_c)} - \frac{j_l'(k_j r_c)}{j_l(k_j r_c)}\right] \quad . \end{aligned}$$

The version which linearizes the logarithmic derivatives is called Linearized augmented plane waves (**LAPW**), and is presently the most accurate electronic structure method available. The expansion of the wave functions in the interstitials in

PW gives a great flexibilitiy because there is no need of empty spheres as in LMTO methods. Forces on the nuclear degrees of freedom can be calculated, and relaxation and/or molecular dynamics simulations become then possible [72, 73].

4.2 Pseudopotential methods

Pseudopotential methods differentiate from each other basically on the basis set which is used in conjunction with them. Four classes of methods have been proposed in the past:

1. Extended basis sets: normally plane waves (**PPW**). This is the most popular approach, and the one we are going to explain in detail here [74, 75, 76].

2. Localized basis sets: the same types of basis sets as in all-electron methods can be used in conjunction with pseudopotentials, e.g. GTO, LCAO, STO. Recently, also pseudo-atomic orbitals (**PAO**), which are atomic-like orbitals but strictly localized inside a cutoff radius, have been introduced by Sankey and Niklevsky [77]. These was then implemented in conjunction with an Order N (linear scaling with the number of atoms) method, into the SIESTA package [78].

3. Mixed basis sets: combination of plane waves and gaussians have been used in the past, particularly in the times when computers were not powerful enough.

4. Projected augmented waves (**PAW**): it is very similar in spirit to the APW all-electron method, but the core electrons are replaced by a pseudopotential. There is an augmentation sphere [79].

4.3 The plane wave basis set

The fact that, according to Bloch theorem, $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_i)$, can be used to introduce the natural (for solid state applications) basis of plane waves (PW). We can always write

$$u_{\mathbf{k}}(\mathbf{r}) = \int e^{i\ \mathbf{g}\cdot\mathbf{r}}\, \tilde{u}_{\mathbf{k}}(\mathbf{g})\ d\mathbf{g} \tag{4.9}$$

but since $u_{\mathbf{k}}(\mathbf{r} + \mathbf{a}_j) = u_{\mathbf{k}}(\mathbf{r})$, then the only allowed values of $\mathbf{g}$ are those that verify $e^{i\ \mathbf{g}\cdot\mathbf{a}_j} = 1$, *i.e.* $\mathbf{g} \cdot \mathbf{a}_j = 2n\pi$ for $j = 1, 2, 3$ the three lattice vectors. This implies that $\mathbf{g} = n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3$, where

$$\mathbf{b}_i = 2\pi \frac{\mathbf{a}_j \times \mathbf{a}_k}{\mathbf{a}_i \cdot (\mathbf{a}_j \times \mathbf{a}_k)} \tag{4.10}$$

and $\mathbf{n} = (n_1, n_2, n_3)$ is a vector of integer numbers. Therefore, the $\mathbf{g}$ vectors in the Fourier transform (4.9) are restricted precisely to the reciprocal lattice vectors $\mathbf{G}$ defined by Eq. (3.3), so that the general expression for the wave function is:

$$\Psi_{\mathbf{k}}(\mathbf{r}) = e^{i\ \mathbf{k}\cdot\mathbf{r}} \sum_{\mathbf{G}=0}^{\infty} C_{\mathbf{k}}(\mathbf{G})\ e^{i\ \mathbf{G}\cdot\mathbf{r}} \tag{4.11}$$

which is the expansion of the wave function in a plane waves $\left\{e^{i\mathbf{G}\cdot\mathbf{r}}\right\}$ basis set. This restriction to the reciprocal lattice vectors implies that PBC are authomatically verified.

Notice that the prefactor $e^{i\mathbf{k}\cdot\mathbf{r}}$ involves a wave vector $\mathbf{k}$ in the first BZ, while the reciprocal lattice vectors $\mathbf{G}$ entering the PW expansion stay always outside the BZ. Wave functions corresponding to different $\mathbf{k}$ vectors obbey different Schrödinger equations. In the case of non-interacting electrons these are completely independent, but in the case of DFT (or any other many-body theory) the equations couple in the sense of self-consistency through the electronic density, which is expressed as the average over the whole BZ (see above).

If we choose the PW basis set, then the basis functions are $\phi_{\mathbf{k}}^{(\mathbf{G})}(\mathbf{r}) = e^{i\ (\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}/\sqrt{\Omega}$, and the matrix elements in (4.5) are very simple to calculate: Firstly, $S_{\mathbf{G},\mathbf{G}'}^{\mathbf{k}} = \delta_{\mathbf{G},\mathbf{G}'}$because the PW are orthogonal, and $H_{\mathbf{G},\mathbf{G}'}^{\mathbf{k}} = T_{\mathbf{G},\mathbf{G}'}^{\mathbf{k}} + V_{\mathbf{G},\mathbf{G}'}^{\mathbf{k}}$ with

$$T_{\mathbf{G},\mathbf{G}'}^{\mathbf{k}} = \left\langle \mathbf{k}+\mathbf{G} \left| -\frac{\hbar^2}{2m}\nabla^2 \right| \mathbf{k}+\mathbf{G}' \right\rangle = \frac{\hbar^2}{2m} \left|\mathbf{k}+\mathbf{G}\right|^2 \delta_{\mathbf{G},\mathbf{G}'} \tag{4.12}$$

$$V_{\mathbf{G},\mathbf{G}'}^{\mathbf{k}} = \left\langle \mathbf{k}+\mathbf{G} \left| V(\mathbf{r}) \right| \mathbf{k}+\mathbf{G}' \right\rangle = \frac{1}{\Omega}\int V(\mathbf{r})\ e^{i\ (\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}} d\mathbf{r} = \tilde{V}(\mathbf{G}-\mathbf{G}') \tag{4.13}$$

where $\tilde{V}(\mathbf{G}-\mathbf{G}')$ is the Fourier transform of the potential, and the kinetic term is diagonal.

4.3.1 Energy cutoff

Summarizing, Bloch's theorem implies that the wave function of an electron in a periodic potential can be expanded in a plane-wave (PW) basis set. The $\mathbf{G}$ vectors allowed in the PW expansion are the reciprocal lattice vectors and, in principle, an infinite number of plane waves is required by the theory. However, the Fourier coefficients $\Psi_{\mathbf{k}}(\mathbf{G})$ of the wave functions decrease with increasing $|\mathbf{k}+\mathbf{G}|$, so that the PW expansion can be effectively truncated at a finite number of terms, e.g. limited to all waves with kinetic energy lower than some particular energy cutoff E_{cut}. The truncation of the basis set leads to an error in the computed physical quantities, but this error can be easily handled by increasing the cutoff. Since this implies to increase the size of the basis set without modifying the hamiltonian, then the energy should decrease variationally with E_{cut}. This is at variance with other types of basis sets (localized, for example), where the fact of increasing the basis size does not necessarily mean that the energy will decrease.

Large $\mathbf{G}$ vectors are associated with the description of short-range features in real space. Therefore, a spatial scale must exist such that the wave functions become so smooth that decreasing the spatial grid spacing does not introduce any relevant information. In that case, it is said that the system is *at convergence* in plane waves. If the BZ of the supercell is sampled with many $\mathbf{k}$-points, then the energy cutoff will depend on $\mathbf{k}$ through the relation: $E_{cut}^{\mathbf{k}} = \left|\mathbf{k}+\mathbf{G}_{cut}\right|^2/2$. Typically the variation of $E_{cut}^{\mathbf{k}}$ with $\mathbf{k}$ is very small because $|\mathbf{k}| < |\mathbf{G}_{\min}|$. It is basically the effect of moving a sphere out of center. The number of basis functions changes discontinuously with E_{cut} because the $\mathbf{G}$ vectors are ordered in shells of equal modulus $|\mathbf{G}|$, so that the number of PW will slightly depend on $\mathbf{k}$.

4.3.2 Advantages and disadvantages of plane waves

The main advantages of using a truncated PW basis set are the following:

1. The Kohn-Sham hamiltonian has a kinetic term that is diagonal in reciprocal space, and a potential term that is local in real space (pseudopotential methods introduce a nonlocal component of the potential, but we shall see this in detail below). This feature can be exploited to speed up the calculations by transforming the wave functions and the density back and forth from real to reciprocal space and viceversa, and calculating the kinetic and potential contributions in the space where they are diagonal. The transformation can be done very efficiently by using Fast Fourier Transform (FFT) techniques.

2. The calculation of the energy, forces on the orbitals, forces on the nuclei, and stresses is very simple in PW.

3. The PW basis set is floating, in the sense that the basis functions are not attached to any particular atom. The basis functions represent with the same accuracy all regions of space. There are no additional forces on the nuclei that arise from the derivation of the basis functions. This is a conseuquence of the fact that Hellmann-Feyman theorem can be strictly applied only when the basis set is very well converged, or when the basis functions do not depend on the nuclear coordinates. Using localized basis sets, the correction for the finiteness of the basis is very important, and gives rise to the so-called Pulay forces.

On the other side, the disadvantages are the following:

1. For molecules, wires and surfaces, a lot of computational effort is used to deal with the vacuum that fills the supercell. This is very different from the case of localized basis.

2. Systems with rapid variations of the wave functions close to the nuclei need a very high energy cutoff (many PW components). Localized basis are much better in this because they are normally tuned to reproduce atomic wave functions. This is important in hydrogen, first-row elements, and transition metals.

4.4 Atomic first-principles pseudopotentials

The electronic states of an atom can be separated into: (1) *core states*, which are highly localized and not affected by the chemical environment, (2) *valence states*, which are extended and responsible for chemical binding, and (3) *semicore states*, which are localized and slightly affected by the environment, but contribute to the chemical binding. The most common pseudopotential approach consists of not allowing to relax core states according to the environment (**frozen core approximation**), although some polarizable core approaches have been proposed. In general, this is a very good approximation that gives total atomic energies within 0.01 eV. Semicore states are often treated as part of the frozen core, but when their contribution to binding is important, they have to be included in the valence.

The valence states, due to orthogonalization with respect to the core states of the same symmetry, show a marked oscillatory behaviour with a number of nodes equal to $n - l - 1$, with n the principal quantum number and l the angular momentum. Nodeless wave functions ($l = n-1$) are not oscillatory but, due to the lack of orthogonalization, the electrons can approach the nucleus with less difficulty and create strongly bound states which are steeply peaked close to the nucleus. This is the case of the $1s$ state in H, the $2p$ states in C, N, O and F, and the $3d$ states in transition metals.

When the basis set choosen is that of plane waves, then the computation of matrix elements requires the use of the Fourier decomposition of the potential. Features like the above ones are very bad for PW, because they need a very large number of basis functions to achieve convergence in the expansion, and this translates in a vast amount of computer time (the dimension of the matrix to diagonalize is too large).

Pseudopotential theory is constructed in two steps:

- Core electrons are removed from the calculation, and the interaction of the valence electrons with the nucleus plus the core states (including orthogonalization) is replaced by an effective, screened potential. The screened potential depends on the angular momentum of the valence electrons because of the different orthogonality conditions. For instance in the C atom, the $2s$ valence state has to be orthogonal to the $1s$ core state, but the $2p$ valence state does not feel the orthogonality constraint (exchange interaction) of the $1s$ state because they have different quantum numbers. Therefore, within the core region, these two states feel very different potentials from the ionic core. Of course, at large distances the potential is $-Z_V/r$ independently of the angular momentum, because the ionic core is seen as a point charge (Z_V is the valence charge, *i.e.* the charge of the ionic core). For each angular momentum l, the pseudopotential must have the atomic valence l-state as the ground state.

Example

	core			valence	
True Si atom	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^2$
Pseudo Si atom		—		$1s^2$	$2p^2$

- The full ion-electron interaction, which includes the orthogonality of the valence wave function to the core states, is replaced by a weaker pseudopotential that acts on a pseudo-wave function rather than the true wave function. The pseudopotential is constructed in such a way that its scattering properties or phase shifts for the pseudo-wave functions are the same as those of the true potential for the true valence wave function, but in such a way that the radial pseudo-wave function has no nodes inside the core region (see below).

It can be shown [80] that a smooth valence wave function Φ_v (not orthogonalized to the core states) constructed as $\Psi_v = \Phi_v - \sum_c \alpha_{cv} \Psi_c$ (where $\alpha_{cv} =< \Psi_c | \Phi_v >$) verifies that $\left[\hat{H} + \sum_c (\varepsilon_v - \varepsilon_c) |\Psi_c >< \Psi_c|\right] \Phi_v = \varepsilon_v \Phi_v$. Therefore, it is possible to find an exact pseudohamiltonian $\hat{H}_{PS} = \hat{H} + \sum_c (\varepsilon_v - \varepsilon_c) |\Psi_c >< \Psi_c|$ with the same eigenvalue as the original hamiltonian, but a smooth, nodeless wave function. From this expression it is clear that the pseudopotential will act differently on wave functions of different angular momentum. The most general form for a pseudopotential of this king is, then, the following:

$$V_{PS}(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \Big| lm > V_{PS}^{l}(r) < lm \Big| \tag{4.14}$$

where $| \, lm >$ are the spherical harmonics and V_{PS}^{l} is the pseudopotential for the angular component l. Acting on the electronic wave function, this operator decomposes it into its spherical harmonics components, each of which is then multiplied by the corresponding pseudpotential. These pseudpotentials are usually called *non-local* becuase of this property of differentiating the angular components, which is a consequence of the non-local exchange with the core. In practice, $V_{PS}^{l}(r)$ is a local operator in the radial coordinate. Therefore, a better name for this type of expression is a *semi-local* pseudopotential. If all the angular components of the pseudopotential are taken to be the same, then it is said to be a *local* pseudopotential. In principle, local versions can be constructed that verify the scattering properties for all angular momenta, but they are not smooth and weak functions. That is why it is computationally more effective to deal with non-local pseudopotentials.

Normally, only a few different angular momenta are present in the core, meaning that for values of $l > l_{\max}$ the pseudopotential acts in the same way. Therefore, the summation in (4.14) can be cut at $l_{\max}$ and recasted in the following form:

$$V_{PS}(\mathbf{r}) = \sum_{l=0}^{\infty} V_{PS}^{loc}(r) \, \hat{P}_l + \sum_{l=0}^{l_{\max}} \left[V_{PS}^{l}(r) \, - V_{PS}^{loc}(r) \right] \hat{P}_l = V_{PS}^{loc}(r) + \sum_{l=0}^{l_{\max}} \Delta V_{PS}^{l}(r) \, \hat{P}_l \tag{4.15}$$

where $\Delta V_{PS}^{l}(r)$ are short-range functions confined to the core region, and $V_{PS}^{loc}(r)$ is an average local potential that contains the screened Coulomb interaction. It is customary to take the local component as the pseudopotential for the first angular momentum that is not represented in the core, *e.g.* $V_{PS}^{p}(r)$ for first-row elements, or $V_{PS}^{d}(r)$ for Si.

4.4.1 How to construct a pseudopotential ?

It is clear that there is an enormous freedom in the way pseudopotentials can be constructed. Empirical pseudopotentials determined by fitting some experimental quantities have been very popular in the past, but they lacked a very important property which is transferability, namely that a pseudopotential constructed for some specific environment would not be useful for the same atomic species in a different environment. The first non-empirical approach to pseudopotentials was the one devised by Phillips and Kleinman [80] (see above). This approach, however, has a problem: the normalized pseudo-wave function has a different amplitude than the all-electron wave function, although outside the core the shapes are the same, and this is not acceptable because an incorrect valence charge distribution leads to errors in the bonding properties. However, this is not a problem of pseudopotentials in general, but of this paticular construction.

The construction of a pseudopotential is actually an inverse problem: given a nodeless psuedo-wave function, which outside some cutoff radius r_c decays exactly as the all-electron wave function, the inversion of the radial Schrödinger equation yields a pseudopotential which has the pseudo-wave function as its eigenfunction, at the correct eigenvalue. This means that the scattering property of the full potential inside r_c is correctly mimicked by the pseudopotential at that particular eigenvalue. This property is called *norm conservation.* Hamann, Schlüter and Chiang [81] leaded a revolution in this field in the late seventies when they proposed a procedure to construct non-local norm-conserving pseudopotentials fitted to first-principles atomic calculations. Moreover, they showed that the norm-conserving condition implies that the pseudopotential and the full potential have the same energy variation to first order when transferred to different environments. This property is known as *transferability*, and it is the one that makes pseudopotentials useful for electronic structure calculations. The norm conservation condition implies the continuity of the logarithmic derivative of the radial wave function — $d \ln R_l(r, \varepsilon)/dr$ — at the cutoff radius r_c, which is simply related to the scattering phase shift.

The norm conservation constraint does not guarantee that the pseudopotential is useful in any energy range, but only in environments such that the eigenvalues does not depart significantly from the eigenvalues used in its construction. For instance, a pseudopotential for H in the H_2 molecule will not be useful for hydrogen at high pressures because the energy ranges are completely different, but a pseudo for Si constructed having in mind the bulk solid will be useful for the Si surface or for liquid Si under similar PT conditions. The property that makes this range wider is the loosely defined concept of smoothness, i.e. the smoother the pseudo, the weaker the energy dependence. The easier recipe for transferability is to reduce the cutoff radius.

The conditions proposed by Hamann, Schlüter and Chiang for the construction of psuedopotentials are the following:

1. Φ_{ps} is nodeless, and it is identical to the all-electron wave function outside a suitably choosen cutoff radius r_c:

$$\Phi_{ps}(r) = \tilde{\Phi}_{ps}(r) \quad \text{for} \quad r < r_c \tag{4.16}$$

   and

$$\Phi_{ps}(r) = \Phi_{ae}(r) \quad \text{for} \quad r \geq r_c \tag{4.17}$$

2. The first and second derivatives of the pseudo-wave function are continuous at r_c.

3. The eigenvalues of the pseudo-wave functions coincide with those of the all-electron wave functions.

4. The norm of the true and pseudo wave functions inside the core region is the same (norm-conservation condition):

$$\int_0^{r_c} \left| r \, \tilde{\Phi}_{ps}(r) \right|^2 dr = \int_0^{r_c} \left| r \, \Phi_{ae}(r) \right|^2 dr \tag{4.18}$$

5. Other conditions to enhance the smoothness of the potentials.

Several schemes have been proposed to generate first-principles pseudopotentials that satisfy the above conditions, differing mainly in functional form of the potentials and the smoothness conditions. The most popular for a long time, due to its simple form suitable for analytic integration both, in PW and Gaussian basis sets calculations, were the ones proposed by Bachelet, Hamann and Schlüter [82], in which the pseudopotentials are fitted to the following form:

$$V_{PS}^{l}(r) = -\frac{Z_V}{r}\left[C \operatorname{erf}\left(\sqrt{\alpha_1^{core}}r\right) + (1-C)\operatorname{erf}\left(\sqrt{\alpha_2^{core}}r\right)\right] + \qquad (4.19)$$

$$+ \sum_{i=1}^{3}(A_{l,i} + r^2 A_{l,i+3}) \exp\left(-\alpha_{l,i}\, r^2\right) \qquad (4.20)$$

4.4.2 Troullier-Martins pseudopotentials

Today, the smoothest norm-conserving pseudopotentials are obtained using the recipe by Troullier and Martins [83], who thoroughly studied the convergence properties of the PW expansion of the pseudopotential. They generalized Kerker [84] scheme by proposing the following analytic form of the wave function inside the cutoff radius:

$$R_l^{PP}(r) = r^l \exp[p(r)]$$

with $p(r) = c_0 + \sum_{i=2}^{n} c_i\, r^i$. The r^l behaviour for small r is to avoid a hard core pseudopotential with a singularity at the origin. In Kerker's scheme (where $n = 4$), the four coefficients of the polynomial are determined by the conditions: (i) charge conservation inside the cutoff radius; (ii)-(iv) continuity of the pseudo-wave function and its two first derivatives at the cutoff radius. Troullier and Martins added variational freedom in the search for smoothness by increasing the order of the polynomial. They realized that the asymptotic, large wave number behaviour of the pseudopotential depends on the values of its odd derivatives at the origin. This implies that a larger degree of smoothness is achieved when all odd coefficients in the polynomial are set to zero. Additionally, they found that pseudopotentials that are flat at the origin are also smoother. With these ingredientes they provided the following practical recipe:

The polynomial is choosen of sixth order in r^2: $p(r) = c_0 + c_2\, r^2 + c_4\, r^4 + c_6\, r^6 + c_8\, r^8 + c_{10}\, r^{10} + c_{12}\, r^{12}$, and the coefficients are determined by the following seven conditions:

1. Norm conservation of the charge within the cutoff radius r_c:

$$2c_0 + \ln\left\{\int_0^{r_c} r^{2(l+1)} \exp\left[2p(r) - 2c_0\right] dr\right\} = \ln\left\{\int_0^{r_c} r^2 \left|R_l^{AE}(r)\right|^2 dr\right\}$$

2. Continuity of the pseudo-wave function and its first 4 derivatives at r_c, which in practice can be written as:

   -
     $$p(r_c) = \ln\left[\frac{P(r_c)}{r_c^{l+1}}\right]$$

- $$p'(r_c) = \frac{P'(r_c)}{P(r_c)} - \frac{l+1}{r_c}$$

- $$p''(r_c) = 2V_{AE}(r_c) - 2\varepsilon_l - \frac{2(l+1)}{r_c} p'(r_c) - \left[p'(r_c)\right]^2$$

- $$p'''(r_c) = 2V'_{AE}(r_c) + \frac{2(l+1)}{r_c^2}\, p'(r_c) - \frac{2(l+1)}{r_c}\, p''(r_c) - 2\, p'(r_c)\, p''(r_c)$$

- $$\begin{aligned} p''''(r_c) &= 2V''_{AE}(r_c) - \frac{4(l+1)}{r_c^3}\, p'(r_c) + \frac{4(l+1)}{r_c^2}\, p''(r_c) - \frac{2(l+1)}{r_c}\, p'''(r_c) - \\ &- 2\left[p''(r_c)\right]^2 - 2\, p'(r_c)\, p'''(r_c) \end{aligned}$$

  where $P(r) = r\, R_l^{AE}(r)$, and $V_{AE}(r)$ is the all-electron atomic screened potential (see below).

3. Zero curvature of the screened pseudopotential at the origin, $V''_{sc,l}(0) = 0$, which translates into: $c_2^2 + c_4(2l+5) = 0$.

The derivatives of the wave function and screened potentials are obtained from the numerical all-electron wave functions and screened potential using seventh-order finite differences.

The general procedure for obtaining a pseudopotential begins by solving the all-electron radial Schrödinger equation:

$$\left\{ -\frac{1}{2}\frac{d}{dr^2} + \frac{l(l+1)}{2r^2} + V[\rho; r] \right\} r\, R_{n,l}^{AE}(r) = \varepsilon_{n,l}\, r\, R_{n,l}^{AE}(r) \tag{4.21}$$

where

$$V[\rho; r] = -\frac{Z}{r} + \int \frac{\rho(r')}{|r - r'|} dr' + \mu_{XC}[\rho]$$

and $\rho(r)$ is the sum of the electronic densities for the occupied wave functions.

Then, the pseudo-wave function constructed according to the above prescription is used to invert (this can always be done because of the nodeless condition) the radial Schrödinger equation for the screened pseudopotential:

$$V_{sc,n,l}^{PP}(r) = \varepsilon_{n,l} - \frac{l(l+1)}{2r^2} + \frac{1}{2rR_{n,l}^{PP}(r)} \frac{d^2}{dr^2}[rR_{n,l}^{PP}(r)]$$

The ionic pseudopotential is finally obtained by subtracting (unscreening) the Hartree and exchange-correlation potentials calculated only for the valence electrons (with the valence pseudo-wave functions):

$$V_{n,l}^{PP}(r) = V_{sc,n,l}^{PP}(r) - \int \frac{\rho_v(r')}{|r-r'|}dr' - \mu_{XC}[\rho_v]$$

with $\rho_v(r) = \sum_{i=nc+1}^{n} \sum_{l=0}^{i-1} \left| r R_{i,l}^{PP}(r) \right|^2$. Relativistic expresions based on Dirac's equation, instead of Schrödinger's, have to be used in the case of heavy atoms. They can be found, *e.g.* in Ref. [82].

4.4.3 Non-linear exchange-correlation core corrections (NLCC)

If there is an overlap between the core and valence charge densities, the unscreening process in the construction of the pseudopotentials leads to some errors because the exchange-correlation potential is not a linear function of the density. A solution to this problem was proposed by Louie, Froyen, and Cohen [85]:

1. Replace the above unscreening expresion by:

$$V_l^{PP}(r) = V_{sc,l}^{PP}(r) - \int \frac{\rho_v(r')}{|r-r'|}dr' - \mu_{XC}[\rho_v + \rho_c]$$

2. In the electronic structure calculations performed with this pseudopotential, compute the exchange-correlation contribution for the whole electronic charge, $\rho_v + \rho_c$, instead of the usual valence charge.

3. Since $\rho_c(r)$ does not converge rapidly in reciprocal space, it is replaced by:

$$\rho_c(r) = A \sin(Br)/r \qquad \text{for} \quad R \leq R_0$$

where the parameters A and B are determined by the continuity condition for ρ_c and its first derivative at the cutoff radius R_0.

4.5 The Pseudopotential Plan Wave (PPW) method

The central, and most computationally intensive issue in electronic structure calculations is the self-consistent solution of Kohn-Sham equations. The first step is to calculate the Kohn-Sham hamiltonian matrix elements. Since quite a large number of density-functional electronic structure calculations, and most first-principles MD simulations up to date, have been carried out within the pseudopotential plane waves (PPW) framework described above, we shall focus in the following on the computation of matrix elements within the PPW scheme.

Kohn-Sham equations in the PW basis set are written: $\sum_{\mathbf{G}'} H_{\mathbf{k+G},\mathbf{k+G}'}^{KS} \; C_{\mathbf{k+G}'}^{(j)} = \varepsilon^{(j)} \; C_{\mathbf{k+G}}^{(j)}$, and the hamiltonian matrix elements are:

$$H_{\mathbf{k+G},\mathbf{k+G}'}^{KS} = \left\langle \mathbf{k}+\mathbf{G} \left| -\frac{\hbar^2}{2m}\nabla^2 + V_H(r) + V_{PS}^{loc}(r) + \sum_{l=0}^{l_{\max}} \Delta V_{PS}^{l}(r)\, \hat{P}_l + \mu_{XC}[\rho] \right| \mathbf{k}+\mathbf{G}' \right\rangle \tag{4.22}$$

where the real-space expression of the basis functions is:

$$\langle \mathbf{r} \mid \mathbf{k}+\mathbf{G}\rangle = \frac{e^{i\ (\mathbf{k}+\mathbf{G}).\mathbf{r}}}{\sqrt{\Omega}} \quad . \tag{4.23}$$

4.5.1 Kinetic term

The expression for the kinetic energy — the first term in (4.22) — is very simple because PW are precisely the solutions of the Laplace equation, which corresponds to free particles (only kinetic energy). This implies that the kinetic operator is diagonal in reciprocal space, and it matrix elements are:

$$\left\langle \mathbf{k}+\mathbf{G} \left| -\frac{\hbar^2}{2m}\nabla^2 \right| \mathbf{k}+\mathbf{G}' \right\rangle = \frac{\hbar^2}{2m}\left|\mathbf{k}+\mathbf{G}\right|^2 \delta_{\mathbf{G},\mathbf{G}'} \quad . \tag{4.24}$$

4.5.2 Local potential

The matrix elements of the local part of the potential, $V^{loc}(r) = V_H(r) + V^{loc}_{PS}(r)$, in the PW basis set are given simply by the Fourier transform:

$$\left\langle \mathbf{k}+\mathbf{G} \left| V^{loc}(r) \right| \mathbf{k}+\mathbf{G}' \right\rangle = \frac{1}{\Omega}\int V^{loc}(r)\exp[i\,(\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}]\ \ d\mathbf{r} = \tilde{V}^{loc}(\mathbf{G}-\mathbf{G}') \quad . \tag{4.25}$$

The local pseudopotential part can be written as:

$$\tilde{V}^{loc}_{PS}(\mathbf{G}) = \sum_{I=1}^{P}\frac{1}{\Omega}\int v^{loc}_{PS}(r-\mathbf{R}_I)\ \ \exp[i\ \mathbf{G}\cdot\mathbf{r}]\ d\mathbf{r} = \sum_{\alpha=1}^{N_\alpha} S_\alpha(\mathbf{G})\,\tilde{v}^{\alpha}_{PS}(G) \tag{4.26}$$

where the sum is over the different atomic species (up to N_α), $\tilde{v}^{\alpha}_{PS}(G)$ is the Fourier transform of the local component of the atomic pseudopotential, and

$$S_\alpha(\mathbf{G}) = \sum_{I=1}^{P_\alpha}\exp[i\ \mathbf{G}\cdot\mathbf{R}^{\alpha}_I] \tag{4.27}$$

is the atomic structure factor for species α, which contains all the information about the positions of the ionic cores. This atomic structure factor should not be confused with what is usually know as structure factor, *i.e.* the function obtained by Fourier transforming the pair correlation function. The latter is a statistical average, while the former corresponds to a single configuration.

The Hartree potential is given by the following convolution:

$$V_H(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}' \tag{4.28}$$

or, equivalently, to the Poisson equation

$$\nabla^2 V_H(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{4.29}$$

The Poisson equation can be solved in many ways, *e.g.* by overrelaxation or multigrid methods. However, when the supercell is periodically repeated, as implicitly emboddied in the PW approach, the most efficient way to solve it is by recalling the fact that the Fourier transform of a convolution is the product of the Fourier transforms:

$$\langle \mathbf{k}+\mathbf{G}\,|V_H|\,\mathbf{k}+\mathbf{G}'\rangle = \tilde{V}_H(\mathbf{G}-\mathbf{G}') = \rho(\mathbf{G}-\mathbf{G}')\,\frac{4\pi}{|\mathbf{G}-\mathbf{G}'|^2} \tag{4.30}$$

where $4\pi/\left|\mathbf{G}\right|^2$ is the Fourier transform of the bare Coulomb potential. The Fourier transformations can be very efficiently done by using the widely spread Fast Fourier Transform technique (FFT), which reduces the computational cost of this part of the calculation from order M^2 to order $M\log M$, with M the number of plane waves in the basis.

Both, the matrix elements of the Hartree potential and of the local component of the pseudopotential diverge as $1/\left|\mathbf{G}\right|^2$ for $|\mathbf{G}|\to 0$. The case of V_H, is it obviously $4\pi ZN/\left|\mathbf{G}\right|^2$, but for $\tilde{V}_{PS}^{loc}$, it is a little more subtle. The bare ion-electron Coulomb interaction would be $-Z/r$, corresponding to $-4\pi(\sum_\alpha Z_\alpha P_\alpha)/\left|\mathbf{G}\right|^2$ in reciprocal space. The local pseudopotential interaction has the same long-range behaviour, since at long distances the ionic cores look like point-like particles, but at short distances the pseudopotential departs from bare Coulomb. However, the $-4\pi(\sum_\alpha Z_\alpha P_\alpha)/\left|\mathbf{G}\right|^2$ divergence is still present because it is a consequence of the long-range Coulomb tail. The sum of the two contributions is $4\pi Z(N-\sum_\alpha Z_\alpha P_\alpha)/\left|\mathbf{G}\right|^2 = 4\pi Q_T/\left|\mathbf{G}\right|^2$, with Q_T the total charge in the supercell. Therefore, in order to avoid the $\mathbf{G}=0$ divergence in the local (Hartree plus pseudo) potential, an electronic structure calculation for a periodic system can only be performed under charge neutrality conditions. This is a consequence of the long range sums of Coulomb-like interactions in infinite systems, and it is not present in calculations that do not verify PBC. Plane waves authomatically imply PBC, so that it is not possible to deal with charged systems within a PW approach in a straightforward way. If one is interested in dealing with charged systems, the customary approach (although not formally justified) is to artificially neutralize the charge in the supercell by adding uniformely distributed charge of the opposite sign. In practice, this simply means to ignore the $\mathbf{G}=0$ term of the potential as usual and to add a background energy to the total energy.

4.5.3 Non-local pseudopotential

The only contribution that is somewhat more complicated is the one for the non-local components of the pseudopotential. For a particular l-component, the matrix elements for an atom located at the origin are:

$$\begin{aligned}
\Delta V^l_{\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}'} &= \left\langle \mathbf{k}+\mathbf{G}\left|\Delta V^l_{PS}\hat{P}_l\right|\mathbf{k}+\mathbf{G}'\right\rangle = \qquad (4.31)\\
&= \sum_{m=-l}^{l}\left\langle \mathbf{k}+\mathbf{G}\left|Y_{lm}\right.\right\rangle \Delta V^l_{PS}(r)\left\langle Y_{lm}\left|\mathbf{k}+\mathbf{G}'\right.\right\rangle = \\
&= 4\pi(2l+1)\,P_l(\cos\theta_{\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}'})\int r^2\, j_l\left(|\mathbf{k}+\mathbf{G}|\,r\right)\,\Delta V_l^{PS}(r)\; j_l\left(|\mathbf{k}+\mathbf{G}'|\,r\right)\,dr
\end{aligned}$$

where $P_l(\cos\theta)$ are Legendre polynomials, and $j_l(x)$ are the spherical Bessel functions. The straightforward calculation of the above matrix elements involves the calculation of $M(M+1)/2$ radial integrals. This was the original way of computing non-local pseudopotential contributions, where the integration was done, *e.g.* via Gauss-Hermite polynomials, to save computer time.

A more efficient way of computing these non-local contributions was devised in 1982 by Kleinman and Bylander [86], who introduced a modified projection procedure onto the different angular momentum states. Instead of the original form $\Delta\hat{V}^l_{PS} = \sum_m \left|Y_{lm}\right\rangle \Delta V^l_{PS}(r) \left\langle Y_{lm}\right|$, they proposed a fully separable projector which uses a single basis state per angular momentum:

$$\Delta\hat{V}^l_{KB} = \sum_{m=-l}^{l} \frac{\left|\Phi^{PS}_{lm}\,\Delta\hat{V}^l_{PS}\right\rangle \left\langle\, \Phi^{PS}_{lm}\,\Delta\hat{V}^l_{PS}\right|}{\left\langle\Phi^{PS}_{lm}\left|\Delta\hat{V}^l_{PS}\right|\Phi^{PS}_{lm}\right\rangle} \tag{4.32}$$

where Φ^{PS}_{lm} are the pseudo-atomic wave functions defined in the preceeding section. When this modified projector is applied to a pseudo-atomic wave function, it gives identical results to the original projector, as can be easily seen from (4.32).

The matrix elements of this new operator are, considering now the full system:

$$\Delta V^{KB,l}_{\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}'} = \sum_{m=-l}^{l}\sum_{I=1}^{P} F^*_{lm,I}(\mathbf{k}+\mathbf{G})\; F_{lm,I}(\mathbf{k}+\mathbf{G}') \tag{4.33}$$

with

$$\begin{aligned} F_{lm,I}(\mathbf{k}+\mathbf{G}) &= \frac{\left\langle\Phi^{PS}_{lm}\left|\Delta\hat{V}^l_{PS}\right|\mathbf{k}+\mathbf{G}\right\rangle}{\sqrt{\left\langle\Phi^{PS}_{lm}\left|\Delta\hat{V}^l_{PS}\right|\Phi^{PS}_{lm}\right\rangle}} = \\ &= \frac{e^{i\,(\mathbf{k}+\mathbf{G})\cdot\mathbf{R}_I}\int r^2\,\Phi^{PS}_{lm}(r)\;\Delta V^{PS}_l(r)\;j_l\left(|\mathbf{k}+\mathbf{G}|\,r\right)\,dr}{\sqrt{\int r^2\,\left|\Phi^{PS}_{lm}(r)\right|^2\,\Delta V^{PS}_l(r)\;dr}} \quad . \end{aligned} \tag{4.34}$$

The advantage of this formulation is obvious: now the calculation of the matrix elements involves the evaluation of only the M integrals $F_{lm}(\mathbf{k}+\mathbf{G})$, instead of the former $M(M+1)/2$ ones. These integrals are then simply multiplied amongst themselves.

When the Kleinman-Bylander form is applied in an environment different from the isolated atom, it does not produce identical results than the original semilocal operator, because the wave function is not projected onto a radially complete set of spherical harmonics. In general this is not a problem, but it sometimes produces some difficulties which are already well-known and under control [87]. These are the so-called *ghost states*, which are visualized as an unphysical divergence in the logarithmic derivative (signature of an eigenstate) of the pseudo-radial wave function at an energy below that of the true valence state. Such divergences are a consequence of the choice of the local potential, and are reflected in a too large value of the quantity $E^{KB}_l = \left[\left\langle\Phi^{PS}_{lm}\left|\Delta\hat{V}^l_{PS}\right|\Phi^{PS}_{lm}\right\rangle\right]^{-1/2}$. A test for ghost states consists of looking at the two lowest eigenvalues of the atomic hamiltonian without the non-local contributions, $E^{loc0}_l$and E^{loc1}_l. If, in the application, $E_l < 0$, then a ghost state exists if and only if $E^{loc0}_l < E_l$, and also if $E_l > 0$, a ghost state exists if and only if $E^{loc1}_l < E_l$ [88].

4.5.4 Exchange-correlation

The exchange-correlation potential is local in real space, so that it is usually calculated by evaluating the appropriate expression for the electronic density (and the gradient of the density in the case of GGA) calculated on the real-space mesh associated with the PW expansion. Then, $\mu_{XC}(\mathbf{r})$ is transformed to reciprocal space using the FFT.

4.5.5 Total energy

The total energy per supercell of the infinite periodic system is given by the Kohn-Sham expression:

$$E_{KS}[\rho] = \frac{1}{N_{cell}} \Big(T_e[\rho] + E_H[\rho] + E_{PS}^{loc} + E_{ii} + E_{PS}^{nl} + E_{XC}[\rho] \Big) \quad . \tag{4.35}$$

where the factor N_{cell} in the denominator stands only for normalizing the energy to one supercell.

The three local electrostatic contributions: E_H, E_{PS}^{loc}, and E_{ii}, diverge when taken individually, in the same way as the local potential operator. The reason for these divergencies, which are problematic from the computational point of view, is that the $\mathbf{G} = \mathbf{0}$ term in the reciprocal space expansion of the energy corresponds to the monopolar term in a multipolar expansion. The Hartree term, for example, takes into account only the electron-electron interaction, so that the coefficient acompanying $4\pi/\left|\mathbf{G}\right|^2$ is $N^2/2\Omega$. The E_{ii} term contributes with $(\sum_\alpha Z_\alpha P_\alpha)^2/2\Omega$ and E_{PS}^{loc} with $-(\sum_\alpha Z_\alpha P_\alpha)N/\Omega$. It is easy to see that the only way to avoid the divergence of the energy is by taking these three contributions together and, exactly as before, the non-divergence condition is that $\sum_\alpha Z_\alpha P_\alpha = N$, *i.e.* charge neutrality in the supercell.

Let us then calculate the contribution of these three terms:

$$\begin{aligned} E_H + E_{PS}^{loc} + E_{ii} &= \frac{1}{2} \int\int \frac{\rho_T(\mathbf{r})\, \rho_T(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\; d\mathbf{r}' + \left(E_{PS}^{loc} - \int\int \frac{\rho(\mathbf{r})\;\rho_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\; d\mathbf{r}' \right) \\ &\quad + \left(E_{ii} - \frac{1}{2} \int\int \frac{\rho_i(\mathbf{r})\, \rho_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\; d\mathbf{r}' \right) \end{aligned} \tag{4.36}$$

where $\rho_i(\mathbf{r})$ is a charge distribution associated with the ionic subsystem, and $\rho_T(\mathbf{r}) = \rho(\mathbf{r}) + \rho_i(\mathbf{r})$ is the total charge density, which is taken to be neutral. The calculation of these double integrals is very expensive, but fortunately this operation can be efficiently done in reciprocal space by using FFTs. The first term is a simple convolution integral, and can be written

$$\frac{1}{N_{cell}} \frac{1}{2} \int\int \frac{\rho_T(\mathbf{r})\, \rho_T(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\; d\mathbf{r}' = \frac{\Omega}{2} \sum_{\mathbf{G}\neq 0} \frac{4\pi}{G^2}\, \tilde{\rho}_T(\mathbf{G})\; \tilde{\rho}_T(-\mathbf{G}) \tag{4.37}$$

The second term ($\tilde{E}_{PS}^{loc}$ for short) can also be reduced to an expression of the same kind by using (4.26), which gives

$$
\begin{aligned}
\frac{1}{N_{cell}}\tilde{E}_{PS}^{loc} &= \frac{1}{N_{cell}}\sum_{\alpha=1}^{N_\alpha}\sum_{I=1}^{P_\alpha}\int \rho(\mathbf{r})\, v_{PS}^{\alpha,loc}(\mathbf{r})\, d\mathbf{r} - \frac{1}{N_{cell}}\int\!\!\int \frac{\rho(\mathbf{r})\,\rho_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\; d\mathbf{r}' = \\
&= \sum_{\mathbf{G}\neq 0}\sum_{\alpha=1}^{N_\alpha}\left[S_\alpha(\mathbf{G})\,\tilde{v}_{PS}^{\alpha}(G) - \frac{4\pi}{G^2}\tilde{\rho}_i(\mathbf{G})\right]\tilde{\rho}(-\mathbf{G})
\end{aligned}
\tag{4.38}
$$

(notice that again the $\mathbf{G} = \mathbf{0}$ term vanishes because $S_\alpha(0) = \tilde{\rho}_i(0) = \sum_\alpha Z_\alpha P_\alpha$, the total ionic charge).

The third term involves the computation of the Coulomb energy of a periodic collection of point-like particles. A real-space evaluation of such an energy is extremely difficult because the Coulomb interaction is long-ranged, and summations are conditionally convergent. Coulomb interactions are also long-ranged in reciprocal space, so that a simple Fourier transformation does not solve the problem. The solution to this problem was given already in 1917 by Ewald [89], who developed a technique that consists of dividing the summation into a real-space and a reciprocal-space part, both of which are rapidly convergent. Ewald's method is based on the following identity:

$$
\begin{aligned}
\sum_{l=-\infty}^{\infty}\frac{1}{|\mathbf{R}_1 + l - \mathbf{R}_2|} &= \frac{2}{\sqrt{\pi}}\sum_{l=-\infty}^{\infty}\int_\eta^\infty \exp\left[|\mathbf{R}_1 + l - \mathbf{R}_2|^2\sigma^2\right] d\sigma + \\
&\quad + \frac{2\pi}{\Omega}\sum_{\mathbf{G}}\int_0^\eta \exp\left[-\frac{|\mathbf{G}|^2}{4\sigma^2}\right]\exp\left[i\left(\mathbf{R}_1 - \mathbf{R}_2\right)\cdot\mathbf{G}\right]\frac{1}{\sigma^3}\, d\sigma
\end{aligned}
\tag{4.39}
$$

which carries to the well-known expression for the *Ewald sums* [74]:

$$
\begin{aligned}
E_{ii} &= \frac{1}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P} Z_I Z_J \sum_{l=-\infty}^{\infty}\frac{1}{|\mathbf{R}_I + l - \mathbf{R}_J|} = \\
&= \frac{1}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P} Z_I Z_J\left[\sum_{l=-\infty}^{\infty}\frac{\operatorname{erfc}\left(|\mathbf{R}_I + l - \mathbf{R}_J|\,\eta\right)}{|\mathbf{R}_I + l - \mathbf{R}_J|}\right] - \frac{\eta}{\sqrt{\pi}}\sum_{I=1}^{P} Z_I^2 - \frac{\pi}{2\eta^2\Omega}\left|\sum_{I=1}^{P} Z_I\right|^2 + \\
&\quad + \frac{1}{\Omega}\sum_{\mathbf{G}\neq 0}\frac{4\pi}{G^2}\exp\left(-\frac{|\mathbf{G}|^2}{4\eta^2}\right)\sum_\alpha S_\alpha(\mathbf{G})\sum_\beta S_\beta(-\mathbf{G}) \quad ,
\end{aligned}
\tag{4.40}
$$

where erfc(x) is the complementary error function, which decays almost like a Gaussian, and $1/\eta$ is a cutoff distance which is appropriately choosen to optimize the convergence properties of the real and reciprocal space sums. This implies that the lattice sums can be cut at some value $l = l_{\max}$, and the reciprocal space sum can also be terminated at some cutoff G_{cut}. In practice, G_{cut} is determined already by the PW expansion of the electronic wave functions, and the value of η is choosen accordingly, to make the reciprocal space sum convergent within G_{cut}. The typical values of $l_{\max}$ are 0 or 1 at most. If η is choosen in such a way that $l_{\max} = 0$, then only the interaction with the closest images of the other atoms are considered. This is known as minimum image convention. When $I = J$ the $l = 0$ term in the first summation should be absent because it would imply that an ion interacts with itself. The terms $\frac{\eta}{\sqrt{\pi}}\sum_{I=1}^{P} Z_I^2$ precisely cancels this self-interaction.

The last term in the second line is the correction due to the fact that the $\mathbf{G} = \mathbf{0}$ term in the reciprocal space sum has been omitted. It appears because the expansion of the Gaussian around $\mathbf{G} = \mathbf{0}$ gives rise to a divergent term $4\pi/G^2$ plus a regular term π/η^2. When expression (4.40) is combined as in the last term of equation (4.36), the second term in parenthesis will cancel the divergence mentioned above. Moreover, if we choose the pseudo-ionic charge distribution $\rho_i(\mathbf{r})$ in the following way:

$$\begin{aligned} \rho_i(\mathbf{r}) &= \frac{\eta^3}{\pi^{3/2}} \sum_{I=1}^{P} Z_I \operatorname{erf}\left(-\eta^2 \left|\mathbf{r} - \mathbf{R}_I\right|^2\right) \\ \tilde{\rho}_i(\mathbf{G}) &= \sum_{\alpha} S_\alpha(\mathbf{G}) \exp\left(-\frac{|\mathbf{G}|^2}{4\eta^2}\right) \quad , \end{aligned} \tag{4.41}$$

the second and last terms in (4.40) exactly cancel with $\frac{1}{2}\int\int \frac{\rho_i(\mathbf{r})\,\rho_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\; d\mathbf{r}'$ in (4.36). The pseudo-ionic charge density can also be viewed as a smearing of the point-like ionic charge into the form of an error function for computational purposes (convergence of the lattice sums), and this is then corrected back.

In conclusion, the local electrostatic energy per supercell is rewritten in the following way:

$$\begin{aligned} \frac{1}{N_{cell}}\left(E_H + E_{PS}^{loc} + E_{ii}\right) &= \frac{\Omega}{2}\sum_{\mathbf{G}\neq 0}\frac{4\pi}{G^2}\,\tilde{\rho}_T(\mathbf{G})\;\tilde{\rho}_T(-\mathbf{G}) - \frac{\eta}{\sqrt{\pi}}\sum_{I=1}^{P} Z_I^2 + \\ &\quad + \sum_{\alpha=1}^{N_\alpha}\sum_{\mathbf{G}\neq 0} S_\alpha(\mathbf{G})\left[\tilde{v}_{PS}^{\alpha}(G) - \exp\left(-\frac{|\mathbf{G}|^2}{4\eta^2}\right)\right]\tilde{\rho}(-\mathbf{G}) + \\ &\quad + \frac{1}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P} Z_I Z_J \left[\sum_{l=-l_{\max}}^{l_{\max}} \frac{\operatorname{erfc}\left(|\mathbf{R}_I + l - \mathbf{R}_J|\,\eta\right)}{|\mathbf{R}_I + l - \mathbf{R}_J|}\right] \quad . \end{aligned} \tag{4.42}$$

The remaining terms in the total energy per supercell are:

- The exchange-correlation energy $E_{XC}[\rho] = \int \rho(\mathbf{r})\, \epsilon_{XC}[\rho]\, d\mathbf{r}$, which is easily calculated by integrating numerically the XC energy density in real-space.

- The kinetic energy $T_e[\rho]$, which can be computed with less effort in reciprocal space, where the kinetic operator is diagonal, *i.e.*

  $$T_e[\rho] = \frac{\hbar^2}{2m}\sum_{\mathbf{k}\in BZ}\omega_{\mathbf{k}}\sum_{i=1}^{N_{\mathbf{k}}} f_i^{\mathbf{k}} \sum_{\mathbf{G}} |\mathbf{k}+\mathbf{G}|^2 \left|\tilde{\Psi}_{\mathbf{k},i}(\mathbf{G})\right|^2 \tag{4.43}$$

  where $f_i^{\mathbf{k}}$ is the occupation number of state i at wave vector $\mathbf{k}$, and $\omega_{\mathbf{k}}$ are the weights of the $\mathbf{k}$-points for the BZ averages. For semiconductors it is customary to use special $\mathbf{k}$-points which exploit the symmetry of the system [64, 65]. For metals, the BZ sums are less rapidly convergent. The two standard alternatives are to use special points [66] in combination with a Fermi surface smearing technique [90, 91], or the linear tetrahedron method [92].

- The non-local pseudopotential energy, which is computed in reciprocal space in the following way:

$$E^{nl}_{PS} = \sum_{\mathbf{k}\in BZ} \omega_{\mathbf{k}} \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \sum_{I=1}^{P} \sum_{i=1}^{N_{\mathbf{k}}} f_i^{\mathbf{k}} \left| \sum_{\mathbf{G}} e^{i\ (\mathbf{k}+\mathbf{G})\cdot\mathbf{R}_I} F_{lm,I}(\mathbf{k}+\mathbf{G}) \tilde{\Psi}_{\mathbf{k},i}(\mathbf{G}) \right|^2 \qquad (4.44)$$

  with $F_{lm}(\mathbf{k}+\mathbf{G})$ given by (4.34).

- The electronic density, which is also needed in the calculation, is more easily calculated in real space:

$$\rho(\mathbf{r}) = \sum_{\mathbf{k}\in BZ} \omega_{\mathbf{k}} \sum_{i=1}^{N_{\mathbf{k}}} f_i^{\mathbf{k}}\ |\Psi_{\mathbf{k},i}(\mathbf{r})|^2 \quad . \qquad (4.45)$$

4.5.6 Forces on the nuclear coordinates

The Helmann-Feynman forces on the nuclear coordinates are needed in order to perform geometry optimizations and MD simulations. Within the PPW approach the forces are very simple to calculate and computationally inexpensive. In particular, the fact that the PW basis set is floating (it does not depend on the nuclear coordinates as in most localized basis sets) implies that the so-called Pulay forces [93], arising from the derivatives of the basis functions with respect to the nuclear coordinates, vanish identically. These forces will also vanish for atom-attached basis sets, but only in the case that the basis set is complete. Otherwise, they have to be explicitly calculated. All-electron methods usually involve the calculation of Pulay forces due to the augmentation spheres that move with the nuclei. Force theorems have, however, been devised by Methfessel and Schilfgaarde in order to avoid the computation of Pulay forces [94]. In the PPW methodology, the only terms that include a dependence on the nuclear coordinates are the pseudopotential (local and non-local parts), and the ion-ion interaction. The forces on the nuclei have the following expression:

$$\begin{aligned}
\mathbf{F}_I \ &= \ \frac{Z_I}{2} \sum_{J\neq I}^{P} Z_J \sum_{l=-l_{\max}}^{l_{\max}} \left[\frac{\mathrm{erfc}\left(|\mathbf{R}_I + l - \mathbf{R}_J|\,\eta\right)}{|\mathbf{R}_I + l - \mathbf{R}_J|^3} + \frac{\eta\ \exp\left(-\eta^2\,|\mathbf{R}_I + l - \mathbf{R}_J|^2\right)}{|\mathbf{R}_I + l - \mathbf{R}_J|} \right] \times \\
&\times \ (\mathbf{R}_I + l - \mathbf{R}_J) - 2 \sum_{\mathbf{k}\in BZ} \omega_{\mathbf{k}} \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \sum_{I=1}^{P} \sum_{i=1}^{N_{\mathbf{k}}} f_i^{\mathbf{k}} \left(\sum_{\mathbf{G}} e^{-i\,\mathbf{G}\cdot\mathbf{R}_I} F^*_{lm,I}(\mathbf{k}+\mathbf{G}) \tilde{\Psi}^*_{\mathbf{k},i}(\mathbf{G}) \right) \times \\
&\times \ \left(\sum_{\mathbf{G}'} i\,(\mathbf{k}+\mathbf{G}')\ e^{i\,\mathbf{G}'\cdot\mathbf{R}_I}\ F_{lm,I}(\mathbf{k}+\mathbf{G}')\tilde{\Psi}_{\mathbf{k},i}(\mathbf{G}') \right) - \\
&- \sum_{\alpha=1}^{N_\alpha} \sum_{\mathbf{G}\neq 0} i\,\mathbf{G}\ e^{i\,\mathbf{G}\cdot\mathbf{R}_I} \left[\tilde{v}^{\alpha}_{PS}(G) - \exp\left(-\frac{|\mathbf{G}|^2}{4\eta^2} \right) \right] \tilde{\rho}(-\mathbf{G}) \qquad (4.46)
\end{aligned}$$

Chapter 5

Electronic self-consistency: minimizing the energy functional

The hamiltonian matrix elements, whose computation within the pseudopotential plane waves (PPW) scheme was thoroughly described in the preceeding chapter, are the crucial ingredient needed to solve the electronic structure problem as emboddied in the self-consistent Kohn-Sham equations. The forces on the nuclear degrees of freedom, also described before, serve to integrate the newtonian equations of motion either as a means for optimizing the nuclear geometry, or to study the dynamical and statistical properties at finite temperature. In this chapter I describe the most commonly used numerical techniques to achieve these goals, once matrix elements and forces are known.

5.1 Minimization of the electronic energy functional

The central problem in electronic structure calculations within the density functional formalism is to minimize the energy as a functional of the electronic density. It can be attacked directly as a minimization problem by using standard techniques like steepest descent or conjugated gradients, or it can be reformulated in terms of the self-consistently Kohn-Sham set of equations. This latter approach is, in fact, the traditional procedure in electronic structure calculations, and it involves a nested procedure of diagonalizing the matrix equation at fixed input density, and constructung the output density with the orbitals that solve the matrix equation, but with the hamiltonian evaluated at the input density:

$$\hat{H}_{KS}[\rho_{in}]\,\varphi_i(\mathbf{r}) = \varepsilon_i\,\varphi_i(\mathbf{r}) \quad .$$

$$\rho_{out}(\mathbf{r}) = \sum_{i=1}^{N} |\varphi_i(\mathbf{r})|^2$$

Since potential and density are univocally connected via ohenberg-Kohn theorem, the above self-consistency condition can also be stated in terms of the Kohn-Sham potential

$$V_{KS}^{out}(\mathbf{r}) = V_{ext}(\mathbf{r}) + \int \frac{\rho_{out}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \mu_{XC}\left[\rho_{out}(\mathbf{r})\right] \quad ,$$

where the output density $\rho_{out}(\mathbf{r})$ corresponds to the solution of Kohn-Sham equations for a trial input potential $V_{KS}^{in}(\mathbf{r})$.

5.1.1 Reaching self-consistency: the problem of the negativity of the kernel.

Reaching the self-consistent solution, *i.e.* $\rho_{out}(\mathbf{r}) = \rho_{in}(\mathbf{r})$, or $V_{KS}^{out}(\mathbf{r}) = V_{KS}^{in}(\mathbf{r})$, is the problem of finding the fixed point of an equation of the type $x = f(x)$. The obvious strategy is to start from some guess x_0, and then to iterate the equation $x_{n+1} = f(x_n)$. If this procedure converges, then the limiting value is the fixed point $\bar{x} = f(\bar{x})$. The success of such a strategy strongly depends on the shape of the function $f(x)$. If the slope of $f(x)$ is too large and negative, this simple iterative solution does not converge. This is precisely our case, because our function is

$$f(\mathbf{r},\mathbf{r}') = \frac{\delta V_{KS}^{out}(\mathbf{r})}{\delta V_{KS}^{in}(\mathbf{r}')} = \int \frac{1}{|\mathbf{r}-\mathbf{r}''|} \frac{\delta \rho_{out}(\mathbf{r}'')}{\delta V_{KS}^{in}(\mathbf{r}')}\, d\mathbf{r}'' + \frac{d\mu_{XC}\left[\rho_{out}(\mathbf{r})\right]}{d\rho_{out}(\mathbf{r})} \frac{\delta \rho_{out}(\mathbf{r})}{\delta V_{KS}^{in}(\mathbf{r}')} \tag{5.1}$$

and the response function $\chi(\mathbf{r},\mathbf{r}') = \delta\rho_{out}(\mathbf{r})/\delta V_{KS}^{in}(\mathbf{r}')$ of the electron gas is negative definite because increasing the potential implies that electrons tend to flow away. The slope may also be very large because the Coulomb kernel $1/|\mathbf{r}-\mathbf{r}'|$ in the reciprocal space representation becomes $4\pi/G^2$, which diverges very strongly for small G.

5.1.2 Mixing schemes

The simplest approach to overcome the convergence problems when a simple *out-in* replacement procedure is ineffective, is to observe that the physical reason for this divergence is that large charge redistributions occur from one iteration to the next. This is the so-call *charge-sloshing* problem. These charge displacements can be damped out by mixing the input and output densities according to some prescription. The simplest strategy is what is know as *simple mixing*:

$$\rho_{in}^{(n+1)}(\mathbf{r}) = \alpha\, \rho_{out}^{(n)}(\mathbf{r}) + (1-\alpha)\, \rho_{in}^{(n)}(\mathbf{r}) \tag{5.2}$$

where α is an empirical parameter adjusted to minimize the number of iterations needed to achieve self-consistency.

This procedure is not always satisfactory. Difficult cases, *e.g.* metallic systems, force the choice of very small value for α — sometimes down to values of the order of 0.01 — in order to avoid the divergence of the iterative procedure. This means that only a tiny fraction of the output density is used to construct the new input density, and implies that a large number of iterations may be needed to achieve self-consistency.

The next natural step is to mix also input and output densities of the preceeding iterations. The simplest scheme along this line was proposed by D. G. Anderson [95, 96], and consists of constructing modified input and output densities by mixing the two last steps:

$$\begin{aligned} \bar{\rho}_{in}^{(n)}(\mathbf{r}) &= \beta\, \rho_{in}^{(n)}(\mathbf{r}) + (1-\beta)\, \rho_{in}^{(n-1)}(\mathbf{r}) \\ \bar{\rho}_{out}^{(n)}(\mathbf{r}) &= \beta\, \rho_{out}^{(n)}(\mathbf{r}) + (1-\beta)\, \rho_{out}^{(n-1)}(\mathbf{r}) \end{aligned} \tag{5.3}$$

and to propose a *guess* for the next iteration of the same form used in the simple mixing scheme:

$$\rho_{in}^{(n+1)}(\mathbf{r}) = \alpha\,\bar{\rho}_{out}^{(n)}(\mathbf{r}) + (1-\alpha)\,\bar{\rho}_{in}^{(n)}(\mathbf{r}) \quad , \tag{5.4}$$

where α is still an empirical mixing parameter but β is choosen in such a way as to minimize the "distance" between $\bar{\rho}_{in}^{(n)}$ and $\bar{\rho}_{out}^{(n)}$. This is a natural criterion for accelerating the convergence of the self-consistent procedure. By minimizing $\left\|\bar{\rho}_{out}^{(n)}(\mathbf{r}) - \bar{\rho}_{in}^{(n)}(\mathbf{r})\right\|^2$ with respect to β, it is easy to show that

$$\beta = \frac{< [\rho_{out}^{(n)} - \rho_{in}^{(n)}] \mid [(\rho_{out}^{(n)} - \rho_{in}^{(n)}) - (\rho_{out}^{(n-1)} - \rho_{in}^{(n-1)}) >}{\left\|(\rho_{out}^{(n)} - \rho_{in}^{(n)}) - (\rho_{out}^{(n-1)} - \rho_{in}^{(n-1)})\right\|^2} \tag{5.5}$$

where $< \cdots \mid \cdots >$ in the numerator is a scalar product. This alternative is extremely simple and effective, allowing to use values of α as large as 0.3 in the difficult cases alluded above, and reducing the number of self-consistency iterations by a factor or 10 or more. More sophisticated schemes that mix more than two iterative steps have also been proposed. The so-called Broyden schemes are described in [96]. A simple generalization of the Anderson scheme to an arbitrary number of iterations has been proposed by Pulay under the name of direct inversion in iterative subspace (DIIS) [97]. Now the guess for the next iteration is constructed exactly as in (5.4), but the modified input and output densities are constructed as

$$\begin{aligned} \bar{\rho}_{in}^{(n)}(\mathbf{r}) &= \sum_{i=1}^{N_{mix}} \beta_i\,\rho_{in}^{(n-N_{mix}+i)}(\mathbf{r}) \\ \bar{\rho}_{out}^{(n)}(\mathbf{r}) &= \sum_{i=1}^{N_{mix}} \beta_i\,\rho_{out}^{(n-N_{mix}+i)}(\mathbf{r}) \end{aligned} \tag{5.6}$$

under the normalization constraint that $\sum_{i=1}^{N_{mix}} \beta_i = 1$. Minimization of the "distance" between $\bar{\rho}_{in}^{(n)}$ and $\bar{\rho}_{out}^{(n)}$ with respect to the N_{mix}coefficients $\{\beta_i\}$ leads to a system of linear equations that can be put in the form of the following matrix equation:

$$\begin{pmatrix} <\delta\rho^{(n-N_{mix}+1)}|\delta\rho^{(n-N_{mix}+1)}> & \cdots & <\delta\rho^{(n-N_{mix}+1)}|\delta\rho^{(n)}> & 1 \\ \vdots & \cdots & \cdots & \vdots \\ <\delta\rho^{(n)}|\delta\rho^{(n-N_{mix}+1)}> & \ddots & <\delta\rho^{(n)}|\delta\rho^{(n)}> & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{N_{mix}} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad , \tag{5.7}$$

where $< \delta\rho^{(i)}|\delta\rho^{(j)} >= \int [\rho_{out}^{(i)}(\mathbf{r}) - \rho_{in}^{(i)}(\mathbf{r})]\,[\rho_{out}^{(j)}(\mathbf{r}) - \rho_{in}^{(j)}(\mathbf{r})]\,d\mathbf{r}$. There is a limit in the number of iterations that can be mixed using the DIIS scheme, because after some iterations the linear system develops a linear dependency and the matrix becomes singular. Before reaching this limit, it is observed that this tendency towards linear dependency makes that mixing more than 4 or 5 iterations does not improve any longer the convergence properties of the algorithm.

5.1.3 Direct minimization strategies

The main drawback of direct diagonalization methods within the PPW approach is that the number of plane waves involved is large (of the order of 100 per atom), and it grows linearly with the volume of the supercell. This implies that the direct diagonalization of the matrix equation at fixed density may become an intractable task due to both, memory storage and CPU time requirements. This was the fact that induced the development of iterative minimization techniques within this field — originally by Car and Parrinello [98].

An alternative to solve the electronic structure problem is to explicitly minimize the energy functional by means of iterative (Krylov) methods. The simplest, altough quite inefficient, such method is known as *steepest descent* (SD). In this procedure, the minimum is reached by a series of downhill steps in the direction of the local gradient of the functional with respect to the single-particle orbitals:

$$\varphi_i^{(n+1)}(\mathbf{r}) = \varphi_i^{(n)}(\mathbf{r}) - \Delta \frac{\delta\Omega_{KS}[\{\varphi_i^{(n)}(\mathbf{r})\}, \mathbf{R}]}{\delta\varphi_i^{*(n)}(\mathbf{r})} \tag{5.8}$$

with $\Omega_{KS} = E_{KS} - \sum_{i,j} \lambda_{ij}(< \varphi_i|\varphi_j > -\delta_{ij})$, and the Kohn-Sham energy functional E_{KS} is given by the usual expression. The second term is to ensure the orthonormalization of the orbitals, as required by the concept of the non-interacting reference system. The functional derivatives in (5.8) are given explicitly in the previous set of notes. There, it can be seen that they represent the action of the Kohn-Sham hamiltonian on the single-particle orbitals, supplemented with the orthonormalization constraint, *i.e.*

$$\varphi_i^{(n+1)}(\mathbf{r}) = \varphi_i^{(n)}(\mathbf{r}) - \Delta \left(\hat{H}_{KS}\, \varphi_i^{(n)}(\mathbf{r}) - \sum_j \lambda_{ij}\, \varphi_j^{(n)}(\mathbf{r}) \right) \tag{5.9}$$

where Δ is a time-like variable (time step) which is adjusted to give the fastest convergence, i.e. the largest possible value that prevents the divergence of the SD procedure. Ths divergence is due to the fact that the SD equations can be thought of as the discrete version of a first-order differential equation: $\dot{\varphi}_i(\mathbf{r}) = -\delta\Omega_{KS}\{\varphi_i^{(n)}(\mathbf{r})\}/\delta\varphi_i^{*(n)}(\mathbf{r})$, whose solution is a combination of exponential decays in the time-like variable. If the time step is too long compared to the inverse of the largest exponent (the fastest decay), then the integration step overshoots and the iterative process diverges. In most applications the value of a constant Δ is estimated by trial and error. Another alternative is to determine it dynamically by performing a line minimization along the direction of the gradient, and to choose $\Delta^{(n)}$ in such a way as to maximize the decrease in energy.

The steepest descent procedure is very inefficient (it may take hundreds of steps to reach the minimum), because it is bound by the fact that the gradient at step $n + 1$ is orthogonal to the gradient at step n, and this reintroduces errors proportional to the previous gradient. In order to avoid this, each minimization step has to be independent from *all* the previous ones. It can be shown (see, *e.g.* [76]) that this condition is equivalent to ask that all the search directions $\mathbf{d}^{(n)}$ should be *conjugated* to each other, *i.e.* $\mathbf{d}^{(n)}\cdot\mathbf{G}\cdot\mathbf{d}^{(m)} = 0$, where $\mathbf{G}$ is the gradient operator.

Algorithms that accomplish the above task are called *conjugated gradients* (CG). It is clear that the above condition does not determine a unique CG algorithm, but a familiy. One of them is given by the following prescription for the search directions:

$$\mathbf{d}^{(n)} = \mathbf{g}^{(n)} + \gamma^{(n)}\, \mathbf{d}^{(n-1)} \tag{5.10}$$

with

$$\gamma^{(n)} = \frac{\mathbf{g}^{(n)} \cdot \mathbf{g}^{(n)}}{\mathbf{g}^{(n-1)} \cdot \mathbf{g}^{(n-1)}} \tag{5.11}$$

where $\mathbf{g}^{(n)} = -\delta\Omega_{KS}\{\varphi_i^{(n)}(\mathbf{r})\}/\delta\varphi_i^{*(n)}(\mathbf{r})$, and $\gamma^{(1)} = 0$.

Since each minimization step is independent from the previous ones, then the dimension of the search space is reduced by one at each iteration. In theory, a CG algorithm should reach the minimum of the target function in a number of steps equal to the dimension of the search space. In practice, however, the number of iterations can be significantly reduced from this value. It is interesting to notice the the well-known Lanczos algorithm for matrix diagonalization is equivalent to a conjugated gradients minimization algorithm.

An alternative to the above relaxation dynamics is to introduce the *annealing* procedure, which consists of a damped second order dynamics:

$$\mu\, \ddot{\varphi}_i(\mathbf{r}) + \eta\, \dot{\varphi}_i(\mathbf{r}) = -\delta\Omega_{KS}\{\varphi_i^{(n)}(\mathbf{r})\}/\delta\varphi_i^{*(n)}(\mathbf{r}) \tag{5.12}$$

with μ a mass-like coefficient and $\eta > 0$ a friction coefficient which ensures that the energy is always scaled down during the dynamical evolution. This second order set of differential equations can be integrated numerically using the following discretized algorithm:

$$\begin{aligned}
\varphi_i^{(n+1)}(\mathbf{r}) \;&=\; \frac{1}{1+\tilde{\Delta}(\eta,\mu)}\left[2\varphi_i^{(n)}(\mathbf{r}) - \left(1-\tilde{\Delta}(\eta,\mu)\right)\varphi_i^{(n-1)}(\mathbf{r})\right] - \\
&\quad -\; \frac{1}{1+\tilde{\Delta}(\eta,\mu)}\frac{\Delta^2}{\mu}\left(\hat{H}_{KS}\;\varphi_i^{(n)}(\mathbf{r}) - \sum_j \lambda_{ij}\,\varphi_j^{(n)}(\mathbf{r})\right)
\end{aligned} \tag{5.13}$$

with $\tilde{\Delta}(\eta,\mu) = \eta\Delta/2\mu$. The relaxation time for this frictional dynamics is $\tau_{anneal} = \mu/\eta$. It is easy to see that the steepest descent algorithm is recovered for $\tilde{\Delta}(\eta,\mu) = 1$, while $\tilde{\Delta}(\eta,\mu) = 0$ (no friction) corresponds to an undamped, conservative dynamics. This latter cannot be used for minimization purposes, but we shall see in the folowing section how it becomes useful in first-principles molecular dynamics simulations. Another possibility is to dynamically adjust the friction coefficient η so as to keep constant, *e.g.* the kinetic energy of the orbitals defined as $K_e = \mu \int |\,\dot{\varphi}_i(\mathbf{r})|^2\, d\mathbf{r}$. This might be useful when the minimization process is difficult for relaxational dynamics, and this happens when the landscape in the orbital space is very smooth.

Convergence acceleration procedures like the DIIS described amongst the mixing schemes can also be used in connection with the minimization of the energy functional in the space of single-particle orbitals [99].

5.1.4 Orthonormalization

The exact integration of the above equations of motion for the orbitals is enough to ensure their mutual orthogonality, provided that they initially were orthogonal. The discretized numerical integration, however, introduces numerical errors such that orthogonality deteriorates very rapidly. It is then necessary to orthogonalize the orbitals at each step in the iterative procedure. In the case of minimization, any orthonormalization algorithm, like Gram-Schmidt, can be used because there is no energy conservation constraint to respect. In a conservative second order dynamics, however, there is only one orthogonalization procedure which is consistent with the equations of motion, and it is given by the following expression for the Lagrange multipliers λ_{ij}[100]:

$$\lambda_{ij} = \frac{(f_i + f_j)}{2} \int \varphi_j^*(\mathbf{r})\, \hat{H}_{KS}\, \varphi_i(\mathbf{r})\, d\mathbf{r} - \mu \int \dot{\varphi}_j^*(\mathbf{r})\, \dot{\varphi}_i(\mathbf{r})\, d\mathbf{r} \quad . \tag{5.14}$$

This is because, even if the orbitals are defined besides a unitary transformation (the electronic density is the only physically relevant quantity), the kinetic-like term, and thus the associated Lagrangian, is not.

An algorithm to force this type of holonomic constraints has been proposed in the context of a geometrically constrained classical molecular dynamics by Ciccotti, Ryckaert and Berendsen [101], under the name of SHAKE. The procedure consists of two steps: first, the equations of motion are integrated as above without including the constraints. This leads to non-orthogonal, unnormalized updated orbitals $\bar{\varphi}_i^{(n+1)}$. Then, a corrective action is applied to these orbitals in the following way:

$$\varphi_i^{(n+1)} = \bar{\varphi}_i^{(n+1)} + \sum_j x_{ij}^*\, \varphi_j^{(n)} \tag{5.15}$$

where $\{\varphi_i^{(n)}\}$ are the orthonormal orbitals at the preceeding iterative or MD step. The matrix $x_{ij}^* = \Delta^2 \lambda_{ij}/\mu$ has to be then determined in such a way that the new orbitals $\{\varphi_i^{(n+1)}\}$ are orthonormal. It is easy to show that the orthonormality condition is the following:

$$A_{ik} + \sum_j x_{ij}\, B_{jk} + \sum_j x_{kj}^*\, B_{ki}^* + \sum_j x_{ij}\, x_{kj}^* = \delta_{ik} \tag{5.16}$$

with $A_{ij} =< \bar{\varphi}_i^{(n+1)} | \bar{\varphi}_j^{(n+1)} >$ and $B_{ij} =< \varphi_i^{(n)} | \bar{\varphi}_j^{(n+1)} >$. In matrix notation, the above equation (5.16) reads: $\mathbf{1} - \mathbf{A} = \mathbf{X}\mathbf{B} + \mathbf{B}^\dagger \mathbf{X}^\dagger + \mathbf{X}\mathbf{X}^\dagger$, where the dagger indicates the hermitian conjugation operation. The solution to this equation can be easily obtained by and iterative procedure which typically converges in less than 10 steps. Once the matrix of the constraints $\mathbf{X}$ has been determined, the new orthonormal orbitals are calculated according to (5.15).

5.1.5 Preconditioning

The maximum integration time step Δ of the equations of motion for the orbitals is determined by the fastest frequency present in the spectrum of the dynamical orbitals, which are given by:

$$\omega_{i,\mathbf{G},\mathbf{G}'}^2 \propto \frac{\delta E_{KS}}{\delta \varphi_i^*(\mathbf{G})\, \delta \varphi_i(\mathbf{G}')} \quad .$$

In general, this tensor can be quite complicated to calculate. However, it is easy to realize that the dominant contribution for large G wave numbers comes from the kinetic energy

$$\frac{\delta T}{\delta\varphi_i^*(\mathbf{G})\,\delta\varphi_i(\mathbf{G}')} = \frac{G^2}{2\Omega}\delta_{\mathbf{G},\mathbf{G}'} \tag{5.17}$$

while the potential energy contributions are dominant for small wave numbers, e.g. for the Hartree potential,

$$\frac{\delta E_H}{\delta\varphi_i^*(\mathbf{G}')\,\delta\varphi_i(\mathbf{G}'')} = \sum_{\mathbf{G}} \frac{4\pi}{G^2}\,\varphi_i(\mathbf{G}'{-}\mathbf{G})\,\varphi_i^*(\mathbf{G}''+\mathbf{G}) \quad . \tag{5.18}$$

The idea of preconditioning methods is based on the fact that for large wave vectors the proper frequencies are basically given by the kinetic kernel (5.17), which assumes a very simple form. In that limit, the force term in the equations of motion becomes

$$\frac{\Delta^2}{\mu}\hat{H}_{KS}\,\varphi_i(\mathbf{G}) \propto \frac{\Delta^2 G^2}{\mu}\varphi_i(\mathbf{G})$$

and then, a larger time step Δ_{prec} can be achieved by defining a G-dependent mass-like coefficient $\mu_{prec}(G) \propto G^2$ for the large G components of the orbitals. This is because this procedure renormalizes down the spectrum in the following form: $\tilde{\omega}_{i,\mathbf{G}} = \omega_{i,\mathbf{G}}/G$. This approach is valid as long as the kinetic kernel is more important than the potential kernel. When the potential terms become relevant, a different preconditioning factor has to be adopted. Even if sophisticated proconditioning schemes that use the G-dependence of the local potential can be devised, the simpler algorithm consisting of using a constant mass-like term μ_0 up to some cutoff vector G_{mass}, and a G-dependent $\mu_{prec}(G) = \mu_0(G/G_{mass})^2$ for $G > G_{mass}$, already gives very satisfactory results at negligible additional cost. In general, the value of G_{mass} will depend on the particular system under study. Systems characterized by a free electron-like behaviour will need a small G_{mass}, while more tightly bound electrons will require a larger G_{mass}. Typical values of $E_{mass} = G^2_{mass}/2$ are between 1 and 4 Ry. The optimal value of E_{mass}, i.e. the one that maximizes the time step Δ_{prec}, can be rapidly obtained by trial and error.

Chapter 6

First-principles Molecular Dynamics

At the end of the first chapter it was mentioned that the set of Newtonian equations of motion:

$$M_I \frac{d^2 R_I(t)}{dt^2} = -\frac{\partial}{\partial R_I} \left\langle \Phi(R) \left| \hat{h}_e(R) \right| \Phi(R) \right\rangle - \frac{\partial V_{nn}(R)}{\partial R_I} \tag{6.1}$$

where

$$\hat{h}_e(R,r) = -\sum_{i=1}^{N} \frac{\hbar^2}{2m} \nabla_i^2 + \frac{e^2}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{1}{\mid r_i - r_j \mid} - e^2 \sum_{I=1}^{P} \sum_{i=1}^{N} \frac{Z_I}{\mid R_I - r_i \mid} \tag{6.2}$$

and

$$V_{nn}(R) = \frac{e^2}{2} \sum_{I=1}^{P} \sum_{J \neq I}^{P} \frac{Z_I Z_J}{\mid R_I - R_J \mid} \, . \tag{6.3}$$

can be integrated numerically to generate realistic physical trajectories in phase space. The fact that these trajectories are realistic is a consequence of the first-principles description of the acting forces, which is achieved at the expenses of introducing explicitly the electronic component in the adiabatic approximation. This avoids the bias that is necessarily introduced when the interatomic interactions are decribed through empirical, classical potentials. Of course, the price is quite high, because now the electronic problem has to be solved every time step of the MD integration, typically amounting to an overload of a factor of 1000 with respect to classical simulations. Therefore, one must be very careful to analyse whether the problem under study really needs a first-principles description or not. Essentially, a first-principles description is necessary when the chemistry of the system plays an important role, *e.g.* when there is making and breaking of chemical bonds, changing environments, variable coordination, etc. If this is not the case, then better put the effort in looking for a suitable classical force field, which can be obtained by fitting the parameters of the potential to the results of some appropriate first-principles calculations or MD simulations. This would allow for much faster and longer simulations of much larger samples, *i.e.* to a significant improvement on the statistical properties.

6.1 Density functional Molecular Dynamics

A feasible first-principles (self-consistent) MD approach can be obtained by solving the electronic problem for the ground state according to density functional theory. In that case we have

$$\begin{aligned}\left\langle \Phi(R) \left| \hat{h}_e(R) \right| \Phi(R) \right\rangle = E_{DFT}[\rho, R] &= T_R[\rho] + \frac{1}{2} \int \frac{\rho(r)\rho(r')}{|r-r'|}\, dr\, dr' + E_{XC}[\rho] + \qquad (6.4) \\ &+ \sum_{I=1}^{P} \int \rho(r)\, v(r - R_I)\, dr + \frac{1}{2} \sum_{I=1}^{P} \sum_{J\neq I}^{P} \frac{Z_I Z_J}{|R_I - R_J|}\end{aligned}$$

and the force on the nuclear coordinates is obtained by simple derivation, noting that only the last two terms have a non-vanishing contribution:

$$F_I = -\frac{\partial E_{DFT}[\rho, R)}{\partial R_I} = -\int \rho(r)\, \frac{\partial v(r - R_I)}{\partial R_I}\, dr + \sum_{J\neq I}^{P} \frac{Z_I Z_J \,(R_I - R_J)}{|R_I - R_J|^3} \quad . \qquad (6.5)$$

At this point, the straightforward procedure towards a computational scheme for a first principles molecular dynamics (FPMD) would be to keep the electronic subsystem *always* in the ground state compatible with the current nuclear configuration. Such a scheme can indeed be devised, but it has to be ensured that the electronic density is very well converged (in the sense of self-consistency) because, otherwise, a systematic perturbation (dragging) is being introduced in the nuclear dynamics.

For some reason, even if this type of scheme was theoretically proposed in the mid eighties, it was not the first one to be realized in the practice. At that time the computational cost of a self-consistent electronic calculation was too high for to the existing facilities, and researchers had in mind that the above plan was not feasible. In 1985, Roberto Car and Michele Parrinello [98] introduced an alternative scheme for a FPMD which did not involve electronic self-consistency at each MD step. They were the first to show that FPMD was possible, and thus opened a completely new field in computational physics with an astonishing impact not only in physics, but also in chemistry, materials science, and biochemistry.

In the preceeding we have used the terms first-principles and density functional as synonyms. It is important to remark that DFT is only one of the possible realizations of a first-principles calculation. One could also think of performing a FPMD simulation in which the eletronic component is described using quantum chemistry methods, *e.g.* Hartree-Fock, MP2, or CI. The advantage of DFT is that its computational cost, at least within standard approximations like the LDA and GGA, is significantly lower (lower than HF, which is the fastest of these methods). In fact, HFMD has been proposed and realized in order to study the dynamics of a molecular system in an excited state, where DFT is not well suited. In order to avoid confusion, we are going to used the term density functional molecular dynamics (DFMD), or Car-Parrinello (CP) method, to distinguish from other possible FPMD schemes, even if these latter are by far the fewest.

6.1.1 The Car-Parrinello lagrangian

Going back to DFT, the solution of Kohn-Sham equations can be thought of as a minimization problem in the many-fold of the single-particle (orthogonal) orbitals of the non-interacting reference system, $\{\varphi_i(r)\}$. These are, in fact, scalar fields which can be numerically represented on a discrete mesh — or in its reciprocal plane waves form — or by expanding them in a basis set like gaussian-type orbitals (GTO), Slater-type orbitals (STO), atomic orbitals (LCAO), Hankel functions (LMTO), etc. The self-consistent solution of Kohn-Sham equations, i.e. the minimization of KS energy functional with respect to $\{\varphi_i(r)\}$, may represent too heavy a computational task. However, it would be possible to avoid such a procedure if the electronic density can be kept sufficiently close to the adiabatic density during the MD simulation, and this plan can be accomplished by introducing a second order *fictitious* dynamics of the KS orbitals. Car and Parrinello introduced this scheme by proposing a dynamical system desribed by the following Lagrangian:

$$\begin{aligned} \mathcal{L}_{CP} &= \frac{1}{2}\sum_{I=1}^{P} M_I\,\dot{R}_I^2 + \mu\sum_{i=1}^{N} f_i \int |\dot{\varphi}_i(r)|^2\,dr - E_{KS}\left[\varphi_i(r), R\right] + \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N} f_i\,\Lambda_{ij}\left(\int \varphi_i^*(r)\,\varphi_j(r)\,dr - \delta_{ij}\right) \end{aligned} \tag{6.6}$$

with

$$\begin{aligned} E_{KS}\left[\varphi_i(r), R\right] &= \sum_{i=1}^{N} f_i \int \varphi_i^*(r)\left(-\frac{\nabla^2}{2} + \sum_{I=1}^{P} v(r-R_I) + \frac{1}{2}\int \frac{\rho(r')}{|r-r'|}\,dr'\right)\varphi_i(r)\,dr + \\ &+ E_{XC}[\rho] + \frac{1}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P} \frac{Z_I Z_J}{|R_I - R_J|} \end{aligned} \tag{6.7}$$

and the density given by

$$\rho(r) = \sum_{i=1}^{N} f_i \int \varphi_i^*(r)\,\varphi_i(r)\,dr \quad . \tag{6.8}$$

The first term in (6.6) is the nuclear kinetic energy, and the third term is the first principles potential as derived from DFT, thought KS equations. The coefficients f_i are occupation numbers corresponding to the orbitals $\{\varphi_i(r)\}$. They assume the values: $f_i = 1$ for $i \leq N$, and $f_i = 0$ for $i > N$, in the case of semiconductors. In the case of metals, $f_i(\varepsilon) = (1 + e^{(\varepsilon-\mu)/k_B T_e})^{-1}$, where $T_e$is an electronic *temperature,* which is typically included not for fundamental reasons (usually electronic Fermi temperatures are much higher than the nuclear temperature), but in order to mimick Brillouin zone integration and to improve the convergence of the self-consistency procedure due to degeneracies at the Fermi level. If the electronic temperature is included as a physical variable, then the assumption implied by the use of the above expression is that the energy exchanges between the electrons are so fast that the equilibrium distributions is always verified. If there is spin degeneracy (LDA instead of LSDA), then the occupation numbers can be multiplied by 2 and the sums carried out up to $N/2$ in place of N.

In order to represent an electronic density arising from a Slater determinant, the orbitals $\{\varphi_i(r)\}$ — which we are going to call dynamical Kohn-Sham orbitals (DKSO), and are different from the true KS orbitals that minimize E_{KS} — must be orthonormal. This is the origin of the last term in the lagrangian. It ensures the orthonormality of the DKSO at every step of the MD. The Lagrange multipliers $\{\Lambda_{ij}\}$ are determined in such a way as to verify this condition, as explained in the preceeding chapter.

The second term is the big innovation of Car and Parrinello. It represents a *fictitious* kinetic energy associated with the DKSO, which are frequently also called electronic degrees of freedom. This doesn't imply a real dynamics of the electrons. It is a just a term which allows for a dynamical evolution of the orbitals, independent of that of the nuclei. It is easy to see that the straightforward DFMD scheme mentioned above is obtained by eliminating the self dynamics of the orbitals ($\mu = 0$ in the second term in the lagrangian).

In the dynamics of the orbitals there are two components: one is their own dynamics, which is controlled by the *fictitious mass* μ, and the other arises as a consequence of a dragging force due to the motion of the nuclei, through E_{KS}. This latter fixed the average trajectory of the orbitals, while the former superimposes independent oscillations. The mass μ controls both, the energy transfer between orbitals and nuclei, which goes like $\mu^{-1/2}$ (this unphysical transfer appears because the DKSO are now trated as dynamical variables, exactly as the nuclear coordinates), and the choice of the integration time step, which is proportional to $\mu^{1/2}$. A compromise is required in order to keep the energy transfer a reasonable low values while preserving an integration time step sufficiently large.

6.1.2 The Car-Parrinello equations of motion

We then have a lagrangian that depends on the nuclear coordinates $\{R\}$ and on the orbitals $\{\varphi_i(r)\}$. The Lagrange equations are obtained in the usual way of classical mechanics:

$$\begin{aligned}
\frac{d}{dt}\left(\frac{\partial \mathcal{L}_{\mathcal{CP}}}{\partial \dot{R}_I}\right) &= -\frac{\partial \mathcal{L}_{\mathcal{CP}}}{\partial R_I} \\
\frac{d}{dt}\left(\frac{\delta \mathcal{L}_{\mathcal{CP}}}{\delta \dot{\varphi}_i^*(r)}\right) &= -\frac{\delta \mathcal{L}_{\mathcal{CP}}}{\delta \varphi_i^*(r)}
\end{aligned} \tag{6.9}$$

The second equation involves functional derivatives because the orbitals are not simple variables but continuous scalar fields. In practice, these fields are defined on a basis set (*e.g.* on a discrete regular grid), and the concept of functional derivation reduces to the standard vectorial derivation with respect to the components of the expansion of the field in the basis. In the case of a discrete grid made of n^3 points in real space, each orbital is described by a set of n^3 degrees of freedom, one for each point in the grid. In reciprocal space, it is equivalent to consider the coefficients of the plane waves expansion. Functional derivation implies to derive with respect to each one of the expansion coefficients. Doing so, we arrive to the Car-Parrinello equations of motion:

$$
\begin{aligned}
M_I \ddot{R}_I &= -\frac{\partial E_{KS}\left[\{\varphi_i(r)\}, R\right]}{\partial R_I} && (6.10) \\
\mu \ddot{\varphi}_i(r,t) &= -\frac{1}{f_i}\frac{\delta E_{KS}\left[\{\varphi_i(r)\}, R\right]}{\delta \varphi_i^*(r)} + \sum_{j=1}^{N} \Lambda_{ij}\varphi_j(r,t) = && (6.11) \\
&= -\hat{h}_{KS}\, \varphi_i(r,t) + \sum_{j=1}^{N} \Lambda_{ij}\varphi_j(r,t) && (6.12)
\end{aligned}
$$

where $\hat{h}_{KS} = -\nabla^2/2 + v(r) + \int [\rho(r')/\left|r - r'\right|]dr' + \mu_{XC}[\rho]$ is the single-particle Kohn-Sham hamiltonian, and Λ_{ij} are the Lagrange multipliers that ensure the orthonormality of the DKSO.

If, for the moment being, we forget about the first member in Eq. (6.12), *i.e.* we concentrate on the stationary solution of this equation (vanishing second derivative), it is obvious that these are just the standard KS equations:

$$
\hat{h}_{KS}\, \varphi_i(r) = \sum_{j=1}^{N} \Lambda_{ij}\varphi_j(r) \quad .
$$

By means of a unitary transformation (it can always be done because $\hat{h}_{KS}$ is hermitic) we can diagonalize the (symmetric) matrix Λ_{ij}:

$$
\mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U} = \epsilon\varphi = \ \mathbf{U}^{-1}\mathbf{\Psi} \Rightarrow
$$

$$
\left(\mathbf{U^{-1}H_{KS}U}\right)\left(\mathbf{U^{-1}}\varphi\right) \ = \ \tilde{\mathbf{H}}_{KS}\, \mathbf{\Psi} = \epsilon\mathbf{\Psi} = \left(\mathbf{U^{-1}\Lambda U}\right)\left(\mathbf{U^{-1}\Psi}\right)
$$

or: $\tilde{H}_{KS}\, \Psi_i(r) = \epsilon_i \Psi_i(r)$. This means that the eigenvalues of matrix Λ are the single-particle energies of Kohn-Sham theory and, in this case, the transformed orbitals $\{\Psi_i(r)\}$ become the original KS orbitals. In general, the dynamical orbitals $\{\varphi_i(r)\}$ are not the solutions of the KS equations, because of their dynamical evolution.

Conclusions:

1. Kohn-Sham orbitals minimize the energy associated with the lagrangian $\mathcal{L}_{\mathcal{CP}}$, at fixed nuclear configuration.

2. These orbitals define the electronic density univocally, appart from a unitary transformation.

3. The time evolution of the dynamical orbitals consists of small oscillations around the Born-Oppenheimer (BO) surface.

In general, the coupled dynamics of nuclei and orbitals is different from true nuclear dynamics, where the electrons are in the ground state (in the Born-Oppenheimer surface). A meaningful Car-Parrinello dynamics is realized only if the dynamical orbitals remain always *close* to the KS orbitals, *i.e.* if the oscillations around the BO surface are small. The conserved quantity in a CP dynamics is: $E_{cons} = K_e + K_n + E_{KS}$, with $K_e = \mu \sum_i f_i \int |\dot{\varphi}_i(r)|^2\, dr$ the fictitious electronic kinetic energy, and $K_n = 1/2 \sum_I M_I \dot{R}_I^2$ the true nuclear kinetic energy. The above condition means that we want to have $K_e \ll K_n + E_{KS}$.

When does Car-Parrinello work properly ?

- when energy exchanges between nuclear and electronic degrees of freedom are minimal.

When it does not work ?

- An energy gap that opens and closes periodically
- An energy gap too small with respect to nuclear vibrations
- A single-particle energy level crossing

Why ?

Classical perturbation theory indicates that motions with two different types of frequencies appear when the system is perturbed out from the BO surface:

$$\omega_{ij}^{(1)} = \sqrt{\frac{f_j(\varepsilon_i^* - \varepsilon_j)}{\mu}} \qquad (6.13)$$

$$\omega_{ij}^{(2)} = \sqrt{\frac{(f_j - f_i)(\varepsilon_i - \varepsilon_j)}{2\mu}} \qquad (6.14)$$

where ε_i^* are the energies of the empty single-particle eigenstates (the eigenvalues higher than N). The most important frequency is the lowest one, because it is the one that will mix better with the much smaller nuclear frequencies. This observation can be formalized in the following way: $\Omega_{\max} \ll \sqrt{2E_g/\mu}$, where $E_g = \varepsilon_i^*(\text{min}) - \varepsilon_j(\text{max})$ is the (single-particle) energy gap, and μ is the fictitious mass of the orbitals. Therefore, if the above relation is not verified, we can always decrease μ in order to stop the energy transfer between electronic and nuclear degrees of freedom. However, this implies an upwards shift of all the frequencies, in particular the highest ones, which are those that fix the integration time step. This means that reducing μ implies also the need to reduce the time step.

If we look at the difference between the forces on the nuclei computed according to the CP prescription (those which drive the MD), and the exact BO forces obtained by self-consistent minimization, we can observe a high frequency component due to the free oscillations of the orbitals, and a second, lower frequency following the nuclear oscillations.

The interesting point is that the errors in the forces are averaged out during the time evolution, namely that the dynamical force oscillates around the BO force.

This dragging must have some effect on the dynamics of the nuclei (the only meaningful dynamics), because it is as if the nuclei *dress up* with the part of the mass of the dynamical orbitals. In the atomic limit this *dressing* effect can be estimated in the following way: $\tilde{M} = M + \frac{4}{3}\mu K_e$,or $\tilde{\omega} = \omega\left(1 + 4\mu K_e/3M\right)^{1/2}$, where K_e is the (fictitious) kinetic energy of the electronic orbitals per atom, expressed in Hartree (atomic units). In other words, the orbitals renormalize the mass of the nuclei. This correction may be important in the case of light atoms (H, He, and first row elements). For instance, in Carbon compounds this correction can amount to a 6% of the calculated frequency.

Reviews on the Car-Parrinello method and recent developments can be found in Ref. [102, 103, 104]

Bibliography

[1] M. Born and J. R. Oppenheimer, Ann. der Phys. **84**, 457 (1927).

[2] A. Messiah, *Quantum Mechanics* (Amsterdam, North-Holland, 1961)

[3] See, e.g., W. H. Zurek, Physics Today (October 1991), p. 36, and references therein.

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

[5] D. R. Hartree, Proc. Cambridge. Philos. Soc. **24**, 89 (1928).

[6] V. Fock, Z. Phys. **61**, 126 (1930).

[7] J. C. Slater, Phys. Rev. **35**, 210 (1930).

[8] For a general overview of quantum chemistry methods see, e.g., A. Szabo and N. S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory* (McGraw-Hill, NY, 1989).

[9] L. H. Thomas, Proc. Cambridge. Philos. Soc. **23**, 542 (1927).

[10] E. Fermi, Z. Phys. **48**, 73 (1928).

[11] The mentor of modern density functional theory, Walter Kohn, has just been awarded the 1998 Nobel prize for chemistry.

[12] Y. Andersson, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. **76**, 103 (1996).

[13] See, e.g., N. H. March in *Theory of the inhomogeneous electron gas*, eds. S. Lundqvist and N. H. March (Plenum, NY, 1983).

[14] C. F. von Weiszäcker, Z. Phys. **96**, 431 (1935).

[15] F. Perrot, J. Phys. Condens. Matter **6**, 431 (1994); L.-W. Wang and M. P. Teter, Phys. Rev. B **45**, 13397 (1992).

[16] E. Smargiassi and P. A. Madden, Phys. Rev. B **49**, 5220 (1994).

[17] M. Foley and P. A. Madden, Phys. Rev. B **53**, 10589 (1996).

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

[19] M. Levy, Phys. Rev. A **26**, 1200 (1982).

[20] A. Theophilou, J. Phys. C **12**, 5419 (1979).

[21] E. K. U. Gross, J. F. Dobson, and M. Petersilka in *Density Functional Theory*, ed. R. F. Nalewajski, Springer Series “Topics in Current Chemistry” (Springer, Berlin, 1996).

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

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

[24] J. F. Janak, Phys. Rev. B **18**, 7165 (1978).

[25] D. C. Langreth and J. P. Perdew, Phys. Rev. B **15**, 2884 (1977).

[26] R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules* (Oxford, 1989).

[27] U. von Barth and L. Hedin, J. Phys. C **12**, 5419 (1979).

[28] S. H. Vosko, L. Wilk, and M. Nusair, Can. J. Phys. **58**, 1200 (1980).

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

[30] See, e.g., G. D. Mahan, *Many Particle Physics* (Plenum, NY, 1990).

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

[32] M. Gell-Mann and K.A. Brückner, Phys. Rev. **106**, 364 (1957).

[33] S. H. Vosko, L. Wilk, and M. Nusair, Can. J. Phys. **58**, 1200 (1980).

[34] J. P. Perdew and Y. Wang, Phys. Rev. B **46**, 12947 (1992).

[35] J. A. Alonso and L. A. Girifalco, Solid State Commun. **24**, 135 (1977); Phys. Rev. B **17**, 3735 (1978).

[36] O. Gunnarsson and R. O. Jones, Phys. Scr. **21**, 394 (1980); J. Chem. Phys. **72**, 5357 (1980); O. Gunnarsson, M. Jonson, and B. I. Lundqvist, Phys. Rev. B **20**, 3136 (1979).

[37] A. R. Denton and N. W. Ashcroft, Phys. Rev. A **39**, 4701 (1989); A. R. Denton, P. Nielaba, K. J. Runge, and N. W. Ashcroft, Phys. Rev. Lett. **64**, 1529 (1990).

[38] J. F. Lutsko and M. Baus, Phys. Rev. Lett. **64**, 761 (1990).

[39] D. J. Singh, Phys. Rev. B **48**, 14099 (1993).

[40] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Phys. Rev. B **44**, 943 (1991).

[41] S.-K. Ma and K. A. Brueckner, Phys. Rev. **165**, 18 (1968). See also, e.g. A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, NY, 1971).

[42] D. C. Langreth and M. J. Mehl, Phys. Rev. Lett. **47**, 446 (1981); Phys. Rev. B **28**, 1809 (1983).

[43] E. K. U. Gross and R. M. Dreizler, Z. Phys. A **302**, 103 (1981).

[44] J. P. Perdew, Phys. Rev. Lett. **55**, 1665 (1985).

[45] S. K. Ghosh and R. G. Parr, Phys. Rev. A **34**, 785 (1986).

[46] C. Filippi, C. J. Umrigar, and M. Taut, J. Chem. Phys. **100**, 1295 (1994).

[47] J. P. Perdew and Y. Wang, Phys. Rev. B, **45**, 13244 (1991).

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

[49] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).

[50] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996); **78**, 1396 (1997)(E).

[51] E. H. Lieb and S. Oxford, Int. J. Quantum Chem. **19**, 427 (1981).

[52] A. D. Becke, J. Chem. Phys. **84**, 4524 (1986).

[53] Y. Zhang and W. Yang, Phys. Rev. Lett. **80**, 890 (1998); J. P. Perdew, K. Burke, and M. Ernzerhof, *ibid* **80**, 891 (1998).

[54] M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).

[55] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993); *ibid* **104**, 1040 (1996); *ibid* **107**, 8554 (1997).

[56] P. S. Svendsen and U. von Barth, Phys. Rev. B **54**, 17402 (1996).

[57] J. P. Perdew, S. Kurth, A. Zupan, and P. Blaha, Phys. Rev. Lett. **82**, 2544 (1999).

[58] M. Levy and J. P. Perdew, Phys. Rev. A **32**, 2010 (1985).

[59] M. Seidl, J. P. Perdew, and M. Levy, Phys. Rev. A **59**, 51 (1999).

[60] A. D. Becke, J. Chem. Phys. **109**, 2092 (1998).

[61] T. van Voorhis and G. E. Scuseria, J. Chem. Phys. **109**, 400 (1998).

[62] C. Adamo, M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **112**, 2643 (2000).

[63] See, e.g., C. Kittel, *Introduction to solid state physics* (Wiley, NY, 1986).

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

[65] D. J. Chadi and M. L. Cohen, Phys. Rev. B **8**, 5747 (1973).

[66] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5189 (1976).

[67] O. K. Andersen, Phys. Rev. B **12**, 3060 (1975).

[68] O. K. Andersen and R. G. Wooley, Mol. Phys. **26**, 905 (1973).

[69] J. Korringa, Physica **13**, 392 (1947); W. Kohn and N. Rostocker, Phys. Rev. **94**, 1111 (1954).

[70] M. Methfessel, C. O. Rodriguez, and O. K. Andersen, Phys. Rev. B **40**, 2009 (1989).

[71] See, e.g., T. L. Loucks, *Augmented Plane Wave Method* (Benjamin, NY, 1967). The original formulation of the APW method was given by J. C. Slater in 1937.

[72] J. M. Soler and A. R. Williams, Phys. Rev. B **42**, 9728 (1990).

[73] R. Yu, D. J. Singh, and H. Krakauer, Phys. Rev. B **43**, 6411 (1991).

[74] M. T. Yin and M. L. Cohen, Phys. Rev. Lett. **45**, 1004 (1980); Phys. Rev. B **25**, 7403 (1982).

[75] W. E. Pickett, Comp. Phys. Reports **9**, 115 (1989).

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

[77] O. Sankey and J. Niklevsky, Phys. Rev. B **40**, 3979 (1989).

[78] P. Ordejón, E. Artacho and J. M. Soler, Phys. Rev. B, **53**, 10441 (1996); D. Sánchez-Portal, P. Ordejón, E. Artacho, and J. M. Soler, Int. J. Quantum Chem. **65**, 453 (1997).

[79] P. Blöchl, Phys. Rev. B, **50**, 17953 (1994).

[80] J. C. Phillips and L. Kleinman, Phys. Rev. **116**, 287 (1959).

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

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

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

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

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

[86] L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1980).

[87] X. Gonze, P. Käckell, and M. Scheffler, Phys. Rev. B **42**, 12264 (1990).

[88] X. Gonze, R. Stumpf, and M. Scheffler, Phys. Rev. B **44**, 8503 (1991).

[89] P. P. Ewald, Ann. Phys. (Leipzig) **54**, 519 (1917); Ann. Phys. (Leipzig) **54**, 557 (1917); Ann. Phys. (Leipzig) **64**, 253 (1921)

[90] C.-L. Fu and K.-M. Ho, Phys. Rev. B **28**, 5480 (1983).

[91] M. Methfessel and A. T. Paxton, Phys. Rev. B **40**, 3613 (1989).

[92] J. Rath and A. J. Freeman, Phys. Rev. B **11**, 2109 (1975).

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

[94] M. Methfessel and M. van Schilfgaarde, Phys. Rev. B **48**, 4937 (1993).

[95] D. G. Anderson, J. Assoc. Comput. Mach. **12**, 547 (1964).

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

[97] P. Pulay, Chem. Phys. Lett. **73**, 393 (1980).

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

[99] J. Hutter, H. P. Lüthi, and M. Parrinello, Comput. Mater. Sci. **2**, 244 (1994).

[100] G. Pastore, E. Smargiassi and F. Buda, Phys. Rev. A **44**, 6334 (1991).

[101] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[102] D. K. Remler and P. A. Madden, Mol. Phys. **70**, 921 (1990).

[103] G. Galli and M. Parrinello in *Computer Simulations in Materials Science*, p. 282, M. Meyer and V. Pontikis (Eds.) (Kluwer, Dodrecht, 1991).

[104] D. Marx and J. Hutter in *Modern Methods and Algorithms of Quantum Chemistry*, pp. 301-449, J. Grotendorst (Ed.) (John von Neumann Institute of Computing, FZ Jülich, 2000). Found also under *http://www.fz-juelich.de/wsqc/proceedings/*