Numerical Methods for Electronic Structure Calculations of Materials*

Yousef Saad†
James R. Chelikowsky‡
Suzanne M. Shontz§

Abstract. The goal of this article is to give an overview of numerical problems encountered when determining the electronic structure of materials and the rich variety of techniques used to solve these problems. The paper is intended for a diverse scientific computing audience. For this reason, we assume the reader does not have an extensive background in the related physics. Our overview focuses on the nature of the numerical problems to be solved, their origin, and the methods used to solve the resulting linear algebra or nonlinear optimization problems. It is common knowledge that the behavior of matter at the nanoscale is, in principle, entirely determined by the Schrödinger equation. In practice, this equation in its original form is not tractable. Successful but approximate versions of this equation, which allow one to study nontrivial systems, took about five or six decades to develop. In particular, the last two decades saw a flurry of activity in developing effective software. One of the main practical variants of the Schrödinger equation is based on what is referred to as density functional theory (DFT). The combination of DFT with pseudopotentials allows one to obtain in an efficient way the ground state configuration for many materials. This article will emphasize pseudopotential-density functional theory, but other techniques will be discussed as well.

Key words. electronic structure, quantum mechanics, Kohn–Sham equation, nonlinear eigenvalue problem, density functional theory, pseudopotentials

AMS subject classification. 81-08

DOI. 10.1137/060651653

Contents

†Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 (saad@cs.umn.edu).

‡Center for Computational Materials, Institute for Computational Engineering and Sciences, Departments of Physics and Chemical Engineering, University of Texas, Austin, TX 78712 (jrc@ices.utexas.edu).

§Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802 (shontz@cse.psu.edu). The majority of this author's research was performed while she was a postdoctoral associate at the University of Minnesota, Minneapolis.

1. Introduction. Some of the most time-consuming jobs of any high-performance computing facility are likely to be those involving calculations related to high-energy physics or quantum mechanics. These calculations are very demanding both in terms of memory and in terms of computational power. They entail computational methods that are characterized by a rich variety of techniques that blend ideas from physics and chemistry with applied mathematics, numerical linear algebra, numerical optimization, and parallel computing. In recent years, the scientific community has dramatically increased its interest in these problems as government laboratories, industrial research labs, and academic institutions have been placing an enormous emphasis on materials and everything related to nanoscience. This trend can be attributed to two converging factors. The first is that the stakes are very high, and the second is that a major practical breakthrough has never been closer because the synergistic forces at play today are making it possible to do calculations with accuracies never anticipated.

Nanoscience may gradually take the forefront of scientific computing in the same way that computational fluid dynamics (CFD) was at the forefront of scientific computing for several decades. The past few decades of scientific computing have been dominated by fluid flow computations, in part because of the needs of the aeronautics and automobile industries (e.g., aerodynamics and turbines). Model test problems for numerical analysts developing new algorithms are often from CFD (such as the model "convection-diffusion equation" or the model "Laplacian"). Similarly, a big part of applied mathematics focuses on error analysis and discretization schemes (finite elements) for fluid flow problems. Today, the need to develop novel or improved methods for CFD is diminishing, though this does not mean in any way that CFD methods are no longer in need of improvements. Yet, a look at recent publications in scientific computing reveals that there is a certain dichotomy between the current trend in nanotechnology and the interests of the scientific community.

The "mega" trend in nanotechnology is only timidly reflected by published articles in scientific computing. Few papers on "algorithms" utilize data sets or examples from standard electronic structure problems, or address problems that are specific to this class of applications. For example, one would expect to see more articles on the problem of computing a very large number of eigenvectors or that of global optimization of very complex functionals.

Part of the difficulty can be attributed to the fact that the problems encountered in quantum mechanics are considerably more complex than those addressed in other areas, e.g., classical mechanics. The approximations and methods used result from several generations of innovations by a community that is much larger and broader than that of mechanical and aerospace engineers. Chemists, chemical engineers, materials scientists, solid state physicists, electrical engineers, and even geophysicists and, more recently, bioengineers all explore materials at the atomic or molecular level, using quantum mechanical models. Therein lies a second difficulty, which is that these different groups have their own notation, constraints, and preferred methods. Chemists have a certain preference for wave-function-based methods (e.g., Hartree–Fock) which are more accurate for their needs but which physicists find too costly, if not intractable. This preference reflects the historical interests of chemists in molecules and the interests of physicists in solids.

Our paper presents an overview of some of the most successful methods used today to study the electronic structures of materials. A large variety of techniques is available and we will emphasize those methods related to pseudopotentials and density functional theory.

One of the greatest scientific achievements of humankind is the discovery, in the early part of the twentieth century, of quantum mechanical laws describing the behavior of matter. These laws make it possible, at least in principle, to predict the electronic properties of matter from the nanoscale to the macroscale. The progress that led to these discoveries is vividly narrated in the book *Thirty Years that Shook Physics* by George Gamov [64]. A series of discoveries, starting with the notion of quantas originated by Max Planck at the end of 1900 and ending roughly in the mid-1920s with the emergence of the Schrödinger wave equation, set the stage for the new physics. Solutions of the Schrödinger wave equation resulted in essentially a complete understanding of the dynamics of matter at the atomic scale. Thus, in 1929, Dirac had this to say:

> The underlying physical laws necessary for the mathematical theory of a
> large part of physics and the whole chemistry are thus completely known,

and the difficulty is only that the exact application of these laws leads to
equations much too complicated to be soluble. It therefore becomes desir-
able that approximate practical methods of applying quantum mechanics
should be developed, which can lead to the explanation of the main features
of complex atomic systems without too much computations.

One could understand atomic and molecular phenomena, formally at least, from
these equations. However, even today, solving the equations in their original form is
nearly impossible, save for systems with a very small number of electrons. In the 76
years that have passed since this statement by Dirac, we continue to strive for a better
explanation of the main features of complex atomic systems "without too much com-
putations." However, Dirac would certainly have been amazed at how much progress
has been achieved in sheer computing power. Interestingly, these gains have been
brought about by a major discovery (the transistor), which can be attributed in large
part to the new physics and a better understanding of condensed matter, especially
semiconductors. The gains made in hardware, on the one hand, and methodology, on
the other, enhance each other to yield huge speed-ups and improvements in compu-
tational capabilities.

When it comes to methodology and algorithms, the biggest steps forward were
made in the 1960s with the advent of two key new ideas. One of them was *density
functional theory* (DFT), which enabled the transformation of the initial problem into
one involving functions of only one space variable instead of N space variables, for N-
particle systems in the original Schrödinger equation. Instead of dealing with functions
in $\mathbb{R}^{3N}$, we only need to handle functions in $\mathbb{R}^3$. The second substantial improvement
came with *pseudopotentials*. In short, pseudopotentials allowed one to reduce the
number of electrons to be considered by constructing special potentials, which would
implicitly reproduce the effect of chemically inert core electrons and explicitly repro-
duce the properties of chemically active valence electrons. With pseudopotentials,
only valence electrons—those on the outer shells of the atom—need be considered;
e.g., a Pb atom is no more complex than a C atom as both have s^2p^2 valence configu-
rations. In addition, the pseudopotential is smooth near the core, whereas the actual
potential is singular. This leads to both substantial gains in memory and a reduction
of computational complexity.

In what follows we often use terminology that is employed by physicists and
chemists. For example, we will speak of "diagonalization" when we in fact mean
"computation of eigenvalues and eigenvectors," i.e., partial diagonalization. We will
use script letters for operators and bold letters for vectors.

2. Quantum Descriptions of Matter. Consider N nucleons of charge Z_n at po-
sitions $\{\mathbf{R}_n\}$ for $n = 1, \ldots, N$ and M electrons at positions $\{\mathbf{r}_i\}$ for $i = 1, \ldots, M$. An
illustration is shown in Figure 2.1. The nonrelativistic, time-independent Schrödinger
equation for the electronic structure of the system can be written as

$$(1) \qquad\qquad\qquad \mathcal{H}\,\Psi = E\,\Psi,$$

where the many-body wave function Ψ is of the form

$$(2) \qquad\qquad\qquad \Psi \equiv \Psi(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots)$$
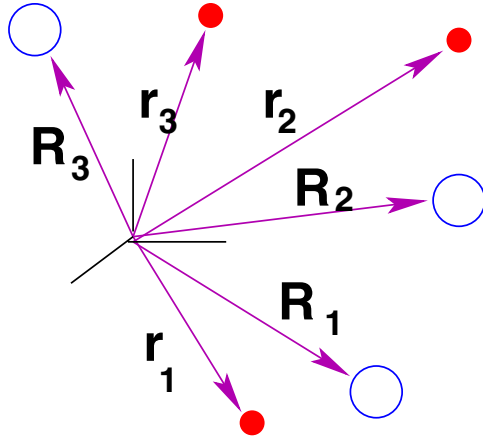
Fig. 2.1 *Atomic and electronic coordinates: filled circles represent electrons, open circles represent nuclei.*

and E is the total electronic energy. The Hamiltonian $\mathcal{H}$ in its simplest form can be written as

$$(3) \quad \mathcal{H}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots; \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = - \sum_{n=1}^{N} \frac{\hbar^2 \nabla_n^2}{2M_n} + \frac{1}{2} \sum_{\substack{n,n'=1, \\ n \neq n'}}^{N} \frac{Z_n Z_{n'} e^2}{|\mathbf{R}_n - \mathbf{R}_{n'}|}$$

$$- \sum_{i=1}^{M} \frac{\hbar^2 \nabla_i^2}{2m} - \sum_{n=1}^{N} \sum_{i=1}^{M} \frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|} + \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \ .$$

Here, M_n is the mass of the nucleus; $\hbar$ is Planck's constant, h, divided by 2π; m is the mass of the electron; and e is the charge of the electron.

The Hamiltonian above includes the kinetic energies for the nucleus (first sum in $\mathcal{H}$) and each electron (third sum), the internuclei repulsion energies (second sum), the nuclei-electronic (Coulomb) attraction energies (fourth sum), and the electron-electron repulsion energies (fifth sum). Each Laplacian ∇_n^2 involves differentiation with respect to the coordinates of the nth nucleus. Similarly, the term ∇_i^2 involves differentiation with respect to the coordinates of the ith electron.

In principle, the electronic structure of any system is completely determined by (1), or, to be exact, by minimizing the energy $\langle \Psi | \mathcal{H} | \Psi \rangle$ under the constraint of normalized wave functions Ψ. Recall that Ψ has a probabilistic interpretation: for the minimizing wave function Ψ,

$$|\Psi(\mathbf{R}_1, \ldots, \mathbf{R}_N; \mathbf{r}_1, \ldots, \mathbf{r}_M)|^2 d^3\mathbf{R}_1 \cdots d^3\mathbf{R}_N d^3\mathbf{r}_1 \cdots d^3\mathbf{r}_M$$

represents the probability of finding electron 1 in volume $|\mathbf{R}_1 + d^3\mathbf{R}_1|$, electron 2 in volume $|\mathbf{R}_2 + d^3\mathbf{R}_2|$, etc.

From a computational point of view, however, the problem is not tractable for systems which include more than just a few atoms and a dozen or so electrons. The main computational difficulty stems from the nature of the wave function, which depends on all coordinates of all particles (nuclei and electrons) simultaneously. For example, if we had just 10 particles and discretized each coordinate using just 100

points for each of the x, y, z directions, we would have 10^6 points for each coordinate for a total of $\left(10^6\right)^{10} = 10^{60}$ variables altogether.

Soon after the discovery of the Schrödinger equation, it was recognized that this equation provided the means of solving for the electronic and nuclear degrees of freedom. Using the variational principle, which states that an approximate (normalized) wave function will always have a less favorable energy than the true ground state wave function, one had an equation and a method to test the solution. One can estimate the energy from

$$(4) \qquad E = \langle\Psi|\mathcal{H}|\Psi\rangle \equiv \frac{\int\ \Psi^*\mathcal{H}\Psi\ d^3\mathbf{R}_1\ d^3\mathbf{R}_2\ d^3\mathbf{R}_3\cdots\ d^3\mathbf{r}_1\ d^3\mathbf{r}_2\ d^3\mathbf{r}_3\cdots}{\int\ \Psi^*\Psi\ d^3\mathbf{R}_1\ d^3\mathbf{R}_2\ d^3\mathbf{R}_3\cdots\ d^3\mathbf{r}_1\ d^3\mathbf{r}_2\ d^3\mathbf{r}_3\cdots}\ .$$

Recall that the wave function Ψ is normalized, since its modulus represents a probability distribution. The state wave function Ψ is an $\mathcal{L}_2$-integrable function in $\mathbb{C}^3 \times \mathbb{C}^3 \times \cdots \times \mathbb{C}^3$. The *bra* (for $\langle|$ and *ket* (for $|\rangle$) notation is common in chemistry and physics. These resemble the notions of outer and inner products in linear algebra and are related to duality. (Duality is defined from a bilinear form $a(x, y)$: The vectors x and y are dual to each other with respect to the bilinear form.)

When applying the Hamiltonian to a state function Ψ, the result is $|\mathcal{H}|\Psi\rangle$, which is another state function Φ. The inner product of this function with another function η is $\langle\eta|\Phi\rangle$, which is a scalar, a complex one in the general setting.

A number of highly successful approximations have been made to compute both the ground state, i.e., the state corresponding to minimum energy E, and excited state energies, or energies corresponding to higher eigenvalues E in (1). The main goal of these approximations is to reduce the number of degrees of freedom in the system as much as possible.

A fundamental and basic approximation is the *Born–Oppenheimer* or adiabatic approximation, which separates the nuclear and electronic degrees of freedom. Since the nuclei are considerably more massive than the electrons, it can be assumed that the electrons will respond "instantaneously" to the nuclear coordinates. This allows one to separate the nuclear coordinates from the electronic coordinates. Moreover, one can treat the nuclear coordinates as classical parameters. For most condensed matter systems, this Born–Oppenheimer approximation is highly accurate [210, 79]. Under this approximation, the first term in (3) vanishes and the second becomes a constant. We can then work with a new Hamiltonian:

$$(5) \quad \mathcal{H}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = \sum_{i=1}^{M}\frac{-\hbar^2\nabla_i^2}{2m}\ -\ \sum_{n=1}^{N}\sum_{i=1}^{M}\frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|}\ +\ \frac{1}{2}\sum_{\substack{i,j=1\\i\neq j}}^{M}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}\ .$$

This simplification essentially removes degrees of freedom associated with the nuclei, but it will not be sufficient to reduce the complexity of the Schrödinger equation to an acceptable level.

2.1. The Hartree Approximation. If we were able to write the Hamiltonian $\mathcal{H}$ as a sum of individual (noninteracting) Hamiltonians, one for each electron, then it is easy to see that the problem would become *separable*. In this case, the wave function Ψ can be written as a product of individual *orbitals*, $\phi_k(\mathbf{r}_k)$, each of which is an eigenfunction of the noninteracting Hamiltonian. This is an important concept and it is often characterized as the "one-electron" picture of a many-electron system.

The eigenfunctions of such a Hamiltonian determine orbitals (eigenfunctions) and energy levels (eigenvalues). For many systems, there is an infinite number of states,

enumerated by quantum numbers. Each eigenvalue represents an "energy" level corresponding to the orbital of interest. For example, in an atom such as hydrogen, an infinite number of bound states exists, each labeled by a set of three discrete integers. In general, the number of integers equals the spatial dimensionality of the system plus spin. In hydrogen, each state can be labeled by three indices (n, l, and m) and s for spin. In the case of a solid, there is essentially an infinite number of atoms and the energy levels can be labeled by quantum numbers, which are no longer discrete but quasi-continuous. In this case, the energy levels form an energy band.

The energy states are filled by minimizing the total energy of the system commensurate with the Pauli principle, i.e., each electron has a unique set of quantum numbers, which label an orbital. The N lowest orbitals account for $2N$ electrons, if one ignores spin (or better occupies an orbital with a spin up and a spin down electron). Orbitals that are not occupied are called "virtual states." The lowest energy orbital configuration is called the "ground state." The ground state can be used to determine a number of properties, e.g., stable structures, mechanical deformations, phase transitions, and vibrational modes. The states above the ground state are known as excited states. They are often used to calculate response functions of the solid, e.g., the dielectric and the optical properties of materials.

In mathematical terms, $\mathcal{H} \equiv \oplus \mathcal{H}^i$, the circled sum being a direct sum meaning that $\mathcal{H}^i$ acts *only* on particle number i, leaving the others unchanged. This not being true in general, Hartree suggested using it as an approximation technique whereby the basis resulting from the calculation is substituted in $\langle \Psi | \mathcal{H} | \Psi \rangle \, / \, \langle \Psi | \Psi \rangle$ to yield an upper bound for the energy.

In order to make the Hamiltonian (5) noninteractive, we must remove the last term in (5), i.e., we assume that the electrons do not interact with each other. Then the *electronic* part of the Hamiltonian becomes

$$(6) \qquad \mathcal{H}_{el} = \mathcal{H}_{el}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = \sum_{i=1}^{M} \frac{-\hbar^2 \nabla_i^2}{2m} \; - \; \sum_{n=1}^{N} \sum_{i=1}^{M} \frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|},$$

which can be cast in the form

$$(7) \qquad \mathcal{H}_{el} = \sum_{i=1}^{M} \left[\frac{-\hbar^2 \nabla_i^2}{2m} + \mathcal{V}_N(\mathbf{r}_i) \right] \equiv \bigoplus_{i=1}^{M} \mathcal{H}^i,$$

where

$$(8) \qquad \mathcal{V}_N(\mathbf{r}_i) = - \sum_{n=1}^{N} \frac{Z_n e^2}{|\mathbf{R}_n - \mathbf{r}_i|} \; .$$

This simplified Hamiltonian is separable and admits eigenfunctions of the form

$$(9) \qquad \psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = \phi_1(\mathbf{r}_1) \phi_2(\mathbf{r}_2) \phi_3(\mathbf{r}_3) \cdots,$$

where the $\phi_i(\mathbf{r})$ orbitals can be determined from the "one-electron" Hamiltonian,

$$(10) \qquad \mathcal{H}^i \phi_i(\mathbf{r}) = E_i \phi_i(\mathbf{r}) \; .$$

The total energy of the system is the sum of the occupied eigenvalues, E_i.

This model is extremely simple, but clearly not realistic. Physically, using the statistical interpretation mentioned above, writing Ψ as the product of ϕ_i's only means

that the electrons have independent probabilities of being located in a certain position in space. This lack of *correlation* between the particles causes the resulting energy to be overstated. In particular, the Pauli principle states that no two electrons can be at the same point in space and have the same quantum number. The solution Ψ computed in (9) is known as the *Hartree* wave function.

The Hartree approximation consists of using the Hartree wave function as an approximation to solve the Hamiltonian *including* the electron-electron interactions. This process starts with the use of the original adiabatic Hamiltonian (5) and forces a wave function to be a product of single-electron orbitals, as in (9). The next step is to minimize the energy $\langle\Psi|\mathcal{H}|\Psi\rangle$ under the constraint $\langle\Psi|\Psi\rangle = 1$ for Ψ in the form (9). This condition is identical to imposing the conditions that the integrals of each $|\psi_k|^2$ be equal to one. If we imposed the equations given by the standard approach of the Lagrange multipliers combined with first order necessary conditions for optimality, we would get

$$\frac{d\langle\psi_k|\mathcal{H}|\psi_k\rangle}{d\psi_k} - \lambda\psi_k = 0.$$

Evaluating $d\langle\psi|\mathcal{H}|\psi\rangle/d\psi$ over functions ϕ of norm 1 is straightforward. The first and second terms are trivial to differentiate. For the simple case when $k = 1$ and $M = 3$, consider the third term, which we denote by $\langle\Psi|\mathcal{V}_e|\Psi\rangle$:

$$(11) \qquad \langle\Psi|\mathcal{V}_e|\Psi\rangle \equiv \frac{1}{2}\sum_{\substack{i,j=1\\i\neq j}}^{M}\int\frac{e^2|\phi_1^2\phi_2^2\phi_3^2|}{|\mathbf{r}_i - \mathbf{r}_j|}d^3\mathbf{r}_1 d^3\mathbf{r}_2 d^3\mathbf{r}_3 \ .$$

Because ϕ_1 is normalized, this is easily evaluated to be a constant independent of ϕ_1 when both i and j are different from $k = 1$. We are left with the differential of the sum over $i = 1, j \neq 1$. Consider only the term $i = 1, j = 2$ (the coefficient e^2 is omitted):

$$\int\frac{|\phi_1^2\phi_2^2\phi_3^2|}{|\mathbf{r}_1 - \mathbf{r}_2|}d^3\mathbf{r}_1 d^3\mathbf{r}_2 d^3\mathbf{r}_3 = \underbrace{\int d^3\mathbf{r}_3|\phi_3(\mathbf{r}_3)|^2}_{=1} \times \int|\phi_1|^2\left[\int\frac{|\phi_2^2|}{|\mathbf{r}_1 - \mathbf{r}_2|}d^3\mathbf{r}_2\right]d^3\mathbf{r}_1 \ .$$

By introducing a variation $\delta\phi_1$ in the above relation, it is easy to see (at least in the case of real variables) that the differential of the above term with respect to ϕ_1 is the functional associated with the integral of $|\phi_2(\mathbf{r}_2)|^2/|\mathbf{r}_2 - \mathbf{r}_1|$. A similar result holds for the term $i = 1, j = 3$. In the end the individual orbitals, $\phi_i(\mathbf{r})$, are solutions of the eigenvalue problem

$$(12) \qquad \left(\frac{-\hbar^2\nabla^2}{2m} + \mathcal{V}_N(\mathbf{r}) + \sum_{\substack{j=1\\j\neq i}}^{M}\int\frac{e^2|\phi_j(\mathbf{r}')|^2}{|\mathbf{r}' - \mathbf{r}|} d^3\mathbf{r}'\right)\phi_i(\mathbf{r}) = E_i\phi_i(\mathbf{r}) \ .$$

The subscripts i, j of the coordinates have been removed as there is no ambiguity. The Hamiltonian related to each particle can be written in the form $\mathcal{H} = \frac{-\hbar^2\nabla^2}{2m} + \mathcal{V}_N + \mathcal{W}_H$, where $\mathcal{V}_N$ was defined earlier and

$$(13) \qquad \mathcal{W}_H \equiv \sum_{\substack{j=1\\j\neq i}}^{M}\int\frac{e^2\phi_j(\mathbf{r})\phi_j(\mathbf{r})^* d^3\mathbf{r}'}{|\mathbf{r}' - \mathbf{r}|} \ .$$

This "Hartree potential," or "Coulomb potential," can be interpreted as the potential seen from each electron by averaging the distribution of the other electrons' $|\phi_j(\mathbf{r})|^2$. It can be obtained by solving the Poisson equation with the charge density $e|\phi_j(\mathbf{r})|^2$ for each electron j. Note that both $\mathcal{V}_N$ and $\mathcal{W}_H$ depend on the electron i. Another important observation is that solving the eigenvalue problem (12) requires knowledge of the other orbitals ϕ_j, i.e., those for $j \neq i$. Also, the electron density of the orbital in question should not be included in the construction of the Hartree potential.

The solution of the problem requires a *self-consistent field* (SCF) iteration. One begins with some set of orbitals and iteratively computes new sets by solving (12), using the most current set of ϕ_j's for $j \neq i$. This iteration is continued until the set of ϕ_i's is self-consistent.

Once the orbitals, $\phi(\mathbf{r})$, which satisfy (12) are computed, the Hartree many-body wave function can be constructed and the total energy determined from (4). The Hartree approximation is useful as an illustrative tool, but it is not an accurate approximation.

As indicated earlier, a key weakness of the Hartree approximation is that it uses wave functions that do not obey one of the major postulates of quantum mechanics, namely, electrons or Fermions must satisfy the Pauli exclusion principle [120]. Moreover, the Hartree equation is difficult to solve. The Hamiltonian is orbitally dependent because the summation in (12) does not include the ith orbital. This means that if there are M electrons, then M Hamiltonians must be considered and (12) solved for each orbital.

2.2. The Hartree–Fock Approximation. Thus far, we have not included *spin* in the state functions Ψ. Spin can be viewed as yet another quantum coordinate associated with each electron. This coordinate can assume two values: spin up or spin down. The exclusion principle states that there can be only two electrons in the same orbit and they must be of opposite spin. Since the coordinates must now include spin, we define $\mathbf{x}_i = \binom{\mathbf{r}_i}{s_i}$, where s_i is the spin of the ith electron. A canonical way to enforce the exclusion principle is to require that a wave function Ψ is an antisymmetric function of the coordinates $\mathbf{x}_i$ of the electrons in that by interchanging any two of these coordinates, the function must change its sign. In the Hartree–Fock approximation, many-body wave functions with antisymmetric properties are constructed, typically cast as *Slater determinants*, and used to approximately solve the eigenvalue problem associated with the Hamiltonian (5).

Starting with one-electron orbitals, $\phi_i(\mathbf{x}) \equiv \phi(\mathbf{r})\sigma(s)$, the following functions meet the antisymmetry requirements:

$$(14) \qquad \Psi(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots) = (M!)^{-1/2} \begin{vmatrix} \phi_1(\mathbf{x}_1) & \phi_1(\mathbf{x}_2) & \cdots & \cdots & \phi_1(\mathbf{x}_M) \\ \phi_2(\mathbf{x}_1) & \phi_2(\mathbf{x}_2) & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \phi_M(\mathbf{x}_1) & \cdots & \cdots & \cdots & \phi_M(\mathbf{x}_M) \end{vmatrix}.$$

The term $(M!)^{-1/2}$ is a normalizing factor. If two electrons occupy the same orbit, two rows of the determinant will be identical and Ψ will be zero. The determinant will also vanish if two electrons occupy the same point in generalized space (i.e., $\mathbf{x}_i = \mathbf{x}_j$), as two columns of the determinant will be identical. Exchanging positions of two particles will lead to a sign change in the determinant. The Slater determinant is a convenient representation, but one should stress that it is an *ansatz*. It is probably the simplest many-body wave function that incorporates the required symmetry properties for Fermions, or particles with noninteger spins.

If one uses a Slater determinant to evaluate the total electronic energy and maintains wave function normalization, the orbitals can be obtained from the following *Hartree–Fock* equations:

$$(15) \quad \mathcal{H}^i \phi_i(\mathbf{r}) = \left(\frac{-\hbar^2 \nabla^2}{2m} + \mathcal{V}_N(\mathbf{r}) + \sum_{j=1}^{M} \int \frac{e^2 \, |\phi_j(\mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|} \, d^3\mathbf{r}' \right) \phi_i(\mathbf{r})$$

$$- \sum_{j=1}^{M} \int \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} \, \phi_j^*(\mathbf{r}') \phi_i(\mathbf{r}') \, d^3\mathbf{r}' \, \delta_{s_i, s_j} \, \phi_j(\mathbf{r}) = E_i \phi_i(\mathbf{r}) \, .$$

It is customary to simplify this expression by defining a probability density, often called electronic-charge density, ρ,

$$(16) \quad \rho(\mathbf{r}) = \sum_{j=1}^{M} |\phi_j(\mathbf{r})|^2,$$

and an orbital-dependent *"exchange-charge density,"* ρ_i^{HF}, for the ith orbital:

$$(17) \quad \rho_i^{HF}(\mathbf{r}, \mathbf{r}') = \sum_{j=1}^{M} \frac{\phi_j^*(\mathbf{r}') \, \phi_i(\mathbf{r}') \, \phi_i^*(\mathbf{r}) \, \phi_j(\mathbf{r})}{\phi_i^*(\mathbf{r}) \, \phi_i(\mathbf{r})} \, \delta_{s_i, s_j} \, .$$

This "density" involves a spin-dependent factor which couples only states (i, j) with the same spin coordinates (s_i, s_j).

With these charge densities defined, it is possible to define corresponding potentials. The *Coulomb* or *Hartree* potential, $\mathcal{V}_H$, is defined by

$$(18) \quad \mathcal{V}_H(\mathbf{r}) = \int \rho(\mathbf{r}) \, \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} \, d^3\mathbf{r}',$$

and an *exchange* potential can be defined by

$$(19) \quad \mathcal{V}_x^i(\mathbf{r}) = - \int \rho_i^{HF}(\mathbf{r}, \mathbf{r}') \, \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} \, d^3\mathbf{r}' \, .$$

This combination results in the following Hartree–Fock equation:

$$(20) \quad \left(\frac{-\hbar^2 \nabla^2}{2m} + \mathcal{V}_N(\mathbf{r}) + \mathcal{V}_H(\mathbf{r}) + \mathcal{V}_x^i(\mathbf{r}) \right) \phi_i(\mathbf{r}) = E_i \phi_i(\mathbf{r}) \, .$$

Once the Hartree–Fock orbitals have been obtained, the total Hartree–Fock electronic energy of the system, E_{HF}, can be obtained from

$$(21) \quad E_{HF} = \sum_{i}^{M} E_i - \frac{1}{2} \int \rho(\mathbf{r}) \mathcal{V}_H(\mathbf{r}) \, d^3\mathbf{r} - \frac{1}{2} \sum_{i}^{M} \int \phi_i^*(\mathbf{r}) \, \phi_i(\mathbf{r}) \mathcal{V}_x^i(\mathbf{r}) \, d^3\mathbf{r} \, .$$

E_{HF} is not a sum of the Hartree–Fock orbital energies, E_i. The factor of one-half in the electron-electron terms arises because the electron-electron interactions have been double counted in the Coulomb and exchange potentials. The Hartree–Fock Schrödinger equation is only slightly more complex than the Hartree equation. Again, the equations are difficult to solve because the exchange potential is orbitally dependent.

There is one notable difference in the Hartree–Fock summations compared to the Hartree summation. The Hartree–Fock sums include the $i = j$ terms in (15). This difference arises because the exchange term corresponding to $i = j$ cancels an equivalent term in the Coulomb summation. The $i = j$ term in both the Coulomb and exchange terms is interpreted as a "self-screening" of the electron. Without a cancellation between Coulomb and exchange terms a "self-energy" contribution to the total energy would occur. Approximate forms of the exchange potential often do not have this property. The total energy then contains a self-energy contribution which one needs to remove to obtain a correct Hartree–Fock energy.

The Hartree–Fock equation is an approximate solution to the true ground-state, many-body wave function. Terms not included in the Hartree–Fock energy are referred to as *correlation* contributions. One definition of the correlation energy, E_{corr}, is to write it as the difference between the exact total energy of the system, E_{exact}, and the Hartree–Fock energies: $E_{corr} = E_{exact} - E_{HF}$. Correlation energies may be included by considering Slater determinants composed of orbitals which represent excited state contributions. This method of including unoccupied orbitals in the many-body wave function is referred to as *configuration interactions* or "CI" [80].

Applying Hartree–Fock wave functions to systems with many atoms is not routine. The resulting Hartree–Fock equations are often too complex to be solved for periodic systems, except in special cases. The number of electronic degrees of freedom grows rapidly with the number of atoms, often prohibiting an accurate solution or even the ability to store the resulting wave function. The Hartree–Fock method scales nominally as N^4 (N being the number of basis functions), though its practical scaling is closer to N^3. As such, it has been argued that a "wave function" approach to systems with many atoms does not offer a satisfactory approach to the electronic structure problem. An alternate approach is based on DFT, which scales nominally as N^3, or close to N^2 in practice.

3. Density Functional Theory. In a few classic papers, Hohenberg, Kohn, and Sham established a theoretical basis for justifying the replacement of the many-body wave function by one-electron orbitals [119, 83, 100]. Their results put the charge density at center stage. The charge density is a distribution of probability, i.e., $\rho(\mathbf{r}_1)d^3\mathbf{r}_1$ represents, in a probabilistic sense, the number of electrons (all electrons) in the infinitesimal volume $d^3\mathbf{r}_1$.

Specifically, the Hohenberg–Kohn results were as follows. The first Hohenberg and Kohn theorem states that *for any system of electrons in an external potential $\mathcal{V}_{ext}$, the Hamiltonian (specifically $\mathcal{V}_{ext}$ up to a constant) is determined uniquely by the ground-state density alone.* Solving the Schrödinger equation would result in a certain ground-state wave function Ψ, to which is associated a certain charge density,

$$(22) \qquad \rho(\mathbf{r}_1) = \sum_{s_1=\uparrow,\downarrow} M \int |\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)| d\mathbf{x}_2 \cdots d\mathbf{x_M} \ .$$

From each possible state function Ψ one can obtain a (unique) probability distribution ρ. This mapping from the solution of the full Schrödinger equation to ρ is trivial. What is less obvious is that the reverse is true: Given a charge density, ρ, it is possible in theory to obtain a unique Hamiltonian and associated ground-state wave function, Ψ. Hohenberg and Kohn's first theorem states that this mapping is one-to-one; i.e., we could get the Hamiltonian (and the wave function) solely from ρ. Remarkably, this statement is easy to prove.

The second Hohenberg–Kohn theorem provides the means for obtaining this reverse mapping: *The ground-state density of a system in a particular external potential can be found by minimizing an associated energy functional.* In principle, there is a certain energy functional that is minimized by the unknown ground-state charge density, ρ. This statement remains at a formal level in the sense that no practical means was given for computing Ψ or a potential, $\mathcal{V}$. From the magnitude of the simplification, one can imagine that the energy functional will not be easy to construct. Indeed, this transformation changes the original problem with a total of $3N$ coordinates plus spin to one with only 3 coordinates, albeit with N orbitals to be determined.

Later, Kohn and Sham provided a workable computational method based on the following result: *For each interacting electron system, with external potential $\mathcal{V}_0$, there is a local potential $\mathcal{V}_{ks}$, which results in a density ρ equal to that of the interacting system.* Thus, the Kohn–Sham energy functional is formally written in the form

$$(23) \qquad \mathcal{H}_{KS} = \frac{\hbar^2}{2m}\nabla^2 + \mathcal{V}_{eff},$$

where the effective potential is defined as for a one-electron potential, i.e., as in (7):

$$(24) \qquad \mathcal{V}_{eff} = \mathcal{V}_N(\rho) + \mathcal{V}_H(\rho) + \mathcal{V}_{xc}(\rho).$$

Note that, in contrast with (7), $\mathcal{V}_{xc}$ is now without an index, as it only holds for one electron. Also note the dependence of each potential term on the charge density ρ, which is implicitly defined from the set of occupied eigenstates ψ_i, $i = 1, \ldots, N$, of (23) by (16).

The energy term associated with the nuclei-electron interactions is $\langle \mathcal{V}_N | \rho \rangle$, while that associated with the electron-electron interactions is $\langle \mathcal{V}_H | \rho \rangle$, where $\mathcal{V}_H$ is the Hartree potential,

$$\mathcal{V}_H = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' .$$

3.1. Local Density Approximation. The Kohn–Sham energy functional is of the following form:

$$E(\rho) = -\frac{\hbar^2}{2m}\sum_{i=1}^{N}\int \phi_i^*(\mathbf{r})\nabla^2\phi_i(\mathbf{r})d\mathbf{r} \;+\; \int \rho(\mathbf{r})\mathcal{V}_{ion}(\mathbf{r})d\mathbf{r}$$
$$(25) \qquad\qquad +\frac{1}{2}\int\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + E_{xc}(\rho(\mathbf{r})).$$

The effective energy, or Kohn–Sham energy, may not represent the true or "experimental" energy, because the Hamiltonian has been approximated.

A key contribution of the work of Kohn and Sham is the *local density approximation* (LDA) [100]. Within LDA, the exchange energy is expressed as

$$(26) \qquad E_x[\rho(\mathbf{r})] = \int \rho(\mathbf{r})\mathcal{E}_x[\rho(\mathbf{r})]\; d^3\mathbf{r},$$

where $\mathcal{E}_x[\rho]$ is the exchange energy per particle of a uniform gas at a density of ρ. Within this framework, the exchange potential in (20) is replaced by a potential determined from the functional derivative of $E_x[\rho]$:

$$(27) \qquad \mathcal{V}_x[\rho] = \frac{\delta E_x[\rho]}{\delta \rho} .$$

One serious issue is the determination of the exchange energy per particle, $\mathcal{E}_x$, or the corresponding exchange potential, $\mathcal{V}_x$. The exact expression for either of these quantities is unknown, save for special cases. From Hartree–Fock theory one can show that the exchange energy is given by

$$(28) \qquad E_{HF}^{FEG} = 2 \sum_{k<k_f} \frac{\hbar^2 k^2}{2m} - \frac{e^2 k_f}{\pi} \sum_{k<k_f} \left(1 + \frac{1-(k/k_f)^2}{2(k/k_f)} \ln\left| \frac{k+k_f}{k-k_f} \right| \right),$$

which is the Hartree–Fock expression for the exchange energy of a *free electron gas*. In this expression, $\mathbf{k}$ is the wave vector for a free electron; it can be related to the momentum by $\mathbf{p} = \hbar\mathbf{k}$. The highest occupied wave vector is given by $\mathbf{k_f}$, where the Fermi energy is given by $E_f = \hbar^2 k_f^2/2m$. One can write

$$(29) \qquad E_x[\rho] = -\frac{3e^2}{4\pi}(3\pi^2)^{1/3} \int [\rho(\mathbf{r})]^{4/3} \, d^3\mathbf{r},$$

and taking the functional derivative, one obtains

$$(30) \qquad \mathcal{V}_x[\rho] = -\frac{e^2}{\pi}(3\pi^2 \rho(\mathbf{r}))^{1/3} \ .$$

In contemporary theories, correlation energies are explicitly included in the energy functionals [119]. These energies have been determined by numerical studies performed on uniform electron gases resulting in local density expressions of the form $\mathcal{V}_{xc}[\rho(\mathbf{r})] = \mathcal{V}_x[\rho(\mathbf{r})] + \mathcal{V}_c[\rho(\mathbf{r})]$, where $\mathcal{V}_c$ represents contributions to the total energy beyond the Hartree–Fock limit [27]. It is also possible to describe the role of spin explicitly by considering the charge density for up and down spins: $\rho = \rho_\uparrow + \rho_\downarrow$. This approximation is called the *local spin density approximation* (LSDA) [119].

3.2. The Kohn–Sham Equation. The *Kohn–Sham equation* [100] for the electronic structure of matter is given by

$$(31) \qquad \left(\frac{-\hbar^2\nabla^2}{2m} + \mathcal{V}_N(\mathbf{r}) + \mathcal{V}_H(\mathbf{r}) + \mathcal{V}_{xc}[\rho(\mathbf{r})] \right) \phi_i(\mathbf{r}) = E_i\phi_i(\mathbf{r}) \ .$$

This equation is nonlinear and it can be solved either as an optimization problem (minimize ground energy with respect to wave functions) or a nonlinear eigenvalue problem. In the optimization approach an initial wave function basis is selected and a gradient-type approach is used to iteratively refine the basis until a minimum energy is reached. In the second approach the Kohn–Sham equation is solved "self-consistently." An approximate charge is assumed to estimate the exchange-correlation potential, and this charge is used to determine the Hartree potential from (18). The output potential is then carefully mixed with the previous input potential(s) and the result is inserted into the Kohn–Sham equation and the total charge density determined as in (16). The "output" charge density is used to construct new exchange-correlation and Hartree potentials. The process is repeated until the input and output charge densities or potentials are identical to within some prescribed tolerance.

Once a solution of the Kohn–Sham equation is obtained, the total energy can be computed from

$$(32) \qquad E_{KS} = \sum_i^M E_i - 1/2 \int \rho(\mathbf{r})\mathcal{V}_H(\mathbf{r}) \, d^3\mathbf{r} + \int \rho(\mathbf{r})\big(E_{xc}[\rho(\mathbf{r})] - \mathcal{V}_{xc}[\rho(\mathbf{r})]\big) \, d^3\mathbf{r},$$

where E_{xc} is a generalization of (26), i.e., the correlation energy density is included. The electronic energy, as determined from E_{KS}, must be added to the ion-ion interactions to obtain the structural energies. This is a straightforward calculation for confined systems. For periodic systems such as crystals, the calculations can be done using Madelung summation techniques [211].

Owing to its ease of implementation and overall accuracy, the LDA is a popular choice for describing the electronic structure of matter. Recent developments have included so-called gradient corrections to the LDA. In this approach, the exchange-correlation energy depends on the local density and the gradient of the density. This approach is called the generalized gradient approximation (GGA) [147].

When first proposed, DFT was not widely accepted within the chemistry community. The theory is not "rigorous" in the sense that it is not clear how to improve the estimates for the ground-state energies. For wave-function-based methods, one can include more Slater determinants, as in a configuration interaction approach. As the accuracy of the wave functions improves, the energy is lowered via the variational theorem. The Kohn–Sham equation is also variational, but owing to the approximate Hamiltonian, the converged energy need not approach the true ground-state energy. This might not be a serious problem provided that one is interested in relative energies, where there is the possibility of some error cancellation. For example, if the Kohn–Sham energy of an atom is 10% too high compared to the true value and the corresponding energy of the atom in a crystal is also 10% too high, the cohesive energies, which involve the difference of the two energies, can be better than the nominal 10% error of absolute energies without any error cancellation. An outstanding fundamental issue of using DFT is obtaining an a priori estimate of the cancellation errors.

In some sense, DFT is an a posteriori theory. Given the transference of the exchange-correlation energies from an electron gas, it is not surprising that errors arise in its application to highly nonuniform electron gas systems, as found in realistic systems. However, the degree of error cancellations is rarely known. Thus, the reliability of DFT has been established by numerous calculations for a wide variety of condensed matter systems. For example, the cohesive energies, compressibility, structural parameters, and vibrational spectra of elemental solids have been calculated within the DFT [34]. The accuracy of the method is best for systems in which the cancellation of errors is expected to be complete. Since cohesive energies involve the difference in energies between atoms in solids and atoms in free space, error cancellations are expected to be significant. This is reflected in the fact that, historically, cohesive energies have presented greater challenges for DFT: the errors between theory and experiment are typically ∼10–20%, depending on the nature of the density functional and the material of interest. In contrast, vibrational frequencies which involve small structural changes within a given crystalline environment are often reproduced to within 1–2%.

3.3. Pseudopotentials. A major difficulty in solving the eigenvalue problem arising from the Kohn–Sham equation is the length and energy scales involved. The inner (core) electrons are highly localized and tightly bound compared to the outer (valence) electrons. A simple basis function approach is frequently ineffectual. For example, a plane wave basis (see next section) might require 10^6 waves to represent converged wave functions for a core electron, whereas only 10^2 waves are required for a valence electron [32]. In addition, the potential is singular near the core, and this causes difficulties in discretizing the Hamiltonian and in representing the wave functions.
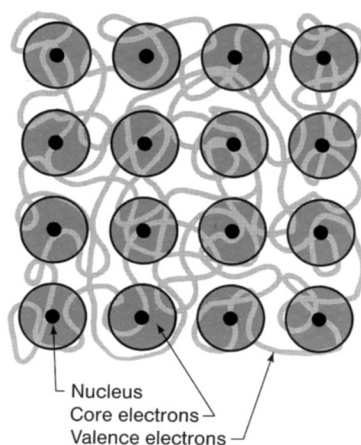
Fig. 3.1 *Standard pseudopotential model of a solid. The ion cores composed of the nuclei and tightly bound core electrons are treated as chemically inert. The pseudopotential model describes only the outer, chemically active, valence electrons.*

The use of pseudopotentials overcomes these problems by removing the core states from the problem and replacing the all-electron potential by one that replicates only the chemically active, valence electron states [32]. It is well known that the physical properties of solids depend essentially on the valence electrons rather than on the core electrons, e.g., the periodic table is based on this premise. By construction, the pseudopotential reproduces exactly the valence state properties such as the eigenvalue spectrum and the charge density outside the ion core. The pseudopotential model treats matter as a sea of valence electrons moving in a background of ion cores (see Figure 3.1).

The cores are composed of nuclei and inert inner electrons. Within this model many of the complexities of an all-electron calculation are avoided. A group IV solid such as C with 6 electrons is treated in a similar fashion to Pb with 82 electrons since both elements have 4 valence electrons.

The pseudopotential approximation takes advantage of this observation by removing the core electrons and introducing a potential that is weaker at the core, which will make the (pseudo-)wave functions behave like the all-electron wave function near the locations of the valence electrons, i.e., beyond a certain radius r_c from the core region. The valence wave functions often oscillate rapidly in the core region because of the orthogonality requirement of the valence states to the core states. This oscillatory or nodal structure of the wave functions corresponds to the high kinetic energy in this region. It is costly to represent these oscillatory functions accurately, no matter what discretization or expansion is used. Pseudopotential calculations center on the accuracy of the valence electron wave function in the spatial region away from the core, i.e., within the "chemically active" bonding region. The smoothly varying pseudowave function should be identical to the appropriate all-electron wave function in the bonding regions. A similar construction was introduced by Fermi in 1934 [54] to account for the shift in the wave functions of high-lying states of alkali atoms subject to perturbations from foreign atoms. In this remarkable paper, Fermi introduced the conceptual basis for both the pseudopotential and the scattering length. In Fermi's

analysis, he noted that it was not necessary to know the details of the scattering potential. Any number of potentials which reproduced the phase shifts of interest would yield similar scattering events.

A variety of methods exists to construct pseudopotentials [122]. Almost all of these methods are based on "inverting" the Kohn–Sham equation. As a simple example, suppose we consider an atom where we know the valence wave function, ψ_v, and the valence energy, E_v. Let us replace the true valence wave function by an approximate pseudowave function, ϕ_v^p. Then the ion core pseudopotential is given by

$$(33) \qquad \mathcal{V}_{ion}^p = \frac{\hbar^2 \phi_v^p}{2m} - \mathcal{V}_H - \mathcal{V}_{xc} + E_v \ .$$

The charge density in this case is $\rho = |\phi_v^p|^2$, from which $\mathcal{V}_H$ and $\mathcal{V}_{xc}$ can be calculated. The key aspect of this inversion is choosing ϕ_v^p to meet several criteria, e.g., $\phi_v^p = \psi_v$ outside the core radius, r_c. Unlike the all-electron potential, pseudopotentials are not simple functions of position. For example, the pseudopotential is state dependent or angular momentum dependent, i.e., in principle one has a different potential for s, p, d, ... states. Details can be found in the literature [122, 32].

For the applied mathematician, the interesting problem here is one approximating a function, given a set of constraints of desirable properties and a few other constraints reflecting mandatory requirements. Viewed from this angle there may be new, interesting mathematical problems to investigate that could lead to improved pseudopotentials. The subareas most likely to have an impact are those related to approximation theory. Recent interest has turned to "ultrasoft" pseudopotentials [199], whereby norm conservation is sacrificed to obtain much softer pseudopotentials, allowing a low kinetic energy cut-off.

Recent research on pseudopotentials has moved away from the norm-conserving requirements of classical pseudopotentials. Ultrasoft pseudopotentials [199] and augmented plane wave methods [17] are two of the more advanced methods in this category. The article [105] compares these approaches and shows that in fact they are strongly related. These techniques typically lead to smaller plane wave bases, but result in a generalized eigenvalue problem which is harder to solve.

4. Discretization. The Kohn–Sham equation must be "discretized" before it can be numerically solved. The term "discretization" is used here in the most inclusive manner and, in agreement with the common terminology of numerical computing, means any method which reduces a continuous problem to one with a finite number of unknowns.

There have been three predominant ways of discretizing the Schrödinger equation. The first uses plane wave bases, the second uses specialized functions such as exponential or Gaussian orbitals, and the third does not use an explicit basis but discretizes the equations in real space.

4.1. Plane Waves. Owing to the use of pseudopotentials, simple basis sets such as a plane wave basis can be quite effective, especially for crystalline matter. For example, in the case of crystalline silicon only 50–100 plane waves need to be used for a well-converged solution. The resulting matrix representation of the Schrödinger operator is dense in Fourier (plane-wave) space, but it is not formed explicitly. Instead, matrix-vector product operations are performed with the help of fast Fourier transforms (FFTs). A plane wave approach is akin to spectral techniques used in solving certain types of partial differential equations [60]. The plane wave basis used
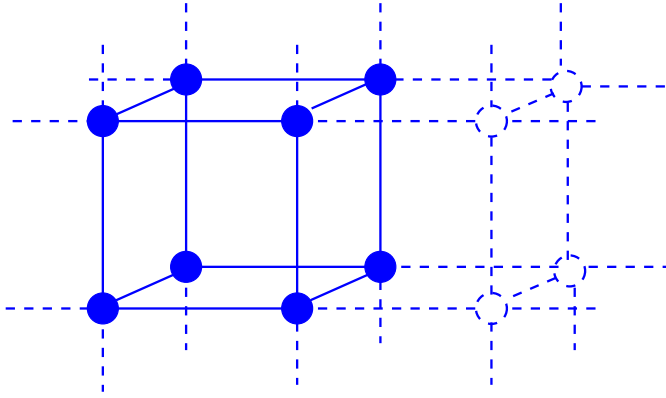
Fig. 4.1 *A simple cubic lattice.*

is of the following form:

$$(34) \qquad \psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} \alpha(\mathbf{k}, \mathbf{G}) \exp\left(i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}\right),$$

where $\mathbf{k}$ is the *wave vector*, $\mathbf{G}$ is a reciprocal lattice vector, and $\alpha(\mathbf{k}, \mathbf{G})$ represent the coefficients of the basis. Thus, plane waves are labeled by the wave vector $\mathbf{k}$, a vector in the "Brillion" zone. The vector parameter $\mathbf{G}$ translates the periodicity of the wave function with respect to a lattice, which along with an atomic basis defines a crystalline structure.

It is interesting to consider the origin of the use of plane waves. As might be guessed, plane wave bases are closely tied to periodic systems. The well-known *Bloch theorem* [98] characterizes the spectrum of the Schrödinger operator $\nabla^2 + \mathcal{V}$ when the potential $\mathcal{V}$ is periodic. It states that eigenfunctions must be of the form $\psi_j^{\mathbf{k}}(\mathbf{r})e^{-i\mathbf{k}.\mathbf{r}}$, where $\mathbf{k}$ is a vector in the "Brillion" zone. For a given lattice, periodicity takes place in three spatial directions; see Figure 4.1. The Hamiltonian is invariant under translation in each of these directions.

Bloch's theorem states that for a periodic potential $\mathcal{V}$, the spectrum of the Schrödinger operator $\nabla^2 + \mathcal{V}$ consists of a countable set of intervals (called energy bands). The eigenvalues are labeled as $\{\epsilon_{j,k}\}$, where k belongs to an interval and $j = 1, 2, \ldots$.

When expressed (i.e., projected) in a plane wave basis, the Hamiltonian is actually a dense matrix. Specifically, the Laplacian term of the Hamiltonian is represented by a diagonal matrix, but the potential term $\mathcal{V}_{tot}^p$ gives rise to a dense matrix.

For periodic systems, where $\mathbf{k}$ is a good quantum number, the plane wave basis coupled to pseudopotentials is quite effective. However, for nonperiodic systems, such as clusters, liquids, or glasses, the plane wave basis is often combined with a *supercell method* [32]. The supercell repeats the localized configuration to impose periodicity to the system. This preserves the "artificial" validity of $\mathbf{k}$ and Bloch's theorem, which (34) obeys.

There is a parallel to be made with spectral methods, which are quite effective for simple periodic geometries but lose their superiority when more generality is required. In addition to these difficulties, the two FFTs performed at each iteration can be costly, requiring $n \log n$ operations, where n is the number of plane waves, versus

$O(N)$ for real space methods, where N is the number of grid points. Usually, the matrix size $N \times N$ is larger than $n \times n$, but only within a constant factor.

4.2. Localized Orbitals. A popular approach to studying the electronic structure of materials uses a basis set of orbitals localized on atomic sites. This is the approach taken, for example, in the SIESTA code [188] where with each atom a is associated a basis set of functions, which combine radial functions around a with spherical harmonics:

$$\phi_{lmn}^a(\mathbf{r}) = \phi_{ln}^a(\mathbf{r}_a)Y_{lm}(\hat{\mathbf{r}}_a),$$

where $\mathbf{r}_a = \mathbf{r} - \mathbf{R}_a$.

The radial functions can be exponentials, Gaussians, or any localized function. Gaussian bases have the special advantage of yielding analytical matrix elements, provided the potentials are also expanded in Gaussians [21, 87, 33, 88]. However, the implementation of a Gaussian basis is not as straightforward as with plane waves. For example, numerous indices must be used to label the state, the atomic site, and the Gaussian orbitals employed. This increases "bookkeeping" operations tremendously. Also, the convergence is not controlled by a single parameter as it is with plane waves; e.g., if atoms in a solid are moved, the basis should be reoptimized for each new geometry. Moreover, it is not always obvious what basis functions are needed, and much testing has to be done to ensure that the basis is complete. On the positive side, a Gaussian basis yields much smaller matrices and requires less memory than plane wave methods. For this reason, Gaussians are especially useful for describing transition metal systems, where large numbers of plane waves are needed.

4.3. Finite Differences in Real Space. An appealing alternative is to avoid explicit bases altogether and work instead in real space, using finite difference discretizations. This approach has become popular in recent years, and has seen a number of implementations [205, 36, 37, 35, 20, 74, 212, 140, 96, 111, 11, 60, 52].

The real-space approach overcomes many of the complications involved with non-periodic systems, and although the resulting matrices can be larger than with plane waves, they are quite sparse, and the methods are easy to implement on sequential and parallel computers. Even on sequential machines, real-space methods can be an order of magnitude faster than methods based on traditional approaches.

The simplest real-space method utilizes finite difference discretization on a cubic grid. There have also been implementations of the method with finite elements [197, 144] and even meshless methods [91]. Finite element discretization methods may be successful in reducing the total number of variables involved, but they are far more difficult to implement. A key aspect to the success of the finite difference method is the availability of *higher order finite difference expansions* for the kinetic energy operator, i.e., expansions of the Laplacian [61]. Higher order finite difference methods significantly improve convergence of the eigenvalue problem when compared with standard, low order finite difference methods. If one imposes a simple, uniform grid on our system, where the points are described in a finite domain by (x_i, y_j, z_k), we approximate $\frac{\partial^2 \psi}{\partial x^2}$ at (x_i, y_j, z_k) by

$$(35) \qquad \frac{\partial^2 \psi}{\partial x^2} = \sum_{n=-M}^{M} C_n \psi(x_i + nh, y_j, z_k) + O(h^{2M+2}),$$

where h is the grid spacing and M is a positive integer. This approximation is accurate to $O(h^{2M+2})$ upon the assumption that ψ can be approximated accurately by a power

Table 4.1 *Finite difference coefficients (Fornberg–Sloan formulas) for $\partial^2/\partial x^2$ for orders 2 to 8.*

ord 2				1	-2	1		
ord 4			$-\frac{1}{12}$	$\frac{4}{3}$	$-\frac{5}{2}$	$\frac{4}{3}$	$-\frac{1}{12}$	
ord 6		$\frac{1}{90}$	$-\frac{3}{20}$	$\frac{3}{2}$	$-\frac{49}{18}$	$\frac{3}{2}$	$-\frac{3}{20}$	$\frac{1}{90}$
ord 8	$-\frac{1}{560}$	$\frac{8}{315}$	$-\frac{1}{5}$	$\frac{8}{5}$	$-\frac{205}{72}$	$\frac{8}{5}$	$-\frac{1}{5}$	$\frac{8}{315}$

ord 8 last column: $-\frac{1}{560}$

series in h. Algorithms are available to compute the coefficients C_n for arbitrary order in h [61]. These are shown for the first few orders in Table 4.1.

With the kinetic energy operator expanded as in (35), one can set up a one-electron Schrödinger equation over a grid. One may assume a uniform grid, but this is not a necessary requirement. $\psi(x_i, y_j, z_k)$ is computed on the grid by solving the eigenvalue problem

$$
-\frac{\hbar^2}{2m} \left[\sum_{n_1=-M}^{M} C_{n_1} \psi_n(x_i + n_1 h, y_j, z_k) + \sum_{n_2=-M}^{M} C_{n_2} \psi_n(x_i, y_j + n_2 h, z_k) \right.
$$
$$
\left. + \sum_{n_3=-M}^{M} C_{n_3} \psi_n(x_i, y_j, z_k + n_3 h) \right] + [\, \mathcal{V}_{ion}(x_i, y_j, z_k) + \mathcal{V}_H(x_i, y_j, z_k)
$$
$$
(36) \quad + \mathcal{V}_{xc}(x_i, y_j, z_k) \,] \psi_n(x_i, y_j, z_k) = E_n \, \psi_n(x_i, y_j, z_k).
$$

If we have L grid points, the size of the full matrix resulting from the above problem is $L \times L$.

A grid based on points uniformly spaced in a three-dimensional cube as shown in Figure 4.2 is typically used. Many points in the cube are far from any atoms in the system, and the wave function on these points may be replaced by zero. Special data structures may be used to discard these points and keep only those having a nonzero value for the wave function. The size of the Hamiltonian matrix is usually reduced by a factor of two to three with this strategy, which is quite important considering the large number of eigenvectors that must be saved. Further, since the Laplacian can be represented by a simple stencil, and since all local potentials sum up to a simple diagonal matrix, the Hamiltonian need not be stored explicitly as a sparse matrix. Handling the ion core pseudopotential is complex, as it consists of a local and a nonlocal term. In the discrete form, the nonlocal term becomes a sum over all atoms, a, and quantum numbers, (l, m), of rank-one updates:

$$
(37) \qquad \mathcal{V}_{ion} = \sum_a \mathcal{V}_{loc,a} + \sum_{a,l,m} c_{a,l,m} \mathbf{U}_{a,l,m} \mathbf{U}_{a,l,m}^T,
$$

where $\mathbf{U}_{a,l,m}$ are sparse vectors which are only nonzero in a localized region around each atom, and $c_{a,l,m}$ are normalization coefficients.

Among research issues for the applied mathematician or the numerical analyst, there is no doubt that much can be gained by investigating alternative bases that could lead to more effective computations. In the simplest case, the finite element approach could be one avenue. High order finite element methods (FEM) may be of interest, but the difficulty of implementation has no doubt been a big impediment. Another approach recently advocated in [200] exploits localized Lagrange polynomials defined
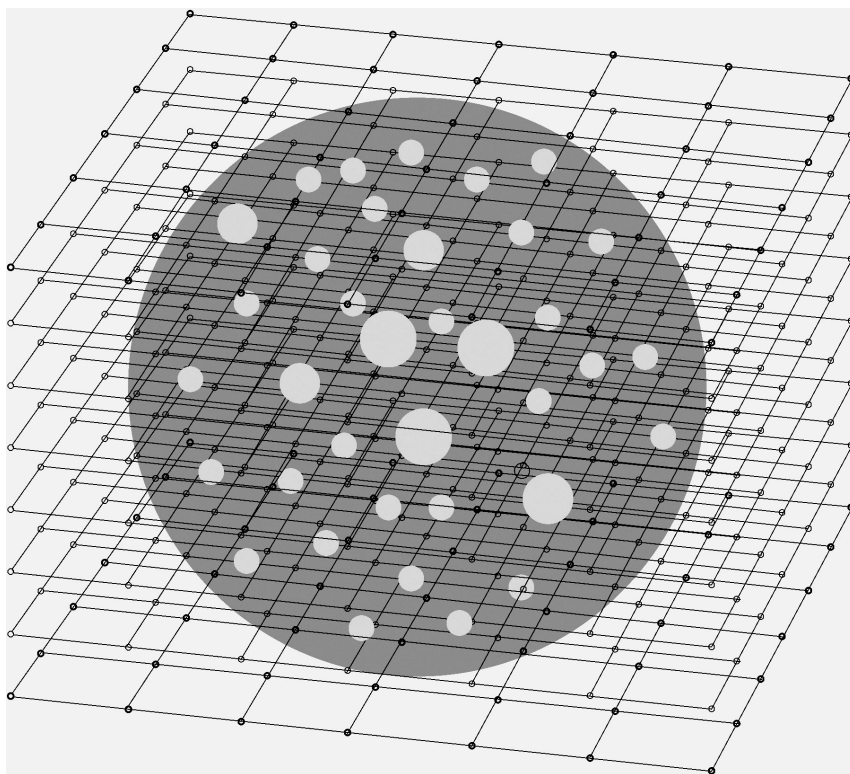
Fig. 4.2 *A uniform grid illustrating a typical configuration for examining the electronic structure of a localized system. The dark gray sphere represents the actual computational domain, i.e., the area where wave functions are allowed to be nonzero. The light spheres within the domain are atoms.*

on a sort of (locally) optimized real-space grid. Wavelet bases hold great promise because of their localized nature in both real space and reciprocal space; see, for example, [69, 182, 65, 84, 38]. Quite recently, a new code named BigDFT, which is based entirely on wavelet basis sets, has been released [15].

5. Diagonalization. There are a number of difficulties which emerge when solving the (discretized) eigenproblems, besides the sheer size of the matrices. The first, and biggest, challenge is that the number of required eigenvectors is proportional to the atoms in the system and can grow to thousands, if not more. In addition to storage, maintaining the orthogonality of these vectors can be very demanding. Usually, the most computationally expensive part of diagonalization codes is orthogonalization. Second, the relative separation of the eigenvalues decreases as the matrix size increases, and this has an adverse effect on the rate of convergence of the eigenvalue solvers. Preconditioning techniques attempt to alleviate this problem. Real-space codes benefit from savings brought about by not needing to store the Hamiltonian matrix, although this may be balanced by the need to store large vector bases.

5.1. Historical Perspective. Large computations on the electronic structure of materials started in the 1970s after the seminal work of Kohn, Hohenberg, and Sham in developing DFT and because of the invention of ab initio pseudopotentials [32]. It

is interesting to note that "large" in the 1970s implied matrices ranging in size from a few hundred to a few thousand. One had to wait until the mid- to late-1980s to see references to calculations with matrices of size around 7,000. For example, the abstract of the paper by Martins and Cohen [121] states, "Results of calculations for molecular hydrogen with matrix sizes as large as 7,200 are presented as an example." Similarly, the well-known Car and Parrinello paper [25], which uses an approach based on simulated annealing, features an example with $16 \times 437 = 6,992$ unknowns. This gives a rough idea of the typical problem sizes common about 20 years ago. The paper by Car and Parrinello [25] is often viewed as a defining moment in the development of computational codes. It illustrated how to effectively combine several ingredients: plane waves, pseudopotentials, the use of FFTs, and especially how to apply pseudopotential methods to molecular dynamics.

From the inception of realistic computations for the electronic structure of materials, the basis of choice has been plane waves. In the early days this contributed to the limitation of the capability because the matrices were treated as dense. The paper [121] (see also [122]) showed how to avoid storing a whole dense matrix by a judicious use of the FFT in plane wave codes and by working essentially in Fourier space. A code called Ritzit, initially published in Algol, was available [165], and this constituted an ideally-suited technique for diagonalization. The method was "preconditioned" by a Jacobi iteration or by direct inversion in the iterative subspace (DIIS) iteration [158].

5.2. Lanczos, Davidson, and Related Approaches. The Lanczos algorithm [107] is one of the best-known techniques [166] for diagonalizing a large sparse matrix $\mathcal{A}$. In theory, the Lanczos algorithm generates an orthonormal basis, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$, via an inexpensive three-term recurrence of the form

$$\beta_{j+1}\mathbf{v}_{j+1} = \mathcal{A}\mathbf{v}_j - \alpha_j\mathbf{v}_j - \beta_j\mathbf{v}_{j-1} .$$

In the above sequence, $\alpha_j = \mathbf{v}_j^H \mathcal{A}\mathbf{v}_j$ and $\beta_{j+1} = \|\mathcal{A}\mathbf{v}_j - \alpha_j\mathbf{v}_j - \beta_j\mathbf{v}_{j-1}\|_2$. So the jth step of the algorithm starts by computing α_j, then proceeds to form the vector $\hat{\mathbf{v}}_{j+1} = \mathcal{A}\mathbf{v}_j - \alpha_j\mathbf{v}_j - \beta_j\mathbf{v}_{j-1}$, and then $v_{j+1} = \hat{\mathbf{v}}_{j+1}/\beta_{j+1}$. Note that for $j = 1$, the formula for $\hat{\mathbf{v}}_2$ changes to $\hat{\mathbf{v}}_2 = \mathcal{A}\mathbf{v}_2 - \alpha_2\mathbf{v}_2$.

Suppose that m steps of the recurrence are carried out, and consider the tridiagonal matrix

$$\mathcal{T}_m = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \beta_3 & \\ & \ddots & \ddots & \ddots \\ & & \beta_m & \alpha_m \end{pmatrix} .$$

Further, denote by $\mathcal{V}_m$ the $n \times m$ matrix $\mathcal{V}_m = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ and by $\mathbf{e}_m$ the mth column of the $m \times m$ identity matrix. After m steps of the algorithm, the following relation holds:

$$\mathcal{A}\mathcal{V}_m = \mathcal{V}_m\mathcal{T}_m + \beta_{m+1}\mathbf{v}_{m+1}\mathbf{e}_m^T .$$

In the ideal situation, where $\beta_{m+1} = 0$ for a certain m, $\mathcal{A}\mathcal{V}_m = \mathcal{V}_m\mathcal{T}_m$, and so the subspace spanned by the $\mathbf{v}_i$'s is invariant under $\mathcal{A}$ and the eigenvalues of $\mathcal{T}_m$ become exact eigenvalues of $\mathcal{A}$. This is the situation when $m = n$, and it may also happen for $m \ll n$, though this situation, called lucky (or happy) breakdown [141], is highly unlikely in practice. In the generic situation, some of the eigenvalues of the

tridiagonal matrix $\mathcal{H}_m$ will start approximating corresponding eigenvalues of $\mathcal{A}$ when m becomes large enough. An eigenvalue $\tilde{\lambda}$ of $\mathcal{H}_m$ is called a Ritz value, and if $\mathbf{y}$ is an associated eigenvector, then the vector $\mathcal{V}_m\mathbf{y}$ is, by definition, the Ritz vector, i.e., the approximate eigenvector of $\mathcal{A}$ associated with $\tilde{\lambda}$. If m is large enough, the process may yield good approximations to the desired eigenvalues $\lambda_1, \ldots, \lambda_s$ of $\mathcal{H}$, corresponding to the occupied states, i.e., all occupied eigenstates.

There are several practical implementations of this basic scheme. All that was said above is what happens in theory. In practice, orthogonality of the Lanczos vectors, which is guaranteed in theory, is lost as soon as one of the eigenvectors starts to converge [141]. As such, a number of schemes have been developed to enforce the orthogonality of the Lanczos vectors; see [185, 109, 186, 108, 206]. The most common method consists of building a scalar recurrence, which parallels the three-term recurrence of the Lanczos vectors and models the loss of orthogonality. As soon as loss of orthogonality is detected, a reorthogonalization step is taken. This is the approach taken in the computational codes PROPACK [108] and PLAN [206]. In these codes, semiorthogonality is enforced, i.e., the inner product of two basis vectors is only guaranteed not to exceed a certain threshold of the order of $\sqrt{\varepsilon}$, where ε is the machine epsilon [72].

Since the eigenvectors are not needed individually, one might think of not computing them but rather just using a Lanczos basis $\mathcal{V}_m = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$ directly. This does not provide a good basis in general. However, a full Lanczos algorithm without partial reorthogonalization can work quite well when combined with a good stopping criterion.

A simple scheme used in [12] is to monitor the eigenvalues of the tridiagonal matrices $\mathcal{T}_i$, $i = 1, \ldots, m$. The cost for computing only the eigenvalues of $\mathcal{T}_i$ is $O(i^2)$. If we were to apply the test at every single step of the procedure, the total cost for all m Lanczos steps would be $O(m^3)$, which can be quite high. This cost can be reduced drastically, to the point of becoming negligible relative to the overall cost, by employing a number of simple strategies. For example, one can monitor the eigenvalues of the tridiagonal matrix $\mathcal{T}_i$ at fixed intervals, i.e., when $MOD(i, s) = 0$, where s is a certain fixed stride. Of course, large values of s will induce infrequent convergence tests, thus reducing the cost from $O(m^3)$ to $O(\frac{m^3}{3s})$. On the other hand, a large stride may inflict unnecessary $O(s)$ additional Lanczos steps before convergence is detected.

Though not implemented in [12], a better strategy is to use the bisection algorithm (see [72], sec. 8.5) to track the latest eigenvalue that has converged, exploiting the important property that the Lanczos procedure is a variational technique in the sense that when an eigenvalue converges, later steps can only improve it. In addition, convergence tends to occur from left to right in the spectrum, meaning that typically the smallest eigenvalue converges first, followed by the second smallest, etc. This suggests many simple procedures based on the bisection algorithm. When convergence has been detected (say, at step $l \leq m$), then the charge densities are approximated as the squares of the norms of the associated eigenvectors. See [12] for details.

Another popular algorithm for extracting the eigenpairs is the Davidson [132] method, which can be viewed as a preconditioned version of the Lanczos algorithm, in which the preconditioner is the diagonal of $\mathcal{A}$. We refer to the generalized Davidson algorithm as a Davidson approach in which the preconditioner is not restricted to being a diagonal matrix. (A detailed description can be found in [168].)

The Davidson algorithm differs from the Lanczos method in the way in which it defines new vectors to add to the projection subspace. Instead of adding just $\mathcal{A}\mathbf{v}_j$, it preconditions a given residual vector $\mathbf{r}_i = (\mathcal{A} - \mu_i\mathcal{I})\mathbf{u}_i$ and adds it to the subspace

(after orthogonalizing it against current basis vectors). The algorithm consists of an "eigenvalue" loop, which computes the desired eigenvalues one by one (or a few at a time), and a "basis" loop, which gradually computes the subspace on which to perform the projection. Consider the eigenvalue loop, which computes the ith eigenvalue and eigenvector of $\mathcal{A}$. If $\mathcal{M}$ is the current preconditioner, and $\mathcal{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ is the current basis, the main steps of the main loop are as follows:

1. Compute the ith eigenpair $(\mu_k, \mathbf{y}_k)$ of $\mathcal{C}_k = \mathcal{V}_k^T \mathcal{A} \mathcal{V}_k$.
2. Compute the residual vector $\mathbf{r}_k = (\mathcal{A} - \mu_k \mathcal{I}) \mathcal{V}_k \mathbf{y}_k$.
3. Precondition $\mathbf{r}_k$, i.e., compute $\mathbf{t}_k = \mathcal{M}^{-1} \mathbf{r}_k$.
4. Orthonormalize $\mathbf{t}_k$ against $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and call $\mathbf{v}_{k+1}$ the resulting vector, so $\mathcal{V}_{k+1} = [\mathcal{V}_k, \mathbf{v}_{k+1}]$.
5. Compute the last column-row of $\mathcal{C}_{k+1} = \mathcal{V}_{k+1}^T \mathcal{A} \mathcal{V}_{k+1}$.

At this point, one needs to decide on the choice of a preconditioner. The original Davidson approach used the diagonal of the matrix as a preconditioner, but this works only for limited cases. For a plane wave basis, it is possible to construct fairly effective preconditioners by exploiting the lower order bases. By this, we mean that if $\mathcal{H}_k$ is the matrix representation obtained by using k plane waves, we can construct a good approximation to $\mathcal{H}_k$ from $\mathcal{H}_m$, with $m \ll k$, by completing it with a diagonal matrix representing the larger (undesirable) modes. Note that these matrices are not explicitly computed as they are dense. This possibility of building lower-dimensional approximations to the Hamiltonian, which can be used to precondition the original matrix, constitutes an advantage of plane-wave-based methods.

Preconditioning techniques in this approach are typically based on filtering ideas and the fact that the Laplacian is an elliptic operator [195]. The eigenvectors corresponding to the few lowest eigenvalues of ∇^2 are smooth functions, and so are the corresponding wave functions. When an approximate eigenvector is known at the points of the grid, a smoother eigenvector can be obtained by averaging the value at every point with the values of its neighboring points. Assuming a Cartesian (x, y, z) coordinate system, the low frequency filter acting on the value of the wave function at the point (i, j, k), which represents one element of the eigenvector, is described by

$$(38) \quad \left[\frac{\psi_{i-1,j,k} + \psi_{i,j-1,k} + \psi_{i,j,k-1} + \psi_{i+1,j,k} + \psi_{i,j+1,k} + \psi_{i,j,k+1}}{12} \right] + \frac{\psi_{i,j,k}}{2} \ \rightarrow \ (\psi_{i,j,k})_{Filtered} \ .$$

Other preconditioners that have been tried result in mixed success. The use of shift-and-invert [141] involves solving linear systems with $\mathcal{A} - \sigma \mathcal{I}$, where $\mathcal{A}$ is the original matrix and the shift σ is close to the desired eigenvalue(s). These methods would be prohibitively expensive in most situations, given the size of the matrix and the number of times that $\mathcal{A} - \sigma \mathcal{I}$ must be factored. Alternatives based on an approximate factorization such as ILUT [167] are ineffective beyond the first few eigenvalues. Methods based on approximate inverse techniques have been somewhat more successful, performing better than filtering at additional preprocessing and storage costs. Preconditioning "interior" eigenvalues, i.e., eigenvalues located well inside the interval containing the spectrum, is still a very hard problem. Current solutions only attempt to dampen the effect of eigenvalues which are far away from those being computed. This is in effect what is achieved by filtering and sparse approximate inverse preconditioners. These techniques do not reduce the number of steps required for convergence in the same way that shift-and-invert techniques do. However, filtering techniques are inexpensive to apply and result in nonnegligible savings in iterations.

In real space it is trivial to operate with the potential term, which is represented by a diagonal matrix, and in Fourier space it is trivial to operate with the Laplacian term, which is also represented by a diagonal matrix. The use of plane wave bases also leads to natural preconditioning techniques which are obtained by simply employing a matrix obtained from a smaller plane wave basis, neglecting the effect of high frequency terms on the potential.

Real-space algorithms avoid the use of FFTs by performing all calculations in real physical space instead of Fourier space. FFTs require global communication; as such, they tend to be harder to implement on message-passing distributed memory multiprocessor systems. The only global operation remaining in *real-space* approaches is that of the inner products. These inner products are required when forming the orthogonal basis used in the generalized Davidson procedure. Inner products will scale well as long as the vector sizes in each processor remain relatively large.

5.3. Diagonalization Methods in Current Computational Codes. Table 5.1 shows a number of computational codes currently available or in development. This list is by no means exhaustive. What is rather remarkable is the time-frame in which these codes have been developed and the speed with which they have been adapted to new computing environments. Most of them have been coded in Fortran90/95 and most offer parallel versions with either MPI or OpenMP. (An interesting account of the impact of new software engineering methods in electronic structure codes can be found in [180].) The middle column of the table shows the type of discretization (basis) used, where PW stands for plane waves, RS for real space, LCAO for linear combination of atomic orbitals, APW for augmented plane waves, Gauss for Gaussian orbitals, and OTH for other. As can be noted, most codes use plane wave bases.

A few of the codes have not been updated in a few years; we only list those for which a website is still maintained. A star next to the code name indicates that the code has restricted distribution (e.g., DoD PW), or that it is still in a development phase. We separated the codes which use the GPL license from the codes which can be downloaded directly. The former are the first 6 listed in the table, and they are separated by a horizontal line from the others. All others require either a fee (e.g., VASP, Wien2K, phi98PP, and CASTEP) or a licensing agreement (no fee).

Not all these codes resort to some form of diagonalization. For example, the CPMD code [1] uses the Car–Parrinello approach, which relies entirely on a statistical approach and molecular dynamics to minimize the energy (see section 6.4). Similarly, the CONQUEST code is a linear-scaling method which uses a density matrix approach (see the next section). In addition, the codes using an LCAO basis obtain dense matrices and resort to standard dense matrix diagonalization.

The earliest electronic structures codes used variants of the subspace iteration algorithm [122]. There should therefore be no surprise that many existing codes propose improved versions of this scheme. For example, ABINIT [73] uses a form of subspace iteration in which an initial subspace is selected and then an iterative scheme is invoked to "improve" the basis vectors individually by some form of preconditioned conjugate gradient (CG) algorithm. In this approach, orthogonality is enforced as a Rayleigh–Ritz procedure and is used once each eigenvector is modified. ABINIT offers a block version of the same algorithm (with parallelism across the different vectors in the block) and an alternative which minimizes residual norms.

The Vienna Ab-initio Simulation Package (VASP) [104, 103] uses three main diagonalization schemes. The first, similar to ABINIT, is a form of subspace iteration in which the wave functions are individually refined by either the CG algorithm or by

Table 5.1 *A few available DFT computational codes for electronic structures.*

Code	Discr./Basis	URL
PWscf	PW	http://www.pwscf.org/
PEtot	PW	http://crd.lbl.gov/~linwang/PEtot/PEtot.html
ABINIT	PW	http://www.abinit.org/
Socorro	PW+RS	http://dft.sandia.gov/Socorro/mainpage.html
PARSEC	RS	http://parsec.ices.utexas.edu/
GPAW	RS	https://wiki.fysik.dtu.dk/gpaw/index.html
fhi98PP	PW	http://www.fhi-berlin.mpg.de/th/fhi98md/fhi98PP/
VASP	PW	http://cms.mpi.univie.ac.at/vasp/vasp/vasp.html
PARATEC	PW	http://www.nersc.gov/projects/paratec/
SeqQuest*	LCAO	http://dft.sandia.gov/Quest/
Wien2K	APW	http://www.wien2k.at/
SIESTA	LCAO	http://www.icmab.es/siesta/
AIMPRO	Gauss	http://aimpro.ncl.ac.uk/
FLEUR	APW	http://www.flapw.de/
CPMD	PW	http://www.cpmd.org/
CONQUEST	RS/OTH	http://www.conquest.ucl.ac.uk/
CASTEP	PW	http://www.tcm.phy.cam.ac.uk/castep/
DoD PW *	PW	http://cst-www.nrl.navy.mil/people/singh/planewave
CRYSTAL	Gauss	http://www.cse.clrc.ac.uk/cmg/CRYSTAL/
Octopus	RS	http://www.tddft.org/programs/octopus/
ASW	Other	http://www.physik.uni-augsburg.de/Neyert/aswhome.shtml
Exciting	APW	http://exciting.sourceforge.net/
ADF	LCAO	http://www.scm.com
NWChem	PW+Gauss	http://www.emsl.pnl.gov/docs/nwchem/capabilities/nwchem_capab.html

a method called the direct inversion in the iterative subspace (DIIS) iteration. The CG method is adapted from a method suggested by Teter, Payne, and Allen [191]. It consists of a preconditioned CG algorithm for minimizing the Rayleigh quotient. The preconditioner is an astutely adjusted diagonal matrix in plane wave space where the kinetic part of the Hamiltonian is diagonal. A few steps of this scheme are applied to each vector of the basis, and once this is done the new basis is orthogonalized in preparation for a Rayleigh–Ritz projection. The process is repeated until self-consistence is achieved. Note that ABINIT uses a variant of this scheme as well. A major drawback of this scheme is the requirement to always orthogonalize the current (preconditioner) residual vectors in CG against all other eigenvectors. This is necessary because the method essentially consists of minimizing the Rayleigh quotient in the space orthogonal to all other eigenvectors. Without it, the method would only compute one eigenvalue, namely, the smallest one.

The second method in VASP avoids this problem by minimizing $\|(\mathcal{A} - \mu(\mathbf{u})\mathcal{I})\mathbf{u}\|_2$ instead of the Rayleigh quotient. This represents the norm of the residual vector, hence the name residual minimization method (RMM). The minimization itself is done with the DIIS method, which is a form of Krylov subspace method dveloped by Pulay [158] in the early 1980s.[1] In the second scheme employed by VASP, an initial subspace is selected and then an iterative scheme is invoked to "improve" the basis vectors individually by minimizing the residual norms. In this approach, there is no need to orthogonalize each vector against all others after each update to the basis

---

[1]It is remarkable to note here that this work parallels the work of many others in scientific computing, working on solving (nonsymmetric) linear systems of equations, who were not aware of this development.

vectors. Finally, the third alternative method proposed in VASP is the Davidson approach with a form of block preconditioning. This is recommended as a more robust alternative in the documentation, though it is also mentioned as being more costly in some cases. This approach will be revisited in the next section.

The PWscf code (part of Espresso) [6] uses similar general methods to VASP. The default diagonalization algorithm is the Davidson method. There are also subspace-type methods offered with CG-like band-by-band diagonalization, or DIIS-like diagonalization.

The Octopus code focuses on time-dependent DFT and can handle excited states. Recall that DFT is only applicable to the ground state. Octopus can also obtain static properties generally computable from DFT codes, such as static polarizabilities and ground-state geometries, but the authors warn that the code is not optimized for this purpose.

GPAW stands for the grid-based projector-augmented wave method and is a relatively recent addition to the list of codes available [133]. It is written in Python and is part of the CAMD Open Software Project (CAMPOS) [24], which is a collection of programs for atomistic-scale simulations. GPAW replaces MIKA (multigrid instead of the K-spAce [196, 81]). GPAW works in the real space and uses a multigrid approach for diagonalization. The methods in GPAW are once more inspired by subspace iteration, the main difference being that CG or DIIS is replaced by a multigrid approach. As the levels are crossed, there is no orthogonalization at each level.

Quite a few papers in the early to mid-1990s were devoted to using the standard CG algorithm in a more elaborate scheme which does not impose the constraint of orthogonality; see, e.g., [2, 189, 188, 191, 49, 148, 124, 62, 125] for a few references. Since these methods are more akin to optimization, we discuss them in the next section. A number of codes, e.g., SIESTA, adopted variants of these schemes.

It was observed by many that the Davidson approach is in fact more robust than methods based on local optimization. For example, the authors of [94] comment that "for relatively small submatrix sizes the Davidson method still gives correct results whereas the latter two frequently fail to do it." The other two methods to which the paper refers are a form of subspace iteration (modification of RITZIT code) with DIIS preconditioning and a form of CG minimization. The observation that Davidson is a more robust approach is not a unanimous viewpoint. For example, developers of PWscf and VASP seem to recommend direct minimization, in spite of its less favorable speed. Clearly, implementation is a key factor. We believe that with proper implementation, a Davidson- or Krylov-based approach should be vastly superior to direct minimization.

6. The Optimization Path: Avoiding the Eigenvalue Problem. From one viewpoint, there is no need to refer to eigenvalues in order to minimize the total energy, and this provided a number of alternative methods used in electronic structures. Indeed, the stationary states of (5) are eigenfunctions of the Hamiltonian, but one can also just consider (4) as an optimization problem in itself.

6.1. Optimization Approaches without Orthogonality. In reading about computational schemes that have been proposed in the literature on DFT, one cannot fail to observe that the most commonly mentioned challenge or burden is that associated with the need to orthogonalize a given basis which approximates the desired eigenbasis. It is therefore only natural that major efforts have been devoted to designing algorithms which do not require orthogonalization or that attempt to reduce its cost. A number of these algorithms have been exploited in the context of order-N ($\mathcal{O}(N)$)

methods as well as in standard methods. The end of the previous section alluded to this approach, which seeks to compute a subspace as represented by a basis. If the basis, call it $\mathcal{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$, is orthogonal, then the problem of minimizing the energy is equivalent to that of minimizing the trace of $\mathcal{V}^* \mathcal{A} \mathcal{V}$. Thus, it is possible to formulate the problem as that of computing a set of vectors such that $tr(\mathcal{V}^* \mathcal{A} \mathcal{V})$ is minimized subject to the constraint $\mathcal{V}^* \mathcal{V} = \mathcal{I}$. Note that an algorithm for explicitly minimizing the trace was developed by Sameh [169] as far back as in 1982, motivated in part by parallelism, but this does not seem to have been noticed by researchers on the applied side.

Many authors have considered a related approach in which the orthogonality constraint is not enforced. In this situation, the problem is equivalent to minimizing $\mathcal{S}^{-1}\mathcal{A}$ or $\mathcal{S}^{-1/2}\mathcal{A}\mathcal{S}^{-1/2}$, where $\mathcal{S}$ is the "overlap" matrix, i.e., the matrix $\mathcal{S} = \mathcal{V}^*\mathcal{V}$, the "overlap" matrix, is only approximately inverted. For example, the simple Neumann series expansion

$$\mathcal{S}^{-1} \approx \sum_{i=0}^{k} \mathcal{S}^i$$

can be used [189, 2, 124, 125, 62].

The paper [49] examines in detail algorithms that minimize energy on Grassmann and Stiefel manifolds, i.e., manifolds of matrices that satisfy orthogonality constraints. In these algorithms, the iterates evolve by following geodesic paths on the manifolds (ideal case). The cost of the ideal case algorithm requires the singular value decomposition (SVD) [72] of matrices of size $n \times p$ (the same size as that of the basis of the desired subspace), and so the authors of [49] show that quadratic convergence can be achieved if the directions used by the algorithms are only approximate. Related practical approaches in DFT, based on minimization with constraints, are discussed in [86] and [204]. Other approaches taken consist of making use of the McWeeny [127] projection, which will be discussed shortly.

6.2. Density Matrix Approaches. As was previously discussed, one can bypass the eigenvalue problem and focus instead on the whole subspace spanned by the occupied eigenvalues as an alternative to computing individual eigenfunctions. We also mentioned this viewpoint in the standard context of eigenvector-based methods when we discussed adapting the Lanczos algorithm for DFT. However, the methods that rely on the density matrix approach go much further by not even referring to eigenspaces. Instead, they attempt to compute directly the eigenprojector associated with these eigenvalues.

Charge-density methods have played a major role since the early days of quantum theory. Such methods were used in particular by Thomas [194] and Fermi [53] as far back as in 1927, who made some of the first realistic attempts at yielding the atomic structure of atoms. They gave qualitatively correct information on the electronic structure of atoms, but were fundamentally flawed as they did not describe the observed shell structure of the atom.

Modern density methods are based on the "density matrix." The density matrix can be viewed as a function of a pair of coordinates, $\rho(\mathbf{r}, \mathbf{r}')$. It is defined by

$$(39) \qquad \rho(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{N} \psi_i(\mathbf{r})^* \psi_i(\mathbf{r}') \ .$$

Note that, after discretization, the function ψ becomes a vector of length N whose ith component is the approximation of ψ at the mesh-point $\mathbf{r}_i$. If we call $\mathcal{V}$ the matrix whose column vectors are the (normalized) eigenvectors ψ_i, $i = 1, \ldots, s$, for the s occupied states, then

$$(40) \qquad \mathcal{P} = \mathcal{V}\mathcal{V}^*$$

is a projector and the charge density at a given point $\mathbf{r}_i$ in space is the ith diagonal element of $\mathcal{P}$. A number of techniques have been developed based on using this observation [7, 113, 181]. Here, we will sketch only the main ideas.

Density matrix methods are prevalent mainly in the $\mathcal{O}(N)$ methods. These methods are based on constructing an approximation to the projector $\mathcal{P}$ in (40) without knowledge of eigenvectors. Denote by p_{ij} the coefficients of the matrix $\mathcal{P}$. A number of properties are exploited for this purpose. First are the following two relations:

$$\mathrm{tr}[\mathcal{P}] = \sum_i p_{ii} \quad = \quad \text{particle number,}$$

$$\mathrm{tr}[\mathcal{P}\mathcal{H}] = \sum_{i,j} p_{ij}\mathcal{H}_{ji} \quad = \quad \text{system energy.}$$

The first relation is a consequence of the fact that each p_{ii} represents the charge density at point $\mathbf{r}_i$, as was mentioned above. The second is a consequence of the fact that $\mathcal{P}\mathcal{H} = \mathcal{P}\mathcal{H}\mathcal{P}$ is the restriction of $\mathcal{H}$ to the invariant subspace associated with eigenvalues corresponding to the occupied states. The trace of $\mathcal{P}\mathcal{H}$ is the sum of these eigenvalues, which is the total energy of the system assuming a "Hartree" approximation, i.e., assuming the total energy is the sum of the occupied eigenvalues.

Another important property that is exploited is the physical fact that entries of $\mathcal{P}$ *decay away from the main diagonal*. Hence, the idea is to try to find a $\mathcal{P}$ whose trace is fixed and which minimizes the trace of $\mathcal{P}\mathcal{H}$. The trace constraint can be avoided by shifting $\mathcal{H}$,

$$\mathrm{tr}[\mathcal{P}(\mathcal{H} - \mu\mathcal{I})] = \mathrm{tr}[\mathcal{P}\mathcal{H}] - \mu N_e.$$

The optimization problem is not yet well posed, since without constraints on $\mathcal{P}$, the minimum can be arbitrarily negative or small. The missing constraint is to force $\mathcal{P}$ to be a projector. This can be achieved by forcing its eigenvalues to be between zero and one [113]. The minimization will yield a matrix $\mathcal{P}$ that has eigenvalues equal to either one or zero and satisfies the desired idempotent constraint automatically.

One strategy that was used in [113] for this purpose is to seek $\mathcal{P}$ in the form

$$\mathcal{P} = 3\mathcal{S}^2 - 2\mathcal{S}^3.$$

If the eigenvalues of $\mathcal{S}$ are in the range $[-0.5, 1.5]$, this transformation will map them into $[0,1]$. The procedure then is to seek a banded $\mathcal{S}$ that minimizes

$$\mathrm{tr}[(3\mathcal{S}^2 - 2\mathcal{S}^3)(\mathcal{H} - \mu\mathcal{I})]$$

using descent-type algorithms. The gradient of the above function is computable. This is referred to as McWeeny purification [127].

Many other linearly scaling methods have been developed. As two examples, we mention the work of Kim, Mauri, and Galli [95], who minimize total energy by using

nonorthogonal localized orbitals, and the work by Challacombe [28, 29] that exploits general sparse matrix tools in the form of sparse approximate inverses developed by Benzi, Meyer, and Tůma [13, 14].

The drawback of this approach is its lack of accuracy. It is also not clear whether a minimum exists, because of the various constraints added, and whether there is a (unique) minimum. In addition, the band required for $\mathcal{S}$ may not be so small for certain systems. Nevertheless, it is interesting to note that explicit eigenvalue calculations are avoided. Some global approximation of the invariant subspace associated with these eigenvalues is extracted, meaning that a global approximation to the set of eigenfunctions is computed via the approximation to the projector $\mathcal{P}$.

In the previous discussion an orthogonal basis set was assumed, though density matrix methods can be also be defined with nonorthogonal basis sets. For example, one of the early implementations of the Hartree–Fock method, known as the Roothan method, involves a self-consistent (fixed point) iteration in which the unknown is the density matrix $\mathcal{P}$ and each wave function is expressed in a basis $\{\chi_k\}$, $k = 1, \ldots, K$,

$$(41) \qquad \phi_k = \sum_{j=1}^{K} c_{jk} \chi_j \ .$$

Formally, the Hartree–Fock energy (15) must be minimized with the constraint that the orbital be of the above form. The χ_j's are convenient and well-selected spatial basis orbitals associated with the atoms (atomic orbitals). For the purpose of simplifying notation we define $\mathcal{H}$ to be the Fock operator in expression (15) for a particular i, say, $i = 1$. Then a Galerkin-type approach would be to write $\mathcal{H}\phi_k = \varepsilon_k \phi_k$ in the space spanned by the χ_j's:

$$\left\langle \chi_i \ \middle| \ \mathcal{H} \ \middle| \sum_{j=1}^{K} c_{kj}\chi_j \right\rangle = \varepsilon_k \left\langle \chi_i \middle| \sum_{j=1}^{K} c_{jk}\chi_j \right\rangle \quad \text{for} \quad i = 1, \ldots, K$$

$$(42) \qquad \rightarrow \sum_{j=1}^{K} \langle \chi_i | \mathcal{H} | \chi_j \rangle c_{jk} = \varepsilon_k \sum_{j=1}^{K} \langle \chi_i | \chi_j \rangle c_{jk} \quad \text{for} \quad k = 1, \ldots, K.$$

If we denote by F the Fock matrix whose entries are $F_{ij} = \langle \chi_i | \mathcal{H} | \chi_j \rangle$ and by S the "overlap" matrix with entries $S_{ij} = \langle \chi_i | \chi_j \rangle$, then it is clear that (42) is a generalized eigenvalue problem of size K. An eigenvector is a column of the $K \times K$ matrix C with entries c_{jk} on the kth column and jth row. In matrix form the problem can be written as

$$FC = CSD,$$

where $D = \text{diag}(\varepsilon_k)$ is the diagonal matrix with entries ε_k. This problem can be solved with standard techniques for dense eigenproblems. The matrix C of eigenvectors is such that $C^T S C = I$. It is also of interest to look at the charge density in this context. The procedure can be written in terms of the density matrix, which is now defined as $P = CC^T$.

Equation (41) can be rewritten in the form

$$\begin{pmatrix} \phi_1(\mathbf{r}) \\ \phi_2(\mathbf{r}) \\ \vdots \\ \phi_N(\mathbf{r}) \end{pmatrix} = C^T \begin{pmatrix} \chi_1(\mathbf{r}) \\ \chi_2(\mathbf{r}) \\ \vdots \\ \chi_N(\mathbf{r}) \end{pmatrix} \ .$$

Since the charge density at a location $\mathbf{r}$ is the 2-norm of the above vector, then clearly

$$\rho(\mathbf{r}) = [\chi_1(\mathbf{r})^*, \chi_2(\mathbf{r})^*, \ldots, \chi_N(\mathbf{r})^*]CC^T \begin{pmatrix} \chi_1(\mathbf{r}) \\ \chi_2(\mathbf{r}) \\ \vdots \\ \chi_N(\mathbf{r}) \end{pmatrix}.$$

Note that the matrix S depends only on the basis $\{\chi_k\}$ selected, while F depends on the solution itself. So the problem is again a nonlinear one which needs to be solved by a self-consistent iteration of the following form:

1. Given $\{\chi_k\}_{k=1,\ldots,K}$ construct S. Get an initial set $\{\phi_j\}_{j=1,\ldots,K}$.
2. Compute the new matrix F.
3. Solve the eigenvalue problem $FSC = SCD$. Compute $P = CC^T$.
4. If P has not converged, then return to 2.

On the implementation side, it is useful to mention that the matrix F consists of two parts, which arise by writing $\mathcal{H}$ as the sum of two operators,

$$\mathcal{H} = \mathcal{H}^{core} + \mathcal{F},$$

where $\mathcal{H}^{core}$ corresponds to the first two terms in (20) and does not involve the ϕ_j's, and $\mathcal{F}$ contains the last two terms which depend on the ϕ_j's. If we write the general matrix term $f_{ij} = \langle \chi_i | \mathcal{H} | \chi_j \rangle$, then we obtain

$$f_{ij} = \langle \chi_i | \mathcal{H}^{core} | \chi_j \rangle + \langle \chi_i | \mathcal{F} | \chi_j \rangle .$$

Thus, the matrix F is the sum of two matrices. The first of these, call it F^{core}, is independent of the eigenfunctions ϕ_i's and needs to be evaluated only once. The second, call it F_ϕ, is a function of the ϕ_i's. As can be seen from the expression of F_ϕ, specifically (17)–(20), this will involve double-electron integrals of the form

$$\int \int d\mathbf{r} d\mathbf{r}' \frac{\chi_i^*(\mathbf{r})\chi_j(\mathbf{r})\chi_k^*(\mathbf{r}')\chi_l(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}.$$

The cost of the procedure is dominated by the fact that there are many such integrals to evaluate.

6.3. The "Car–Parrinello" Viewpoint. Car and Parrinello [25] took an approach which combined molecular dynamics with pseudopotentials and DFT by proposing a scheme that exploits heuristic optimization procedures to achieve the minimum energy. Specifically, they used simulated annealing [97] as a conceptual approach (and a simulated annealing strategy based on molecular dynamics in practice) to minimize an energy functional, which they write in the form

$$\mathcal{E}[\{\psi_i\}, \{\mathbf{R}_I\}, \{\alpha_v\}] = \sum_i \int_\Omega d^3\mathbf{r}_i \ \psi_i^*[-(\hbar^2/2m)\nabla^2]\psi_i(\mathbf{r}_i) + \mathcal{V}[\rho(\mathbf{r}_i), \{\mathbf{R}_I\}, \{\alpha_v\}],$$

where $\{\alpha_v\}$ are the external constraints imposed on the system such as the volume Ω or the strain $\varepsilon_{\mu\nu}$. $\mathcal{V}$ contains the internuclear Coulomb repulsion and the effective electronic potential energy, which includes the external nuclear, Hartree, and exchange and correlation terms.

They then use a Lagrangian formula to generate trajectories for ionic and electronic degrees of freedom via a coupled set of equations of motion. The idea is to

propagate the electronic wave functions, i.e., the Kohn–Sham orbitals, along the motions of the atoms. To this end, they consider the parameters $\{\psi_i\}$, $\{\mathbf{R}_I\}$, $\{\alpha_v\}$ in the energy functional to be time dependent and introduce the Lagrangian, $\mathcal{L}$, which is the difference in the kinetic and potential energies of the system:

$$\mathcal{L} = \sum_i \frac{1}{2}\mu \int_\Omega d^3\mathbf{r}_i \ |\dot{\psi}_i|^2 + \sum_I \frac{1}{2}M_I\dot{\mathbf{R}}_I^2 + \sum_v \frac{1}{2}\mu_v\dot{\alpha}_v^2 - \mathcal{E}[\{\psi_i\},\{\mathbf{R}_I\},\{\alpha_v\}],$$

where M_I are the physical ionic masses, μ and μ_v are arbitrary parameters having the appropriate units, and the ψ_i are subject to an orthonormality constraint.

The Lagrangian generates dynamics for the parameters through the following equations of motion:

$$(43) \qquad \mu\ddot{\psi}_i(\mathbf{r}_i,t) = -\delta\mathcal{E}/\delta\psi_i^*(\mathbf{r}_i,t) + \sum_k \Lambda_{ik}\psi_k(\mathbf{r_i},t),$$

$$(44) \qquad M_I\ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R_I}}\mathcal{E},$$

$$(45) \qquad \mu_v\ddot{\alpha}_v = -(\partial\mathcal{E}/\partial\alpha_v),$$

where the Lagrange multipliers, Λ_{ik}, have been introduced to satisfy the orthonormality constraint. Only the ion dynamics have real physical meaning; the rest are fictitious and are employed by the simulated annealing algorithm.

The Lagrangian formula defines both potential and kinetic energies for the system. The equilibrium value of the kinetic energy can be calculated as the temporal average over the trajectories generated by the equations of motion. By varying the velocities, the temperature of the system can be reduced; as $T \to 0$, the equilibrium state of the DFT energy functional is reached. At equilibrium, $\ddot{\psi}_i = 0$, and (43) corresponds to the Kohn–Sham equation through a unitary transformation. At this temperature, the eigenvalues of Λ agree with the occupied Kohn–Sham eigenvalues, and the Lagrangian describes a real physical system whose representative point lies on the Born–Oppenheimer surface.

The main advantage of this approach is that diagonalization, self-consistency, ionic relaxation, and volume and strain relaxation are achieved *simultaneously* rather than separately. However, there are also several disadvantages of this method due to the coupling between the real and fictitious energies. First, the computational time step must be rather small since the equations of motion involve both fast (electronic) and slow (nuclear) degrees of freedom [16]. Second, the method does not scale well. In particular, a single iteration scales as $\tilde{} MN^2$, where M is the size of the electronic basis and N is the number of particles; this is required to impose the holonomic orbital orthogonality constraint [16]. Third, the method can only handle systems in the ground state, which means the electronic degrees of freedom must always reside near the Born–Oppenheimer surface [16]. Fourth, energy can slosh between the electronic and nuclear variables [99].

Pastore, Smargiassi, and Buda investigated the theoretical basis of the Car–Parrinello method in [145]. There they showed how the classical dynamics generated by the Car–Parrinello Lagrangian approximated efficiently the quantum adiabatic evolution of a system, and they discussed the role played by the spectrum of the eigenvalues of the Kohn–Sham Hamiltonian matrix.

The Car–Parrinello method is one of several ab initio molecular dynamics (AIMD) methods. A discussion of AIMD methods is beyond the scope of this paper. However, the interested reader is referred to [146, 63, 198, 142] for descriptions of AIMD methods

and their diverse applications, which include the melting of silicon and the study of
molecular crystals and liquids.

The Car–Parrinello method has been used extensively in materials science, physics,
chemistry, and biology since its invention. The Car–Parrinello Molecular Dynamics
Consortium website [192] lists numerous papers that have been published on this
method since 1994; however, this list is not exhaustive. To give an idea of the wide
range of applications studied by this method, we list several papers in materials sci-
ence [41, 90, 123, 208, 126], physics [55, 161, 26, 19, 172], chemistry [207, 67, 102,
154, 31], and biology [130, 66, 149, 164, 92] that employ this method. For information
on recent advances in chemistry and materials science with Car–Parrinello molecu-
lar dynamics methods, see [1]. For a nice review of the first 15 years of the use of
the Car–Parrinello methods in physics, chemistry, and biology, the reader is referred
to [163].

6.4. Use of Orthogonal Polynomials. Approximation theory, and especially the
theory of orthogonal polynomials, has been extensively used in DFT. A typical ap-
proach is to write the projector (40) as

$$\mathcal{P} = h(\mathcal{H}),$$

where $h(\lambda)$ is the shifted Heaviside step function which has value one for $\lambda \leq E_F$
and zero elsewhere. The Heaviside function can now be approximately expanded into
orthogonal polynomials,

$$h(\lambda) \approx \sum_{i=1}^{n} \alpha_i p_i(\lambda) \ .$$

The most common orthogonal polynomials that are used in this context are the Cheby-
shev polynomials of the first class. If a localized basis is used, such as Gaussian
orbitals, then it is known that the density matrix has a finite decay range; i.e., it
will be represented by a banded matrix when its smaller entries are neglected.[2] It is
therefore possible to construct a good approximation to $\mathcal{P}$ efficiently in this case; see,
e.g., [115, 89].

Another approach which has been used focuses not on $\mathcal{P}$ directly, but rather on
the basis $\mathcal{V}$; see, e.g., [85]. Here a number of trial vectors $\xi_1, \ldots, \xi_m$ are selected and
then a basis for the desired space span($\mathcal{V}$) is built by computing approximations to
the vectors

$$w_i = h(\mathcal{H})\xi_i$$

with the help of the polynomial expansion. Note that each of these vectors is a
member of the desired subspace and collectively the set of w_i's will constitute a basis
of the space under generic conditions. This set is then orthonormalized to get $\mathcal{V}$.

The above approaches attempt to extract the charge density $\rho(\mathbf{r}, \mathbf{r})$, which is
represented by the diagonal of the operator $\mathcal{P}$. One can easily imagine that while the
techniques should scale well for large systems, the prefactor in the cost function can
be very high. This was observed in [89]. There are, however, situations where the
use of orthogonal polynomials can be very cost-effective. In [162] the *density of states*

---

[2]In a plane wave basis the density matrix has a sparse (but not banded) structure when its small
entries are dropped. This fact does not seem to have been exploited in the literature so far.

(DOS) is computed using this strategy. One starts with a density of eigenvalues in $\mathbb{R}^n$ which is in the form of a sum of Dirac functions:

$$\eta(\lambda) = \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \ .$$

This is the distribution (in a mathematical sense) that is sought. The classical moment problem is to find this distribution from a knowledge of its classical moments $\mu_k = \int \lambda^k \eta(\lambda) d\lambda$.

A numerically viable alternative is to use the modified or generalized moments,

$$(46) \qquad \mu_k = \int t_k(\lambda)\eta(\lambda)d\lambda,$$

where $\{t_k\}$ is a sequence of orthogonal polynomials on the interval containing the eigenvalues of $\mathcal{H}$. Typically, the problem is shifted and scaled so that the action takes place in the interval $[-1, 1]$. In addition, the polynomials t_k are just the Chebyshev polynomials, though other polynomials can be used as well. Assuming that the moments have been computed for the Chebyshev polynomials, then the expression of the distribution in the interval $(-1, 1)$ is given by

$$(47) \qquad \eta(\lambda) = \frac{2}{\pi\sqrt{1 - \lambda^2}} \sum_{k=0}^{\infty} \frac{\mu_k}{1 + \delta_{k0}} \ t_k(\lambda) \ .$$

Here δ_{k0} is the Kronecker symbol. Of course, the sum is truncated at M terms, resulting in a certain function $\eta_M(\lambda)$. Another notable approximation lies in the computation of the moments μ_k. The sequence μ_k is not readily available and can be only approximated [203, 162]. The various methods proposed in the literature consist of using probabilistic arguments for this purpose. Specifically, μ_k can be written as

$$(48) \qquad \mu_k = \int_{-1}^{1} t_k(\lambda)\eta(\lambda)d\lambda = \sum_{i=1}^{N} t_k(\lambda_i) = \text{tr}[t_k(\mathcal{H})],$$

so one needs only compute the traces of the sequence of operators $t_k(\mathcal{H})$. These traces are typically computed with a Monte-Carlo-type technique. A sequence of n_r random vectors $\mathbf{r_j}$ is generated and μ_k is approximated by

$$\mu_k \approx \frac{1}{n_r} \sum_{j=1}^{n_r} \langle r_j | t_k(\mathcal{H}) | \mathbf{r_j} \rangle \ .$$

The three-term recurrence of the Chebyshev polynomials is exploited to reduce the memory and computational cost of the calculations.

This type of calculation for the DOS can only be of interest in cases where the geometry is fixed and the Hamiltonian can be well approximated without a self-consistent iteration. A related, but more complex, technique allows us to obtain *optical absorption spectra* (OAS) [32, 203]. The calculation can be viewed as an extension of the problem discussed above to two variables. What is sought is the 2-variable distribution

$$(49) \qquad \eta(\lambda_1, \lambda_2) = \sum_{i,j} |\sigma_{i,j}|^2 \delta(\lambda_1 - \lambda_i)\delta(\lambda_2 - \lambda_i) \quad \text{with} \quad \sigma_{i,j} = \langle \phi_i | i\hbar\nabla | \phi_j \rangle,$$

from which the OAS can be obtained by computing a double integral [203]. Physically, $|\sigma_{i,j}|^2$ represents the transition probability between states i and j. To compute the function (49) requires the generalization of the 1-variable moments defined by (48) to two 2-variable moments:

$$\mu_{k,l} = \sum_{i,j} \ t_k(\lambda_j)t_k(\lambda_i)t_l(\lambda_j)|\sigma_{i,j}|^2 = \int_{-1}^{1} d\lambda_1 \int_{-1}^{1} d\lambda_2 \ \ t_k(\lambda_1)t_l(\lambda_2)\eta(\lambda_1, \lambda_2) \ .$$

A probabilistic technique that is similar to the one for the DOS is used to recover an approximation to the function (49). In this case, approximating $\mu_{k,l}$ consists of averaging terms of the form $\langle \mathbf{r}|\mathbf{i}\hbar\nabla\mathbf{t_k}(\mathcal{H}).\mathbf{i}\hbar\nabla\mathbf{t_k}(\mathcal{H})|\mathbf{r}\rangle$, where $\mathbf{r}$ is a random vector as before [203].

7. Geometry Optimization. The composition and structure of a material (including its geometry) determine many of its physical and chemical properties. For example, the reactivity, polarity, phase, color, and magnetism of a material are determined, in large part, by the material's geometry [128]. Thus, it is important to determine the geometry of the material in a stable state.

7.1. The Geometry Optimization Problem. The *geometry optimization problem* (referred to as the *structural relaxation problem* by materials scientists and computational physicists) is to find a low-energy configuration of the material. This is done by repositioning the atoms of the material and subsequently evaluating its energy at various places on the potential energy surface. The potential energy surface is a hypersurface and is a plot of the potential energy, $\mathcal{E}$, vs. the atomic coordinates, $\mathbf{r}$. Here $\mathbf{r}$ is a vector of length $3K$ containing the atomic coordinates for the K atoms, i.e., $\mathbf{r} = (\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, \ldots, \mathbf{x}_K, \mathbf{y}_K, \mathbf{z}_K)^T$. There have been many reviews (see, e.g., [159, 173, 174, 176, 175]) written on this topic in recent years.

There are some applications which require the lowest-energy conformation of the system of interest, i.e., a global minimum of the potential energy surface [176]. Finding a global minimum is a very difficult optimization problem; often it is impossible or impractical to find a global minimum. However, there are many applications where it is enough to find a local minimum. For example, local minima can be used as starting points for global minimization algorithms [8, 9, 30, 101, 112, 114, 150, 151, 171, 170, 201].

Even though finding a local minimum is an easier problem than finding a global minimum, it can be quite difficult. One problem that may occur is that the optimization algorithm may become trapped at a saddle point, for example. Because the gradient is zero at all critical points, the Hessian must be used to determine whether or not the critical point is a local minimum. At a local minimum, the Hessian matrix is positive definite. A second problem is that the optimization algorithm may not converge from all starting points, especially not from those points corresponding to very high potential energies. The methods we review in this paper focus on finding a local minimum (as opposed to a global minimum).

The geometry optimization problem is a nonlinear, unconstrained optimization problem. In optimization terms, the objective function and gradient are given by the potential energy and forces, respectively. There are four important qualities which serve to characterize the optimization problem. First, the objective function is highly nonlinear. For example, a simple model that is often given for the problem is the Lennard–Jones potential, which describes the potential energy of two inert gas atoms

in locations $\mathbf{r_1}$ and $\mathbf{r_2}$:

$$\mathcal{E}(\mathbf{r}_1, \mathbf{r}_2) = 4\epsilon \left[\left(\frac{\sigma}{|\mathbf{r}_1 - \mathbf{r}_2|} \right)^{12} - \left(\frac{\sigma}{|\mathbf{r}_1 - \mathbf{r}_2|} \right)^{6} \right].$$

Here, ϵ is the well-depth and σ is a hard sphere radius.

Second, it is very expensive to evaluate the energy and forces of the material for a particular geometry. For the PARSEC package [106, 143] on which the authors work, the self-consistent field iteration, corresponding to the solution of the Kohn–Sham equations, must be executed each time the energy and forces are evaluated. This corresponds to solving the nonlinear eigenvalue problem in (31). For this reason, it is impractical to compute the Hessian via finite differencing, and it is not possible to compute the Hessian via automatic differentiation due to the structure of the nonlinear eigenvalue problem. Third, the energy and force fields often contain inaccuracies, as it is difficult to know the potential energy surface exactly. Finally, there can be many local minima, any of which can be considered acceptable solutions to the geometry optimization problem.

7.2. Minimization Algorithms. Many different minimization algorithms are employed by electronic structures packages to solve the geometry optimization problem. Table 7.1 gives the type of minimization algorithm used by each DFT electronic structure package listed in Table 5.1. These minimization algorithms fall into six main categories: the steepest descent method, quasi-Newton methods, truncated Newton methods, CG methods, iterative subspace methods, and molecular dynamics methods. In this section, we describe the first five classes of methods and review algorithms from each category. We give just a brief description of one method in the sixth category,

Table 7.1 *Minimization algorithms for a few electronic structures codes.*

Code	Minimization algorithm
PWscf	damped dynamics
PEtot	BFGS
ABINIT	molecular dynamics (Numerov, Verlet)
Socorro	steepest descent, CG, quenched minimization
PARSEC	L-BFGS
GPAW	molecular dynamics (Verlet), CG, quasi-Newton (in CAMPOS Group)
fhi98PP	ionic relaxation with damped Verlet
VASP	CG, RMM-DIIS
PARATEC	direct minimization of total energy
SeqQuest	modified Broyden, damped dynamics, steepest descent, accelerated steepest descent
Wien2K	geometry optimization details not given
SIESTA	CG, molecular dynamics algorithms (including standard Verlet)
AIMPRO	CG
FLEUR	geometry optimization details not given
CPMD	GDIIS, L-BFGS, P-RFO, RFO, BFGS, steepest descent
CONQUEST	geometry optimization details not given
CASTEP	BFGS, damped molecular dynamics, delocalized internal coordinates
DoD PW	damped dynamics
CRYSTAL	Berny (modified CG) algorithm
Octopus	steepest descent
ASW	geometry optimization details not given
Exciting	geometry optimization details not given
ADF	quasi-Newton method
NWChem	CG and BFGS algorithm

molecular dynamics methods, since such methods are beyond the scope of this review paper.

7.2.1. The Steepest Descent Method. Steepest descent methods exploit the direction, $\mathbf{d}$, that yields the fastest decrease in $\mathcal{E}$ from $\mathbf{r}$. Mathematically, this direction results from the solution of the optimization problem

$$(50) \qquad \min_{\mathbf{d} \in \mathbb{R}^{3K}} \nabla \mathcal{E}(\mathbf{r})^T \mathbf{d} \quad \text{subject to } \|\mathbf{d}\| = 1 .$$

The solution to the above minimization problem is given by $\mathbf{d} = -\nabla \mathcal{E}(\mathbf{r}) / \|\nabla \mathcal{E}(\mathbf{r})\|_2$ in the l_2-norm. This is the *steepest descent direction.*

Iteration k of Cauchy's classical steepest descent algorithm [47] is as follows:

1. Find the lowest point of $\mathcal{E}$ in the direction $-\nabla \mathcal{E}(\mathbf{r}_k)$ from $\mathbf{r}_k$, i.e., find λ_k that solves $\min_{\lambda_k > 0} \mathcal{E}(\mathbf{r}_k - \lambda_k \nabla \mathcal{E}(\mathbf{r}_k))$.
2. Update $\mathbf{r}_{k+1} = \mathbf{r}_k - \lambda_k \nabla \mathcal{E}(\mathbf{r}_k)$.

Clearly, this is only a theoretical algorithm in that the first step requires the solution of a one-dimensional minimization problem. In order to implement the algorithm, an inexact minimization must be performed. Goldstein [71] showed that under mild conditions, the steepest descent algorithm converges to a local minimum or a saddle point of $\mathcal{E}$. However, the convergence is only linear.

Several electronic structure codes (e.g., Socorro, SeqQuest, CPMD, and Octopus) employ the steepest descent algorithm for the geometry optimization problem because it is easy to implement. The ease of implementation comes at the cost of slow convergence. More sophisticated minimization algorithms usually yield a better convergence rate and are more cost effective.

7.2.2. Newton's Method. An example of an optimization algorithm with a higher rate of convergence is Newton's method. It enjoys quadratic convergence in the best case. The goal of Newton's method is to find a point $\mathbf{r}_{k+1}$ such that $\nabla \mathcal{E}(\mathbf{r}_{k+1}) = \mathbf{0}$. For such a point, $\mathbf{r}_{k+1}$ satisfies a necessary condition for being a minimizer of $\mathcal{E}$. In order to do this, a quadratic model, m_k, of the function is created. This model is given by $m_k(\mathbf{r}_k + \mathbf{d}) = \mathcal{E}(\mathbf{r}_k) + \nabla \mathcal{E}(\mathbf{r}_k)^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \nabla^2 \mathcal{E}(\mathbf{r}_k) \mathbf{d}$. Then the point $\mathbf{r}_{k+1} = \mathbf{r}_k + \mathbf{d}_k^N$ is determined so that $\nabla m_k(\mathbf{r}_{k+1}) = \mathbf{0}$, making $\mathbf{r}_{k+1}$ a critical point of m_k. The vector $\mathbf{d}_k^N$ is called the Newton step.

Formally, iteration k of Newton's method is written as the following two-step procedure:

1. Solve $\nabla^2 \mathcal{E}(\mathbf{r}_k) \mathbf{d}_k^N = -\nabla \mathcal{E}(\mathbf{r}_k)$.
2. Update $\mathbf{r}_{k+1} = \mathbf{r}_k + \mathbf{d}_k^N$.

There are many difficulties associated with this simple version of Newton's method. First, the Newton direction might be a direction of negative curvature, i.e., it might not be a descent direction. Second, if the Hessian matrix is ill-conditioned, the rate of convergence might be decreased. Third, Newton's method is not globally convergent, in general. However, if the Hessian matrices are positive definite with a uniformly bounded condition number and if the step lengths satisfy the Wolfe conditions, then Newton's method is globally convergent because the assumptions of Zoutendijk's theorem (which gives sufficient conditions for linesearch algorithms to be globally convergent) are then satisfied for $\lambda_k = 1$ [139].

Another major difficulty associated with using Newton's method for electronic structure calculations is that, in most cases, an analytic Hessian is not available. Newton's method (as written) is not a viable option for electronic structure codes. Therefore, we turn to quasi-Newton methods, which employ approximate Hessians.

7.2.3. Quasi-Newton Methods. Quasi-Newton methods are among the most successful approaches for the geometry optimization problem [173, 174, 176, 202]. The PEtot, SeqQuest, CPMD, CASTEP, and GPAW electronic structure packages (among others) all employ quasi-Newton methods, which are modified Newton methods in which the actual Hessian is replaced by an approximation to it. Typically, the approximation is obtained by updating an initial Hessian which may be a positive multiple of the identity, or it may come from experimental results or from optimizing the geometry at a lower level of theory. For example, if one is interested in performing a geometry optimization for the DFT level of theory, it may be possible to initialize the Hessian using the Hessian from a geometry optimization with a semiempirical force field. Another way of initializing the Hessian may be to use the Hessian from the geometry optimization of a related model problem. For example, in [209], Zhao, Wang, and Meza proposed a motif-based technique for estimating the initial Hessian matrix for the BFGS method to be described below. Their idea is to develop motifs based on just a few atoms representative of local bonding types and to calculate Hessian matrix elements for each motif which are then assembled together to approximate the Hessian matrix for the nanostructure.

The generic quasi-Newton method is the same as Newton's method except that $\nabla^2 \mathcal{E}(\mathbf{r}_k)$ is replaced by $\mathcal{B}_k \approx \nabla^2 \mathcal{E}(\mathbf{r}_k)$ in the computation of the Newton step. One way in which the many quasi-Newton methods differ is in their techniques for updating the Hessian. One major class of Hessian update formulas is that of the secant updates which enforce the quasi-Newton condition $\mathcal{B}_{k+1}\mathbf{s}_k = \mathbf{y}_k$, where $\mathbf{s}_k = \mathbf{r}_{k+1} - \mathbf{r}_k$ and $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$, where $\mathbf{g}_k = \nabla \mathcal{E}(\mathbf{r}_k)$. This condition is used to create low-rank updates for use in approximating the Hessian.

One of the most successful updates from this class has been the BFGS [23, 56, 70, 183] update, which was discovered independently by Broyden, Fletcher, Goldfarb, and Shanno in 1970. It is given by

$$\mathcal{B}_{k+1} = \mathcal{B}_k + \frac{\mathbf{y}_k\,\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k} - \frac{\mathcal{B}_k\,\mathbf{s}_k\,\mathbf{s}_k^T\mathcal{B}_k}{\mathbf{s}_k^T\mathcal{B}_k\,\mathbf{s}_k}.$$

This is a rank-two, symmetric secant update. In addition, $\mathcal{B}_{k+1}$ is positive definite if $\mathbf{y}_k^T\mathbf{s}_k > 0$ and $\mathcal{B}_k$ is positive definite. (The former condition is the only one to be concerned with, as $\nabla^2 \mathcal{E}_0$ is usually a positive multiple of the identity. The update is normally skipped whenever $\mathbf{y}_k^T\mathbf{s}_k \leq 0$.) This is a desirable feature of a Hessian update since the Hessian matrix is positive definite at a minimum. Thus, we seek positive definite Hessian update methods for minimizations of a quadratic model. One difficulty with the BFGS update is that, within the linesearch framework (to be discussed below), it only converges to the true Hessian if accurate linesearches are used [47]. Owing to inaccuracies in the potential energies and forces, this is often not the case in geometry optimization problems.

A competitive update which converges to the true Hessian on a quadratic surface without an exact linesearch is the symmetric rank-one (SR1) update by Murtagh and Sargent [134]. This update is given by

$$\mathcal{B}_{k+1} = \mathcal{B}_k + \frac{(\mathbf{y}_k - \mathcal{B}_k\mathbf{s}_k)(\mathbf{y}_k - \mathcal{B}_k\mathbf{s}_k)^T}{(\mathbf{y}_k - \mathcal{B}_k\mathbf{s}_k)^T\mathbf{s}_k}.$$

Unfortunately, this is not a positive-definite update, and sometimes the denominator becomes vanishingly small. However, the diagonal of the approximate Hessian may be perturbed to make the approximate Hessian positive definite at a given step.

A third update which has been useful for geometry optimization problems is to take a specific convex combination of the SR1 update shown above and the BFGS update. Specifically, this update is given by

$$\mathcal{B}_{k+1} = \mathcal{B}_k + \phi\, \frac{(\mathbf{y}_k - \mathcal{B}_k\, \mathbf{s}_k)(\mathbf{y}_k - \mathcal{B}_k\, \mathbf{s}_k)^T}{(\mathbf{y}_k - \mathcal{B}_k\, \mathbf{s}_k)^T \mathbf{s}_k} + (1 - \phi)\left(-\frac{\mathcal{B}_k\, \mathbf{s}_k\, \mathbf{s}_k^T \mathcal{B}_k}{\mathbf{s}_k^T \mathcal{B}_k\, \mathbf{s}_k} + \frac{\mathbf{y}_k\, \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} \right),$$

where ϕ is given by $\sqrt{t_1/t_2 t_3}$, where $t_1 = ((\mathbf{y}_k - \mathcal{B}_k\, \mathbf{s}_k)^T \mathbf{s}_k)^2$, $t_2 = (\mathbf{y}_k - \mathcal{B}_k\, \mathbf{s}_k)^T (\mathbf{y}_k - \mathcal{B}_k\, \mathbf{s}_k)$, and $t_3 = \mathbf{s}_k^T \mathbf{s}_k$. This update is due to Farkas and Schlegel [50] and is based on an idea of Bofill's [18] for locating transition-state structures. This update strives to take advantage of the positive definiteness of the BFGS update and the greater accuracy of the SR1 update. Other Hessian updates are also possible; see [47] for a description of several others.

The second way in which quasi-Newton methods differ is in their techniques for controlling the Newton step. Linesearch methods [139, 47, 93, 68, 57] attempt to choose a step length, λ_k, such that the quasi-Newton step given by $\mathbf{r}_{k+1} = \mathbf{r}_k + \lambda_k \mathbf{d}_k$ satisfies sufficient decrease and curvature conditions. One of the most successful linesearch codes is the limited memory BFGS code, L-BFGS, which was implemented by Liu and Nocedal [117, 138]. It is intended for large-scale optimization problems. In the L-BFGS code, quasi-Newton update vectors, rather than the full Hessian matrix, are stored. When the available storage has been depleted, the oldest correction vector is removed to make room for a new one. The step length in this code is determined by the sophisticated Moré–Thuente linesearch [131].

In contrast with linesearch methods, trust-region methods choose the direction and step length by minimizing the quadratic model subject to an elliptical constraint. The constrained minimization problem they solve is given by $\min m_k(\mathbf{r}_k + \mathbf{d}_k) = \mathcal{E}(\mathbf{r}_k) + \nabla\mathcal{E}(\mathbf{r}_k)^T \mathbf{d}_k + \frac{1}{2}\, \mathbf{d}_k^T\, \mathcal{B}_k\, \mathbf{d}_k$ subject to $\|\mathbf{d}_k\|_2 \le \delta_k$, where δ_k is the trust-region radius. The resulting step is of the form $\mathbf{d}_k = -(\mathcal{B}_k + \lambda_k \mathcal{I})^{-1} \nabla\mathcal{E}(\mathbf{r}_k)$. The trust-region radius is adjusted based on how well the quadratic model approximates the function. See [139, 47, 93, 68, 57, 39] for more details on trust-region methods.

The rational function optimization (RFO) method [187, 5] is related to trust-region methods in that it seeks to compute a step in a direction that will improve the convergence of the method. In this method, the quadratic model found in Newton's method is replaced with a rational function approximation. In particular,

$$\Delta\mathcal{E} = \mathcal{E}(\mathbf{r}_k + \mathbf{d}_k) - \mathcal{E}(\mathbf{r}_k) \approx \frac{\frac{1}{2}\begin{pmatrix} 1 & \mathbf{d}_k^T \end{pmatrix} \begin{pmatrix} 0 & \mathbf{g}_k^T \\ \mathbf{g}_k & B_k \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{d}_k \end{pmatrix}}{\begin{pmatrix} 1 & \mathbf{d}_k^T \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathcal{S}_k \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{d}_k \end{pmatrix}},$$

where $\mathcal{S}_k$ is a symmetric matrix that is normally taken to be the identity. Observe that the numerator in the above formula is the quadratic model used in the quasi-Newton method. The displacement vector, $\mathbf{d}_k$, is computed so as to minimize $\Delta\mathcal{E}$. For further details on solving this optimization problem, see [155].

Recent years have seen the development of hybrid methods for optimization based upon quasi-Newton methods. One such example is the hybrid method by Morales and Nocedal [129] that interlaces iterations of L-BFGS with a Hessian-free Newton method. The performance of this method is compared with the L-BFGS method and a truncated Newton method at the end of the next section.

7.2.4. Truncated Newton Methods. If the exact Hessian is available, then it is possible to use a truncated Newton method. Truncated Newton methods are a

subclass of Newton methods which are used in the context of large-scale optimization. Here an iterative method is used to compute the search direction, $\mathbf{d}_k$, using a linesearch or trust-region method. They are based on the idea that when far away from the solution, it does not make sense to compute an exact solution to the Newton equation, as this may be very computationally intensive and a descent direction may suffice. However, near a minimum, a more exact solution is desired. At each outer iteration, it is required that the residual, $\mathbf{r}_k$, satisfy $\mathbf{r}_k = \|\nabla^2 \mathcal{E}_k\, \mathbf{d}_k + \mathbf{g}_k\| \leq \eta_k \|\mathbf{g}_k\|$, where η_k is the forcing sequence. The methods are called truncated Newton methods, as they are stopped (or truncated) when the above convergence criterion is met. For appropriately chosen η_k, asymptotic quadratic convergence of the method is achieved as $\|\nabla \mathcal{E}_k\| \to 0$ [46]. One appropriate choice of η_k given in [176] is

$$\eta_k = \min\left\{c_r/k, \|\nabla \mathcal{E}_k\|\right\}, \quad 0 < c_r \leq 1.$$

See [136] for an excellent survey of truncated Newton methods.

Schlick and Overton developed the idea for a truncated Newton method which was used for potential energy minimization in [179]. The resulting Fortran package, TNPACK [177, 178], written by Schlick and Fogelson, was later incorporated into the CHARMM [22] molecular mechanics package [48]. The user of TNPACK is required to implement a sparse preconditioner which alters the clustering of the eigenvalues and enhances convergence. Automatic preconditioning is included in an implementation by Nash [135] that makes it readily portable.

Das, Meirovitch, and Navon [44] tested the performance of the Morales-Nocedal hybrid method (which was discussed in the quasi-Newton section), the Liu–Nocedal L-BFGS method, and the truncated Newton method with the preconditioner of Nash on the protein bovine pancreatic trypsin inhibitor (BPTT) and a loop of protein ribonuclease A. Their results showed that the hybrid approach is usually two times more efficient in terms of CPU time and function/gradient evaluations than the other two methods [44].

7.2.5. Conjugate Gradient Methods. Nonlinear conjugate gradient (NLCG) algorithms [77, 184] form another important class of methods used in electronic structure packages (e.g., Socorro, VASP, SIESTA, AIMPRO, and CRYSTAL) for solving the geometry optimization problem. For an excellent survey paper on NLCG methods, see [77].

In the 1960s, Fletcher and Reeves generalized the CG algorithm to nonlinear problems [59] by building upon earlier work by Davidon [45] and Fletcher and Powell [58]. The NLCG algorithms were developed by combining the linear CG algorithm with a linesearch. The NLCG and linear CG algorithms are related in the following way: if the objective function is convex and quadratic and an exact linesearch is used, then the nonlinear algorithm reduces to the linear one. This reduction is important since the linear CG algorithm requires at most $3K$ steps in exact arithmetic. This is because the search vectors span the entire $3K$-dimensional space after $3K$ steps.

NLCG algorithms are of the form

$$\mathbf{r}_{k+1} = \mathbf{r}_k + \alpha_k \mathbf{d}_k,$$

$$\mathbf{d}_k = -\mathbf{g}_k + \beta_{k-1}\mathbf{d}_{k-1}, \ \mathbf{d}_0 = -\mathbf{g}_0,$$

where α_k is obtained via a linesearch and β_{k-1} is a scalar that determines how much the previous direction is used in the computation of the current one. NLCG methods

differ in their choice of β_k; many different formulas have been proposed. Seven of these methods are the Hestenes and Stiefel method (HS) [82], the Fletcher–Reeves method (FR) [59], Daniel's method (D) [43], the method by Polak, Ribière, and Polyak (PRP) [152, 153], the conjugate descent method by Fletcher (CD) [57], the Liu and Storey method (LS) [118], and the method by Dai and Yuan (DY) [42]; their formulas for computing β_k are as follows:

$$\beta_{k-1}^{HS} = \frac{\mathbf{g}_k^T \mathbf{y}_{k-1}}{\mathbf{d}_{k-1}^T \mathbf{y}_{k-1}},$$

$$\beta_{k-1}^{FR} = \frac{\|\mathbf{g}_k\|^2}{\|\mathbf{g}_{k-1}\|^2},$$

$$\beta_{k-1}^{D} = \frac{\mathbf{g}_k^T \, \nabla^2 \mathcal{E}(\mathbf{r}_{k-1}) \, \mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^T \, \nabla^2 \mathcal{E}(\mathbf{r}_{k-1}) \, \mathbf{d}_{k-1}},$$

$$\beta_{k-1}^{PRP} = \frac{\mathbf{g}_k^T \mathbf{y}_{k-1}}{\|\mathbf{g}_{k-1}\|^2},$$

$$\beta_{k-1}^{CD} = \frac{\|\mathbf{g}_k\|^2}{-\mathbf{g}_{k-1}^T \mathbf{d}_{k-1}},$$

$$\beta_{k-1}^{LS} = \frac{\mathbf{g}_k^T \mathbf{y}_{k-1}}{-\mathbf{g}_{k-1}^T \mathbf{d}_{k-1}},$$

and

$$\beta_{k-1}^{DY} = \frac{\|\mathbf{g}_k\|^2}{\mathbf{d}_{k-1}^T \mathbf{y}_{k-1}},$$

where $\mathbf{y}_{k-1} = \mathbf{g}_k - \mathbf{g}_{k-1}$ and $\| \cdot \|$ is the l_2-norm.

The most popular formulas from the above list are FR, PRP, and HS. The FR method converges if the starting point is sufficiently near the desired minimum. On the other hand, PRP cycles infinitely in rare cases; this undesirable behavior can be remedied by restarting the method whenever $\beta_k^{PRP} < 0$. It is often the case that PRP converges more quickly than the FR method and is the method most often implemented in codes.

Recently Hager and Zhang [75, 76] developed a new NLCG method, CG_DESCENT, with guaranteed descent and an inexact linesearch. Their formula for computing β_k is given by

$$\beta_{k-1}^{HZ} = \left(\mathbf{y}_{k-1} - 2\mathbf{d}_{k-1} \frac{\|\mathbf{y}_{k-1}\|^2}{\mathbf{d}_{k-1}^T \mathbf{y}_{k-1}} \right) \frac{\mathbf{g}_k}{\mathbf{d}_{k-1}^T \mathbf{y}_{k-1}}.$$

Numerical comparisons in [75] showed that CG_DESCENT outperformed L-BFGS and several other NLCG methods on a set of 113 problems from the CUTEr (Constrained and Unconstrained Testing Environment, revisited) test set [193] with dimensions ranging from 50 to 10,000. Thus, CG_DESCENT should be seriously considered for the geometry optimization problem.

Some of the best-performing NLCG methods today are hybrid methods [77]. These methods dynamically adjust the formula for β_k as the iterations evolve. In [76],

several numerical experiments were performed which compared hybrid NLCG methods with CG_DESCENT and L-BFGS. The top performers relative to CPU time were CG_DESCENT, a code based upon a hybrid DY/HS scheme, and Liu and Nocedal's L-BFGS code. Thus, the hybrid DY/HS scheme also has excellent potential for use on the geometry optimization problem.

Baysal, Meirovitch, and Navon [10] studied the performance of several particular minimization algorithms as applied to models of peptides and proteins. In particular, they compared the performance of Liu and Nocedal's L-BFGS code with the performances of the truncated Newton method with the automatic preconditioner of Nash and the NLCG algorithm (SP) of Shanno and Phua. Their results [10] show that for one potential energy formulation, the truncated Newton method outperformed L-BFGS and SP by a factor of 1.2 to 2. With another potential energy formulation, L-BFGS outperformed TN by a factor of 1.5 to 2.5 and SP by a larger factor.

7.2.6. Iterative Subspace Methods. The final class of optimization methods we review are those that employ iterative subspace optimization. Electronic structure packages which employ iterative subspace methods include VASP and CPMD. One algorithm in this class is the DIIS method [157, 158], which is also referred to as residual minimization method-DIIS (RMM-DIIS). DIIS is the same as a Krylov subspace method in the case of solving a linear system without preconditioning. The relationship between the methods in the nonlinear case is more complicated and is described in [78].

DIIS was first used to accelerate self-consistent field calculations before it was extended to the geometry optimization problem and to charge mixing. The name of the method that has been specifically tailored for the geometry optimization problem is geometry optimization by direction inversion in the iterative subspace (GDIIS) [40].

GDIIS is different from quasi-Newton methods in that it assumes a linear connection between the coordinate and gradient changes; this is similar to using a quadratic approximation to the potential energy surface. However, in the quasi-Newton case, the linear connection is between the Hessian matrix and the gradient.

We now give the derivation for the GDIIS method from [51]. The development of the GDIIS method is based on a linear interpolation (and extrapolation) of previous molecular geometries, r_i, that minimizes the length of an error vector. The formula for the interpolation/extrapolation is given by

$$\mathbf{r}^* = \sum c_i \ \mathbf{r}_i, \ \text{where} \ \sum c_i = 1.$$

An error vector, $\mathbf{e}_i$, is created for each molecular geometry using a quadratic model of the potential energy surface. First, a simple relaxation step, $\mathbf{r}_i^*$, is computed using a Newton step, i.e.,

$$\mathbf{r}_i^* = \mathbf{r}_i - \nabla^2 \mathcal{E}^{-1} \mathbf{g}_i.$$

Then, the corresponding error vector, $\mathbf{e}_i$, is taken to be the displacement from the atomic structure:

$$\mathbf{e}_i = \mathbf{r}_i^* - \mathbf{r}_i = -\nabla^2 \mathcal{E}^{-1} \mathbf{g}_i.$$

The error (or residual) vector for $\mathbf{r}^*$ is the linear combination of the individual error vectors and is given by

$$\mathbf{z} = \sum c_i \ \mathbf{e}_i = \mathbf{e}^*.$$

Next, the coefficients are obtaining by solving the least-squares problem which corresponds to minimizing $\|\mathbf{z}\|^2$ subject to the constraint that $\sum c_k = 1$. Finally, the next atomic geometry in the optimization procedure is given by

$$\mathbf{r}_{k+1} = \mathbf{r}^* + \mathbf{z} = \sum c_i \ \mathbf{r}_i^*.$$

According to [51], this version of the GDIIS method is quite efficient in the quadratic vicinity of a minimum. However, further away from the minimum, the method is not as reliable and can fail in three major ways: convergence to a nearby critical point of higher order, oscillation around an inflection point on the potential energy surface, and numerical instability problems in determining the GDIIS coefficients. In [51], Farkas and Schlegel give an improved GDIIS method which overcomes these issues and performs as well as a quasi-Newton RFO method on a test set of small molecules. On a system with a large number of atoms, their improved GDIIS algorithm outperformed the quasi-Newton RFO method.

7.2.7. Molecular Dynamics Algorithms. Molecular dynamics algorithms are also used for geometry optimization despite the fact that they rely on statistical approaches and dynamics to minimize the energy. In section 6.3, we described the Car–Parrinello molecular dynamics algorithm.

Here, we will describe a second order damped molecular dynamics algorithm presented by Tassone, Mauri, and Car in [190]. Their method was designed for energy minimization within the context of Car–Parrinello simulations. Thus, their approach is also based on a fictitious dynamics of the electronic degrees of freedom and the introduction of Lagrangian multipliers to ensure the orthonormalization of the electronic orbitals during dynamic evolution. The resulting equation for the second order dynamics is as follows:

$$(51) \qquad \mu \ddot{\psi}_i(\mathbf{r}_i, t) = -\frac{1}{2} \delta \mathcal{E} / \delta \psi_i(\mathbf{r}_i, t) + \sum_k \Lambda_{ik} \psi_k(\mathbf{r_i}, t),$$

where the wave functions ψ_i have been assumed to be real and the indices are over the occupied states only. The Lagrangian multipliers given in the symmetric matrix Λ_{ik} are used to enforce the orthonormality condition, i.e., $\psi_i^* \psi_k = \delta_{ik}$. In [190], the authors then introduced a damping term to reduce the number of iterations required by the second order molecular dynamics algorithm. Then (51) becomes

$$(52) \qquad \mu \ddot{\psi}_i(\mathbf{r}_i, t) = -\frac{1}{2} \delta \mathcal{E} / \delta \psi_i(\mathbf{r}_i, t) - 2\gamma \mu \dot{\psi}_i(\mathbf{r}_i, t) + \sum_k \Lambda_{ik} \psi_k(\mathbf{r_i}, t),$$

where γ is set to obtain the maximal rate of convergence.

It was shown in [190] that this second order damped dynamics method has better efficiency than that of the steepest descent method. In particular, it has efficiency comparable to CG methods when used for electronic minimization. The difficulty of this method lies in how to determine an appropriate damping rate and timestep. In [156], Probert presented a new damped dynamics approach which yields a much faster rate of convergence than other molecular dynamics algorithms. He also demonstrated how to use the algorithm to dynamically select the optimal timestep or to precondition the ionic masses for the modified dynamics which yields additional efficiency gains near convergence [156].

7.3. Practical Recommendations. We conclude this section on geometry optimization with some practical recommendations. First, it is important to *consider different starting points*. A given optimization algorithm might not be globally convergent, although Zoutendijk's theorem can be used to modify linesearch algorithms to make them globally convergent (if they are not already so). Another problem is that the geometry optimization might converge to another type of critical point such as a local maximum or a saddle point. The user can distinguish the type of critical point by calculating the eigenvalues at the solution. One example of a system in the literature where global convergence was not achieved with an optimization algorithm is the biphenyl molecule. When started from anything other than a flat geometry, the geometry optimization produced an acceptable result. However, when starting with the flat geometry, it produced a flat ring geometry which corresponds to a maximum [116].

Second, the user can *try different algorithmic parameters, approximate initial Hessian matrices* (in the case of quasi-Newton methods), and *termination criterion*, for example, as these can dramatically affect the algorithms' convergence, as well. It can also be helpful to *try using various optimization algorithms* on one's geometry optimization problem. Different optimization algorithms perform better on some problems and not as well on other problems, as was demonstrated in this paper. The algorithm that will produce the best results for a given problem depends on several characteristics of the problem such as deviation of the objective function from quadratic, condition number of the Hessian (or approximate Hessian) matrices, convexity, and eigenvalue structure. See [137] for a numerical study which compares the performances of the L-BFGS method, a truncated Newton method, and the Polak–Ribière CG method on a set of test problems and analyzes the results in terms of these quantities.

Finally, it may be worth considering using *a different coordinate system*. In [3], Baker studied the use of Cartesian and natural internal coordinates (based upon the bonds and angles in the material) for geometry optimization; he concluded that for good starting geometries and initial Hessian approximations, geometry optimizations performed in Cartesian coordinates are as efficient as those using natural internal coordinates. Thus, the standard Cartesian coordinates are recommended for finding local minima. However, for the case of no initial Hessian information, natural internal coordinates were more effective. Thus, natural internal coordinates are recommended for finding a global minimum. See [160, 4] for alternative coordinate systems.

8. Open Issues and Concluding Remarks. Though significant progress has been made in recent years in developing effective practical methods for studying electronic structure of materials, there are from an algorithmic point of view many challenges remaining. Developing effective techniques for solving large eigenvalue problems in the case of a very large number of eigenvectors still remains an important issue. Interestingly, the large and *dense* eigenvalue problem will gain importance as systems become larger. This is because most methods solve a dense eigenvalue problem which arises from projecting the Hamiltonian into some subspace. As the number of states increases, this dense problem can reach a size in the tens of thousands. Because of the cubic scaling of standard eigenvalue methods for dense matrices, these calculations may become a bottleneck.

In the same vein, as systems become larger, eigenfunction-free methods may start playing a major role. Although there has been much work done in this area, linear scaling methods in existence today have limited applicability and it becomes important to explore their generalizations.

There are also challenging issues related to the optimization of the geometry. The difficulty here is that existing optimization methods or software packages do not yet address the specificity of this problem. Though the problem has a moderate number of variables (coordinates of atoms), the corresponding function is extremely expensive to evaluate as it requires one converged SCF iteration, and it can also be inaccurate. A desirable goal is to reduce the number of iterations by taking advantage of information gathered during the iterative process, but the inaccuracy of the function evaluation can make this an impossible task.

Many questions also remain to be explored on a more theoretical front; see, e.g., [110] for an overview. Work needs to be done, for example, in gaining a better understanding of the relation between the choice of the exchange-correlation functional and the nature of the resulting nonlinear eigenvalue problem. Thus, the self-consistent iteration is slow to converge in many cases (e.g., metallic compounds). It is known that such problems are intrinsically harder due to several factors, including the small gap between the eigenvalues of the occupied states and the others. In situations like these, it is intuitive that the solution will be more sensitive to small changes in the equations than in other cases. In particular, the solutions may depend more critically on the functional used for the exchange-correlation energy.

REFERENCES

[1] W. ANDREONI AND A. CURIONI, *New advances in chemistry and material science with CPMD and parallel computing*, Parallel Comput., 26 (2000), pp. 819–842.

[2] T. A. ARIAS, M. C. PAYNE, AND J. D. JOANNOPOULOS, *Ab initio molecular dynamics: Analytically continued energy functionals and insights into iterative solutions*, Phys. Rev. Lett., 69 (1992), pp. 1077–1080.

[3] J. BAKER, *Techniques for geometry optimization: A comparison of Cartesian and natural internal coordinates*, J. Comput. Chem., 14 (1993), pp. 1085–1100.

[4] J. BAKER AND P. PULAY, *Geometry optimization of atomic microclusters using inverse-power distance coordinates*, J. Chem. Phys., 105 (1996), pp. 11100–11107.

[5] A. BANERJEE, N. ADAMS, J. SIMONS, AND R. SHEPARD, *Search for stationary points on surfaces*, J. Phys. Chem., 89 (1985), pp. 52–57.

[6] S. BARONI, S. DAL CORSO, S. DE GIRONCOLI, AND P. GIANNOZZI, *PWSCF and PHONON: Plane-wave pseudo-potential codes*. Available online from http://www.pwscf.org.

[7] S. BARONI AND P. GIANNOZZI, *Towards very large scale electronic structures calculations*, Europhys. Lett., 17 (1992), pp. 547–552.

[8] C. BAYSAL AND H. MEIROVITCH, *Efficiency of the local torsional deformations method for identifying the stable structures of cyclic molecules*, J. Phys. Chem. A, 101 (1997), pp. 2185–2191.

[9] C. BAYSAL AND H. MEIROVITCH, *Determination of the stable microstates of a peptide from NOE distance constraints and optimization of atomic solvation parameters*, J. Amer. Chem. Soc., 120 (1998), pp. 800–812.

[10] C. BAYSAL, H. MEIROVITCH, AND I. M. NAVON, *Performance of efficient minimization algorithms as applied to models of peptides and proteins*, J. Comput. Chem., 20 (1999), pp. 354–364.

[11] T. L. BECK, *Real-space mesh techniques in density functional theory*, Rev. Modern Phys., 74 (2000), pp. 1041–1080.

[12] C. BEKAS, Y. SAAD, M. L. TIAGO, AND J. R. CHELIKOWSKY, *Computing charge densities with partially reorthogonalized Lanczos*, Comput. Phys. Comm., 171 (2005), pp. 175–186.

[13] M. BENZI, C. D. MEYER, AND M. TŮMA, *A sparse approximate inverse preconditioner for the conjugate gradient method*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.

[14] M. BENZI AND M. TŮMA, *A robust incomplete factorization preconditioner for positive definite matrices*, Numer. Linear Algebra Appl., 10 (2003), pp. 385–400.

[15] *BigDFT website*, http://www-drfmc.cea.fr/sp2m/L_Sim/BigDFT/, 2008.

[16] E. R. BITTNER AND D. S. KOSOV, *Car-Parrinello molecular dynamics on excited state surfaces*, J. Chem. Phys., 110 (1999), pp. 6645–6656.

[17] P. E. BLÖCHL, *Projector augmented-wave method*, Phys. Rev. B, 50 (1994), pp. 17953–17979.

[18] J. M. BOFILL, *Updated Hessian matrix and the restricted step method for locating transition structures*, J. Comput. Chem., 15 (1994), pp. 1–11.

[19] B. BRENA, D. NORDLUND, M. ODEIUS, H. OGASAWARA, AND L. G. M. PETTERSSON, *Ultrafast molecular dissociation of water in ice*, Phys. Rev. Lett., 93 (2004), article 148302.

[20] E. L. BRIGGS, D. J. SULLIVAN, AND J. BERNHOLC, *Large-scale electronic-structure calculations with multigrid acceleration*, Phys. Rev. B, 52 (1995), pp. R5471–R5474.

[21] A. BRILEY, M. R. PEDERSON, K. A. JACKSON, D. C. PATTON, AND D. V. POREZAG, *Vibrational frequencies and intensities of small molecules: All-electron, pseudopotential, and mixed-potential methodologies*, Phys. Rev. B, 58 (1997), pp. 1786–1793.

[22] B. R. BROOKS, R. E. BRUCCOLERI, B. D. OLAFSON, D. J. STATES, S. SWAMINATHAN, AND M. KARPLUS, *A program for macromolecular energy, minimization, and dynamics calculations*, J. Comput. Chem., 4 (1983), pp. 187–217.

[23] C. G. BROYDEN, *The convergence of a class of double-rank minimization algorithms. 2. The new algorithm.*, J. Inst. Math. Appl., 6 (1970), pp. 222–231.

[24] *CAMD Open Software Project (CAMPOS)*, http://www.camp.dtu.dk/English/Software.aspx, 2008.

[25] R. CAR AND M. PARRINELLO, *Unified approach for molecular dynamics and density functional theory*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.

[26] M. CAVALLERI, M. ODELIUS, A. NILSSON, AND L. G. M. PETTERSSON, *X-ray absorption spectra of water within a plane-wave Car-Parrinello molecular dynamics framework*, J. Chem. Phys., 12 (2004), pp. 10065–10075.

[27] D. M. CEPERLEY AND B. J. ALDER, *Ground state of the electron gas by a stochastic method*, Phys. Rev. Lett., 45 (1980), pp. 566–569.

[28] M. CHALLACOMBE, *A simplified density matrix minimization for linear scaling self-consistent field theory*, J. Chem. Phys., 110 (1999), pp. 2332–2342.

[29] M. CHALLACOMBE, *A general parallel sparse-blocked matrix multiply for linear scaling SCF theory*, Comput. Phys. Comm., 128 (2000), pp. 93–107.

[30] G. CHANG, W. C. GUIDA, AND W. C. STILL, *An internal-coordinate Monte Carlo method for searching conformational space*, J. Amer. Chem. Soc., 111 (1989), pp. 4379–4386.

[31] T. CHARPENTIER, S. ISPAS, M. PROFETA, F. MAURI, AND C. J. PICKARD, *First-principles calculation of ^{17}O, Si-29, and Na-23 NMR spectra of sodium silicate crystals and glasses*, J. Phys. Chem. B, 108 (2004), pp. 4147–4161.

[32] J. R. CHELIKOWSKY AND M. L. COHEN, *Pseudopotentials for semiconductors*, in Handbook of Semiconductors, 2nd ed., T. S. Moss and P. T. Landsberg, eds., Elsevier, Amsterdam, 1992, pp. 219–255.

[33] J. R. CHELIKOWSKY AND S. G. LOUIE, *First-principles linear combination of atomic orbitals method for the cohesive and structural properties of solids: Application to diamond*, Phys. Rev. B, 29 (1984), pp. 3470–3481.

[34] J. R. CHELIKOWSKY AND S. G. LOUIE, EDS., *Quantum Theory of Materials*, Kluwer, Dordrecht, The Netherlands, 1996.

[35] J. R. CHELIKOWSKY, N. TROULLIER, X. JING, D. DEAN, N. BIGGELI, K. WU, AND Y. SAAD, *Algorithms for the structural properties of clusters*, Comput. Phys. Comm., 85 (1995), pp. 325–335.

[36] J. R. CHELIKOWSKY, N. TROULLIER, AND Y. SAAD, *The finite-difference-pseudopotential method: Electronic structure calculations without a basis*, Phys. Rev. Lett., 72 (1994), pp. 1240–1243.

[37] J. R. CHELIKOWSKY, N. TROULLIER, K. WU, AND Y. SAAD, *Higher order finite difference pseudopotential method: An application to diatomic molecules*, Phys. Rev. B, 50 (1994), pp. 11355–11364.

[38] G. N. CHUEV AND M. V. FEDOROV, *Density functional method based on wavelets for quantum classical systems*, Internat. J. Quantum Chem., 100 (2004), pp. 539–547.

[39] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[40] P. CSÁSZÁR AND P. PULAY, *Geometry optimization by direct inversion in the iterative subspace*, J. Mol. Struct., 114 (1984), pp. 31–34.

[41] A. CURIONI AND W. ANDREONI, *Computer simulations for organic light-emitting diodes*, IBM J. Res. Dev., 45 (2001), pp. 101–113.

[42] Y. H. DAI AND Y. YUAN, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim., 10 (1999), pp. 177–182.

[43] J. W. DANIEL, *The conjugate gradient method for linear and nonlinear operator equations*, SIAM J. Numer. Anal., 4 (1967), pp. 10–26.

[44] B. Das, H. Meirovitch, and I. M. Navon, *Performance of hybrid methods for large-scale unconstrained optimization as applied to models of proteins*, J. Comput. Chem., 24 (2003), pp. 1222–1231.

[45] W. C. Davidon, *Variable Metric Method for Minimization*, Technical Report ANL-5990, Argonne National Laboratory, Argonne, IL, 1959.

[46] R. S. Dembo and T. Steihaug, *Truncated Newton algorithms for large-scale unconstrained optimization*, Math. Programming, 26 (1983), pp. 190–212.

[47] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[48] P. Derreumaux, G. Zhang, B. Brooks, and T. Schlick, *A truncated-Newton minimizer adapted for CHARMM and biomolecular applications*, J. Comput. Chem., 15 (1994), pp. 532–552.

[49] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[50] O. Farkas and H. B. Schlegel, *Methods for optimizing large molecules.* II. *Quadratic search*, J. Chem. Phys., 111 (1999), pp. 10806–10814.

[51] O. Farkas and H. B. Schlegel, *Methods for optimizing large molecules.* III. *An improved algorithm for geometry optimization using direct inversion in the iterative subspace (GDIIS)*, Phys. Chem. Chem. Phys., 4 (2002), pp. 11–15.

[52] J.-L. Fattebert and J. Bernholc, *Towards grid-based O(N) density-functional theory methods: Optimized nonorthogonal orbitals and multigrid acceleration*, Phys. Rev. B, 62 (2000), pp. 1713–1722.

[53] E. Fermi, *Application of statistical gas methods to electronic systems*, Atti Accad. Naz. Lincei, 6 (1927), pp. 602–607.

[54] E. Fermi, *Tentativo di una teoria dei raggi beta*, Nuovo Cimento, 2 (1934), pp. 1–19.

[55] D. Fischer, A. Curioni, S. R. Billeter, and W. Andreoni, *Effects of nitridation on the characteristics of silicon dioxide: Dielectric and structural properties from ab initio calculations*, Phys. Rev. Lett., 92 (2004), article 236405.

[56] R. Fletcher, *A new approach to variable metric algorithms*, Comput. J., 13 (1970), pp. 317–322.

[57] R. Fletcher, *Practical Methods of Optimization, Vol.* 1: *Unconstrained Optimization*, John Wiley & Sons, Chichester, UK, 1987.

[58] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.

[59] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.

[60] C. Y. Fong, *Topics in Computational Materials Science*, World Scientific, Singapore, 1998.

[61] B. Fornberg and D. M. Sloan, *A review of pseudospectral methods for solving partial differential equations*, Acta Numer., 94 (1994), pp. 203–268.

[62] G. Galli and M. Parrinello, *Large scale electronic structure calculation*, Phys. Rev. Lett., 69 (1992), pp. 3547–3550.

[63] G. Galli and A. Pasquarello, *First-principles molecular dynamics*, in Computer Simulation in Chemical Physics, M. P. Allen and D. J. Tildesley, eds., Kluwer, Dordrecht, The Netherlands, 1993, pp. 261–313.

[64] G. Gamov, *Thirty Years That Shook Physics: The Story of Quantum Theory*, Dover, New York, 1966.

[65] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, and R. Schneider, *Daubechies wavelets as a basis set for density functional pseudopotential calculations*, J. Chem. Phys., 129 (2008), article 014109.

[66] F. L. Gervasio, R. Chelli, P. Procacci, and V. Schettino, *The nature of intermolecular interactions between aromatic amino acid residues*, Proteins, 48 (2002), pp. 117–125.

[67] F. L. Gervasio, A. Laio, M. Iannuzzi, and M. Parrinello, *Influence of the DNA structure on the reactivity of the guanine radical cation*, Chem. European J., 10 (2004), pp. 4846–4852.

[68] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, New York, 1981.

[69] S. Goedecker and O. V. Ivanov, *Frequency localization properties of the density matrix and its resulting hypersparsity in a wavelet representation*, Phys. Rev. B, 59 (1999), pp. 7270–7273.

[70] D. Goldfarb, *A family of variable metric methods derived by variational means*, Math. Comp., 24 (1970), pp. 23–26.

[71] A. A. Goldstein, *Constructive Real Analysis*, Harper and Row, New York, 1967.

[72] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[73] X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G.-M. Rignanese, L. Sindic, M. Verstraete, G. Zerah, F. Jollet, M. Torrent, A. Roy, M. Mikami, P. Ghosez, J.-Y. Raty, and D. Allan, *First-principles computation of material properties: The ABINIT software project*, Comput. Materials Sci., 25 (2002), pp. 478–492.

[74] F. Gygi and G. Galli, *Real-space adaptive-coordinate electronic-structure calculations*, Phys. Rev. B, 52 (1995), pp. R2229–R2232.

[75] W. W. Hager and H. Zhang, *A new conjugate gradient method with guaranteed descent and an efficient linesearch*, SIAM J. Optim., 16 (2005), pp. 170–192.

[76] W. W. Hager and H. Zhang, *Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent*, ACM Trans. Math. Software, 32 (2006), pp. 113–137.

[77] W. W. Hager and H. Zhang, *A survey of nonlinear conjugate gradient methods*, Pacific J. Optim., 2 (2006), pp. 35–58.

[78] R. J. Harrison, *Krylov subspace accelerated inexact Newton method for linear and nonlinear equations*, J. Comput. Chem., 25 (2004), pp. 328–334.

[79] A. Haug, *Theoretical Solid State Physics*, Pergamon Press, Oxford, 1972.

[80] M. Head-Gordon, *Quantum chemistry and molecular processes*, J. Phys. Chem., 100 (1996), article 13213.

[81] M. Heikanen, T. Torsti, M. J. Puska, and R. M. Nieminen, *Multigrid method for electronic structure calculations*, Phys. Rev. B, 63 (2001), article 245106.

[82] M. R. Hestenes and E. L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards Sect. 5, 49 (1952), pp. 409–436.

[83] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Phys. Rev., 136 (1964), pp. B864–B871.

[84] T. Höynälänmaa, T. T. Rantala, and K. Ruotsalainen, *Solution of atomic orbitals in an interpolating wavelet basis*, Phys. Rev. E, 70 (2004), article 066701.

[85] Y. Huang, D. J. Kouri, and D. K. Hoffman, *Direct approach to density functional theory: Iterative treatment using a polynomial representation of the Heaviside step function operator*, Chem. Phys. Lett., 243 (1995), pp. 367–377.

[86] S. Ismail-Beigi and T. Arias, *New algebraic formulation of density functional calculation*, Comput. Phys. Comm., 128 (2000), pp. 1–15.

[87] K. A. Jackson, M. R. Pederson, D. V. Porezag, Z. Hajnal, and T. Fraunheim, *Density-functional-based predictions of Raman and IR spectra for small Si clusters*, Phys. Rev. B, 55 (1997), pp. 2549–2555.

[88] R. W. Jansen and O. F. Sankey, *Ab initio linear combination of pseudo-atomic-orbital scheme for the electronic properties of semiconductors: Results for ten materials*, Phys. Rev. B, 36 (1987), pp. 6520–6531.

[89] L. O. Jay, H. Kim, Y. Saad, and J. R. Chelikowsky, *Electronic structure calculations using plane wave codes without diagonalization*, Comput. Phys. Comm., 118 (1999), pp. 21–30.

[90] R. O. Jones and P. Ballone, *Density functional calculations for polymers and clusters—progress and limitations*, Comput. Materials Sci., 22 (2001), pp. 1–6.

[91] S. Jun, *Meshfree implementation for the real-space electronic-structure calculation of crystalline solids*, Internat. J. Numer. Methods Engrg., 59 (2004), pp. 1909–1923.

[92] R. Kaschner and D. Hohl, *Density functional theory and biomolecules: A study of glycine, alanine, and their oligopeptides*, J. Phys. Chem. A, 102 (1998), pp. 5111–5116.

[93] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.

[94] H. Kim and J. Ihm, *Ab initio pseudopotential plane-wave calculations of the electronic structure of $YBa_2Cu_3O_7$*, Phys. Rev. B, 51 (1995), pp. 3886–3892.

[95] J. Kim, F. Mauri, and G. Galli, *Total-energy global optimizations using nonorthogonal localized orbitals*, Phys. Rev. B, 52 (1995), pp. 1640–1648.

[96] Y. H. Kim, I. H. Lee, and R. M. Martin, *Object-oriented construction of a multigrid electronic structure code with Fortran*, Comput. Phys. Comm., 131 (2000), pp. 10–25.

[97] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[98] C. Kittel, *Introduction to Solid State Physics*, 7th ed., Wiley, Hoboken, NJ, 1996.

[99] J. Kohanoff, *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods*, Cambridge University Press, Cambridge, UK, 2006.

[100] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965), pp. A1133–A1138.

[101] I. KOLOSSVÁRY AND G. M. KESERÜ, *Hessian-free low-mode conformational search for large-scale protein loop optimization: Application to c-jun N-terminal kinase JNK3*, J. Comput. Chem., 22 (2001), pp. 21–30.

[102] M. KONOPKA, R. ROUSSEAU, I. STICH, AND D. MARX, *Detaching thiolates from copper and gold clusters: Which bonds to break?*, J. Amer. Chem. Soc., 126 (2004), pp. 12103–12111.

[103] G. KRESSE AND J. J. FURTHMÜLLER, *Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set*, Phys. Rev. B, 54 (1996), pp. 11169–11186.

[104] G. KRESSE AND J. HAFNER, *Ab initio molecular dynamics for liquid metals*, Phys. Rev. B, 47 (1993), pp. 558–561.

[105] G. KRESSE AND D. JOUBERT, *From ultrasoft pseudopotentials to the projector augmented-wave method*, Phys. Rev. B, 59 (1999), pp. 1758–1775.

[106] L. KRONIK, A. MAKMAL, M. L. TIAGO, M. M. G. ALEMANY, M. JAIN, X. HUANG, Y. SAAD, AND J. R. CHELIKOWSKY, *PARSEC—the pseudopotential algorithm for real-space electronic structure calculations: Recent advances and novel applications to nano-structure*, Phys. Status Solidi (B), 243 (2006), pp. 1063–1079.

[107] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45 (1950), pp. 255–282.

[108] R. M. LARSEN, *Lanczos Bidiagonalization with Partial Reorthogonalization for Large Sparse SVD and Linear Least Squares Calculations*, Technical Report, Computer Science Department, Aarhus University, 1998; available online from http://soi.stanford.edu/~rmunk/PROPACK/.

[109] R. M. LARSEN, *Efficient Algorithms for Helioseismic Inversion*, Ph.D. thesis, Department of Computer Science, University of Aarhus, Aarhus, Denmark, 1998.

[110] C. LE BRIS AND P. L. LIONS, *From atoms to crystals: A mathematical journey*, Bull. Amer. Math. Soc. (N.S.), 42 (2005), pp. 291–363.

[111] I. H. LEE, Y. H. KIM, AND R. M. MARTIN, *One-way multigrid method in electronic-structure calculations*, Phys. Rev. B, 61 (2000), pp. 4397–4400.

[112] J. LEE, H. A. SCHERAGA, AND S. RACKOVSKY, *New optimization method for conformational energy calculations on polypeptides: Conformational space annealing*, J. Comput. Chem., 18 (1997), pp. 1222–1232.

[113] X.-P. LI, R. W. NUNES, AND D. VANDERBILT, *Density-matrix electronic-structure method with linear system-size scaling*, Phys. Rev. B, 47 (1993), pp. 10891–10894.

[114] Z. LI AND H. A. SCHERAGA, *Monte Carlo-minimization approach to the multiple-minima problem in protein folding*, Proc. Natl. Acad. Sci., 84 (1987), pp. 6611–6615.

[115] W. Z. LIANG, C. SARAVANAN, Y. SHAO, A. BELL, AND H. HEAD-GORDON, *Improved Fermi operator expansion methods for fast electronic structure calculations*, J. Chem. Phys, 119 (2003), pp. 4117–4125.

[116] K. B. LIPKOWITZ, *Abuses of molecular mechanics: Pitfalls to avoid*, J. Chem. Educ., 72 (1995), pp. 1070–1075.

[117] D. C. LIU AND J. NOCEDAL, *On the limited memory BFGS method for large-scale optimization*, Math. Programming, 45 (1989), pp. 503–528.

[118] Y. LIU AND C. STOREY, *Efficient generalized conjugate gradient algorithms, Part* I: *Theory*, J. Optim. Theory Appl., 69 (1991), pp. 129–137.

[119] S. LUNDQVIST AND N. H. MARCH, EDS., *Theory of the Inhomogeneous Electron Gas*, Plenum, New York, 1983.

[120] O. MADELUNG, *Introduction to Solid State Theory*, Springer-Verlag, Berlin, 1996.

[121] J. L. MARTINS AND M. COHEN, *Diagonalization of large matrices in pseudopotential band-structure calculations: Dual-space formalism*, Phys. Rev. B, 37 (1988), pp. 6134–6138.

[122] J. L. MARTINS, N. TROULLIER, AND S.-H. WEI, *Pseudopotential plane-wave calculations for ZnS*, Phys. Rev. B, 43 (1991), pp. 2213–2217.

[123] R. MARTONAK, C. MOLTENI, AND M. PARRINELLO, *A new constant-pressure ab initio/classical molecular dynamics method: Simulation of pressure-induced amorphization in a $Si_{35}H_{36}$ cluster*, Comput. Materials Sci., 20 (2001), pp. 293–299.

[124] F. MAURI AND G. GALLI, *Electronic-structure calculations and molecular-dynamics simulation with linear system size scaling*, Phys. Rev. B, 50 (1994), pp. 4316–4326.

[125] F. MAURI, G. GALLI, AND R. CAR, *Orbital formulation for electronic-structure calculations with linear system size scaling*, Phys. Rev. B, 47 (1993), pp. 9973–9976.

[126] D. G. MCCULLOCH, A. R. MERCHANT, N. A. MARKS, N. C. COOPER, P. FIRZHENRY, M. M. M. BILEK, AND D. R. MCKENZIE, *Wannier function analysis of tetrahedral amorphous networks*, Diamond and Related Materials, 12 (2003), pp. 2026–2031.

[127] R. MCWEENY, *Some recent advances in density matrix theory*, Rev. Modern Phys., 32 (1960), pp. 335–369.

[128] *Molecular geometry*, http://en.wikipedia.org/wiki/Molecular_form, 2008.

[129] J. L. MORALES AND J. NOCEDAL, *Enriched methods for large-scale unconstrained optimization*, Comput. Optim. Appl., 21 (2002), pp. 143–154.

[130] T. MORDASINI, A. CURIONI, R. BURSI, AND W. ANDREONI, *The binding mode of progesterone to its receptor deduced from molecular dynamics simulations*, ChemBioChem, 4 (2003), pp. 155–161.

[131] J. J. MORÉ AND D. J. THUENTE, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 268–307.

[132] R. B. MORGAN AND D. S. SCOTT, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Comput., 7 (1986), pp. 817–825.

[133] J. J. MORTENSEN, L. B. HANSEN, AND K. W. JACOBSEN, *Real-space grid implementation of the projector augmented wave method*, Phys. Rev. B, 71 (2005), article 035109.

[134] B. MURTAGH AND R. W. H. SARGENT, *Computational experience with quadratically convergent minimisation methods*, Comput. J., 13 (1970), pp. 185–194.

[135] S. G. NASH, *User's guide for TN/TNBC: FORTRAN Routines for Nonlinear Optimization*, Tech. Report 397, Mathematical Sciences Department, Johns Hopkins University, Baltimore, MD, 1984.

[136] S. G. NASH, *A survey of truncated-Newton methods*, J. Comput. Appl. Math., 124 (2000), pp. 45–59.

[137] S. G. NASH AND J. NOCEDAL, *A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization*, SIAM J. Optim., 1 (1991), pp. 358–372.

[138] J. NOCEDAL, *Updating quasi-Newton matrices with limited storage*, J. Math. Comput., 35 (1980), pp. 773–782.

[139] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, 2nd ed., Springer-Verlag, New York, 2006.

[140] T. ONO AND K. HIROSE, *Timesaving double-grid method for real-space electronic structure calculations*, Phys. Rev. Lett., 82 (1999), pp. 5016–5019.

[141] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[142] M. PARRINELLO, *From silicon to RNA: The coming of age of ab initio molecular dynamics*, Solid State Commun., 102 (1997), pp. 107–120.

[143] *PARSEC website*, http://www.ices.utexas.edu/parsec/index.html, 2005.

[144] J. E. PASK, B. M. KLEIN, C. Y. FONG, AND P. A. STERNE, *Real-space local polynomial basis for solid-state electronic-structure calculations: A finite-element approach*, Phys. Rev. B, 59 (1999), pp. 12352–12358.

[145] G. PASTORE, E. SMARGIASSI, AND F. BUDA, *Theory of ab initio molecular-dynamics calculations*, Phys. Rev. A, 44 (1991), pp. 6334–6347.

[146] M. C. PAYNE, M. P. TETER, D. C. ALLAN, T. A. ARIAS, AND J. D. JOANNOPOULOS, *Iterative minimization techniques for ab-initio total energy calculations: Molecular dynamics and conjugate gradients*, Rev. Modern Phys., 64 (1992), pp. 1045–1097.

[147] J. P. PERDEW, K. BURKE, AND Y. WANG, *Generalized gradient approximation for the exchange-correlation hole of a many-electron system*, Phys. Rev. B, 54 (1996), pp. 16533–16539.

[148] B. G. PFROMMER, J. DEMMEL, AND H. SIMON, *Unconstrained energy functionals for electronic structure calculations*, J. Comput. Phys., 150 (1999), pp. 287–298.

[149] S. PIANA, P. CARLONI, AND M. PARRINELLO, *Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease*, J. Mol. Biol., 319 (2002), pp. 567–583.

[150] J. PILLARDY, Y. A. ARNAUTOVA, C. CZAPLEWSKI, K. D. GIBSON, AND H. A. SCHERAGA, *Conformation-family Monte Carlo: A new method for crystal structure prediction*, Proc. Natl. Acad. Sci., 98 (2001), pp. 12351–12356.

[151] J. PILLARDY, C. CZAPLEWSKI, A. LIWO, J. LEE, D. R. RIPOLL, R. KAZMIERKIEWICZ, S. OLDZIEJ, W. J. WEDEMEYER, K. D. GIBSON, Y. A. ARNAUTOVA, J. SAUNDERS, Y. J. YE, AND H. A. SCHERAGA, *Recent improvements in prediction of protein structure by global optimization of a potential energy function*, Proc. Natl. Acad. Sci., 98 (2001), pp. 2329–2333.

[152] E. POLAK AND G. RIBIÈRE, *Note sur la convergence de méthodes de directions conjugées*, Rev. Française Informat. Recherche Opérationnelle, 16 (1969), pp. 35–43.

[153] B. T. POLYAK, *The conjugate gradient method in extreme problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 94–112.

[154] V. POPHRISTIC, M. L. KLEIN, AND M. N. HOLERCA, *Modeling small aluminum chlorohydrate polymers*, J. Phys. Chem. A, 108 (2004), pp. 113–120.

[155] X. Prat-Resina, *Study of the Reaction Mechanism in Mandelate Racemase Enzyme: Reaction Path and Dynamical Sampling Approaches*, Ph.D. thesis, Universitat Autonóma de Barcelona, Spain, 2004.

[156] M. I. J. Probert, *Improved algorithm for geometry optimisation using damped molecular dynamics*, J. Comput. Phys., 191 (2003), pp. 130–146.

[157] P. Pulay, *Convergence acceleration of iterative sequences. The case of SCF iteration*, Chem. Phys. Lett., 73 (1980), pp. 393–398.

[158] P. Pulay, *Improved SCF convergence acceleration*, J. Comput. Chem., 3 (1982), pp. 556–560.

[159] P. Pulay, *Analytical derivative methods in quantum chemistry*, Adv. Chem. Phys., 69 (1987), pp. 241–286.

[160] P. Pulay and G. Fogarasi, *Geometry optimization in redundant internal coordinates*, J. Chem. Phys., 96 (1992), pp. 2856–2860.

[161] L. M. Ramaniah, M. Boero, and M. Laghate, *Tantalum-fullerene clusters: A first principles study of static properties and dynamical behaviour*, Phys. Rev. B, 70 (2004), pp. 35411–35424.

[162] H. Roder, R. Silver, D. Drabold, and J. Dong, *The kernel polynomial method for nonorthogonal electronic structure calculation of amorphous diamond*, Phys. Rev. B, 55 (1997), pp. 10582–10586.

[163] U. Rothlisberger, 15 *years of Car-Parrinello simulations in physics, chemistry and biology*, in Computational Chemistry: Reviews of Current Trends, Vol. 6, J. Leszczynski, ed., World Scientific, Singapore, 2001, pp. 33–68.

[164] C. Rovira, B. Schulze, M. Eichinger, J. D. Evanseck, and M. Parrinello, *Influence of the heme pocket conformation on the structure and vibrations of the Fe-CO bond in myoglobin: A QM/MM density functional study*, Biophys. J., 81 (2001), pp. 435–445.

[165] H. Rutishauser, *Simultaneous iteration for symmetric matrices*, in Handbook for Automatic Computations (Linear Algebra), J. Wilkinson and C. Reinsch, eds., 1971, Springer-Verlag, New York, pp. 202–211.

[166] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.

[167] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelpha, PA, 2003.

[168] Y. Saad, A. Stathopoulos, J. Chelikowsky, K. Wu, and S. Öğüt, *Solution of large eigenvalue problems in electronic structure calculations*, BIT, 36 (1996), pp. 563–578.

[169] A. H. Sameh and J. A. Wisniewski, *A trace minimization algorithm for the generalized eigenvalue problem*, SIAM J. Numer. Anal., 19 (1982), pp. 1243–1259.

[170] M. Saunders, *Searching for conformers of nine- to twelve-ring hydrocarbons on the MM2 and MM3 energy surfaces: Stochastic search for interconversion pathways*, J. Comput. Chem., 12 (1991), pp. 645–663.

[171] M. Saunders, K. N. Houk, Y.-D. Wu, W. C. Still, M. Lipton, G. Chang, and W. C. Guida, *Conformations of cycloheptadecane. A comparison of methods for conformational searching*, J. Amer. Chem. Soc., 112 (1990), pp. 1419–1427.

[172] F. Schautz, F. Buda, and C. Filippi, *Excitations in photoactive molecules from quantum Monte Carlo*, J. Chem. Phys., 121 (2004), pp. 5836–5844.

[173] H. B. Schlegel, *Geometry optimization on potential energy surfaces*, in Modern Electronic Structure Theory, D. R. Yarkony, ed., World Scientific, Singapore, 1995, pp. 459–500.

[174] H. B. Schlegel, *Geometry optimization* 1, in Encyclopedia of Computational Chemistry, Vol. 2, P. v. R. Schleyer, N. L. Allinger, P. A. Kollman, T. Clark, H. Schaefer, III, J. Gasteiger, and P. R. Schreiner, eds., Wiley, Chichester, UK, 1998, pp. 1136–1142.

[175] H. B. Schlegel, *Exploring potential energy surfaces for chemical reactions: An overview of practical methods*, J. Comput. Chem., 24 (2003), pp. 1514–1527.

[176] T. Schlick, *Geometry optimization* 2, in Encyclopedia of Computational Chemistry, Vol. 2, P. v. R. Schleyer, N. L. Allinger, P. A. Kollman, T. Clark, H. Schaefer, III, J. Gasteiger, and P. R. Schreiner, eds., Wiley, Chichester, UK, 1998, pp. 1142–1157.

[177] T. Schlick and A. Fogelson, *TNPACK—A truncated Newton minimization package for large scale problems:* I. *Algorithm and usage*, ACM Trans. Math. Software, 18 (1992), pp. 46–70.

[178] T. Schlick and A. Fogelson, *TNPACK—A truncated Newton minimization package for large scale problems:* II. *Implementation examples*, ACM Trans. Math. Software, 18 (1992), pp. 71–111.

[179] T. Schlick and M. Overton, *A powerful truncated Newton method for potential energy minimization*, J. Comput. Chem., 8 (1987), pp. 1025–1039.

[180] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, and M. C. Payne, *First-principles simulation: Ideas, illustrations and the CASTEP code*, J. Phys. Condens. Matter, 14 (2002), pp. 2717–2744.

[181] A. P. Seitsonen, M. J. Puska, and R. M. Nieminen, *Real-space electronic-structure calculations: Combination of the finite-difference and conjugate-gradient methods*, Phys. Rev. B, 51 (1995), pp. 14057–14061.

[182] H. Sekino, Y. Maeda, T. Yanai, and R. J. Harrison, *Basis set limit Hartree–Fock and density functional theory response property evaluation by multiresolution multiwavelet basis*, J. Chem. Phys., 129 (2008), article 034111.

[183] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647–656.

[184] J. R. Shewchuk, *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*, Report CMU-CS-94-125, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1994.

[185] H. D. Simon, *Analysis of the symmetric Lanczos algorithm with reorthogonalization methods*, Linear Algebra Appl., 61 (1984), pp. 101–132.

[186] H. D. Simon, *The Lanczos algorithm with partial reorthogonalization*, Math. Comp., 42 (1984), pp. 115–142.

[187] J. Simons, P. Jorgensen, H. Taylor, and J. Ozment, *Walking on potential energy surfaces*, J. Phys. Chem., 87 (1983), pp. 2745–2753.

[188] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *The SIESTA method for ab-initio order-N materials simulation*, J. Phys. Condens. Matter, 14 (2002), pp. 2745–2779.

[189] I. Stich, R. Car, M. Parrinello, and S. Baroni, *Conjugate gradient minimization of the energy functional: A new method for electronic structure calculation*, Phys. Rev. B, 39 (1989), pp. 4997–5004.

[190] F. Tassone, F. Mauri, and R. Car, *Acceleration schemes for ab-initio molecular dynamics and electronic structure calculations*, Phys. Rev. B, 50 (1994), pp. 10561–10573.

[191] M. P. Teter, M. C. Payne, and D. C. Allen, *Solution of Schrödinger's equation for large systems*, Phys. Rev. B, 40 (1989), pp. 12255–12263.

[192] *The CPMD Consortium publication website*, http://www.cpmd.org/cpmd_publications.html, 2005.

[193] *The CUTEr Test Problem Set*, http://cuter.rl.ac.uk/cuter-www/Problems/mastsif.html, 2005.

[194] L. H. Thomas, *The calculation of atomic fields*, Proc. Cambridge Philos. Soc., 23 (1927), pp. 542–548.

[195] C. H. Tong, T. F. Chan, and C. C. J. Kuo, *Multilevel filtering preconditioners: Extensions to more general elliptic problems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 227–242.

[196] T. Torsti, M. Heiskanen, M. Puska, and R. Nieminen, *MIKA: A multigrid-based program package for electronic structure calculations*, Internat. J. Quantum Chem., 91 (2003), pp. 171–176.

[197] E. Tsuchida and M. Tsukada, *Electronic-structure calculations based on the finite-element method*, Phys. Rev. B, 52 (1995), pp. 5573–5578.

[198] M. E. Tuckerman, P. J. Ungar, T. von Rosenvinge, and M. L. Klein, *Ab initio molecular dynamics simulations*, J. Phys. Chem., 100 (1996), pp. 12878–12887.

[199] D. Vanderbilt, *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism*, Phys. Rev. B, 41 (1990), pp. 7892–7895.

[200] K. Varga, Z. Zhang, and S. T. Pantelides, *Lagrange functions: A family of powerful basis sets for real-space order-n electronic structure calculations*, Phys. Rev. Lett., 93 (2004), article 176403.

[201] M. Vásquez and H. A. Scheraga, *Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone enkephalin*, Biopolymers, 24 (1985), pp. 1437–1447.

[202] D. J. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, UK, 2003.

[203] L. W. Wang, *Calculating the density of states and optical-absorption spectra of large quantum systems by the plane-wave moments method*, Phys. Rev. B, 49 (1994), pp. 10154–10158.

[204] V. Weber, J. VandeVondele, J. Hutter, and A. M. N. Niklasson, *Direct energy functional minimization under orthogonality constraints*, J. Chem. Phys., 128 (2008), article 084113.

[205] S. R. White, J. W. Wilkins, and M. P. Teter, *Finite-element method for electronic structure*, Phys. Rev. B, 39 (1989), pp. 5819–5833.

[206] K. Wu and H. Simon, *A Parallel Lanczos Method for Symmetric Generalized Eigenvalue Problems*, Tech. Report 41284, Lawrence Berkeley National Laboratory, 1997; available online at http://www.nersc.gov/research/SIMON/planso.html.

[207] S. Yoo, J. J. Zhao, J. L. Wang, and X. C. Zeng, *Endohedral silicon fullerenes Si-N (27 <= N <= 39)*, J. Amer. Chem. Soc., 126 (2004), pp. 13845–13849.

[208] X. Zhao, Y. Liu, S. Inoue, T. Suzuki, R. O. Jones, and Y. Ando, *Smallest carbon nanotube is 3 Åin diameter*, Phys. Rev. Lett., 92 (2004), article 125502.

[209] Z. Zhao, L. Wang, and J. Meza, *Motif based Hessian matrix for ab initio geometry optimization of nanostructures*, Phys. Rev. B, 73 (2006), article 193309.

[210] J. M. Ziman, *Electrons and Phonons*, Oxford University Press, Oxford, 1960.

[211] J. M. Ziman, *Principles of the Theory of Solids*, 2nd ed., Cambridge University Press, Cambridge, UK, 1986.

[212] G. Zumbach, N. A. Modine, and E. Kaxiras, *Adaptive coordinate, real-space electronic structure calculations on parallel computers*, Solid State Commun., 99 (1996), pp. 57–61.