Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients

M. C. Payne

Cavendish Laboratory, Madingley Road, Cambridge, CB3 0HE, United Kingdom

M. P. Teter and D. C. Allan

Applied Process Research, Corning Incorporated, Corning, New York 14831

T. A. Arias and J. D. Joannopoulos

Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

This article describes recent technical developments that have made the total-energy pseudopotential the most powerful *ab initio* quantum-mechanical modeling method presently available. In addition to presenting technical details of the pseudopotential method, the article aims to heighten awareness of the capabilities of the method in order to stimulate its application to as wide a range of problems in as many scientific disciplines as possible.

CONTENTS

Copyright ©1992 The American Physical Society

I. INTRODUCTION

There is little doubt that most of low-energy physics, chemistry, and biology can be explained by the quantum mechanics of electrons and ions. The limits of applicability and even the interpretation of the predictions of modern quantum theory are lively areas of debate amongst philosophers. Questions such as "How do we interpret the probabilistic nature of wave functions?," "What constitutes a measurement?," "How much can we ever know about the state of a system?," and "Can quantum mechanics describe consciousness?" are of fundamental importance. Despite the fact that these questions are still debated, it is clear that whether or not a more complete description of the world is possible, those things that modern quantum theory does predict are predicted with incredible accuracy. One outstanding example of this accuracy is the calculation of the gyromagnetic ratio of the electron, which agrees with the experimental result to the limit of the measurement, some 10 significant figures. Quantum theory has also proven correct and provided fundamental understanding for a wide variety of phenomena, including the energy levels of atoms, the covalent bond, and the distinction between metals and insulators. Further, in every instance of its application to date, the equations of quantum mechanics have yet to be shown to fail. There is, therefore, every reason to believe that an understanding of a great number of phenomena can be achieved by continuing to solve these equations.

As we shall soon see, the ability of quantum mechanics to predict the total energy of a system of electrons and nuclei, enables one to reap a tremendous benefit from a quantum-mechanical calculation. In fact this entire article is dedicated to just this one type of quantum-mechanical calculation, the foundation for which is quite strong. The quantum-mechanical rules, or Hamiltonians, for calculating the total energy of simple one-atom systems have provided some of the most precise tests of the theory, and the rules for calculating the energies of more complicated systems are simple, straightforward extensions of these atomic Hamiltonians. It is therefore eminently reasonable to expect quantum mechanics to predict accurately the total energies of aggregates of atoms as well. So far, this expectation has been confirmed time and time again by experiment.

A few moments' thought shows that nearly all physical properties are related to total energies or to differences between total energies. For instance, the equilibrium lattice constant of a crystal is the lattice constant that minimizes the total energy; surfaces and defects of solids adopt the structures that minimize their corresponding total energies. If total energies can be calculated, any physical property that can be related to a total energy or to a difference between total energies can be determined computationally. For example, to predict the equilibrium lattice constant of a crystal, a series of total-energy calculations are performed to determine the total energy as a function of the lattice constant. As shown in Fig. 1, the results are then plotted on a graph of energy versus lattice constant, and a smooth curve is constructed through the points. The theoretical value for the equilibrium lattice constant is the value of the lattice constant at the minimum of this curve. Total-energy techniques also have been successfully used to predict with accuracy equilibrium lattice constants, bulk moduli, phonons, piezoelectric constants, and phase-transition pressures and temperatures (for reviews see Cohen, 1984; Joannopoulos, 1985; Pickett, 1989).

Part of our aim in this article is to introduce the usefulness of quantum total-energy techniques to a broad

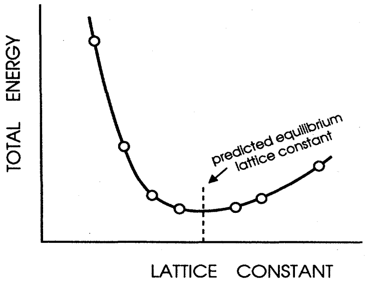

FIG. 1. Theoretical determination of an equilibrium lattice constant. Calculations (open circles) at various possible lattice constants are performed and a smooth function is fitted through the points. The predicted lattice constant is determined by the minimum in the curve.

range of scientists, including, for example, chemists, biologists, and geophysicists, who can at last benefit from these techniques. It is often suggested that quantum mechanics was primarily developed to describe events on the atomic scale, raising the question, "Of what use are quantum-mechanical calculations in a science not concerned directly with events on the atomic scale?" Since our world is composed of and defined by the interactions of atoms and molecules, a detailed and fundamental understanding of the world must ultimately rest on comprehending these interactions. The good news that this article brings is that the search for this kind of understanding is no longer a mere idle philosophical musing but rather a practical methodology.

There are many examples of connections between atomic and macroscopic levels being made every day. The "lock and key" mechanism in biological systems has led to the development of "designer drugs" whose shapes correspond to the "key" in the relevant biological reaction. In turn the shapes of the drugs are known and understood only because of the quantum-mechanical description of covalent bonds. Although no drug has yet been designed by determining the shape of the molecule by solving the Schrödinger equation, this particular application of quantum mechanics is not far away. There are also examples of the application of quantum mechanics beyond the atomic scale in materials science, for instance in the onset of failure in a material. The failure of a material starts on the atomic scale when one bond is stressed beyond its yield-stress and breaks. Thus it is obvious that where and when a material actually starts to fail is determined by quantum mechanics. Though there are many examples in materials science of significant progress having been made without any need for quantum-mechanical modeling, this progress is often limited as one pushes forward and encounters the atomic world. For instance, the understanding of the properties and behavior of dislocations comes from classical elasticity theory, but even in these cases very little is known about the core of a dislocation, precisely because this part of the dislocation requires detailed quantum-mechanical modeling. Moreover, even classical elasticity theory, which can only be applied on a macroscopic scale, is directly related to the atomic world through the elastic constants, the parameters of elasticity theory determined by the quantum-mechanical behavior of the material. Though materials constants such as the elastic constants may often be simply measured in the laboratory, the geophysicist modeling a continental drift does not have the luxury of performing experiments to bypass quantum-mechanical calculations. It is not possible, at present, to generate geophysical pressures in the laboratory. Therefore the relevant high-pressure parts of the phase diagrams of the materials that constitute the earth are unknown. Geophysicists will greatly benefit from quantum-mechanical modeling, which can provide them with the parameters needed to pursue their research.

When should a scientist consider quantum-mechanical modeling? The examples given above suggest that quantum-mechanical modeling be considered in situations where experiment is impossible. This principle is not limited, as in the geophysics example, to situations that are completely inaccessible to experiment, but also includes the performance of "computational experiments," which afford far greater "experimental" control than their physical counterparts. For instance, one can "reach into" a theoretical chemical calculation and, at will, bend bonds at experimentally unstable and inaccessible angles to gain insight into the processes controlling chemical reactions. Or, one can study the properties of isolated defects in a material in which segregation of impurities towards those defects tends to cloud the experimental results.

Another relevant consideration is what can be calculated quantum mechanically and at what cost. The boundary of feasible quantum-mechanical calculations has shifted significantly, to the extent that it may now be more cost effective to employ quantum-mechanical modeling even when experiments do offer an alternative. Moreover, many fields of science, not just physics and basic materials science, may benefit. It is true that the original modeling of the covalent bond was not quantitatively accurate, and it did not give the correct value for the energy associated with the bond. However, chemists solving the Schrödinger equation nowadays accurately calculate the energy of covalent bonds as well as their equilibrium lengths, force constants, and polarizabilities. Physicists have developed many methods that can be used to calculate a wide range of physical properties of materials, such as lattice constants and elastic constants. These methods, which require only a specification of the ions present (by their atomic number), are usually referred to as *ab initio* methods.

Many of the *ab initio* methods that physicists and chemists use have existed for more than a decade, and it is not the purpose of this article to describe or compare the wide range of methods that exist. All of the *ab initio* methods have been continuously refined over recent years and all have benefited from the availability of increasingly powerful computers. A decade ago most *ab initio* methods were capable only of modeling systems of a few atoms, and hence applicability to real-world systems at that point was extremely limited. Most methods can now model systems that contain some tens of atoms and are used to study a small but significant range of interesting problems. Of all the methods, one, the *total-energy pseudopotential method*, stands alone. A decade ago this method was also capable of modeling only few-atom systems. Now, however, this method can model thousand-atom systems, and it is already clear that this number will increase by at least a factor of 10 within the next five years. Pushing back the limits of quantum-mechanical modeling to this extent, the total-energy pseudopotential method opens up a wide range of interesting problems to quantum-mechanical calculation, and the future should bring the application of quantum mechanics to many new fields of science. The increase in the number of atoms that can be handled is directly due to an increase in computational efficiency of the *ab initio* pseudopotential method, which also means that there is an increasing class of problems for which it is more cost effective to use quantum-mechanical modeling than experiment to determine the physical parameters of systems. One purpose of this article is to heighten awareness amongst scientists in a range of *scientific* disciplines that the world is quantum mechanical and that there now exists an *ab initio* method that allows the quantum mechanics to be solved and incorporated into everyday science.

There is an economy of scale to *ab initio* total-energy calculations because so many physical properties are related to total energies. While just one piece of theoretical "apparatus" is needed to calculate *all* the physical properties that are related to total energies, completely different sets of experimental apparatus are required to measure *each class* of physical property of a material. This represents an enormous advantage of quantum-mechanical modeling over experimental measurements. Comparing the decreasing cost of computers with the cost of a large number of different pieces of experimental apparatus needed to carry out the same functions, one sees that the cost effectiveness of quantum-mechanical modeling methods over physical experimentation will continue to increase with time. This, then, is the time for researchers in a wide range of scientific disciplines to consider very seriously whether quantum-mechanical modeling may be applied in a cost-effective way to their own field of research, even if the field of research is far removed from what is usually assumed to be the quantum-mechanical world.

Total-energy pseudopotential calculations do require significant amounts of computer time, even for systems containing a few atoms in the unit cell, and the computational time required to perform the calculations always increases with the number of atoms in the unit cell. Thus, for large systems containing hundreds of atoms in the unit cell, it is essential to use the most efficient numerical algorithms. In the following pages we shall discuss in detail the latest methods for doing this. Among the methods we shall discuss, two have revolutionized the field of *ab initio* total-energy calculation, each increasing the number of atoms in a calculable system by more than an order of magnitude over previously existing techniques. As we shall discuss in detail, the molecular-dynamics method developed by Car and Parrinello (1985) transformed the way in which we viewed quantum-mechanical calculations and hence total-energy pseudopotential calculations; instead of finding a coupled self-consistent solution to a descretized partial differential equation through matrix techniques, Car and Parrinello minimized a single function through simulated annealing. The Car-Parrinello method can be used to perform calculations for systems containing on the order of one hundred atoms in the unit cell. However, severe difficulties are encountered in certain cases when one attempts to use this method to perform calculations on much larger systems. Recently, conjugate-gradients methods have been developed (Teter *et al.* 1989; Gillan, 1989) that overcome the difficulties encountered with the molecular-dynamics technique. Conjugate-gradients methods have again transformed total-energy pseudopotential calculations by replacing simulated annealing minimization with a direct, completely self-consistent second-order search for the minimum. Using these methods, one can perform calculations for systems containing many hundreds, and soon thousands, of atoms.

The molecular-dynamics and conjugate-gradients methods allow pseudopotential calculations to be performed for much larger systems than was possible using conventional matrix diagonalization methods. They also allow, for the first time, tractable *ab initio* quantum-mechanical simulations to be performed for systems at *nonzero temperatures*. While these capabilities offer the obvious advantage of permitting more complex systems to be studied, there is yet another benefit to be gained by using these new computational methods. By increasing the efficiency of the total-energy pseudopotential technique, they have greatly extended the range of application of this technique by allowing, for the first time, the inclusion of noble and transition-metal atoms and first-row elements such as oxygen in large pseudopotential calculations. Until recently it was widely believed that computations including such elements would be completely intractable with pseudopotentials in a plane-wave representation. Recent work (Allan and Teter, 1987; Bar-Yam *et al.*, 1989; Rappe *et al.*, 1990; Vanderbilt, 1990; Trouillier and Martins, 1991) has shown that pseudopotential calculations can be performed for systems containing these atoms by employing a substantial but manageable number of plane waves in the basis set. With their increased efficiency, molecular-dynamics and conjugate-gradients methods can handle very large plane-wave basis sets and therefore permit large total-energy pseudopotential calculations to be performed with these new pseudopotentials, thus opening the way for study of a larger variety of chemical systems than was previously possible.

This article provides a detailed description of the total-energy pseudopotential method, the molecular-dynamics method, and conjugate-gradient minimization. The article also discusses related techniques and developments in a number of areas that have played a role in increasing the computational efficiency of these methods. It is hoped that the information presented here is sufficiently detailed and at the leading edge of the work being done to be useful to scientists working both in and outside the field of *ab initio* quantum-mechanical calculations.

II. TOTAL-ENERGY PSEUDOPOTENTIAL CALCULATIONS

This section describes the total-energy pseudopotential method. An extremely useful review of the pseudopotential method can be found in articles by Ihm *et al.* (1979) and Denteneer and van Haeringen (1985). These articles are essential reading for anyone intending to implement codes for total-energy pseudopotential calculations. Total-energy calculations can only be performed if a large number of simplifications and approximations are used. These simplifications and approximations are described in the following sections.

A. Overview of approximations

Prediction of the electronic and geometric structure of a solid requires calculation of the quantum-mechanical total energy of the system and subsequent minimization of that energy with respect to the electronic and nuclear coordinates. Because of the large difference in mass between the electrons and nuclei and the fact that the forces on the particles are the same, the electrons respond essentially instantaneously to the motion of the nuclei. Thus the nuclei can be treated adiabatically, leading to a separation of electronic and nuclear coordinates in the many-body wave function—the so-called Born-Oppenheimer approximation. This "adiabatic principle" reduces the many-body problem to the solution of the dynamics of the electrons in some frozen-in configuration of the nuclei.

Even with this simplification, the many-body problem remains formidable. Further simplifications, however, can be introduced that allow total-energy calculations to be performed accurately and efficiently. These include *density-functional theory* to model the electron-electron interactions, *pseudopotential theory* to model the electron-ion interactions, *supercells* to model systems with aperiodic geometries, and *iterative minimization* techniques to relax the electronic coordinates.

Very briefly, the essential concepts are the following:

(i) Density-functional theory (Hohenberg and Kohn, 1964; Kohn and Sham, 1965) allows one, in principle, to map exactly the problem of a strongly interacting electron gas (in the presence of nuclei) onto that of a single particle moving in an effective nonlocal potential. Although this potential is not known precisely, local approximations to it work remarkably well. At present, we have no *a priori* arguments to explain why these approximations work. Density-functional theory was revitalized in recent years only because theorists performed total-energy calculations using these potentials and showed that they reproduced a variety of ground-state properties within a few percent of experiment. Thus the acceptance of local approximations to density-functional theory has only emerged, *a posteriori,* after many successful investigations of many types of materials and systems. Generally, total-energy differences between related structures can be believed to within a few percent and structural parameters to at least within a tenth of an Å. Cohesive energies, however, can be in error by more than 10%.

(ii) Pseudopotential theory (Phillips, 1958; Heine and Cohen, 1970) allows one to replace the strong electron ion potential with a much weaker potential—a pseudopotential—that describes all the salient features of a valence electron moving through the solid, including relativistic effects. Thus the original solid is now replaced by pseudo valence electrons and pseudo-ion cores. These pseudoelectrons experience exactly the same potential outside the core region as the original electrons but have a much weaker potential inside the core region. The fact that the potential is weaker is crucial, however, because it makes the solution of the Schrödinger equation much simpler by allowing expansion of the wave functions in a relatively small set of plane waves. Use of plane waves as basis functions makes the accurate and systematic study of complex, low-symmetry configurations of atoms much more tractable.

(iii) The supercell approximation allows one to deal with aperiodic configurations of atoms within the framework of Bloch's theorem (see Ashcroft and Mermin, 1976). One simply constructs a large unit cell containing the configuration in question and repeats it periodically throughout space. By studying the properties of the system for larger and larger unit cells, one can gauge the importance of the induced periodicity and systematically filter it out. This approach has been successfully tested against "exact" Koster-Slater Green's-function methods (see Baraff and Schluter, 1979), which are only tractable for very-high-symmetry configurations.

(iv) Finally, new iterative diagonalization approaches (Car and Parrinello, 1985; Payne *et al.*, 1986; Williams and Soler, 1987; Gillan, 1989; Stich *et al.*, 1989; Teter *et al.*, 1989) can be used to minimize the total-energy functional. These are much more efficient than the traditional diagonalization methods. These new methods allow expedient calculation of ionic forces and total energies and significantly raise the level of modern total-energy calculations. These methods are the subject of Secs. III, IV, and V.

B. Electron-electron interactions

The most difficult problem in any electronic structure calculation is posed by the need to take account of the effects of the electron-electron interaction. Electrons repel each other due to the Coulomb interaction between their charges. The Coulomb energy of a system of electrons can be reduced by keeping the electrons spatially separated, but this has to balanced against the kinetic-energy cost of deforming the electronic wave functions in order to separate the electrons. The effects of the electron-electron interaction are briefly described below.

1. Exchange and correlation

The wave function of a many-electron system must be antisymmetric under exchange of any two electrons because the electrons are fermions. The antisymmetry of the wave function produces a spatial separation between electrons that have the same spin and thus reduces the Coulomb energy of the electronic system. The reduction in the energy of the electronic system due to the antisymmetry of the wave function is called the exchange energy. It is straightforward to include exchange in a total-energy calculation, and this is generally referred to as the Hartree-Fock approximation.

The Coulomb energy of the electronic system can be reduced below its Hartree-Fock value if electrons that have opposite spins are also spatially separated. In this case the Coulomb energy of the electronic system is reduced at the cost of increasing the kinetic energy of the electrons. The difference between the many-body energy of an electronic system and the energy of the system calculated in the Hartree-Fock approximation is called the *correlation energy* (see Fetter and Walecka, 1971). It is extremely difficult to calculate the correlation energy of a complex system, although some promising steps are being taken in this direction using quantum Monte Carlo simulations of the electron-gas dynamics (Fahy *et al.*, 1988; Li *et al.*, 1991). At present these methods are not tractable in total-energy calculations of systems with any degree of complexity, and alternative methods are required to describe the effects of the electron-electron interaction.

2. Density-functional theory

Density-functional theory, developed by Hohenberg and Kohn (1964) and Kohn and Sham (1965), provided some hope of a simple method for describing the effects of exchange and correlation in an electron gas. Hohenberg and Kohn proved that the total energy, including exchange and correlation, of an electron gas (even in the presence of a static external potential) is a unique functional of the electron density. The minimum value of the total-energy functional is the ground-state energy of the system, and the density that yields this minimum value is the exact single-particle ground-state density. Kohn and Sham then showed how it is possible, formally, to replace the many-electron problem by an exactly equivalent set of self-consistent one-electron equations. For more details about density-functional theory see von Barth (1984), Dreizler and da Providencia (1985), Jones and Gunnarson (1989), and Kryachko and Ludena (1990).

a. The Kohn-Sham energy functional

The Kohn-Sham total-energy functional for a set of doubly occupied electronic states ψ_i can be written

$$E[\{\psi_i\}]=2\sum_i\int\psi_i\left[-\frac{\hbar^2}{2m}\right]\nabla^2\psi_i d^3\mathbf{r}+\int V_{\text{ion}}(\mathbf{r})n(\mathbf{r})d^3\mathbf{r}+\frac{e^2}{2}\int\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3\mathbf{r}\,d^3\mathbf{r}'+E_{XC}[n(\mathbf{r})]+E_{\text{ion}}(\{\mathbf{R}_I\}) , \tag{2.1}$$

where E_{ion} is the Coulomb energy associated with interactions among the nuclei (or ions) at positions $\{\mathbf{R}_I\}$, V_{ion} is the static total electron-ion potential, $n(\mathbf{r})$ is the electronic density given by

$$n(\mathbf{r})=2\sum_i|\psi_i(\mathbf{r})|^2 , \tag{2.2}$$

and $E_{XC}[n(\mathbf{r})]$ is the exchange-correlation functional.

Only the minimum value of the Kohn-Sham energy functional has physical meaning. At the minimum, the Kohn-Sham energy functional is equal to the ground-state energy of the system of electrons with the ions in positions $\{\mathbf{R}_I\}$.

b. Kohn-Sham equations

It is necessary to determine the set of wave functions ψ_i that minimize the Kohn-Sham energy functional. These are given by the self-consistent solutions to the Kohn-Sham equations (Kohn and Sham, 1965):

$$\left[\frac{-\hbar^2}{2m}\nabla^2+V_{\text{ion}}(\mathbf{r})+V_H(\mathbf{r})+V_{XC}(\mathbf{r})\right]\psi_i(\mathbf{r})=\varepsilon_i\psi_i(\mathbf{r}) , \tag{2.3}$$

where ψ_i is the wave function of electronic state i, ε_i is the Kohn-Sham eigenvalue, and V_H is the Hartree potential of the electrons given by

$$V_H(\mathbf{r})=e^2\int\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathbf{d}^3\mathbf{r}' \ . \tag{2.4}$$

The exchange-correlation potential, V_{XC}, is given formally by the functional derivative

$$V_{XC}(\mathbf{r})=\frac{\delta E_{XC}[n(\mathbf{r})]}{\delta n(\mathbf{r})} \ . \tag{2.5}$$

The Kohn-Sham equations represent a mapping of the interacting many-electron system onto a system of noninteracting electrons moving in an effective potential due to all the other electrons. If the exchange-correlation energy functional were known exactly, then taking the functional derivative with respect to the density would produce an exchange-correlation potential that included the effects of exchange and correlation exactly.

The Kohn-Sham equations must be solved self-consistently so that the occupied electronic states generate a charge density that produces the electronic potential that was used to construct the equations. The sum of the single-particle Kohn-Sham eigenvalues does not give the total electronic energy because this overcounts the effects of the electron-electron interaction in the Hartree energy and in the exchange-correlation energy. The Kohn-Sham eigenvalues are not, strictly speaking, the energies of the single-particle electron states, but rather the derivatives of the total energy with respect to the occupation numbers of these states (Janak, 1978). Nevertheless, the highest occupied eigenvalue in an atomic or molecular calculation is nearly the unrelaxed ionization energy for that system (Perdew *et al.*, 1982).

The Kohn-Sham equations are a set of eigenequations, and the terms within the brackets in Eq. (2.3) can be regarded as a Hamiltonian. The bulk of the work involved in a total-energy pseudopotential calculation is the solution of this eigenvalue problem once an approximate expression for the exchange-correlation energy is given.

c. Local-density approximation

The Hohenberg-Kohn theorem provides some motivation for using approximate methods to describe the exchange-correlation energy as a function of the electron density. The simplest method of describing the exchange-correlation energy of an electronic system is to use the *local-density approximation* (LDA; Kohn and Sham, 1965), and this approximation is almost universally used in total-energy pseudopotential calculations. In the local-density approximation the exchange-correlation energy of an electronic system is constructed by assuming that the exchange-correlation energy per electron at a point $\mathbf{r}$ in the electron gas, $\varepsilon_{XC}(\mathbf{r})$, is equal to the exchange-correlation energy per electron in a homogeneous electron gas that has the same density as the electron gas at point $\mathbf{r}$. Thus

$$E_{XC}[n(\mathbf{r})]=\int\varepsilon_{XC}(\mathbf{r})n(\mathbf{r})d^3\mathbf{r} \tag{2.6a}$$

and

$$\frac{\delta E_{XC}[n(\mathbf{r})]}{\delta n(\mathbf{r})}=\frac{\partial[n(\mathbf{r})\varepsilon_{XC}(\mathbf{r})]}{\partial n(\mathbf{r})} \tag{2.6b}$$

with

$$\varepsilon_{XC}(\mathbf{r})=\varepsilon_{XC}^{\text{hom}}[n(\mathbf{r})] \ . \tag{2.6c}$$

The local-density approximation assumes that the exchange-correlation energy functional is purely local. Several parametrizations exist for the exchange-correlation energy of a homogeneous electron gas (Wigner, 1938; Kohn and Sham, 1965; Hedin and Lundqvist, 1971; Vosko *et al.*, 1980; Perdew and Zunger, 1981), all of which lead to total-energy results that are very similar. These parametrizations use interpolation formulas to link exact results for the exchange-correlation energy of high-density electron gases and calculations of the exchange-correlation energy of intermediate and low-density electron gases.

The local-density approximation, in principle, ignores corrections to the exchange-correlation energy at a point $\mathbf{r}$ due to nearby inhomogeneities in the electron density. Considering the inexact nature of the approximation, it is remarkable that calculations performed using the LDA have been so successful. Recent work has shown that this success can be partially attributed to the fact that the local-density approximation gives the correct sum rule for the exchange-correlation hole (Harris and Jones, 1974; Gunnarsson and Lundqvist, 1976; Langreth and Perdew, 1977). A number of attempts to improve the LDA, for instance by using gradient expansions, have not shown any improvement over results obtained using the simple LDA. One of the reasons why these "improvements" to the LDA do so poorly is that they do not obey the sum rule for the exchange-correlation hole. Methods that do enforce the sum rule appear to offer a consistent improvement over the LDA (Langreth and Mehl, 1981, 1983).

The LDA appears to give a single well-defined global minimum for the energy of a non-spin-polarized system of electrons in a fixed ionic potential. Therefore any energy minimization scheme will locate the global energy minimum of the electronic system. For magnetic materials, however, one would expect to have more than one local minimum in the electronic energy. If the energy functional for the electronic system had many local minima, it would be extremely costly to perform total-energy calculations because the global energy minimum could only be located by sampling the energy functional over a large region of phase space.

C. Periodic supercells

In the preceding section it was demonstrated that certain observables of the many-body problem can be mapped into equivalent observables in an effective single-particle problem. However, there still remains the formidable task of handling an infinite number of noninteracting electrons moving in the static potential of an infinite number of nuclei or ions. Two difficulties must be overcome: a wave function must be calculated for each of the infinite number of electrons in the system, and, since each electronic wave function extends over the entire solid, the basis set required to expand each wave function is infinite. Both problems can be surmounted by performing calculations on periodic systems and applying Bloch's theorem to the electronic wave functions.

1. Bloch's theorem

Bloch's theorem states that in a periodic solid each electronic wave function can be written as the product of a cell-periodic part and a wavelike part (see Ashcroft and Mermin, 1976),

$$\psi_i(\mathbf{r})=\exp[i\mathbf{k}\cdot\mathbf{r}]f_i(\mathbf{r}) \; . \tag{2.7}$$

The cell-periodic part of the wave function can be expanded using a basis set consisting of a discrete set of plane waves whose wave vectors are reciprocal lattice vectors of the crystal,

$$f_i(\mathbf{r})=\sum_{\mathbf{G}} c_{i,\mathbf{G}}\exp[i\mathbf{G}\cdot\mathbf{r}] \; , \tag{2.8}$$

where the reciprocal lattice vectors $\mathbf{G}$ are defined by $\mathbf{G}\cdot\boldsymbol{l}=2\pi m$ for all $\boldsymbol{l}$ where $\boldsymbol{l}$ is a lattice vector of the crystal and m is an integer. Therefore each electronic wave function can be written as a sum of plane waves,

$$\psi_i(\mathbf{r})=\sum_{\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}}\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}] \; . \tag{2.9}$$

2. k-point sampling

Electronic states are allowed only at a set of **k** points determined by the boundary conditions that apply to the bulk solid. The density of allowed **k** points is proportional to the volume of the solid. The infinite number of electrons in the solid are accounted for by an infinite number of **k** points, and only a finite number of electronic states are occupied at each **k** point. The Bloch theorem changes the problem of calculating an infinite number of electronic wave functions to one of calculating a finite number of electronic wave functions at an infinite number of **k** points. The occupied states at each **k** point contribute to the electronic potential in the bulk solid so that, in principle, an infinite number of calculations are needed to compute this potential. However, the electronic wave functions at **k** points that are very close together will be almost identical. Hence it is possible to represent the electronic wave functions over a region of **k** space by the wave functions at a single **k** point. In this case the electronic states at only a finite number of **k** points are required to calculate the electronic potential and hence determine the total energy of the solid.

Methods have been devised for obtaining very accurate approximations to the electronic potential and the contribution to the total energy from a filled electronic band by calculating the electronic states at special sets of **k** points in the Brillouin zone (Chadi and Cohen, 1973; Joannopoulos and Cohen, 1973; Monkhorst and Pack, 1976; Evarestov and Smirnov, 1983). Using these methods, one can obtain an accurate approximation for the electronic potential and the total energy of an insulator or a semiconductor by calculating the electronic states at a very small number of **k** points. The electronic potential and total energy are more difficult to calculate if the system is metallic because a dense set of **k** points is required to define the Fermi surface precisely.

The magnitude of any error in the total energy due to inadequacy of the **k**-point sampling can always be reduced by using a denser set of **k** points. The computed total energy will converge as the density of **k** points increases, and the error due to the **k**-point sampling then approaches zero. In principle, a converged electronic potential and total energy can always be obtained provided that the computational time is available to calculate the electronic wave functions at a sufficiently dense set of **k** points. The computational cost of performing a very dense sampling of **k** space can be significantly reduced by using the $\mathbf{k}\cdot\mathbf{p}$ total-energy method (Robertson and Payne, 1990, 1991). In this technique solutions on the dense set of **k** points are generated from the solutions on a much coarser grid of **k** points using $\mathbf{k}\cdot\mathbf{p}$ perturbation theory.

3. Plane-wave basis sets

Bloch's theorem states that the electronic wave functions at each **k** point can be expanded in terms of a discrete plane-wave basis set. In principle, an infinite plane-wave basis set is required to expand the electronic wave functions. However, the coefficients $c_{i,\mathbf{k}+\mathbf{G}}$ for the plane waves with small kinetic energy $(\hbar^2/2m)|\mathbf{k}+\mathbf{G}|^2$ are typically more important than those with large kinetic energy. Thus the plane-wave basis set can be truncated to include only plane waves that have kinetic energies less than some particular cutoff energy. If a continuum of plane-wave basis states were required to expand each electronic wave function, the basis set would be infinitely large no matter how small the cutoff energy. Application of the Bloch theorem allows the electronic wave functions to be expanded in terms of a discrete set of plane waves. Introduction of an energy cutoff to the discrete plane-wave basis set produces a finite basis set.

The truncation of the plane-wave basis set at a finite cutoff energy will lead to an error in the computed total energy. However, it is possible to reduce the magnitude of the error by increasing the value of the cutoff energy. In principle, the cutoff energy should be increased until the calculated total energy has converged, but it will be shown later that it is possible to perform calculations at lower cutoff energies.

One of the difficulties associated with the use of plane-wave basis sets is that the number of basis states changes discontinuously with cutoff energy. In general these discontinuities will occur at different cutoffs for different **k** points in the **k**-point set. (In addition, at a fixed-energy cutoff, a change in the size or shape of the unit cell will cause discontinuation in the plane-wave basis set.) This problem can be reduced by using denser **k**-point sets, so that the weight attached to any particular plane-wave basis state is reduced. However, the problem is still present even with quite dense **k**-point samplings. It can be handled by applying a correction factor which accounts approximately for the difference between the number of states in a basis set with infinitely large number of **k** points and the number of basis states actually used in the calculation (Francis and Payne, 1990).

4. Plane-wave representation of Kohn-Sham equations

When plane waves are used as a basis set for the electronic wave functions, the Kohn-Sham equations assume a particularly simple form. Substitution of Eq. (2.9) into (2.3) and integration over **r** gives the secular equation

$$\sum_{\mathbf{G}'}\left[\frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2\delta_{\mathbf{GG}'}+V_{\text{ion}}(\mathbf{G}-\mathbf{G}')+V_H(\mathbf{G}-\mathbf{G}')+V_{XC}(\mathbf{G}-\mathbf{G}')\right]c_{i,\mathbf{k}+\mathbf{G}'}=\varepsilon_i c_{i,\mathbf{k}+\mathbf{G}}\,. \qquad (2.10)$$

In this form, the kinetic energy is diagonal, and the various potentials are described in terms of their Fourier transforms. Solution of Eq. (2.10) proceeds by diagonalization of a Hamiltonian matrix whose matrix elements $H_{\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}'}$ are given by the terms in the brackets above. The size of the matrix is determined by the choice of cutoff energy $(\hbar^2/2m)|\mathbf{k}+\mathbf{G}_c|^2$, and will be intractably large for systems that contain both valence and core electrons. This is a severe problem, but it can be overcome by use of the pseudopotential approximation, as discussed in Sec. II.D.

5. Nonperiodic systems

The Bloch theorem can be applied neither to a system that contains a single defect nor in the direction perpendicular to a crystal surface. A continuous plane-wave basis set would be required for the defect calculation, and, although the plane-wave basis set for the surface calculation would be discrete in the plane of the surface, it would be continuous in the direction perpendicular to the surface. Hence an infinite number of plane-wave basis states would be required for both of these calculations, no matter how small the cutoff energy chosen for the basis set. Calculations using plane-wave basis sets can only be performed on these systems if a periodic supercell is used. The supercell for a point-defect calculation is illustrated schematically in Fig. 2. The supercell contains the defect surrounded by a region of bulk crystal. Periodic boundary conditions are applied to the supercell so that the supercell is reproduced throughout space, as implied in the figure. Therefore the energy per unit cell of a crystal containing an array of defects is calculated, rather than the energy of a crystal containing a single defect. It is essential to include enough bulk solid in the supercell to prevent the defects in neighboring cells from interacting appreciably with each other. The independence of defects in neighboring cells can be checked by increasing the volume of the supercell until the computed defect energy has converged. It can then be assumed that defects in neighboring unit cells no longer interact.

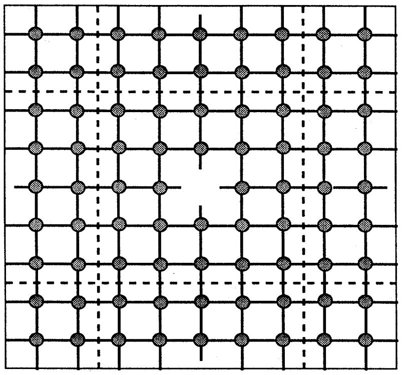

FIG. 2. Schematic illustration of a supercell geometry for a point defect (i.e., vacancy) in a bulk solid. The supercell is the area enclosed by the dashed lines.

A surface may have periodicity in the plane of the surface, but it cannot have periodicity perpendicular to the surface. The supercell for a surface calculation is illustrated schematically in Fig. 3. The supercell contains a

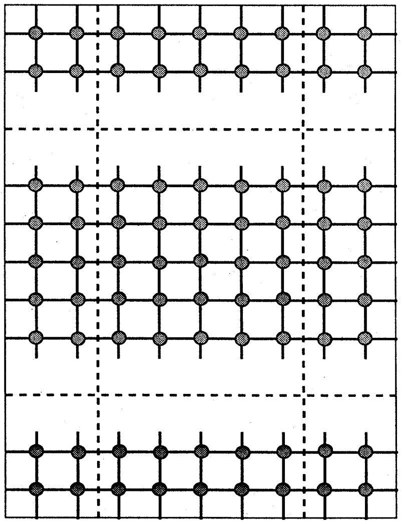

FIG. 3. Schematic illustration of a supercell geometry for a surface of a bulk solid. Same convention as in Fig. 2.

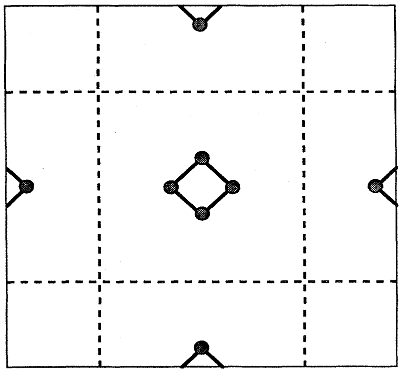

FIG. 4. Schematic illustration of a supercell geometry for a molecule. Same convention as in Fig. 2.

crystal slab and a vacuum region. The supercell is repeated over all space, so the total energy of an array of crystal slabs is calculated. To ensure that the results of the calculation accurately represent an isolated surface, the vacuum regions must be wide enough so that faces of adjacent crystal slabs do not interact across the vacuum region, and the crystal slab must be thick enough so that the two surfaces of each crystal slab do not interact through the bulk crystal.

Finally, even molecules can be studied in this fashion (Joannopoulos *et al.*, 1991), as illustrated in Fig. 4. Again, the supercell needs to be large enough so that the interactions between the molecules are negligible.

D. Electron-ion interactions

1. Pseudopotential approximation

Although Bloch's theorem states that the electronic wave functions can be expanded using a discrete set of plane waves, a plane-wave basis set is usually very poorly suited to expanding electronic wave functions because a very large number of plane waves are needed to expand the tightly bound core orbitals and to follow the rapid oscillations of the wave functions of the valence electrons in the core region. An extremely large plane-wave basis set would be required to perform an all-electron calculation, and a vast amount of computational time would be required to calculate the electronic wave functions. The pseudopotential approximation (Phillips, 1958; Heine and Cohen, 1970; Yin and Cohen, 1982a) allows the electronic wave functions to be expanded using a much smaller number of plane-wave basis states.

It is well known that most physical properties of solids are dependent on the valence electrons to a much greater extent than on the core electrons. The pseudopotential approximation exploits this by removing the core electrons and by replacing them and the strong ionic potential by a weaker pseudopotential that acts on a set of pseudo wave functions rather than the true valence wave functions. An ionic potential, valence wave function and the corresponding pseudopotential and pseudo wave function are illustrated schematically in Fig. 5. The valence wave functions oscillate rapidly in the region occupied by the core electrons due to the strong ionic potential in this region. These oscillations maintain the orthogonality between the core wave functions and the valence wave functions, which is required by the exclusion principle. The pseudopotential is constructed, ideally, so that its scattering properties or phase shifts for the pseudo wave functions are identical to the scattering properties of the ion and the core electrons for the valence wave functions, but in such a way that the pseudo wave functions have no radial nodes in the core region. In the core region, the total phase shift produced by the ion and the core electrons will be greater by π, for each node that the valence functions had in the core region, than the phase shift produced by the ion and the valence electrons. Outside the core region the two potentials are identical, and the scattering from the two potentials is indistinguishable. The phase shift produced by the ion core is different for each angular momentum component of the valence wave function, and so the scattering from the pseudopotential must be angular momentum dependent. The most general form for a pseudopotential is

$$V_{NL}=\sum_{lm}|lm\rangle V_l\langle lm| \ , \tag{2.11}$$

where $|lm\rangle$ are the spherical harmonics and V_l is the pseudopotential for angular momentum l. Acting on the electronic wave function with this operator decomposes the wave function into spherical harmonics, each of which is then multiplied by the relevant pseudopotential V_l.

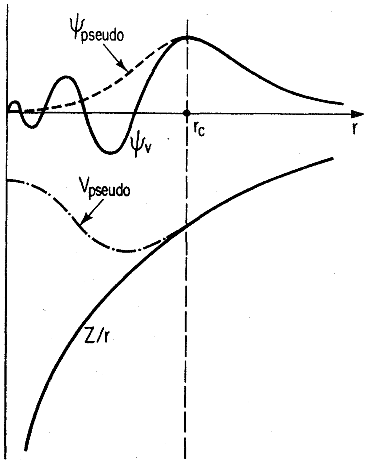

FIG. 5. Schematic illustration of all-electron (solid lines) and pseudoelectron (dashed lines) potentials and their corresponding wave functions. The radius at which all-electron and pseudoelectron values match is designated r_c.

A pseudopotential that uses the same potential for all the angular momentum components of the wave function is called a local pseudopotential. A local pseudopotential is a function only of the distance from the nucleus. It is possible to produce arbitrary, predetermined phase shifts for each angular momentum state with a local potential, but there are limits to the amount that the phase shifts can be adjusted for the different angular momentum states, while maintaining the crucial smoothness and weakness of the pseudopotential. Without a smooth, weak pseudopotential it becomes difficult to expand the wave functions using a reasonable number of plane-wave basis states.

a. Norm conservation

In total-energy calculations, the exchange-correlation energy of the electronic system is a function of the electron density. If the exchange-correlation energy is to be desired accurately, it is necessary that outside the core regions the pseudo wave functions and real wave functions be identical, not just in their spatial dependences but also in their absolute magnitudes, so that the two wave functions generate identical charge densities. Adjustment of the pseudopotential to ensure that the integrals of the squared amplitudes of the real and the pseudo wave functions inside the core regions are identical guarantees the equality of the wave function and pseudo wave function outside the core region. One of the first attempts to construct pseudopotentials of this type was by Starkloff and Joannopoulos (Joannopoulos *et al.* 1977, Starkloff and Joannopoulos 1977). They introduced a class of local pseudopotentials that described the valence energies and wave functions of many heavy atoms accurately.

Of course, in general, the scattering from the ion core is best described by a nonlocal pseudopotential that uses a different potential for each angular momentum component of the wave function. Various groups (Redondo *et al.*, 1977; Hamann *et al.*, 1979; Zunger and Cohen, 1979; Kerker, 1980; Bachelet *et al.*, 1982; Shirley *et al.*, 1989) have now introduced nonlocal pseudopotentials of this type that work extremely well. Moreover, as pointed out by Hamann, Schluter, and Chiang (1979), a match of the pseudo and real wave functions outside the core region also assures that the first-order energy dependence of the scattering from the ion core is correct, so that the scattering is accurately described over a wide range of energy. A method for the construction of pseudopotentials that corrects even the higher-order energy dependence of the scattering has recently been introduced by Shirley *et al.* (1989). Local and nonlocal pseudopotentials of these types are currently termed *ab initio* or *norm conserving* and are capable of describing the scattering due to the ion in a variety of atomic environments, a property referred to as *transferability.*

b. Generation procedure

The typical method for generating an ionic pseudopotential for an atom of species α, v_α is illustrated in Fig. 6 and proceeds as follows. All-electron calculations are performed for an isolated atom in its ground state and some excited states, using a given form for the exchange-correlation density functional. This provides valence electron eigenvalues and valence electron wave functions for the atom. A parametrized form for the ionic pseudopotential is chosen. The parameters are then adjusted, so that a pseudoatom calculation using the same form for exchange-correlation as in the all-electron atom gives both pseudowave functions that match the valence wave functions outside some cutoff radius r_c and pseudoeigenvalues that are equal to the valence eigenvalues. The ionic pseudopotential obtained in this fashion is then used, without further modification, for any environment of the atom. The electronic density in any new environment of the atom is then determined using both the ionic pseudopotential obtained in this way and the same form of exchange-correlation functional as employed in the construction of the ionic pseudopotential. A generalization of this pseudopotential construction procedure for solutions of the atom that are not normalizable has recently been introduced by Hamann (1989).

Finally, it should be noted that ionic pseudopotentials are constructed with r_c ranging typically from one to two times the value of the core radius. It should also be noted that, in general, the smaller the value of r_c, the more "transferable" the potential. (The entire procedure for solving the problem of a solid, given an ionic pseudopotential, is outlined in Sec. II.F.)

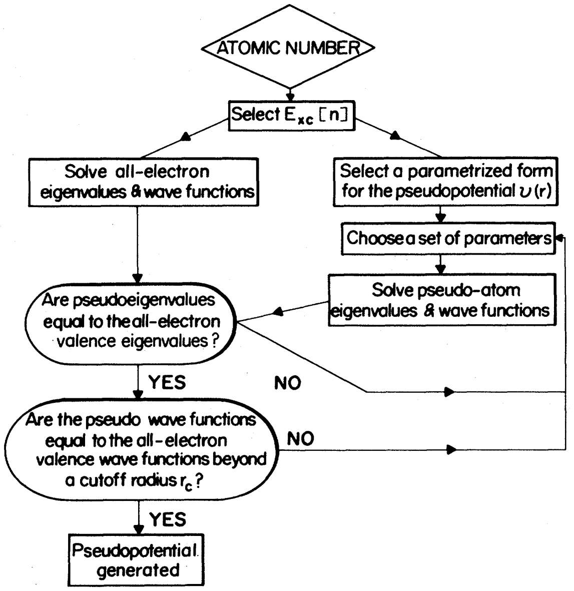
ATOMIC NUMBER
Select $E_{xc}[n]$
Solve all-electron eigenvalues & wave functions
Select a parametrized form for the pseudopotential $\upsilon(r)$
Choose a set of parameters
Solve pseudo-atom eigenvalues & wave functions
Are pseudoeigenvalues equal to the all-electron valence eigenvalues?
YES
NO
Are the pseudo wave functions equal to the all-electron valence wave functions beyond a cutoff radius r_c?
NO
YES
Pseudopotential generated

FIG. 6. Flow chart describing the construction of an ionic pseudopotential for an atom.

c. Convergence properties

The replacement of the true ionic potential by a weaker pseudopotential allows the electronic wave functions to be expanded using far fewer plane-wave basis states than would be needed to expand the wave functions in a full ionic potential. The rapid oscillations of the valence wave functions in the cores of the atoms have been removed, and the small core electron states are no longer present. The pseudopotential approximation has a number of other advantages in addition to reducing the number of plane-wave basis states needed to expand the electronic wave functions. The removal of the core electrons means that fewer electronic wave functions have to be calculated. More importantly, the total energy of the valence electron system is typically a thousand times smaller than the energy of the all-electron system. The difference between the electronic energies of different ionic configurations appears almost totally in the energy of the valence electrons, so that the accuracy required to determine energy differences between ionic configurations in a pseudopotential calculation is much smaller than the accuracy required in an all-electron calculation. The energy differences are just as large when only the valence electrons are included in the calculation, but the total energies are typically a thousand times smaller in the pseudopotential calculation than in the all-electron calculation. But, of course, the total energy is no longer meaningful. Only differences have meaning.

d. Plane-wave basis sets

One property of a pseudopotential that is not incorporated into the standard generation procedure is the value of the cutoff energy required for the plane-wave basis sets. Obviously, the smaller this value, the smaller the basis set required for any particular calculation, and the faster the calculation. A number of approaches to this problem have been adopted. Some authors add additional constraints in the process of generating the pseudopotential which are intended to produce a more rapidly convergent potential (Trouillier and Martins, 1991). Alternatively, a separate optimization procedure can be applied after generating a pseudopotential using one of the standard techniques (Rappe *et al.*, 1990; Rappe and Joannopoulos, 1991). A rather more radical approach has been suggested by Vanderbilt (Vanderbilt, 1990; Laasonen *et al.*, 1991), which involves relaxing the norm conservation of the pseudopotential. This approach will be described more fully in Sec. IX.B.

2. Structure factor

The total ionic potential in a solid is obtained by placing an ionic pseudopotential at the position of every ion in the solid. The information about the positions of the ions is contained in the structure factor. The value of the structure factor at wave vector **G** for ions of species α is given by

$$S_\alpha(\mathbf{G})=\sum_I \exp[i\mathbf{G}\cdot\mathbf{R}_I] \ , \tag{2.12}$$

where the sum is over the positions of all the ions of species α in a single unit cell.

The periodicity of the system restricts the nonzero components of the ionic potential to reciprocal-lattice vectors. Hence it is only necessary to calculate the structure factor at the set of reciprocal-lattice vectors.

3. Total ionic potential

The total ionic potential V_{ion} is obtained by summing the product of the structure factor and the pseudopotential over all the species of ions. For example, for a local potential V_{ion} is given by

$$V_{\text{ion}}(\mathbf{G})=\sum_\alpha S_\alpha(\mathbf{G})v_\alpha(\mathbf{G}) \ . \tag{2.13}$$

At large distances the pseudopotential is a pure Coulomb potential of the form Z/r, where Z is the valence of the atom. On taking the Fourier transform, one finds that the pseudopotential diverges as Z/G^2 at small wave vectors. Therefore the total ionic potential at $\mathbf{G}=\mathbf{0}$ is infinite, so the electron-ion energy is infinite. However, there are similar divergences in the Coulomb energies due to the electron-electron interactions and the ion-ion interactions. The Coulomb $\mathbf{G}=\mathbf{0}$ contributions to the total energy from the three interactions cancel exactly. This is not surprising because there is no Coulomb potential at $\mathbf{G}=\mathbf{0}$ in a charge-neutral system, and so there cannot be a contribution to the total energy from the $\mathbf{G}=\mathbf{0}$ component of the Coulomb potential.

The pseudopotential is, however, not a pure Coulomb potential and therefore not exactly proportional to Z/G^2 for small G. There is a constant contribution to the pseudopotential at small G, equal to the integral of the difference between the pure Coulomb Z/r potential and the pseudopotential. This constant for species α is

$$v_{\alpha,\text{core}}=\int [Z/r-v_\alpha^0(r)]4\pi r^2 dr \ , \tag{2.14}$$

where v_α^0 is the pseudopotential for the $l=0$ angular momentum state. This integral is nonzero only within the core region because the potentials are identical outside the core region.

There is no contribution to the total energy from the Z/G^2 component of the pseudopotential at $\mathbf{G}=\mathbf{0}$ because of the cancellation of the infinities in the electron-ion, electron-electron, and ion-ion energies. However, the non-Coulomb part of the pseudopotential at $\mathbf{G}=\mathbf{0}$ does contribute to the total energy. The contribution is equal to

$$N_{\text{el}}\Omega^{-1}\sum_\alpha N_\alpha v_{\alpha,\text{core}} \ , \tag{2.15}$$

where N_{el} is the total number of electrons in the system,

N_α is the total number of ions of species α, and Ω is the volume of the unit cell.

E. Ion-ion interactions

It is extremely difficult to compute the Coulomb energy of the ionic system using a direct real-space summation because the Coulomb interaction is long ranged. The Coulomb interaction is also long ranged in reciprocal space, so the problem is not solved by performing the summation in reciprocal space. Ewald (1917a, 1917b, 1921) developed a rapidly convergent method for performing Coulomb summations over periodic lattices.

Ewald's method is based on the following identity:

$$\sum_l \frac{1}{|\mathbf{R}_1+l-\mathbf{R}_2|} = \frac{2}{\sqrt{\pi}} \sum_l \int_\eta^\infty \exp[-|\mathbf{R}_1+l-\mathbf{R}_2|^2\rho^2]d\rho + \frac{2\pi}{\Omega}\sum_{\mathbf{G}}\int_0^\eta \exp\left[-\frac{|\mathbf{G}|^2}{4\rho^2}\right] \times \exp[i(\mathbf{R}_1-\mathbf{R}_2)\cdot\mathbf{G}]\frac{1}{\rho^3}d\rho\ , \qquad (2.16)$$

where l are lattice vectors, $\mathbf{G}$ are reciprocal-lattice vectors, and Ω is the volume of the unit cell. This identity provides a method for rewriting the lattice summation for the Coulomb energy due to the interaction between an ion positioned at $\mathbf{R}_2$ and an array of atoms positioned at the points $\mathbf{R}_1+l$. The identity holds for all positive values of η.

At first sight, the infinite Coulomb summation on the left-hand side of Eq. (2.16) has been replaced by two infinite summations, one over lattice vectors and the other over reciprocal-lattice vectors. However, if one chooses an appropriate value of η the two summations become rapidly convergent in their respective spaces. Then the real and reciprocal-space summations can be computed with only a few lattice vectors and a few reciprocal-lattice vectors.

As mentioned in the preceding section, the contributions to the total energy from the electron-ion, ion-ion, and electron-electron interactions at $\mathbf{G}=\mathbf{0}$ cancel exactly, and so the $\mathbf{G}=\mathbf{0}$ contribution to the Coulomb energy of the ionic system must be removed in order to compute the correct total energy. In the Ewald summations the $\mathbf{G}=\mathbf{0}$ contribution to the Coulomb energy has been divided between the real-space and the reciprocal-space summations, so that it is not sufficient simply to omit the $\mathbf{G}=\mathbf{0}$ term in the reciprocal-space Ewald summation. The $\mathbf{G}=\mathbf{0}$ term in the reciprocal-space summation should be omitted and two terms added to the Ewald energy to give the correct total energy. The correct form for the total energy is (Yin and Cohen, 1982b)

$$E_{\text{ion}}=\frac{1}{2}\sum_{I,J} Z_I Z_J e^2 \left\{ \sum_l \frac{\operatorname{erfc}(\eta|\mathbf{R}_1+l-\mathbf{R}_2|)}{|\mathbf{R}_1+l-\mathbf{R}_2|} - \frac{2\eta}{\sqrt{\rho}}\delta_{IJ} + \frac{4\pi}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}} \frac{1}{|\mathbf{G}|^2}\exp\left[-\frac{|\mathbf{G}|^2}{4\eta^2}\right]\cos[(\mathbf{R}_1-\mathbf{R}_2)\cdot\mathbf{G}] - \frac{\pi}{\eta^2\Omega}\right\}\ , \qquad (2.17)$$

where Z_I and Z_J are the valences of ions I and J, respectively, and erfc is the complementary error function.

An ion does not interact with its own Coulomb charge, so the $l=\mathbf{0}$ term must be omitted from the real-space summation when $I=J$. This is indicated by the l in the first summation in Eq. (2.17).

F. Computational procedure with conventional matrix diagonalization

The sequence of steps required to carry out a total-energy pseudopotential calculation with conventional matrix diagonalization techniques is shown in the flow diagram in Fig. 7. The procedure requires an initial guess for the electronic charge density, from which the Hartree potential and the exchange-correlation potential can be calculated. The Hamiltonian matrices for each of the $\mathbf{k}$ points included in the calculation must be constructed, as in Eq. (2.10), and diagonalized to obtain the Kohn-Sham eigenstates. These eigenstates will normally generate a different charge density from the one originally used to construct the electronic potentials, and hence a new set of Hamiltonian matrices must be constructed using the new electronic potentials. The eigenstates of the new Hamiltonians are obtained, and the process is repeated until the solutions are self-consistent. In practice the new electronic potential is taken to be a combination of the electronic potentials generated by the old and the new eigenstates, since this speeds the convergence to self-consistency. To complete the total-energy calculation, tests should be performed to ensure that the total energy is converged both as a function of the number of $\mathbf{k}$ points and as a function of the cutoff energy for the plane-wave basis set. Very few total-energy calculations are taken to absolute convergence. For most calculations, the difference in energy between different ionic configurations is more important than the absolute energies of the individual configurations, and the calculations are performed using an energy cutoff and number of $\mathbf{k}$ points at which the energy differences have converged rather than an energy cutoff and number of $\mathbf{k}$ points at which the absolute energies have converged.

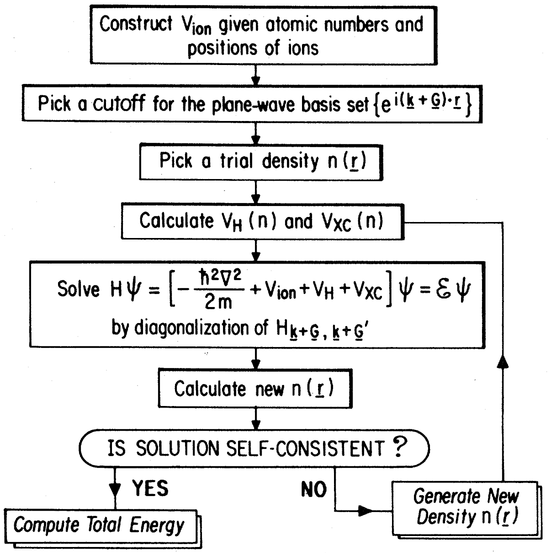
Construct V_{ion} given atomic numbers and positions of ions
Pick a cutoff for the plane-wave basis set $\{e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\}$
Pick a trial density $n(\mathbf{r})$
Calculate $V_H(n)$ and $V_{XC}(n)$
Solve $H\psi = \left[-\frac{\hbar^2\nabla^2}{2m} + V_{ion} + V_H + V_{XC}\right]\psi = \mathcal{E}\psi$ by diagonalization of $H_{\mathbf{k}+\mathbf{G},\,\mathbf{k}+\mathbf{G}'}$
Calculate new $n(\mathbf{r})$
IS SOLUTION SELF-CONSISTENT ?
YES
NO
Compute Total Energy
Generate New Density $n(\mathbf{r})$

FIG. 7. Flow chart describing the computational procedure for the calculation of the total energy of a solid, using conventional matrix diagonalization.

G. Drawbacks of conventional procedure

Pseudopotential calculations with a plane-wave basis are not very well suited to conventional matrix diagonalization techniques. In a total-energy pseudopotential calculation there are typically 100 plane-wave basis states for each atom in the system. The cost of matrix diagonalization increases as the third power of the number of plane-wave basis states, and the memory required to store the Hamiltonian matrix increases as the square of the number of basis states. As a result, conventional matrix diagonalization techniques are restricted to the order of 1000 plane-wave basis states, and this in turn restricts the number of atoms in the unit cell to the order of 10. Using conventional matrix diagonalization methods, the Kohn-Sham eigenvalues of all of the electronic states are calculated, although only those of the lowest occupied states are required to compute the total energy. Furthermore, considerable effort is expended to compute the eigenvalues to machine accuracy, even when the electronic potential is far from self-consistency.

H. Alternative methods

It has been demonstrated that the total-energy pseudopotential technique gives accurate and reliable values for total energies of solids. However, as described above, the power of the pseudopotential method is severely restricted when using conventional matrix diagonalization techniques to solve for the Kohn-Sham eigenstates. In Secs. III, IV, and V, descriptions are given of alternative methods for performing total-energy pseudopotential calculations. These methods are alternative techniques for minimizing the Kohn-Sham energy functional and lead to the same self-consistent Kohn-Sham eigenstates and eigenvalues as conventional matrix diagonalization. However, they are much better suited to performing total-energy pseudopotential calculations because the computational time and memory requirements scale more slowly with the size of the system, allowing calculations on larger and more complex systems than can be studied using conventional matrix diagonalization techniques.

III. THE MOLECULAR-DYNAMICS METHOD

Eigenvalue problems may be solved by successively "improving" a trial wave function. A simple illustration of this process is given in Sec. III.A. Although the Car-Parrinello method (Car and Parrinello, 1985) should be regarded primarily as a scheme for performing *ab initio* dynamical simulations, the molecular-dynamics treatment of the electronic degrees of freedom introduced in the Car-Parrinello method can be used to calculate directly the self-consistent Kohn-Sham eigenstates of a system. In this case the method operates by carrying out a series of iterations that "improve" a set of trial wave functions until they eventually converge to the Kohn-Sham eigenstates. The total energy can be easily computed once the self-consistent Kohn-Sham eigenstates have been determined. In this section we describe the molecular-dynamics treatment of the electronic degrees of freedom and show how it provides a very efficient technique for finding the electronic ground state for a fixed ionic configuration. In Sec. III.A we begin this discussion with a description of a simple scheme for the iterative solution of an eigenvalue problem based on the variational principle. The molecular-dynamics-based method is not as transparent as the example presented here, but it has the common feature of varying trial wave functions until they become eigenstates.

A. Eigenvalue solution by successive "improvement" of a trial wave function

The variational theorem gives an upper bound for the expectation value of a Hamiltonian H for any arbitrary normalized trial wave function ψ. The expectation value is greater than or equal to the energy of the lowest-energy eigenstate of the Hamiltonian. Hence

$$\langle \psi | H | \psi \rangle \geq \lambda_0 \,, \tag{3.1}$$

where λ_0 is the energy of the lowest-energy eigenstate of the Hamiltonian H.

If ψ is expanded using a set of arbitrary orthonormal basis functions $\{\phi\}$,

$$\psi = \sum_n c_n \phi_n \,. \tag{3.2}$$

Substitution of Eq. (3.2) into (3.1) gives

$$\sum_{m,n} c_m^* c_n \langle \phi_m | H | \phi_n \rangle \geq \lambda_0 \ , \tag{3.3}$$

and the constraint of normalization requires that

$$\sum_n |c_n|^2 = 1 \ . \tag{3.4}$$

The values of the coefficients c_n can be varied subject to the constraint of normalization until the minimum value for the expectation value of the Hamiltonian is reached. This minimum value gives an upper bound for the ground-state energy of the Hamiltonian.

The variational theorem gives an upper bound to the ground-state energy of the Hamiltonian. However, the difference between the minimum value of the expectation value and the true ground-state energy of any given Hamiltonian is due to the lack of completeness in the basis set $\{\phi\}$. The eigenstate and the eigen-energy obtained by using the variational theorem are exact in the space of the basis set. Diagonalization of the Hamiltonian matrix in the Hilbert space of the same basis states would yield an identical solution for the lowest-energy eigenstate.

The variational principle can be applied to obtain an estimate for the energy of the next-lowest-energy eigenstate of the Hamiltonian by using a trial wave function that is orthogonal to the ground-state wave function. The eigenstate and eigen-energy obtained for the second eigenstate will again be identical to those calculated by diagonalizing the Hamiltonian matrix in the Hilbert space of the same basis functions. A third eigenstate can be obtained by using a trial wave function that is orthogonal to the ground state and to the first excited state. This process can be repeated until all of the eigenstates have been obtained. The essential point is that the variational principle can be used to obtain eigenstates that are exact in the Hilbert space of the basis set used in the calculation. The molecular-dynamics method is essentially a dynamical method for applying the variational principle, in which the eigenstates of all the lowest-energy electronic states are determined simultaneously.

B. Molecular-dynamics procedure

The molecular-dynamics method will be introduced by a description of its application to a system in which the positions of the ions and the size of the unit cell remain fixed. The calculation can then be directly compared to a total-energy calculation performed using conventional matrix diagonalization techniques. In the traditional molecular-dynamics approach a system of classical particles with coordinates $\{\mathbf{X}_i\}$ interact through an interaction potential $V(\{\mathbf{X}_i\})$. If the configuration of minimum energy is required, the system is started at a high temperature, and the temperature is gradually reduced until the particles reach a configuration $\{\mathbf{X}_i\}_0$ that minimizes V. This procedure is illustrated schematically in Fig. 8.

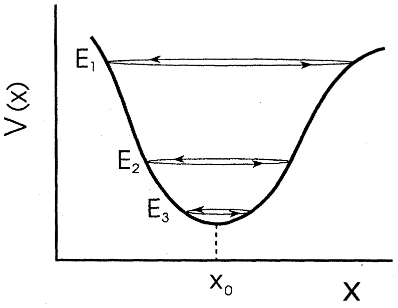

FIG. 8. Schematic illustration of annealing procedure in molecular dynamics. The system is started at a high temperature with total energy E_1. The trajectory at this energy allows the system to sample a large amount of phase space. As the system is gradually cooled to E_2, E_3, etc., it settles down to a minimum energy configuration.

In the Car-Parrinello scheme the Kohn-Sham energy functional $E[\{c_i\}]$ is a function of the set of coefficients of the plane-wave basis set $\{c_i\}$. Each coefficient c_i can be regarded as the coordinate of a classical "particle." To minimize the Kohn-Sham energy functional, these "particles" are given a kinetic energy, and the system is gradually cooled until the set of coordinates reaches the values $\{c_i\}_0$ that minimize the functional. Thus the problem of solving for the Kohn-Sham eigenstates is reduced to one of solving for a set of classical equations of motion. It should be emphasized, however, that the Kohn-Sham energy functional is physically meaningful *quantum mechanically* only when the coefficients take the values $\{c_i\}_0$.

1. Molecular-dynamics Lagrangian

Car and Parrinello formulated their method in the language of molecular dynamics. Their essential step was to treat the electronic wave functions as dynamical variables. A Lagrangian is defined for the electronic system as follows:

$$L = \sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle - E[\{\psi_i\},\{\mathbf{R}_I\},\{\alpha_n\}] \ , \tag{3.5}$$

where μ is a fictitious mass associated with the electronic wave functions, E is the Kohn-Sham energy functional, $\mathbf{R}_I$ is the position of ion I, and α_n define the size and shape of the unit cell. The kinetic-energy term in the Lagrangian is due to the fictitious dynamics of the electronic degrees of freedom. The Kohn-Sham energy functional takes the place of the potential energy in a conventional Lagrangian formulation.

2. Constraints

The electronic wave functions are subject to the constraints of orthonormality,

$$\int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})d^3\mathbf{r}=\delta_{ij} \ . \tag{3.6}$$

These constraints are incorporated in the molecular-dynamics Lagrangian by using the method of Lagrange multipliers. The molecular-dynamics Lagrangian becomes

$$L=\sum_i \mu\langle \dot{\psi}_i|\dot{\psi}_i\rangle - E[\{\psi_i\},\{\mathbf{R}_I\},\{\alpha_n\}] + \sum_{i,j}\Lambda_{ij}\left[\left\{\int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})d^3\mathbf{r}\right\}-\delta_{ij}\right] \ . \tag{3.7}$$

The Lagrange multipliers Λ_{jj} ensure that wave functions remain normalized, while the Lagrange multipliers Λ_{ij} $(i\neq j)$ ensure that the wave functions remain orthogonal. As described below, the Lagrange multipliers may be thought of as providing additional forces acting on the wave functions, which ensure that the wave functions remain orthonormal.

C. Molecular-dynamics equations of motion

The equations of motion for the electronic states are derived from the Lagrange equations of motion,

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\psi}_i^*}\right]=\frac{\partial L}{\partial \psi_i^*} \ , \tag{3.8}$$

which give

$$\mu\ddot{\psi}_i=-H\psi_i+\sum_j \Lambda_{ij}\psi_j \ , \tag{3.9}$$

where H is the Kohn-Sham Hamiltonian.

The force $-(H\psi_i)$ is the gradient of the Kohn-Sham energy functional at the point in the Hilbert space that corresponds to the wave function ψ_i. The Lagrange multipliers add forces $\Lambda_{ij}\psi_j$ to the force $-(H\psi_i)$. These forces ensure that the electronic wave functions remain orthonormal as they propagate along their molecular-dynamics trajectories. A general discussion of the consequences of various orthonormalization schemes is given by Broughton and Khan (1989).

1. Unconstrained equations of motion

The constraints of orthonormality play a crucial role in the evolution of the electronic states in the molecular-dynamics method. To illustrate the importance of these constraints, we consider the evolution of the electronic states in the absence of any constraints:

$$\mu\ddot{\psi}_i=-[H-\sigma]\psi_i \ , \tag{3.10}$$

where H is the Kohn-Sham Hamiltonian, and σ is an energy shift that defines the zero of energy.

If ψ_i is expanded in the basis set of the eigenstates of Hamiltonian H,

$$\psi_i=\sum_n c_{i,n}\xi_n \ , \tag{3.11}$$

and if Eq. (3.11) is substituted into (3.10), the following equation of motion for the coefficient of the eigenstate ξ_n is obtained:

$$\mu\ddot{c}_{i,n}=-[\varepsilon_n-\sigma]c_{i,n} \ , \tag{3.12}$$

where ε_n is the eigenvalue corresponding to eigenstate ξ_n. Integration of these equations of motion, under the assumption that the velocities of the coefficients are initially zero, gives the coefficients at time t as

$$c_{i,n}(t)=c_{i,n}(0)\cos\{[(\varepsilon_n-\sigma)/\mu]^{1/2}t\}, \quad \varepsilon_n>\sigma \ , \tag{3.13a}$$

$$c_{i,j}(t)=c_{i,n}(0)\cosh[(|\varepsilon_n-\sigma|/\mu)^{1/2}t], \quad \varepsilon_n<\sigma \ . \tag{3.13b}$$

Here $c_{i,n}(0)$ are the initial values of the coefficients.

It can be seen that the amplitudes of the coefficients of the eigenstates with energies greater than σ oscillate with time, while the amplitudes of the coefficients of the eigenstates with energies less than σ increase with time. If σ is chosen to be larger than the lowest-energy eigenvalue, then all the electronic states that have $c_{i,0}(0)\neq 0$ will converge to the lowest-energy eigenstate ξ_0, since the coefficient $c_{i,0}$ will increase faster than any other coefficient. Therefore, under the unconstrained equations of motion, the electronic wave functions remain neither orthogonal nor normalized. The initial wave functions will only converge to different eigenstates when the constraints of orthogonality are imposed.

2. Constrained equations of motion

The constrained molecular-dynamics equations of motion for the electronic states,

$$\mu\ddot{\psi}_i=-H\psi_i+\sum_j \Lambda_{ij}\psi_j \ , \tag{3.14}$$

ensure that the electronic wave functions remain orthonormal at every instant in time. The molecular-dynamics evolution of the electronic wave functions under these equations of motion would also conserve the total energy in the electronic degrees of freedom for the system of fixed ions we assume for this section. However, to ensure these properties, the values of the Lagrange multipliers must vary continuously with time, and so implementation of this form of the molecular-dynamics equations requires that the Lagrange multipliers be evaluated at infinitely small time separations. To make the calculations tractable, variation of the Lagrange multipliers during a time step is neglected and the Lagrange multipliers are approximated by a constant value during the time step. In this case the wave functions will not be exactly orthonormal at the end of the time step, and a separate orthonormalization step is needed in the calculation.

3. Partially constrained equations of motion

Since a separate orthonormalization step is required at the end of each time step, it is possible to remove the constraints of orthogonality from the equation of motion and use a *partially* constrained equation of motion. The constraints of orthogonality are then imposed after the equations of motion have been integrated, and the Lagrange multipliers for the constraints of normalization Λ_{ii} are approximated by the expectation values of the energies of the states, λ_i, where

$$\lambda_i = \langle \psi_i | H | \psi_i \rangle \ . \tag{3.15}$$

This leads to an equation of motion that has the form

$$\mu \ddot{\psi}_i = -[H - \lambda_i]\psi_i \ . \tag{3.16}$$

With this equation of motion, the acceleration of an electronic state is always orthogonal to that state, a necessary requirement to maintain normalization, and the acceleration becomes zero when the wave function is an exact eigenstate.

D. Integration of equations of motion

Once the accelerations of the coefficients have been calculated, the equations of motion for the coefficients of the plane-wave basis states have to be integrated. Car and Parrinello used the Verlet algorithm (Verlet, 1967) to integrate the equations of motion.

1. The Verlet algorithm

The Verlet algorithm is derived from the simplest second-order difference equation for the second derivative. It gives the value of the ith electronic state at the next time step, $\psi_i(\Delta t)$, as

$$\psi_i(\Delta t) = 2\psi_i(0) - \psi_i(-\Delta t) + \Delta t^2 \ddot{\psi}_i(0) \ , \tag{3.17a}$$

where Δt is the length of the time step, $\psi_i(0)$ is the value of the state at the present time step, and $\psi_i(-\Delta t)$ is the value of the state at the last time step. Substitution of Eq. (3.16) into (3.17a) then gives

$$\psi_i(\Delta t) = 2\psi_i(0) - \psi_i(-\Delta t) - \frac{\Delta t^2}{\mu}[H - \lambda_i]\psi_i(0) \ . \tag{3.17b}$$

The Verlet algorithm introduces an error of order Δt^4 into the integration of the equations of motion. A more sophisticated finite-difference algorithm could be used to integrate the equations of motion and hence reduce the error in the integration to a higher order of Δt. In principle, for a given level of accuracy this would allow a longer time step to be used in the integration of the equations of motion and hence reduce the total computational time by reducing the number of time steps required to perform the calculation. The maximum stable time step, however, is not significantly increased with a higher-order difference scheme. A more sophisticated finite-difference equation would also require the values of the coefficients or the corresponding accelerations from a larger number of time steps in order to integrate the equations of motion. It requires a large amount of memory to store the wave-function coefficients and accelerations for each time step in a total-energy pseudopotential calculation. If the extra coefficients and accelerations did not fit into core memory, the computation could become I/O bound and the total time required for the calculation may actually increase.

2. Stability of the Verlet algorithm

A general performance measure of algorithms of the Verlet type is the rate at which they converge to minimum-energy state. A given problem normally requires a certain amount of real time to converge, and the computational effort is then determined by the size of the time step Δt. In what follows it is demonstrated that the largest Δt allowed for *stability* is related to the difference between the largest and smallest eigenvalues of the system.

Given the assumption that ψ_i is near the lowest-energy eigenstate ξ_0, the state ψ_i is expanded as in Eq. (3.11),

$$\psi_i = \xi_0 + \sum_{\alpha \neq 0} \delta_\alpha(t)\xi_\alpha \ , \tag{3.18}$$

where the δ_α represent infinitesimal coefficients. Substitution of Eq. (3.18) into (3.17b) gives to first order in δ_α

$$\delta_\alpha(\Delta(t)) = 2\delta_\alpha(0) - \delta_\alpha(-\Delta(t)) - \frac{(\Delta t)^2}{\mu}[\varepsilon_\alpha - \varepsilon_0]\delta_\alpha(0) \ . \tag{3.19}$$

In the standard stability analysis (see Mathews and Walker, 1970), a constant growth factor g is introduced at each time step, so that

$$\delta_\alpha(n\Delta t) = g\delta_\alpha((n-1)\Delta t) \ . \tag{3.20}$$

Substitution of Eq. (3.20) into (3.19) then gives

$$g^2 - 2g + 1 + \frac{(\Delta t)^2}{\mu}(\varepsilon_\alpha - \varepsilon_0)g = 0 \ , \tag{3.21}$$

and the real part of g can become greater than 1 if

$$\Delta t > \frac{2\mu^{1/2}}{(\varepsilon_\alpha - \varepsilon_0)^{1/2}} \ . \tag{3.22a}$$

Therefore the largest possible Δt that is allowed for stability must be

$$\Delta t \approx \frac{2\mu^{1/2}}{(\varepsilon_{\max} - \varepsilon_0)^{1/2}} \ , \tag{3.22b}$$

where $\varepsilon_{\max}$ is the largest eigenvalue of the problem.

For a Hamiltonian representation in a plane-wave basis, the largest eigenvalue is primarily determined by the cutoff kinetic energy of the basis set. Thus the Verlet algorithm will require the time step to be reduced as the cutoff energy is increased. This problem is addressed again in Sec. IV.A.

In the discussion above, it has been tacitly assumed that the Hamiltonian remains fixed during the time evolution of the system. New instabilities can arise, however, when the Hamiltonian is not allowed to vary when it must vary, as in the case of self-consistency. These difficulties are discussed in Sec. III.J.

E. Orthogonalization of electronic wave functions

After one integrates the partially constrained equations of motion for the coefficients of all the basis states for each electronic state, the wave functions will no longer be orthogonal. Car and Parrinello use an iterative technique to orthogonalize the wave functions, repeating the application of the following algorithm to generate a new set of wave functions ψ_i' from a normalized set of wave functions ψ_i:

$$\psi_i' = \psi_i - \tfrac{1}{2}\sum_{j\neq i} \langle \psi_j | \psi_i \rangle \psi_j \ . \tag{3.23}$$

The electronic wave functions can be made orthogonal to any desired accuracy by a repeated application of this algorithm. If the algorithm is applied to two wave functions, these wave functions will be exactly orthogonal after a single application of the algorithm. However, applying the algorithm to orthogonalize each of the new wave functions to a third wave function will make the two wave functions nonorthogonal. The number of iterations of this algorithm required to orthogonalize a set of wave functions to a particular accuracy increases with the number of wave functions and with the initial degree of nonorthogonality. The algorithm does not maintain the normalization of the wave functions, so they must be normalized after each application of the algorithm. Various methods for imposing orthogonality have been compared by Broughton and Khan (1989).

F. Damping

When damping is applied to the motions of the coefficients $c_{i,n}$, the coefficients will evolve to the values that minimize the Kohn-Sham energy functional. This is illustrated schematically in Fig. 9. The damping of the coefficients can be applied in a number of ways: a damping term of the form $-\gamma\dot{\psi}_i$ can be added to the equation of motion for the wave function ψ_i, or the velocities of the coefficients can be reduced at the end of a time step by replacing the value of each coefficient at the previous time step by a value lying between the values of the coefficient at the previous and present timesteps.

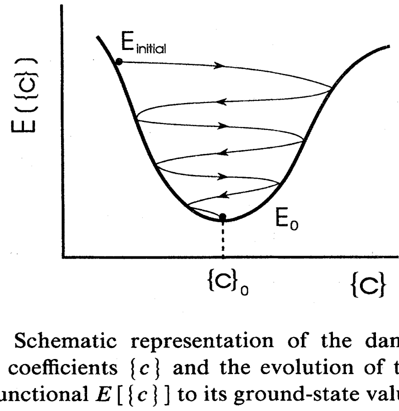

FIG. 9. Schematic representation of the damping of wave-function coefficients $\{c\}$ and the evolution of the Kohn-Sham energy functional $E[\{c\}]$ to its ground-state value E_0.

G. Self-consistency

The accelerations of the wave functions in the molecular-dynamics equations of motion are governed by the Kohn-Sham Hamiltonian. As described in Sec. II, the Hartree potential and the exchange-correlation potential contribute to the Kohn-Sham Hamiltonian and depend on the charge density generated by the electronic wave functions. As the wave functions evolve under the molecular-dynamics equations of motion, these potentials vary. The potentials are recalculated at the end of each time step (when a new set of wave functions has been generated) and lead to a new Kohn-Sham Hamiltonian. Thus the evolution of the coefficients to their stationary values is accompanied by an evolution of the Kohn-Sham Hamiltonian to self-consistency. This is illustrated in Fig. 10. Each of the solid-lines shown passing through the open circles corresponds to a *static* Kohn-Sham Hamiltonian with a *fixed* charge density. During a molecular-dynamics time step, the coefficients $\{c\}$ move along a trajectory that lies between two open circles. At the end of each time step, the Kohn-Sham Hamiltonian is updated with the new charge density, and the trajectory shifts to a new solid line. This shift is indicated by the dotted-line trajectory. The final time step leads to a self-consistent solution of the Kohn-Sham Hamiltonian and the determination of the minimum in the total energy.

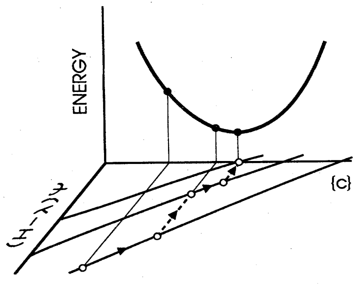

FIG. 10. Schematic representation of the evolution of coefficients $\{c\}$, Kohn-Sham Hamiltonian H, and the total energy, in the final two time steps of the molecular-dynamics method. Solid-line trajectories between open circles correspond to H with a *fixed* charge density. The final time step leads to a self-consistent solution of H and a simultaneous determination of the minimum total energy. Note that, if a trajectory along a solid line were followed all the way to the $\{c\}$ axis, this would correspond to conventional matrix diagonalization of H with a *fixed* charge density.

H. Kohn-Sham eigenstates

The Kohn-Sham energy functional is minimized by any set of wave functions that are a linear combination of the lowest-energy Kohn-Sham eigenstates. These wave functions will be stationary under the molecular-dynamics equations of motion and subsequent orthogonalization. Therefore, in the molecular-dynamics method, each electronic wave function will, in general, converge to a linear combination of the lowest-energy Kohn-Sham eigenstates. This is not a problem for systems with a gap, but it can be a severe problem for metallic systems in which the occupancy of a state depends on its eigenvalue. Some ideas for handling metals in Car-Parrinello dynamical simulations have been proposed by Fernando *et al.* (1989), Woodward *et al.* (1989), Benedek *et al.* (1991), and Pederson and Jackson (1991). The actual Kohn-Sham eigenvalues can be found by diagonalization of the matrix whose matrix elements are given by

$$O_{ij} = \langle \psi_i | H | \psi_j \rangle \ . \tag{3.24}$$

A simpler approach, which guarantees convergence to Kohn-Sham eigenstates, is presented later in Sec. IV.B.

I. Computational procedure with molecular dynamics

The procedure for performing a total-energy pseudopotential calculation using the molecular-dynamics technique is shown in the flow diagram of Fig. 11. The procedure requires an initial set of trial wave functions from which the Hartree potential and the exchange-correlation potential can be calculated. The Hamiltonian matrices for each of the **k** points included in the calculation are constructed, and from these the accelerations of the wave functions are calculated. The equations of motion for the electronic states are integrated, and the wave functions are orthogonalized and normalized. The charge density generated by the new set of wave functions is then calculated. This charge density is used to construct a new set of Hamiltonian matrices, and a further set of wave functions is obtained by integration of the equations of motion and orthonormalization of the resultant wave functions. These iterations are repeated until the wave functions are stationary. The wave functions are then linear combinations of the Kohn-Sham eigenstates. The Kohn-Sham energy functional is minimized, and its value gives the total energy of the system. The solution is identical to the solution that would be obtained by using matrix diagonalization techniques with the same basis states. The convergence tests described in Sec. II.F must be performed to ensure that the calculated total energy has converged both as a function of the number of **k**-points included in the calculation and as a function of the cutoff energy for the plane-wave basis set.

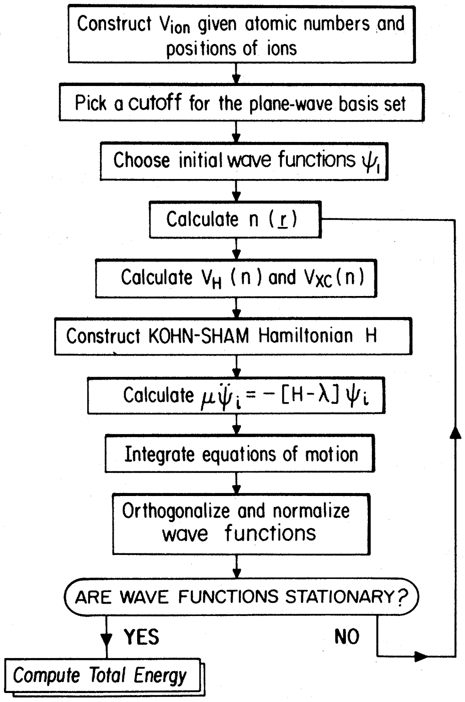
Construct $V_{\rm ion}$ given atomic numbers and positions of ions
Pick a cutoff for the plane-wave basis set
Choose initial wave functions ψ_i
Calculate $n(\underline{r})$
Calculate $V_H(n)$ and $V_{XC}(n)$
Construct KOHN-SHAM Hamiltonian H
Calculate $\mu\ddot{\psi}_i = -[H-\lambda]\psi_i$
Integrate equations of motion
Orthogonalize and normalize wave functions
ARE WAVE FUNCTIONS STATIONARY?
YES
NO
Compute Total Energy

FIG. 11. Flow chart describing the computational procedure for the calculation of the total energy of a solid with molecular dynamics.

J. Instabilities and fluctuations in the Kohn-Sham energy

In Sec. III.D.2 above a criterion was derived that provides an upper bound to the largest stable time step in the Verlet algorithm. There are, however, additional limitations to this time step that are much more subtle. These limitations are related to instabilities in the Kohn-Sham energy and can arise when the Hamiltonian is allowed to evolve under the equations of motion, as required by self-consistency. These instabilities are caused by charge fluctuations and are commonly referred to as *charge sloshing.*

As discussed in Sec. III.G above, the Hartree and exchange-correlation potentials of the Kohn-Sham Hamiltonian depend on the electronic density and must change after each time step. If these changes are too large, the problem becomes unstable and the time step must be reduced. This instability is indicated schematically in Fig. 12. The trajectories in this figure should be contrasted with the trajectories shown in Fig. 10.

The major difficulty lies with the Hartree potential, $V_H(\mathbf{G})$, in Eq. (2.10). $V_H(\mathbf{G})$ is proportional to $n(\mathbf{G})/|\mathbf{G}|^2$, where $n(\mathbf{G})$ is the Fourier transform of the charge density. Therefore, at small reciprocal-lattice vectors, a small change in $n(\mathbf{G})$ will produce large changes in the potential. Since these changes are not taken into account during an elemental time step, they can lead to large increases in energy if the time step is too large. This is particularly true for systems with sufficiently small reciprocal-lattice vectors. The smallest (nonzero) reciprocal-lattice vector is inversely proportional to the

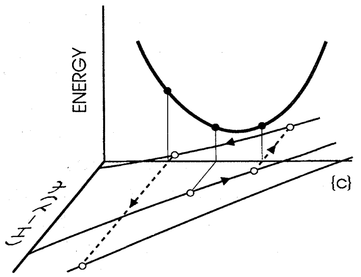

FIG. 12. Illustration of instability in the molecular-dynamics method and Kohn-Sham Hamiltonian as a consequence of a large time step. The convention is the same as in Fig. 10. Note that each iteration drives the coefficients further from their equilibrium value.

longest length of the supercell. Therefore, as supercells become physically larger, the stability of the problem eventually becomes dominated by "charge sloshing" considerations, and the time step must be reduced to keep the fluctuations small.

A related difficulty, which leads to a similar phenomenon, arises if many states with nearly the same eigenvalue exist in the neighborhood of the Fermi energy. Under these circumstances, macroscopic oscillations in electron density can occur with very little change in total energy. The problem is particularly serious for metals. Since small energy changes may yield large differences in electron density, and hence in forces, close convergence of the electrons to their ground state is necessary in metallic systems.

Thus the largest stable time step in the Car-Parrinello algorithm is dominated by either the maximum kinetic energy in the problem (as discussed in Sec. III.D.2) or the need to limit *charge sloshing*. One of these two considerations will place a practical limit on the size and type of problem that can be attacked with the Car-Parrinello algorithm. Despite these limitations, however, there are many problems for which the method has outstanding utility.

Finally, it should be noted that, even for an infinitesimal time step, the Kohn-Sham energy will not always decrease monotonically during the evolution of the electronic system to its ground state. Insufficient damping of the wave-function coefficients in the equations of motion, for example, will result in fluctuations of the Kohn-Sham energy during this evolution. This is simply the expected dynamical behavior of an underdamped system and will not prevent the electrons from *eventually* reaching their ground state.

K. Computational cost of molecular dynamics

The molecular-dynamics method provides a general technique for solving eigenvalue problems. However, this method was developed specifically in the context of total-energy pseudopotential calculations, and in this section the computational cost of using the molecular-dynamics method to perform total-energy pseudopotential calculations will be calculated. An important feature of any computational method is the rate at which the computational time increases with the size of the system, and so particular attention will be paid to the operations that increase fastest as the size of the system increases.

In a total-energy pseudopotential calculation the wave functions ψ_i are expanded in a plane-wave basis set so that

$$\psi_i(\mathbf{r})=\sum_{\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}}\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}] \ . \tag{3.25}$$

A feature of total-energy pseudopotential calculations is that the number of plane-wave basis states used in the calculations is always much larger than the number of occupied bands. The ratio is typically 100:1. This ratio is independent of the size of the system: increasing the size of the unit cell will increase the number of atoms in the unit cell. However, the lengths of the reciprocal-lattice vectors will be reduced, so that there will be more plane waves with energies below the cutoff energy for the plane-wave basis set. The ratio of the number of plane waves to the number of occupied bands is considerably larger when the system contains any transition-metal atoms or first-row atoms. These atoms have strong pseudopotentials, and much larger basis sets are needed to expand the electronic wave functions. Only the parts of the total-energy calculation that scale faster than linearly with the size of the system will be considered in the following discussion. The parts of the calculation that scale linearly with the size of the system do not have large "prefactors" associated with their computational cost, so the computational time required for these parts of the calculation is negligible. A calculation for a system that has N_B occupied bands in which the electronic wave functions are expanded in a basis set containing N_{PW} plane waves will be considered. The operations described below scale linearly with the number of $\mathbf{k}$ points so that a calculation for a single $\mathbf{k}$ point will be considered.

1. Calculation of charge density

The charge density is most cheaply computed in real space because it is simply the square of the magnitude of the wave function. This requires that the wave functions be Fourier transformed from reciprocal space to real space. However, the charge density has components with wave vectors up to twice the cutoff wave vector for the electronic wave functions. Hence, to maintain a faithful representation of the charge density, one must compute it on a Fourier grid twice as dense in each spatial direction as the grid required to provide a faithful representation of the wave function. Furthermore, the plane-wave components of the wave function, which lie within a sphere in reciprocal space, must be placed in an orthorhombic grid of plane waves to perform the fast Fourier transformation (FFT). Fourier transformation of a single wave function from reciprocal space to real space can be performed in $N_{\text{FFT}} = N_{RS}\ln(N_{RS})$ operations using a fast Fourier transform algorithm, where N_{RS} is the number of points in the real-space grid. For the reasons given above, N_{RS} is of the order of 16 times the number of plane waves in the wave function. Therefore the cost of performing the fast Fourier transform for a single band requires $N_{\text{FFT}} = 16N_{PW}\ln(N_{PW})$ operations; the factor 16 is irrelevant in the logarithm. The total charge density can then be calculated in $N_B N_{\text{FFT}}$ operations.

2. Construction of Hamiltonian matrix

The Kohn-Sham Hamiltonian is a matrix of dimension N_{PW}. The total number of elements in the matrix is N_{PW}^2. The number of operations to construct the matrix and the storage required by the matrix both increase as N_{PW}^2.

3. Accelerations of the coefficients

The molecular-dynamics equation of motion for the coefficient of the plane-wave basis state with wave vector $\mathbf{k}+\mathbf{G}$ is obtained by substitution of Eq. (3.25) into (3.16) and use of (2.3) for H. Integration over $\mathbf{r}$ then gives

$$\mu\ddot{c}_{i,\mathbf{k}+\mathbf{G}} = -\left[\frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2-\lambda_i\right]c_{i,\mathbf{k}+\mathbf{G}} - \sum_{\mathbf{G}'} V_H(\mathbf{G}-\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'} - \sum_{\mathbf{G}'} V_{XC}(\mathbf{G}-\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'} - \sum_{\mathbf{G}'} V_{\text{ion}}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'} \, , \tag{3.26}$$

with λ_i defined as in Eq. (3.15).

If Eq. (3.26) is used to calculate the accelerations of the coefficients of the plane-wave basis states, N_{PW}^2 operations are required to calculate the accelerations for a single wave function. This is the number of operations required to multiply the wave function ψ_i by the Kohn-Sham Hamiltonian H. The accelerations of each of the N_B occupied bands must be calculated, so that a total of $N_B N_{PW}^2$ operations are required to compute the accelerations of the wave functions.

4. Integration of equations of motion

Integration of the equations of motion for the electronic states requires $N_B N_{PW}$ operations.

5. Orthogonalization

The number of operations required to orthogonalize the wave functions using Car and Parrinello's algorithm is proportional to $N_B^2 N_{PW}$.

6. Comparison to conventional matrix diagonalizations

The total computational time for the processes described above is dominated by the $N_B N_{PW}^2$ operations required to calculate the accelerations of the wave functions. This should be compared to the computational cost of conventional matrix diagonalization techniques for total-energy pseudopotential calculations, which is dominated by the N_{PW}^3 operations required to diagonalize the Kohn-Sham Hamiltonian. As the number of occupied bands in a total-energy pseudopotential calculation is so much smaller than the number of plane waves included in the basis set, it appears that the molecular-dynamics method offers a considerable increase in computational speed over matrix diagonalization techniques. However, the number of time steps required to integrate the equations of motion and converge the electron states in the molecular-dynamics method is usually *larger* than the number of iterations required to achieve self-consistency in matrix diagonalization calculations. Hence the equation of motion given in (3.26) does not provide a significantly faster method for obtaining the self-consistent eigenstates than conventional matrix diagonalization techniques. It should also be noted that the full Kohn-Sham Hamiltonian has to be stored, which requires N_{PW}^2 words of memory. Therefore, even if the molecular-dynamics method were faster than conventional matrix diagonalization methods, the memory requirement would make it difficult to perform calculations that required extremely large plane-wave basis sets.

7. Local pseudopotentials

The problems described above can be overcome by replacing the nonlocal ionic pseudopotential by a local pseudopotential. A local pseudopotential is a function only of the distance from the nucleus or, equivalently, is a function only of the difference between the wave vectors of the plane-wave basis states. Therefore use of the same procedure as before to derive Eq. (3.26) results in the equation of motion for the coefficient of the plane-wave basis state at wave vector $\mathbf{k}+\mathbf{G}$ with a local pseudopotential

$$\mu\ddot{c}_{i,\mathbf{k}+\mathbf{G}} = -\left[\frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2-\lambda_i\right]c_{i,\mathbf{k}+\mathbf{G}} - \sum_{\mathbf{G}'} V_T(\mathbf{G}-\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'} \, , \tag{3.27}$$

where $V_T(\mathbf{G})$ is the total potential given by

$$V_T(\mathbf{G}) = V_{\text{ion}}(\mathbf{G}) + V_H(\mathbf{G}) + V_{XC}(\mathbf{G}) \, . \tag{3.28}$$

The equation of motion can more usefully be written as

$$\mu \ddot{c}_{i,\mathbf{k}+\mathbf{G}} = -\left[\frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2 - \lambda_i\right] c_{i,\mathbf{k}+\mathbf{G}} - \int [V_T(\mathbf{r})\psi_i(\mathbf{r})]\exp(i[\mathbf{k}+\mathbf{G}]\cdot\mathbf{r})d^3\mathbf{r} \ . \qquad (3.29)$$

This form of the equation of motion shows that the multiplication of the wave function by the Kohn-Sham Hamiltonian can be divided into a part that is diagonal in reciprocal space and a part that is diagonal in real space. However in order to maintain a faithful representation of the potentials and the product $V_T(\mathbf{r})\psi_i(\mathbf{r})$, one must perform the real-space multiplication on a double-density Fourier transform grid (see Sec. III.K.1). This separation procedure leads to a multiplication that can be performed in just $17N_{PW}$ operations: $16N_{PW}$ for the multiplication of the wave function by the potential in real space and N_{PW} for the multiplication of the wave function by the kinetic-energy operator in reciprocal space. The computational cost of calculating the accelerations of the wave functions is dominated by the N_{FFT} operations required to Fourier-transform the wave function from reciprocal space to real space and the N_{FFT} operations required to transform the contribution to the acceleration that is calculated in real space back to reciprocal space. The accelerations of all the plane-wave coefficients for the N_B occupied electron states can be calculated in $2N_B N_{\mathrm{FFT}}$ operations. Reduction of the number of operations required to evaluate the accelerations of the coefficients from $N_B N_{PW}^2$ to $2N_B N_{\mathrm{FFT}}$ makes the molecular-dynamics method very much faster than conventional matrix diagonalization techniques for any system, although the saving in time is obviously much greater for large systems. The same technique can be used with any iterative matrix diagonalization technique, since all iterative matrix diagonalization techniques involve the multiplication of trial wave functions by the Hamiltonian matrix.

Another advantage of using local pseudopotentials in total-energy pseudopotential calculations is that the Hamiltonian "matrix" can be stored in $17N_{PW}$ words of memory by storing the potential-energy operator in real space in $16N_{PW}$ words and the kinetic-energy operator in reciprocal space in N_{PW} words. This allows extremely large "matrices" to be stored with ease.

Unfortunately, it is not possible to describe all atoms with local pseudopotentials. It is important to use an efficient scheme for applying nonlocal pseudopotentials if the computational speed of the molecular-dynamics method is to be retained. Efficient schemes have been developed for applying nonlocal pseudopotentials in the molecular-dynamics method and other iterative methods. These schemes will be described in Sec. IX, but it should be noted that the most efficient of these schemes requires $16N_{PW}N_B$ operations to apply the operators for each component of the nonlocal pseudopotential. This is less than the cost of Fourier-transforming the wave functions and is less than the cost of orthogonalizing the wave functions in any but the smallest of systems, and so these operations dominate the computational cost irrespective of whether local or nonlocal pseudopotentials are used.

IV. IMPROVEMENTS IN ALGORITHMS

In this section a number of relatively simple modifications to the molecular-dynamics-based method are introduced which offer significant improvements over the original approach when calculating the electronic ground state for a fixed ionic configuration. These improvements include methods that increase the computational speed of the calculations and methods that permit the electrons to converge to the exact Kohn-Sham eigenstates. As discussed in Sec. III, the possibility of obtaining the latter is important, for it paves the way for studies of *metallic* systems. However, it should be noted that these modifications are not generally useful when performing dynamical simulations of the ionic system, for reasons that will be discussed in Sec. VII. The remaining problems in the calculation of the electronic ground state for a static ionic configuration, problems that cannot be overcome with the improvements described below, are addressed in Sec. IV.D.

A. Improved integration

It has already been pointed out in Sec. III.D that an attempt to improve on the Verlet algorithm by using a more sophisticated finite-difference technique for integrating the equations of motion may not produce any increase in computational speed. This is because a more sophisticated finite-difference equation typically requires information from a large number of time steps in order to integrate the equations of motion. A large amount of memory is generally needed to store this information. Therefore, if sufficient core memory is not available, the computation can involve an excessive number of input and output operations.

An alternative technique for improving the integration of the equations of motion, which does not require any additional storage, is to calculate the change in the magnitude of the acceleration of the coefficient during the time step, *as the values of the coefficients change*. [Recall that simple use of the Verlet algorithm (3.17) requires the coefficients to remain fixed during a time step.] At first thought, it might appear that the determination of the time dependence of the coefficients during a time step actually necessitates a solution of the equations of motion in the first place! This, however, turns out not to be the case. An examination and manipulation of the form of the equations of motion (3.26) and (3.27) reveals that it is indeed possible to obtain a good approximation to the time dependence of the coefficients during a time step. The argument is as follows (Payne *et al.*, 1986).

With the assumption of a local pseudopotential for simplicity, the equation of motion (3.27) can be rewritten as

$$\mu\ddot{c}_{i,\mathbf{k}+\mathbf{G}} = -\left[\frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2 + V_T(\mathbf{G}=\mathbf{0}) - \lambda_i\right] c_{i,\mathbf{k}+\mathbf{G}} - \sum_{\mathbf{G}'\neq\mathbf{G}} V_T(\mathbf{G}-\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'} \; . \tag{4.1}$$

If one now introduces the definitions

$$\omega^2_{i,\mathbf{k}+\mathbf{G}} = \left[\frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2 + V_T(\mathbf{G}=\mathbf{0}) - \lambda_i\right] \Big/ \mu \tag{4.2a}$$

and

$$B_{i,\mathbf{k}+\mathbf{G}} = \left[\sum_{\mathbf{G}'\neq\mathbf{G}} V_T(\mathbf{G}-\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'}\right] \Big/ \mu \; , \tag{4.2b}$$

equation (4.1) becomes

$$\ddot{c}_{i,\mathbf{k}+\mathbf{G}} = -\omega^2_{i,\mathbf{k}+\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}} - B_{i,\mathbf{k}+\mathbf{G}} \; . \tag{4.3}$$

This shows that the equation of motion for the coefficient of each plane-wave basis state is essentially an *oscillator* equation. This permits an immediate determination of the largest acceptable time step for use with the Verlet algorithm. For plane-wave basis states at large reciprocal-lattice vectors, the oscillation frequency of the coefficient increases roughly linearly with the magnitude of the wave vector of the plane-wave basis state or as the square root of its kinetic energy. To integrate the oscillator equations *stably,* using the Verlet algorithm, one must restrict the length of the time step so that $(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t) < 1$ for all of the plane-wave basis states. This means that the length of the time step is restricted by the plane-wave basis states that have the highest kinetic energies, which is consistent with the discussion in Sec. III.D. The highest kinetic-energy basis states, however, provide the least important contributions to the wave functions, since the coefficients of these basis states are small if the cutoff energy for the basis set is large enough to converge the total energy. It is unsatisfactory that the least physically significant basis states restrict the length of the time step used to integrate the equations of motion. One method of overcoming this restriction is to perform an analytic integration of the equation of motion. A direct comparison between the Verlet algorithm and the analytic integration scheme is presented in Sec. IV.C below.

1. Analytic integration of second-order equations of motion

It is extremely difficult to integrate the equation of motion for the coefficient $c_{i,\mathbf{k}+\mathbf{G}}$ [Eq. (4.3)] because the term $B_{i,\mathbf{k}+\mathbf{G}}$ depends on the values of all the other coefficients $c_{i,\mathbf{k}+\mathbf{G}'}$. However, if the variation of $B_{i,\mathbf{k}+\mathbf{G}}$ is ignored, then the equation of motion (4.3) can be easily integrated analytically over a time step. The particular integral is

$$c_{i,\mathbf{k}+\mathbf{G}} = -B_{i,\mathbf{k}+\mathbf{G}}/\omega^2_{i,\mathbf{k}+\mathbf{G}} \; , \tag{4.4}$$

and the complementary function is

$$c_{i,\mathbf{k}+\mathbf{G}}(t) = A_1\exp(i\omega_{i,\mathbf{k}+\mathbf{G}}t) + A_2\exp(-i\omega_{i,\mathbf{k}+\mathbf{G}}t) \; . \tag{4.5}$$

The coefficients A_1 and A_2 are determined by the values of the coefficient at the present and previous time steps to give

$$c_{i,\mathbf{k}+\mathbf{G}}(\Delta t) = 2\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)c_{i,\mathbf{k}+\mathbf{G}}(0) - c_{i,\mathbf{k}+\mathbf{G}}(-\Delta t) - 2[1-\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)]B_{i,\mathbf{k}+\mathbf{G}}/\omega^2_{i,\mathbf{k}+\mathbf{G}} \; , \tag{4.6}$$

where $c_{i,\mathbf{k}+\mathbf{G}}(0)$ is the value of the coefficient at the present time step and $c_{i,\mathbf{k}+\mathbf{G}}(-\Delta t)$ is the value of the coefficient at the previous time step.

In Sec. III.F, the importance of damping the electronic equations of motion was discussed. One of various possible choices is to apply the damping directly to the equation of motion. If a term of the form $\gamma\dot{c}_{i,\mathbf{k}+\mathbf{G}}$ is added to Eq. (4.3), the coefficient at the next time step is now given by

$$\begin{aligned} c_{i,\mathbf{k}+\mathbf{G}}(\Delta t) = {} & 2\exp(-\gamma_{i,\mathbf{k}+\mathbf{G}}\Delta t)\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)c_{i,\mathbf{k}+\mathbf{G}}(0) - \exp(-2\gamma_{i,\mathbf{k}+\mathbf{G}}\Delta t)c_{i,\mathbf{k}+\mathbf{G}}(-\Delta t) \\ & -[1+\exp(-2\gamma_{i,\mathbf{k}+\mathbf{G}}\Delta t) - 2\exp(-\gamma_{i,\mathbf{k}+\mathbf{G}}\Delta t)\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)]B_{i,\mathbf{k}+\mathbf{G}}/\omega^2_{i,\mathbf{k}+\mathbf{G}} \end{aligned} \tag{4.7}$$

where

$$\gamma_{i,\mathbf{k}+\mathbf{G}} = D|\omega_{i,\mathbf{k}+\mathbf{G}}| \tag{4.8}$$

and D is the damping factor applied to the motion of the coefficients.

If the expectation value of the energy of the state $\lambda_{i,\mathbf{k}+\mathbf{G}}$ is larger than $|\hbar^2|\mathbf{k}+\mathbf{G}|^2/2m + V(\mathbf{G}=\mathbf{0})]$, the value of $\omega_{i,\mathbf{k}+\mathbf{G}}$ in Eq. (4.6) becomes imaginary. This can occur for the basis states at small reciprocal-lattice vectors in the higher-energy electronic states. Nevertheless, the analytic expression for the value of the coefficient of the plane-wave basis state at time Δt is still valid, provided the argument of the cosine function is taken as imaginary.

It might appear to be unduly expensive to use Eq. (4.6) to compute the value of the coefficient of the plane-wave basis state because several function calls are required to evaluate the expression. A square root is needed to obtain $\omega_{i,\mathbf{k}+\mathbf{G}}$, and then a cosine or hyperbolic cosine must be evaluated to obtain $\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)$. These function calls are required for every coefficient of the basis states for each electronic state. However, the heavy computational cost of the function calls can be avoided by an interpolation of the value of $\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)$ from a data array that contains $\cos(\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t)$ as a function of $\omega_{i,\mathbf{k}+\mathbf{G}}$. If the damping is applied directly in the equation of motion, the function $\exp(-\lambda_{i,\mathbf{k}+\mathbf{G}}\Delta t)$ can be interpolated in a similar manner.

2. Analytic integration of first-order equations of motion

Some authors (Williams and Soler, 1987) have suggested a change to a first-order equation of motion for the electronic degrees of freedom. Thus Eq. (4.3) becomes

$$\dot{c}_{i,\mathbf{k}+\mathbf{G}} = -\omega^2_{i,\mathbf{k}+\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}} - B_{i,\mathbf{k}+\mathbf{G}} , \qquad (4.9)$$

with $\omega^2_{i,\mathbf{k}+\mathbf{G}}$ and $B_{i,\mathbf{k}+\mathbf{G}}$ defined as in Eq. (4.2). This would have adverse consequences for following ion motion, as described later in Sec. VIII.C, but is completely appropriate for fixed-ion positions. Again assuming that $B_{i,\mathbf{k}+\mathbf{G}}$ does not vary during the time step, Eq. (4.9) can be analytically integrated to give

$$c_{i,\mathbf{k}+\mathbf{G}}(\Delta t) = -\frac{B_{i,\mathbf{k}+\mathbf{G}}}{\omega^2_{i,\mathbf{k}+\mathbf{G}}} + \left[c_{i,\mathbf{k}+\mathbf{G}}(0) + \frac{B_{i,\mathbf{k}+\mathbf{G}}}{\omega^2_{i,\mathbf{k}+\mathbf{G}}} \right] \exp(-\omega^2_{i,\mathbf{k}+\mathbf{G}}\Delta t) . \qquad (4.10)$$

This first-order equation gives roughly the same asymptotic convergence rate as the corresponding second-order equation of (4.6). But even given the same convergence rate, the first-order method is more efficient, since it requires only half the storage or half the input/output operations for the large number of wave-function degrees of freedom.

B. Orthogonalization of wave functions

When analytic expressions for the integration of the equations of motion (4.6) and (4.10) are used to calculate the coefficients, the length of the time step Δt is no longer restricted by the requirement $\omega_{i,\mathbf{k}+\mathbf{G}}\Delta t \ll 1$ for all of the plane-wave basis states, and a longer time step can be used. Car and Parrinello's iterative scheme for orthogonalization of the wave functions, described in Sec. III.E, becomes increasingly expensive to apply as the time steps increase in length. This is because the wave functions become more nonorthogonal during the time step, and more iterations of the algorithm are required to orthogonalize the wave functions. Since each iteration of the orthogonalization scheme requires $N_B^2 N_{PW}$ operations, it is sensible to adopt an alternative orthogonalization scheme that makes full use of the increase in computational speed obtained by integrating the equations of motion using a longer time step.

1. The Gram-Schmidt scheme

The simplest and most computationally efficient orthogonalization technique for long time steps is the Gram-Schmidt scheme (see Lin and Segel, 1974). In this approach, a set of orthonormal wave functions $\{\psi_i'\}$ is easily obtained from a set of linearly independent wave functions $\{\psi_i\}$ by use of the following algorithm:

$$\psi_i'' = \psi_i - \sum_{j<i} \langle \psi_j' | \psi_i \rangle \psi_j' \qquad (4.11a)$$

with

$$\psi_i' = \psi_i'' / |\psi_i''| . \qquad (4.11b)$$

The Gram-Schmidt scheme also has the advantage of breaking spurious symmetries that may occur in the choice of initial conditions for the electrons. These symmetries can propagate through the equations of motion and keep the electrons from reaching their true ground state. This is a rather subtle and technical point, addressed later in Sec. VI.A.

The most significant difference, however, between the Gram-Schmidt scheme and the Car and Parrinello orthogonalization method (3.23) lies in the convergence of electrons to Kohn-Sham eigenstates.

2. Convergence to Kohn-Sham eigenstates

As discussed earlier in Sec. III.H., the Kohn-Sham energy functional is minimized by any set of wave functions that are linear combinations of the lowest-energy Kohn-Sham eigenstates. The orthogonalization method of Car and Parrinello (3.23) generates orthogonal wave functions by a procedure that tends to intermix the wave functions, thereby preventing the Kohn-Sham eigenstates from being singled out. As a consequence, the final wave functions are, in general, linear combinations of the Kohn-Sham eigenstates.

In contrast, the Gram-Schmidt procedure orthogonalizes wave functions in a definite order. All of the higher-energy wave functions are forced to be orthogonal to the lowest-energy wave function, and so on. This in turn *forces* each state to converge to its lowest possible energy under the constraint that it be orthogonal to all states below it. The set of lowest possible single-particle levels under these constraints comprises the Kohn-Sham eigenstates.

The ability to converge to Kohn-Sham eigenstates is very important for metallic systems. After each time step it is necessary to know the correct ordering of the energy levels in order to fill states properly up to the Fermi level. Without Kohn-Sham eigenstates, the Fermi level of a metallic system cannot be defined.

C. Comparison between algorithms

The simple modifications to the molecular-dynamics-based method described in Secs. III.A and III.B above can produce significant increases in computational speed. This is illustrated in Fig. 13, where the evolution of the total energy of an 8-atom cubic supercell of germanium in the diamond structure is shown. Here the ions are held fixed and the electrons are being iterated to convergence. A local pseudopotential of the Starkloff-Joannopoulos type is used with a basis set of 4096 plane waves. The open circles show the results obtained by using the Verlet algorithm to integrate the equations of motion, while the filled circles show the results obtained by using analytic integration of the equation of motion. The use of the analytic expression to integrate the equation of motion allows the time step to be increased to six times the value at which the Verlet algorithm becomes unstable and gives convergence of the total energy in one-tenth of the number of time steps.

D. Remaining difficulties

The molecular-dynamics method and the algorithm improvements described above all become ineffective as the size of the system increases. For example, in silicon, only systems with supercell length scales less than about 50 Å benefit from these modifications to the molecular-dynamics method. For larger length scales the maximum stable time step is completely dominated by the need to suppress fluctuations in the charge density, or *charge sloshing*. As discussed earlier in Sec. III.J, this is a consequence of instabilities in the Kohn-Sham energy Hamiltonian which arise for very small reciprocal-lattice vectors and which require intractably small time steps to overcome. These instabilities present severe obstacles to future studies of large and more complex systems.

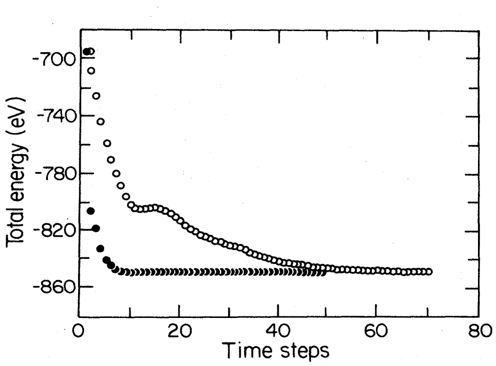

FIG. 13. Evolution of the total energy for an eight-atom cell of germanium in the diamond structure. Open circles represent the original scheme. The filled circles correspond to an analytic integration of the equations of motion as described in the text.

Clearly, some different and novel methods are required to surmount the problems encountered in the regime of large systems. One method that overcomes these difficulties involves a *direct* minimization of the Kohn-Sham total energy with a conjugate-gradients approach. This is the subject of the next section.

V. DIRECT MINIMIZATION OF THE KOHN-SHAM ENERGY FUNCTIONAL

At the end of Sec. IV, it was stated that all the algorithms described so far in this article encounter difficulties when performing calculations on large systems. The difficulties encountered can be attributed to the *discontinuous* changes in the Kohn-Sham Hamiltonian from iteration to iteration. There are an infinite number of Kohn-Sham Hamiltonians, each of which has a different set of eigenstates. One of these sets of eigenstates, the set generated by the self-consistent Kohn-Sham Hamiltonian, minimizes the Kohn-Sham energy functional. All the methods for performing total-energy pseudopotential calculations described so far involve an *indirect* search for the self-consistent Kohn-Sham Hamiltonian. The search procedure used in the molecular-dynamics method was illustrated in Fig. 10. The discontinuous evolution of the Hamiltonian can be clearly seen in this figure. In Sec. III.J it was shown that use of too long a time step in the molecular-dynamics method can lead to an unstable evolution of the Kohn-Sham Hamiltonian. This results in wave functions that move further from the self-consistent Kohn-Sham eigenstates at each time step, as illustrated in Fig. 12. Unfortunately, the value of the critical time step at which the instability occurs decreases as the size of the system increases. It is always possible to ensure stable evolution of the electronic configuration using the molecular-dynamics method, but this is at the cost of using smaller time steps and hence more computational time as the size of the system increases. This problem is also present in all of the algorithms described in Sec. III, since all of these employ an *indirect* search for the self-consistent Kohn-Sham Hamiltonian.

To perform a total-energy pseudopotential calculation it is necessary to find the electronic states that minimize the Kohn-Sham energy functional. As discussed above, to perform this process *indirectly*, by searching for the self-consistent Kohn-Sham Hamiltonian, can lead to instabilities. These instabilities would not be encountered if the Kohn-Sham energy functional were minimized *directly* because the Kohn-Sham energy functional normally has a single well-defined energy minimum. A search for this energy minimum cannot lead to instabilities in the evolution of the electronic configuration. In this section, a computational method is introduced that allows direct minimization of the Kohn-Sham energy functional in a tractable and efficient manner. In Sec. V.A, an introductory discussion is provided of two general methods that can be used for the minimization of any function. Of these general methods, the conjugate-gradients method is shown to be particularly promising. The modifications and extensions to the conjugate-gradients method that are needed in order to perform total-energy pseudopotential calculations tractably and stably for large supercells and large plane-wave kinetic-energy cutoffs are described in Sec. V.B.

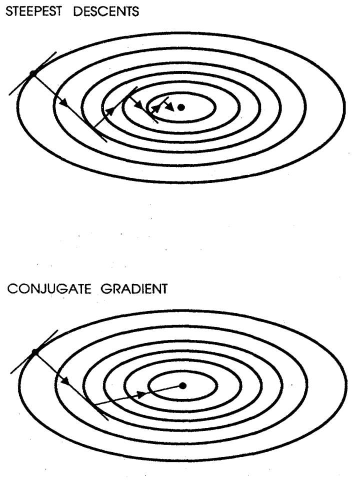

FIG. 14. Schematic illustration of two methods of convergence to the center of an anisotropic harmonic potential. Top: steepest-descents method requires many steps to converge. Bottom: Conjugate-gradients method allows convergence in two steps.

A. Minimization of a function

In this section two general methods are described that can be used to locate the minimum of function $F(\mathbf{x})$, where $\mathbf{x}$ is a vector in the multidimensional space [a detailed description of the techniques outlined in this section can be found in the book by Gill, Murray, and Wright (1981, p. 144)]. It will be assumed that the function $F(\mathbf{x})$ has a single minimum. If the function had several minima the methods described here would locate the position of the minimum "closest" to the initial sampling point (strictly speaking, it would locate the minimum in whose basin of attraction the initial sampling point lies).

1. The method of steepest descents

In the absence of any information about the function $F(\mathbf{x})$, the optimum direction to move from the point $\mathbf{x}^1$ to minimize the function is just the steepest-descent direction $\mathbf{g}^1$ given by

$$\mathbf{g}^1 = -\frac{\partial F}{\partial \mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}^1} . \tag{5.1}$$

It will be assumed that the direction of steepest descent at the point $\mathbf{x}^1$ can be obtained from the negative of a gradient operator G acting on the vector $\mathbf{x}^1$ so that

$$\mathbf{g}^1 = -G\mathbf{x}^1 . \tag{5.2}$$

To reduce the value of the function $F(\mathbf{x})$ one should move from the point $\mathbf{x}^1$ in the steepest-descent direction $\mathbf{g}^1$ to the point $\mathbf{x}^1 + b^1\mathbf{g}^1$, where the function is a minimum. This can be done by sampling the function $F(\mathbf{x})$ at a number of points along the line $\mathbf{x}^1 + b\mathbf{g}^1$ in order to determine the value of b at which $F(\mathbf{x}^1 + b\mathbf{g}^1)$ is a minimum. Alternatively, if the gradient operator G is accessible, the minimum value of the function along the line $\mathbf{x}^1 + b\mathbf{g}^1$ can be found by locating the point where the gradient of the function is orthogonal to the search direction, so that $\mathbf{g}^1 \cdot G(\mathbf{x}^1 + b^1\mathbf{g}^1) = 0$. It should be noted that this process minimizes only the value of the function along a particular line in the multidimensional space. To find the absolute minimum of the function $F(\mathbf{x})$, one must perform a series of such line minimizations. Thus the vector $\mathbf{x}^1 + b^1\mathbf{g}^1$ is used as the starting vector for the next iteration of the process. This next point is conventionally labeled $\mathbf{x}^2$. (The superscripts label the iterations of the minimization process.) The steps described above can be repeated to generate a series of vectors $\mathbf{x}^m$ such that the value of the function $F(\mathbf{x})$ decreases at each iteration. Hence $F(\mathbf{x}^l) < F(\mathbf{x}^k)$ for $l > k$. Each iteration reduces the value of the function $F(\mathbf{x})$ and moves the trial vector $\mathbf{x}^m$ towards the vector that minimizes the function. This process is illustrated schematically in the top panel of Fig. 14.

Although each iteration of the steepest-descents algorithm moves the trial vector towards the minimum of the function, there is no guarantee that the minimum will be reached in a finite number of iterations. In many cases a very large number of steepest-descents iterations is needed to get close to the minimum of the function. The method of steepest descents performs particularly poorly when the minimum of the function $F(\mathbf{x})$ lies in a long narrow valley such as the one illustrated in Fig. 14. The reason for the poor performance in this case is that each steepest-descent vector is orthogonal to the steepest-descent vector of the previous iteration. If the initial steepest-descent vector does not lie at right angles to the axis of the valley, successive vectors will point across rather than along the valley, so that a large number of iterations will be needed to move along the valley to the minimum of the function. This problem is overcome by using the conjugate-gradients technique.

2. The conjugate-gradients technique

It might seem surprising that there can be a better method of minimizing a function than to move in the direction in which the function decreases most rapidly. The rate of convergence of the steepest-descents method is limited by the fact that, after a minimization is performed along a given gradient direction, a subsequent minimization along the new gradient reintroduces errors proportional to the previous gradient. If the only information one has about the function $F(x)$ is its value and gradient at a set of points, the optimal method would allow one to combine this information, so that each minimization step is independent of the previous ones. To accomplish this, one must first derive the condition that makes one minimization step independent of another.

For simplicity, consider a symmetric and positive-definite function of the form

$$F(x)=\tfrac{1}{2}\mathbf{x}\cdot G\cdot\mathbf{x}\ , \tag{5.3}$$

where G is the gradient operator defined in Eq. (5.2). Consider now the minimization of $F(\mathbf{x})$ along some direction $\mathbf{d}^1$ from some point $\mathbf{x}^1$. The minimum will occur at $\mathbf{x}^2=\mathbf{x}^1+b^1\mathbf{d}^1$ where b^1 satisfies

$$(\mathbf{x}^1+b^1\mathbf{d}^1)\cdot G\cdot\mathbf{d}^1=0\ . \tag{5.4a}$$

This is obtained by differentiation of Eq. (5.3) with respect to b^1 at $\mathbf{x}^2$. A subsequent minimization along some direction $\mathbf{d}^2$ will then yield $\mathbf{x}^3=\mathbf{x}^2+b^2\mathbf{d}^2$, where b^2 satisfies

$$(\mathbf{x}^1+b^1\mathbf{d}^1+b^2\mathbf{d}^2)\cdot G\cdot\mathbf{d}^2=0\ . \tag{5.4b}$$

However, the best choice of b^1 and b^2, for the minimization of $F(x)$ along $\mathbf{d}^1$ and $\mathbf{d}^2$, is obtained from the differentiation of Eq. (5.3) with respect to both b^1 and b^2 at $\mathbf{x}^3$. This gives

$$(\mathbf{x}^1+b^1\mathbf{d}^1+b^2\mathbf{d}^2)\cdot G\cdot\mathbf{d}^1=0 \tag{5.5a}$$

and

$$(\mathbf{x}^1+b^1\mathbf{d}^1+b^2\mathbf{d}^2)\cdot G\cdot\mathbf{d}^2=0\ . \tag{5.5b}$$

It is clear that in order for Eqs. (5.4) and (5.5) to be consistent, and consequently for the minimization along $\mathbf{d}^1$ and $\mathbf{d}^2$ to be independent, one must require that

$$\mathbf{d}^1\cdot G\cdot\mathbf{d}^2=\mathbf{d}^2\cdot G\cdot\mathbf{d}^1=0\ . \tag{5.6}$$

This is the condition that the directions $\mathbf{d}^1$ and $\mathbf{d}^2$ be *conjugate* to each other (see Gill *et al.*, 1981) and can be immediately generalized to

$$\mathbf{d}^n\cdot G\cdot\mathbf{d}^m=0 \quad \text{for } n\neq m\ . \tag{5.7}$$

The conjugate-gradients technique provides a simple and effective procedure for implementation of such a minimization approach. The initial direction is taken to be the negative of the gradient at the starting point. A subsequent conjugate direction is then constructed from a linear combination of the new gradient and the previous direction that minimized $F(x)$. In a two-dimensional problem, it is clear that one would need only two conjugate directions, and this would be sufficient to span the space and arrive at the minimum in just two steps, as shown at the bottom of Fig. 14. It is less clear, however, that the current gradient and the previous direction vector would maintain all of the information necessary to include minimization over *all* previous directions in a multidimensional space. The proof that directions generated in this manner are indeed conjugate is the important result of the conjugate-gradients derivation. The precise search directions $\mathbf{d}^i$ generated by the conjugate-gradients method are obtained from the following algorithm:

$$\mathbf{d}^m=\mathbf{g}^m+\gamma^m\mathbf{d}^{m-1}\ , \tag{5.8}$$

where

$$\gamma^m=\frac{\mathbf{g}^m\cdot\mathbf{g}^m}{\mathbf{g}^{m-1}\cdot\mathbf{g}^{m-1}}\ , \tag{5.9}$$

and $\gamma^1=0$.

Since minimizations along the conjugate directions are independent, the dimensionality of the vector space explored in the conjugate-gradients technique is reduced by 1 at each iteration. When the dimensionality of the function space has been reduced to zero, there are no directions left in which to minimize the function, so the trial vector must be at the position of the minimum. Therefore the exact location of the minimum of a quadratic function will be found, conservatively speaking, in a number of iterations that is equal to the dimensionality of the vector space. In practice, however, it is usually possible to perform the calculations so that far fewer iterations are required to locate the minimum.

Another way of thinking about the difference between the conjugate-gradients technique and the method of steepest descents is that, in the method of steepest descents, each direction is chosen only from information about the function at the present sampling point. In contrast, in the conjugate-gradients technique the search direction is generated using information about the function obtained from all the sampling points along the conjugate-gradients path.

The conjugate-gradients technique provides an efficient method for locating the minimum of a general function. This suggests that it should be a good technique for locating the minimum of the Kohn-Sham energy functional. It is important, however, to implement the conjugate-gradients technique in such a way as to maximize computational speed, so that each iteration of the method is not significantly more expensive than alternative techniques, and to minimize the memory requirement so that calculations are not limited by the available memory. A conjugate-gradients method that fulfills these criteria has been developed by Teter *et al.* (1989). This method is described in Sec. V.B below. Stich *et al.* (1989) have used a conjugate-gradients method with the *indirect* approach in order to search for the eigenstates of the Kohn-Sham *Hamiltonian.* While this is an efficient method for obtaining the eigenstates of the Hamiltonian, it is expected, for reasons discussed earlier, that this method will encounter difficulties when calculations are performed on very large systems. The authors point out that in these cases it will be necessary to use a direct minimization method. Gillan (1989) has developed a conjugate-gradients technique for directly minimizing the

Kohn-Sham energy functional. His method has many similarities to that described in the following section, but it does have a significantly larger memory requirement.

B. Application of the conjugate-gradients method

In this section, a computational technique is introduced that overcomes the instabilities described earlier, which are associated with large supercell sizes and large plane-wave kinetic-energy cutoffs. This technique uses the conjugate-gradients approach, with the proper preconditioning, to minimize *directly* the Kohn-Sham energy functional.

The abstract description of the conjugate-gradients technique presented above considered a function F of the vector $\mathbf{x}$ where the gradient of the function could be calculated using a gradient operator G. In the case of total-energy calculations, the Kohn-Sham energy functional E takes the place of the function F, the wave functions $\{\psi_i\}$ take the place of the vector $\mathbf{x}$, and the Kohn-Sham Hamiltonian H is the relevant gradient operator G.

1. The update of a single band

A conjugate-gradients iteration can be used to update all of the electronic wave functions simultaneously. The only drawback to this approach is that a large amount of data has to be stored between iterations to ensure the conjugacy of the search directions. To determine the conjugate direction requires, among other things, the previous conjugate direction and the present steepest-descent vector. If all of the wave functions are updated simultaneously, both of these constitute an array of the same size as the wave-function array, so that, in total, *three* arrays of the size of the wave-function array are needed to perform the conjugate-gradients calculation. The memory requirement in molecular-dynamics calculations is dominated by the storage of the wave-function arrays. Increasing the number of arrays that must be stored could limit the size of system for which calculations can be performed. Ideally all of the advantages of the conjugate-gradients technique should be retained without increasing the memory requirement. This can be achieved by updating a single band at a time.

The steepest-descent direction for a single band is given by

$$\zeta_i^m = -(H-\lambda_i^m)\psi_i^m \ , \tag{5.10}$$

where

$$\lambda_i^m = \langle \psi_i^m | H | \psi_i^m \rangle \ . \tag{5.11}$$

It should be noted that the superscript m labels the iteration number and the subscript i labels the band.

2. Constraints

A total-energy calculation differs from the conjugate-gradients minimization described previously in that the electronic wave functions are constrained to be orthogonal. The orthogonality constraints can be maintained by ensuring that the steepest-descent vector is orthogonal to all the other bands. The steepest-descent direction calculated as

$$\zeta_i'^m = \zeta_i^m - \sum_{j\neq i} \langle \psi_j | \zeta_i^m \rangle \psi_j \tag{5.12}$$

is the direction of steepest descent allowed by the orthogonality constraints. There is no iteration index on the wave functions ψ_j in Eq. (5.12) because these wave functions do not vary during iterations for band i.

If the search direction for band i were not orthogonal to the wave functions of all the other bands, all of the wave functions would have to change during each iteration in order to maintain the constraints of orthogonality. Since all the wave functions would change, the charge density from all of the bands would have to be computed at each iteration. By ensuring that the search direction is orthogonal to all the other bands, this technique requires computation only of the change in the charge density from the single band at each iteration.

3. Preconditioning

Successive steps along the conjugate-gradient directions will reduce the magnitude of the error in the wave function. From Eq. (5.10) it is clear that a measure of the error in the wave function ψ_i is contained in the steepest-descent vector ζ_i. Ideally, if the steepest-descent vector were a simple multiple of the error in the wave function, then moving the correct distance along the steepest-descent direction would entirely eliminate the error in the wave function. The actual relation between the error in the wave function, $\delta\psi_i$, and the steepest-descent direction ζ_i is most easily demonstrated by expansion of $\delta\psi_i$ in terms of the eigenstates of the Kohn-Sham Hamiltonian,

$$\delta\psi_i = \sum_\alpha c_{i,\alpha}\xi_\alpha \ . \tag{5.13}$$

The steepest-descent vector is obtained by substitution of Eq. (5.13) into (5.10), which gives

$$\zeta_i = -[H-\lambda_i]\sum_\alpha c_{i,\alpha}\xi_\alpha \tag{5.14}$$

$$= -\sum_\alpha (\varepsilon_\alpha - \lambda_i) c_{i,\alpha}\xi_\alpha \ , \tag{5.15}$$

where ε_α is the eigenvalue associated with the eigenstate ξ_α.

It can be seen that the steepest-descent vector ζ_i is only a multiple of the error vector $\delta\psi_i$ if all the unoccupied eigenstates of the Kohn-Sham Hamiltonian are degenerate; however, the Kohn-Sham Hamiltonian has a broad spectrum of eigenvalues, which extends up to the cutoff energy for the plane-wave basis set and leads to poor convergence in a conjugate-gradient calculation. Each step tends to remove components of the error vector that correspond to eigenstates in a particular energy range. The rate of convergence will be improved if some method is used to conjugate for the weighting factors $\varepsilon_\alpha - \lambda_i$ that distinguish the error vector and the steepest-descent vector. The technique of *preconditioning* can be used to achieve this approximately (see Gill *et al.*, 1981).

The technique of preconditioning involves multiplying the steepest-descent vector by a preconditioning matrix K to produce a preconditioned steepest-descent vector η that more accurately represents the error vector, as illustrated in Fig. 15. In principle, a preconditioning matrix exists that perfectly preconditions the steepest-descent vector so that the preconditioned vector is parallel to the error vector. However, this preconditioning matrix will be a full $N_{PW}xN_{PW}$ matrix, which would then require N_{PW}^2 operations to precondition the steepest-descent vector for each band, thus making the computational scheme prohibitively expensive. In practice, it is extremely expensive to construct an exact preconditioning matrix, and the cost of this operation would make the cost of the calculation even more unfavorable. It is always more efficient to use an approximation preconditioning matrix and a succession of conjugate-gradient minimizations rather than to attempt to compute and apply exact preconditioning.

The broad eigenvalue spectrum of the Kohn-Sham Hamiltonian in pseudopotential calculations that use plane-wave basis sets is associated with the wide range of energies of the basis states. The higher-energy eigenstates of the Hamiltonian are dominated by plane-wave basis states whose high kinetic energies lie close to the eigenvalue of the state. To make those states whose eigenvalues are dominated by their kinetic energy nearly degenerate, one must remove the effect of the kinetic-energy operator in the Hamiltonian. This is easily achieved by multiplication of a diagonal preconditioning matrix, which is essentially the inverse of the kinetic-energy operator. This argument breaks down for the lower-energy eigenstates because the potential and kinetic energies are similar, and so the potential is strong enough to mix a range of different energy plane-wave basis states into the eigenstates. Therefore the elements of the preconditioning matrix should become a constant for the plane-wave basis states of low energy rather than varying as the inverse of the kinetic energy.

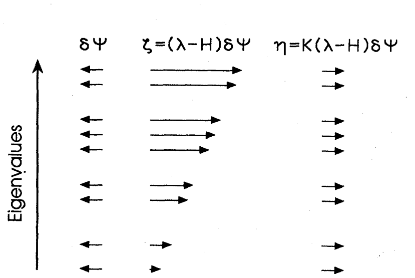

FIG. 15. Spectral representation of error in wave function $\delta\psi$, the gradient of the wave function ζ, and the preconditioned gradient η. Note that the error in the wave function can be eliminated completely, in a single step, only if the preconditioned gradient is added to the wave function ψ.

It has been found that *preconditioned* steepest-descent vectors that accurately represent the errors in the wave functions can be obtained by multiplication of the steepest-descent vectors by a preconditioning matrix K whose matrix elements are given by the following expression:

$$K_{\mathbf{G},\mathbf{G}'} = \delta_{\mathbf{G},\mathbf{G}'} \frac{27+18x+12x^2+8x^3}{27+18x+12x^2+8x^3+16x^4} , \quad (5.16)$$

where

$$x = \frac{(\hbar^2|\mathbf{k}+\mathbf{G}|^2)/2m}{T_i^m} ,$$

and $T_i^m = \langle \psi_i^m | (-\hbar^2/2m)\nabla^2 | \psi_i^m \rangle$ is the kinetic energy of the state ψ_i^m. The matrix elements $K_{\mathbf{G},\mathbf{G}'}$ have the following attractive properties. As x approaches zero, the $K_{\mathbf{G},\mathbf{G}'}$ approach unity, with zero first, second, and third derivatives. This guarantees that the small wave-vector components of the steepest-descent vector remain unchanged. Above $x=1$, the $K_{\mathbf{G},\mathbf{G}'}$ asymptotically approach $1/[2(x-1)]$ with an asymptotic expansion correct to fourth order in $1/x$.

This factor thus causes all of the large wave-vector components to converge at nearly the same rate. This preconditioning procedure should be compared to the analytic integration technique outlined in Sec. IV.A.1. In the molecular-dynamics method the diagonal dominance of the Hamiltonian due to the kinetic energy causes a rapid oscillation of the wave-function coefficients, and the analytic integration technique provides a method for allowing a time step that is not restricted by these oscillations. In the conjugate-gradients method, the same diagonal dominance of the Hamiltonian produces a steepest-descent vector that is biased towards plane-wave components with large wave vectors, and the preconditioning directly reduces these components. Although both problems have the same basic cause, their solutions are rather different in the two methods.

The preconditioned steepest-descent vector η_i^m is

$$\eta_i^m = K\zeta_i'^m . \quad (5.17)$$

The preconditioned steepest-descent vector is not orthogonal to all the bands. For the computational reasons given above, any change to a band vector must leave all the other bands unaffected. The preconditioned steepest-allowed-descent vector that is orthogonal to all the bands is calculated as

$$\eta_i'^m = \eta_i^m - \langle \psi_i^m | \eta_i^m \rangle \psi_i^m - \sum_{j \neq i} \langle \psi_j | \eta_i^m \rangle \psi_j \ . \tag{5.18}$$

4. Conjugate directions

The conjugate-gradient direction is constructed out of steepest-descent vectors as indicated in Eq. (5.8). With the inclusion of preconditioning as described above, the preconditioned conjugate directions φ_i^m are given by

$$\varphi_i^m = \eta_i'^m + \gamma_i^m \varphi_i^{m-1} \tag{5.19}$$

where

$$\gamma_i^m = \frac{\langle \eta_i'^m | \zeta_i'^m \rangle}{\langle \eta_i'^{m-1} | \zeta_i'^{m-1} \rangle} \tag{5.20}$$

and $\gamma_i^1 = 0$.

The conjugate direction generated by Eq. (5.19) will be orthogonal to all the other bands because it is constructed from preconditioned steepest-descent vectors that are orthogonal to all the other bands. However, the conjugate direction will not be orthogonal to the wave function of the present band. A further orthogonalization to the present band should be performed and a normalized conjugate direction $\varphi_i'^m$ calculated as

$$\varphi_i''^m = \varphi_i^m - \langle \psi_i^m | \varphi_i^m \rangle \psi_i^m \ , \tag{5.21}$$

$$\varphi_i'^m = \frac{\varphi_i''^m}{\langle \varphi_i''^m | \varphi_i''^m \rangle^{1/2}} \ . \tag{5.22}$$

5. Search for the energy minimum

The steps outlined above yield a preconditioned conjugate direction described by the normalized vector $\varphi_i'^m$, which is orthogonal to all the bands. The following combination of the present wave function ψ_i^m and the preconditioned conjugate vector $\varphi_i'^m$,

$$\psi_i^m \cos\theta + \varphi_i'^m \sin\theta \quad (\theta \text{ real}) \ , \tag{5.23}$$

is a normalized vector that is orthogonal to all the other bands ψ_j $(j \neq i)$. Therefore any vector described by Eq. (5.23) obeys the constraints of orthonormality required for the electronic wave functions.

The conjugate-gradients technique requires that the value of θ that minimizes the Kohn-Sham energy functional be found. The search for the position of minimum energy could be performed by the calculation of the Kohn-Sham energy for various values of θ until the minimum is located. If the approximate location of the minimum is not known, this would be a relatively expensive search technique. As an alternative method for locating the minimum, the Kohn-Sham energy could be written as a general function of θ,

$$E(\theta) = E_{\text{avg}} + \sum_{n=1}^{\infty} [A_n \cos(2n\theta) + B_n \sin(2n\theta)] \ . \tag{5.24}$$

One piece of information, such as a total energy or a gradient of the total energy, is required to evaluate each term in this expression. As the summation in Eq. (5.24) is infinite, any attempt to locate the minimum of the Kohn-Sham energy functional using this expression would be even more costly than searching for the minimum by calculation of the Kohn-Sham energy for various values of θ. However, it is found that the variation of the Kohn-Sham energy with θ is very accurately reproduced by just the $n=1$ term in the summation in Eq. (5.24). The accuracy of the fit can be seen in Fig. 16. Therefore the following expression for the Kohn-Sham energy is sufficient to locate the minimum of the Kohn-Sham energy functional:

$$E(\theta) = E_{\text{avg}} + A_1 \cos(2\theta) + B_1 \sin(2\theta) \ . \tag{5.25}$$

Three pieces of information are required to evaluate the three unknowns in this expression. The value of the total energy at $\theta = 0$, $E(0)$, is already known. The gradient of the energy with respect to θ at $\theta = 0$ is given by

$$\left. \frac{\partial E}{\partial \theta} \right|_{\theta=0} = \langle \varphi_i'^m | H | \psi_i^m \rangle + \langle \psi_i^m | H | \varphi_i'^m \rangle$$
$$= 2\,\mathrm{Re}(\langle \varphi_i'^m | H | \psi_i^m \rangle) \ . \tag{5.26}$$

Since $H|\psi_i^m\rangle$ has been computed to determine the steepest-descent vector η_i^m, the value of $(\partial E/\partial \theta)|\theta=0$ can be computed cheaply. Therefore just one further piece of information is required to determine all the parameters in Eq. (5.25). This could be the second derivative of the Kohn-Sham energy functional at $\theta = 0$, or the Kohn-Sham energy, or its derivative at any other value of θ. Calculation of the second derivative of the Kohn-Sham energy functional would require additional programming effort in most pseudopotential codes, whereas no additional programming effort is required to calculate the Kohn-Sham energy or its derivative at a second value of θ. Therefore we shall first describe the method that uses the Kohn-Sham energy at a second value of θ.

The second value of θ should be chosen so that the

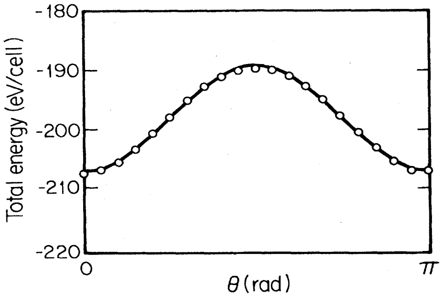

FIG. 16. The total energy of an 8-atom silicon supercell plotted as a function of the parameter θ, which determines the proportions of wave function and its gradient as described in the text. The dots represent the exact calculation and the line is the lowest harmonic contribution.

sampling point is far enough from $\theta=0$ to avoid rounding errors but not so far from the origin that the estimate of the curvature of the Kohn-Sham energy functional at $\theta=0$ becomes inaccurate. In the later stages of the calculation, shorter and shorter steps will be taken along the conjugate directions, and so it is important that the curvature of the Kohn-Sham energy functional at $\theta=0$ be accurately determined in order to locate the position of the minimum to high precision. It has been found that computing the Kohn-Sham energy at the point $\theta=\pi/300$ gives reliable results. If the value of the Kohn-Sham energy at the point $\theta=\pi/300$, $E(\pi/300)$, is computed, the three unknowns in Eq. (5.25) are calculated as

$$E_{\rm avg}=\frac{E\left[\frac{\pi}{300}\right]-\frac{1}{2}\frac{\partial E}{\partial\theta}\bigg|_{\theta=0}-E(0)\cos\left[\frac{2\pi}{300}\right]}{1-\cos\left[\frac{2\pi}{300}\right]}\ , \tag{5.27}$$

$$A_1=\frac{E(0)-E\left[\frac{\pi}{300}\right]+\frac{1}{2}\frac{\partial E}{\partial\theta}\bigg|_{\theta=0}}{1-\cos\left[\frac{2\pi}{300}\right]}\ , \tag{5.28}$$

$$B_1=\frac{1}{2}\frac{\partial E}{\partial\theta}\bigg|_{\theta=0}\ . \tag{5.29}$$

Once the parameters $E_{\rm avg}$, A_1, and B_1 have been determined, the value of θ that minimizes the Kohn-Sham energy function can be calculated. The stationary points of the function (5.25) occur at the points

$$\theta_S=\frac{1}{2}\tan^{-1}\frac{B_1}{A_1}\ . \tag{5.30}$$

The value of θ_S that lies in the range $0<\theta<\pi/2$ is the required value, $\theta_{\rm min}$.

The calculation of an analytic second derivative of the Kohn-Sham energy at $\theta=0$ provides an elegant way of determining the optimum step length, even though it does require some additional programming. The required expression for the second derivative is

$$\frac{\partial^2 E}{\partial^2\theta}\bigg|_{\theta=0}=2(\langle\varphi_i'^m|H|\varphi_i'^m\rangle-\langle\psi_i^m|H|\psi_i^m\rangle)+f[\text{Hartree term}+\text{exchange-correlation term}]\ , \tag{5.31}$$

where f is the product of the occupation number and the $\mathbf{k}$-point weight, the Hartree term is

$$\frac{1}{\Omega^2}\int_{\text{unit cell}}v(\mathbf{r})\cdot 2\,\mathrm{Re}[\varphi_i'^m(\mathbf{r})^*\psi_i^m(\mathbf{r})]d^3\mathbf{r}\ , \tag{5.32}$$

where

$$v(\mathbf{r})=\frac{e^2}{4\pi\varepsilon_0}\int\frac{2\,\mathrm{Re}[\varphi_i'^m(\mathbf{r}')^*\psi_i^m(\mathbf{r}')]}{|\mathbf{r}-\mathbf{r}'|}d^3\mathbf{r}'\ . \tag{5.33}$$

Alternatively, the Hartree term can be written

$$\frac{1}{\Omega}\sum_{\mathbf{G}\neq 0}\frac{e^2}{\varepsilon_0}\frac{|h(\mathbf{G})|^2}{G^2}\ , \tag{5.34}$$

where

$$h(\mathbf{G})=\frac{1}{\Omega}\int_{\text{unit cell}}\exp(i\mathbf{G}\cdot\mathbf{r})\times 2\,\mathrm{Re}[\varphi_i'^m(\mathbf{r})^*\psi_i^m(\mathbf{r})]d^3\mathbf{r}\ . \tag{5.35}$$

The exchange-correlation term is

$$\frac{1}{\Omega^2}\int_{\text{unit cell}}(2\,\mathrm{Re}[\varphi_i'^m(\mathbf{r})^*\psi_i^m(\mathbf{r})])^2\frac{\partial V_{XC}}{\partial n(\mathbf{r})}d^3\mathbf{r}\ . \tag{5.36}$$

The cost of computing the analytic second derivative is the same as the cost of calculating the Kohn-Sham energy at a trial value of θ.

The required value of θ, $\theta_{\rm min}$, is determined using

$$\theta_{\rm min}=\frac{1}{2}\tan^{-1}\left[-\frac{\frac{\partial E}{\partial\theta}\Big|_{\theta=0}}{\frac{1}{2}\frac{\partial^2 E}{\partial^2\theta}\Big|_{\theta=0}}\right]\ . \tag{5.37}$$

The wave function used to start the next iteration of the conjugate-gradients procedure, ψ_i^{m+1}, is

$$\psi_i^{m+1}=\psi_i^m\cos(\theta_{\rm min})+\varphi_i'^m\sin(\theta_{\rm min})\ . \tag{5.38}$$

The new wave function generates a different charge density from that generated by the previous wave function, and so the electronic potentials in the Kohn-Sham Hamiltonian must be updated before commencing the next iteration.

6. Calculational procedure

The flow diagram in Fig. 17 illustrates the steps involved in an update of the wave function of a single band using the conjugate-gradients method. Eventually the wave functions of all of the bands must be updated. There is no point in converging a single band exactly if large errors remain in the bands that have not yet been updated. Thus no more than three or four conjugate-gradients iterations should be performed on one band before moving to the next band. Once all of the bands have been updated, conjugate-gradients iterations are started again on the lowest band. Rather than perform a fixed number of conjugate-gradients iterations on each band, one can perform conjugate-gradients iterations on one band until the total energy changes by less than a particular value or by less than a given fraction of the change of energy in the first conjugate-gradients iteration. Then iterations are started on the next band. The convergence criterion should be changed as the system moves towards the minimum of the Kohn-Sham energy functional.

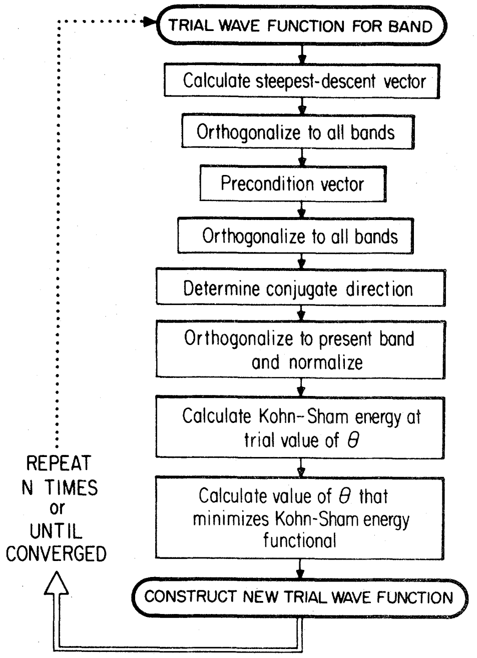
TRIAL WAVE FUNCTION FOR BAND
Calculate steepest-descent vector
Orthogonalize to all bands
Precondition vector
Orthogonalize to all bands
Determine conjugate direction
Orthogonalize to present band and normalize
Calculate Kohn-Sham energy at trial value of θ
Calculate value of θ that minimizes Kohn-Sham energy functional
CONSTRUCT NEW TRIAL WAVE FUNCTION
REPEAT N TIMES or UNTIL CONVERGED

FIG. 17. Flow diagram for the update of a single band in the (*direct-minimization*) conjugate-gradients method.

After each sweep through the bands, the change in the total energy at which conjugate-gradients iterations on a band are stopped should be reduced. The reasons for applying the convergence criteria outlined in this section and a discussion of the optimum choice of the parameters is given by Arias *et al.* (1991).

7. Computational cost

The computational cost of performing a conjugate-gradients iteration on a single band is higher than the cost of performing a molecular-dynamics time step on a single band. In a molecular-dynamics calculation the computational cost per iteration is dominated by the three Fourier transforms for each band and the orthogonalization. The number of operations required for each band at each time step is $3N_{\mathrm{FFT}}$ for the Fourier transforms and $N_B N_{PW}$ for the orthogonalization, where N_{PW} is the number of plane-wave basis states, $N_{\mathrm{FFT}} = 16 N_{PW} \ln N_{PW}$, and N_B is the number of occupied bands. The computational cost of a conjugate-gradients iteration is also dominated by the cost of performing Fourier transforms and the cost of orthogonalizing the wave functions. Only the steps in the conjugate-gradients iteration that involve Fourier transforms or orthogonalizations will be considered in this section. All of the other steps in the conjugate-gradients update of a single band require only N_{PW} operations and constitute a negligible part of the computational effort.

The calculation of the steepest-descent vector in the conjugate-gradients method is identical to the calculation of the accelerations of the wave functions in the molecular-dynamics method. Hence the computational cost is dominated by the cost of performing two Fourier transformations which require $2N_{\mathrm{FFT}}$ operations. If preconditioning is applied, one orthogonalization of the steepest-descent vector to all the bands and one orthogonalization of the preconditioned steepest-descent vector to all the bands must be performed. These two orthogonalizations require $2N_B N_{PW}$ operations. Only a single orthogonalization would be needed if preconditioning were not applied. To calculate the Kohn-Sham energy at the trial value of θ, one must transform the trial wave function to real space, so that the charge density can be computed, and then transform the charge density to reciprocal space so that the Hartree energy can be calculated. These two Fourier transforms require $2N_{\mathrm{FFT}}$ operations. After the wave function is updated, the new Kohn-Sham Hamiltonian must be calculated. The new charge density can be computed directly from the previous wave function and the trial wave function, both of which have already been transformed to real space, so that no extra Fourier transforms are required for this operation. However, the new charge density must be transformed to reciprocal space so that the new Hartree potential can be computed, and the new Hartree potential must be transformed back to real space. These two Fourier transforms require $2N_{\mathrm{FFT}}$ operations. Therefore the total number of operations required to perform a conjugate-gradients update on a single band is $6N_{\mathrm{FFT}}$ operations for the Fourier transforms and $2N_B N_{PW}$ operations for the orthogonalizations. Hence each conjugate-gradients iteration requires twice the number of operations required by a molecular-dynamics time step for a single band.

C. Speed of convergence

In this section the speed of convergence of methods that *directly* minimize the Kohn-Sham Hamiltonian is compared with the speed of convergence of the molecular-dynamics method. The systems used for these calculations were specifically chosen to highlight the problems associated with the use of conventional methods to perform total-energy pseudopotential calculations. However, it is important to appreciate that these systems are representative of systems for which molecular-dynamics calculations are currently being performed.

1. Large energy cutoff

Figure 18 shows the error in the total energy against iteration number for a calculation on an 8-atom unit cell of silicon with a cutoff energy for the plane-wave basis set of 32 rydberg. Although this cutoff energy is much higher than the value actually needed for calculations on silicon, it is typical of the cutoff energies required for the

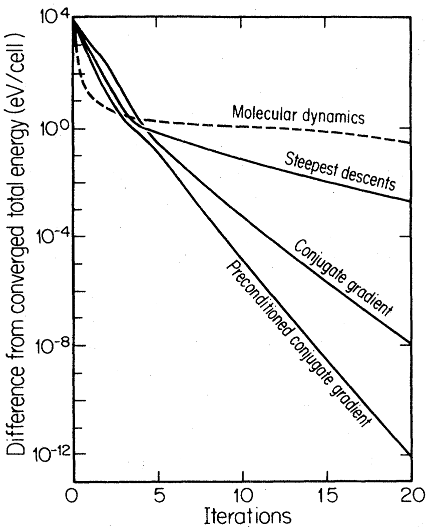

FIG. 18. Error in the total energy of an 8-atom silicon supercell with a 32-Ry kinetic-energy cutoff vs iteration number for *indirect-minimization* (dashed line) and *direct-minimization* (solid line) methods. Note that the curve labeled "molecular dynamics" involves a first-order equation of motion, and the number of iterations associated with this curve *has been divided by five* to allow comparison at the same level of computational effort as discussed in the text.

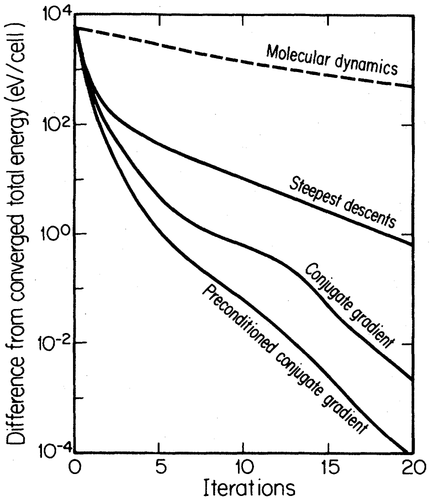

FIG. 19. Error in the total energy of a row of 12 silicon unit cells with an 8-Ry kinetic-energy cutoff vs iteration number for *indirect-minimization* (dashed line) and *direct-minimization* (solid line) methods. Same scaling as in Fig. 18.

first-row elements and for transition metals. The dashed curve represents results obtained with the molecular-dynamics method and a first-order equation of motion to evolve the electronic wave functions. This also includes the algorithmic improvements discussed in Sec. IV. All of the other curves show results obtained with methods that *directly* minimize the Kohn-Sham energy functional, methods for which the iteration number in Fig. 18 labels sweeps through all the bands. Each sweep through the bands involves a number of conjugate-gradients steps for each band. In contrast, each iteration has been taken to represent five time steps in the molecular-dynamics method. This number has been chosen so that the computational time required for each "iteration" is similar for all the schemes. The same scaling is used in the other examples presented in this section.

Figure 18 shows that all the schemes that *directly* locate the minimum of the Kohn-Sham energy functional converge in a smaller amount of computational time than the molecular-dynamics method. The improved performance of the conjugate-gradients method over the method of steepest descents is clearly demonstrated. It can be seen that the preconditioning of the conjugate gradients significantly increases the speed of convergence for this system. This is expected because the cutoff energy for the plane-wave basis set is very high, so the spectrum of eigenvalues of the Kohn-Sham Hamiltonian is particularly broad, and preconditioning is particularly beneficial.

2. Long supercells

The difficulties associated with instability due to the discontinuous evolution of the Kohn-Sham Hamiltonian are most apparent for large systems, when one or more of the unit cell vectors becomes very long. The performance of different computational methods for such a system has been tested by performing calculations on a long unit cell containing 24 silicon atoms. The results are shown in Fig. 19. The time step used in the molecular-dynamics method had to be drastically reduced to maintain stability in the calculation. The consequences of using a very small time step are clearly revealed by the slow rate of convergence shown in the figure. In contrast, all of the methods that *directly* minimize the Kohn-Sham energy functional perform extremely well. These methods do not suffer from the instabilities associated with an *indirect* minimization of the Kohn-Sham energy functional. Comparing the relative speeds of convergence of the directly minimization methods shows that the speed of convergence improves on changing from steepest-descent to conjugate-gradients methods and then to the preconditioned conjugate-gradients method, as expected.

3. A real system

To demonstrate that the results shown in Figs. 18 and 19 are representative of calculations on real systems, Fig. 20 shows the results of calculations on a twelve-atom unit cell of silicon dioxide in the α-critobalite structure. For test purposes a cutoff energy of 32 rydberg for the plane-

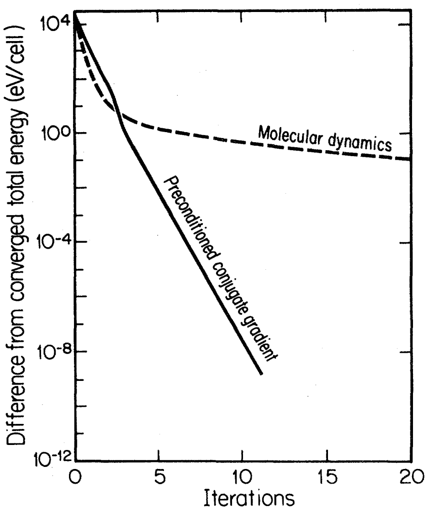

FIG. 20. Error in the total energy in eV of a 12-atom cell of the α-cristobalite form of SiO_2 with a 32-Ry kinetic-energy cutoff vs iteration number for *indirect-minimization* (dashed line) and *direct-minimization* (solid line) methods. Same scaling as in Fig. 18.

wave basis set was used for this calculation. The figure compares results obtained by use of the molecular-dynamics method and the preconditioned conjugate-gradients method, with the letter giving an improved speed of convergence, as it did in the previous examples.

VI. CHOICE OF INITIAL WAVE FUNCTIONS FOR THE ELECTRONIC STATES

A. Convergence to the ground state

The most important consideration when choosing the initial electronic states for any of the iterative schemes described in the previous sections is to choose a set of wave functions that allows the electronic configuration to converge to its ground state. Two factors that can prevent a set of initial states from converging to the ground state are described in this section.

1. Spanning the ground state

The most obvious reason why an electronic configuration might not converge to the ground state is that the initial states do not span the ground state. In this case the electronic wave functions relax to a self-consistent set of Kohn-Sham eigenstates but not to the set that forms the ground state.

The simplest choice of initial wave functions for the electronic states is a single plane wave for each state, and the most obvious choice of initial states is the set of plane waves with the lowest kinetic energies. However, there is considerable danger that these initial states may not span the ground state of the electronic configuration. For example, when the lowest-energy plane waves are used as initial states for a calculation of the electronic structure of an 8-atom cubic cell of any of the tetrahedral semiconductors, the electronic configuration does not converge to the ground state. This is shown schematically in Fig. 21. Germanium, silicon, and carbon each have four valence electrons. An 8-atom unit cell of any of these materials contains 32 electrons, so 16 doubly occupied electronic states are required to accommodate the electrons.

Consider a calculation for the **k** point (0,0,0). The 16 lowest-energy plane-wave basis states at the (0,0,0) **k** point are the plane wave with wave vector (0,0,0), the six plane waves with wave vectors in the $\{1,0,0\}$ star of wave vectors, and any nine of the twelve plane waves with wave vectors in the $\{1,1,0\}$ star. At least six of the nine plane waves chosen from the $\{1,1,0\}$ wave-vector star can be paired, so that the wave vectors of the plane waves in each pair are separated by reciprocal-lattice vectors in the $\{2,2,0\}$ set. For instance, the plane waves with wave vectors (1,1,0) and $(\bar{1},\bar{1},0)$ are connected by the (2,2,0) and $(\bar{2},\bar{2},0)$ reciprocal-lattice vectors. The band gap in the tetrahedral semiconductors is formed by the potentials at the $\{2,2,0\}$ reciprocal-lattice vectors. The electronic states on either side of the band gap are bonding and antibonding combinations of the plane waves connected by the $\{2,2,0\}$ reciprocal-lattice vectors. If the plane waves with wave vectors (1,1,0) and $(\bar{1},\bar{1},0)$ are among the initial states used in the calculation for the 8-atom unit cell, one of these initial filled states will relax to a valence-band state based on the combination $[(1,1,0)+(\bar{1},\bar{1},0)]$, and the other will relax to a conduction-band state based on the combination $[(1,1,0)-(\bar{1},\bar{1},0)]$, which is clearly not occupied in the physical ground state. Hence the choice of the lowest-

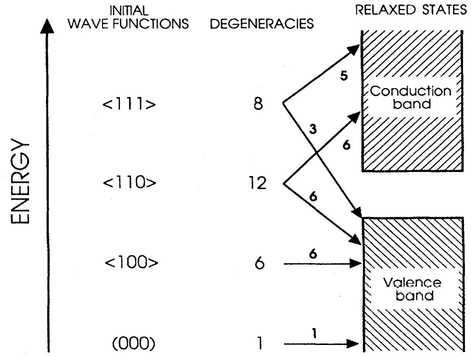

FIG. 21. Schematic energy-level diagram for the lowest states of an 8-atom cubic supercell of the diamond structure at the (0,0,0) **k** point. The figure shows that three plane waves from the ⟨111⟩ set must be included in the initial conditions in order for the electronic configuration to relax to the 16 valence-band states that constitute the ground state. Note that, for the system to relax to the ground state using the lowest-energy plane waves as initial conditions, 22 bands must be included, of which 6 will end up in the conduction band.

energy plane waves as the initial states for a calculation for the tetrahedral semiconductors will not yield the electronic ground state. In order to span the ground state, the initial electronic configuration must not contain any pairs of plane waves from the $\{1,1,0\}$ star of wave vectors connected by $\{2,2,0\}$ reciprocal-lattice vectors. Only when three plane waves from the $\{1,1,1\}$ star are included among the initial states will the electronic states relax to the required 16 valence-band states. If the choice of lowest-energy plane waves as the initial states were retained, 22 electronic states would have to be included in the calculation before the electronic configuration could relax to the ground state, and of these, 6 states would end up being unoccupied.

The computational cost of all of the iterative schemes described in the previous sections increases at least linearly with the number of electronic states included in the calculation, and the cost of orthogonalizing the electronic wave functions increases as the square of the number of bands. Including unoccupied states in the electronic relaxation reduces the advantage in computational speed of the molecular-dynamics and conjugate-gradients methods over conventional matrix diagonalization techniques. It is sensible, therefore, to attempt to use the smallest possible number of electronic states in the calculation. However, it is essential that the initial electronic states be able to converge to the ground state.

The problem with the calculation for an 8-atom cubic cell of the diamond-structure materials is that there are many reciprocal-lattice vectors at which the structure factor is zero. Hence there is no lattice potential associated with these reciprocal-lattice vectors. If the ionic potential is nonzero at every reciprocal-lattice vector, any choice of plane waves for the initial electronic states will span the ground state. The most common reason for the structure factor to be zero at some reciprocal-lattice vectors is symmetry. If the ionic configuration has no symmetry, any choice of initial electronic states should converge to the ground state. Only if the system has some symmetry must precautions be taken to ensure that the initial electronic states span the ground state.

2. Symmetry conservation

There is another reason why a particular set of initial states may not relax to the ground state when molecular-dynamics equations of motion are used to evolve the electronic configuration. This is a constraint on the evolution of the electronic states that arises from symmetry conservation. The effect is described in detail by Payne *et al.* (1988), and only a brief outline of the problem will be presented here.

The molecular-dynamics equations of motion and the related first-order equations of motion for the electronic wave functions will conserve any symmetry that is shared by the Hamiltonian and the initial electronic configuration. This symmetry can be broken when the electronic wave functions are orthogonalized, but this depends on the orthogonalization scheme used. In the Gram-Schmidt orthogonalization scheme, the symmetry is broken, whereas in many others, including the iterative scheme by Car and Parrinello, it is not. Since the electronic ground-state configuration must have the same symmetry as the Hamiltonian, one might not expect conservation of symmetry in the electronic configuration to cause any problems. However, the initial electronic configuration may not be able to propagate to the ground-state configuration without breaking this symmetry. If this is the case, the electronic configuration will not reach the ground state unless a symmetry-breaking orthogonalization scheme is used. Symmetry breaking is not necessary if random initial conditions are applied to the electronic wave functions, as described in Secs. VI.C.3 and VI.C.4.

B. Rate of convergence

Once it has been ensured that the initial electronic wave functions can relax to the ground state, the next most important consideration in choosing the wave functions is to maximize the rate of convergence. In all iterative matrix diagonalization techniques, new wave functions are generated by successive improvements to the previous wave functions. The closer the initial wave functions are to the self-consistent Kohn-Sham eigenstates, the fewer the number of iterations required to reach the ground state of the electronic configuration. However, it is extremely difficult to estimate the form of all the Kohn-Sham eigenstates of a complex system. Hence it will be necessary to use fairly poor approximations to the eigenstates as the initial states for most calculations.

C. Specific initial configurations

Selection of a set of initial wave functions is straightforward when the system to be studied is similar to a system that has been studied previously. In this case the converged wave functions for that system should be used as the initial states. This is particularly useful when testing for **k**-point convergence or for convergence as the cutoff energy for the plane-wave basis set is increased. The wave functions that were calculated with the previous set of **k** points or with the previous energy cutoff can be used as the initial electronic states.

1. Reduced basis sets

It is possible to obtain approximations to the Kohn-Sham eigenstates by using conventional matrix diagonalization techniques to find the eigenstates of the Kohn-Sham Hamiltonian with a small number of basis states. For most of the calculations that have been performed using the molecular-dynamics and conjugate-gradients methods, this technique would yield a relatively large matrix to diagonalize, even if a basis set consisting of only a few plane waves per atom were used. The Hamiltonian matrix would have to be diagonalized several times to achieve approximate self-consistency. Using this method to obtain a set of initial states for a molecular-dynamics or conjugate-gradients calculation might not save any computational effort over a method that used much poorer approximations to the self-consistent Kohn-Sham eigenstates but applied iterative matrix diagonalization methods throughout. To avoid the cost of diagonalizing the Hamiltonian matrix using conventional matrix diagonalization techniques, one could use iterative techniques to find the initial states with the smaller basis set. However, the computational cost of the iterative techniques increases only linearly with the number of basis states, so that there may not be much to be gained by using a reduced number of basis states initially. If most of the computation has to be performed using the complete basis set, then there will be an insignificant saving in computational time if the calculation is started with a reduced basis set.

2. Symmetrized combinations of plane waves

The choice of single plane waves for the initial electronic states in a molecular-dynamics or conjugate-gradients calculation does not exploit the symmetry of the system. The computational cost of such a calculation could be reduced by using symmetrized combinations of plane waves as the initial states for the electronic relaxation. Factoring the Hamiltonian matrix into submatrices of different symmetry, which can be solved independently, drastically reduces the computational time from that required using conventional matrix diagonalization techniques. However, there is an insignificant time saving to be gained by exploiting the symmetry of the system in the molecular-dynamics or conjugate-gradients methods. A significant saving in computational time could only be achieved if the fast Fourier-transform routines were rewritten for each symmetrized combination of basis states, and this requires a considerable investment of programming effort. Computational time can be saved in the orthogonalization routine by using symmetrized combinations of plane waves as the initial states. States of different symmetry are automatically orthogonal, and only states with the same symmetry have to be orthogonalized. If there are N_1 states with one symmetry, N_2 of another, and so on, orthogonalization requires $(N_1^2+N_2^2+\cdots)N_{PW}$ operations rather than the $N_B^2N_{PW}$ operations required if symmetrized combinations of plane waves are not used. However, rounding errors will tend to destroy the symmetry of the wave functions, and so additional computational effort will be required to resymmetrize them periodically.

The relatively small reduction in computational speed for systems with low symmetry is one of the strengths of the molecular-dynamics and conjugate-gradients methods. The use of symmetry is contrary to the spirit of these methods. In all of the calculations performed using these techniques, the positions of the ions have been allowed to vary. If the positions of the ions vary during the calculation, the ionic configuration will spend most of the time in regions of the phase space that have very low symmetry. The use of symmetry was essential in the days of primitive computers, when conventional matrix diagonalization techniques were used to solve for the Kohn-Sham eigenstates. Imposing symmetry onto a system adds fictitious constraints to the motions of the ions and restricts the relaxation of the ionic configuration. There is no point in imposing symmetry in molecular-dynamics or conjugate-gradients calculations because this does not significantly increase the computational speed of these techniques.

3. Random initial configurations

The initial electronic states for a molecular-dynamics or conjugate-gradients calculation can be generated by choosing random values for the coefficients of the plane-wave basis states. This method ensures that the ground-state is spanned by the initial states and that there is no conserved symmetry in the initial electronic configuration that might prevent the wave functions from relaxing to the Kohn-Sham eigenstates. It is sensible to give nonzero values only to the coefficients of plane-wave basis states that have small energies, so that the initial states do not have very high energies. With this precaution the electronic states are unlikely to be significantly further from eigenstates than simple plane waves, and very few extra iterations will be required to converge to the ground state.

4. Random initial velocities

In the molecular-dynamics method there are two degrees of freedom available for choosing the initial electronic configuration: the initial wave functions and their velocities can be chosen arbitrarily. Adding random velocities to the coefficients of the initial electronic states avoids the problem of the initial configuration's not spanning the ground-state and not relaxing to the ground state due to a conserved symmetry. It is sensible to limit the kinetic energy of the initial wave functions so that there is not too much excess energy in the electronic system.

VII. RELAXATION OF THE IONIC SYSTEM

Up to this point, the relaxation of the electronic configuration to its ground state has been considered, while the ionic positions and the size and shape of the unit cell have been held fixed. Once these additional degrees of freedom are allowed to relax to equilibrium, new features emerge. This procedure is much simpler than a full dynamical simulation of the ionic system because only the final state of the system (ions *and* electrons in their minimum energy configurations) is of interest, and the path towards this state is irrelevant. Hence errors can be tolerated along the relaxation path. This is not the case with a full dynamical simulation, where errors must be carefully controlled at all points along the ionic trajectories. The problems associated with full dynamical simulations of the ionic system will be discussed in Sec. VIII. Here we describe how ionic relaxation is easily incorporated into a molecular-dynamics-based method.

A. The Car-Parrinello Lagrangian

The positions of the ions and the coordinates that define the size and shape of the unit cell can be included as dynamical variables in the molecular-dynamics Lagrangian. The resulting Lagrangian is usually referred to as the "Car-Parrinello Lagrangian" and takes the form

$$L=\sum_i \mu\langle\dot{\psi}|\dot{\psi}\rangle+\sum_I \tfrac{1}{2}M_I\dot{R}_I^2 +\sum_\nu \tfrac{1}{2}\beta\dot{\alpha}\nu^2-E[\{\psi_i\},\{\mathbf{R}_I\},\{\alpha\nu\}]\ , \qquad (7.1)$$

where M_I is the mass of ion I and β is a fictitious mass associated with the dynamics of the coordinates that define the unit cell, $\{\alpha\nu\}$.

B. Equations of motion

Two further sets of equations of motion can be obtained from the Lagrangian (7.1), the first for the positions of the ions,

$$M_I\ddot{\mathbf{R}}_I=-\frac{\partial E}{\partial \mathbf{R}_I}\ , \qquad (7.2)$$

which simply relates the acceleration of the ions to the forces acting on them. The second set of equations is for the coordinates of the unit cell,

$$\beta\ddot{\alpha}_\nu=-\frac{\partial E}{\partial \alpha_\nu}\ . \qquad (7.3)$$

These equations relate the rate of acceleration of the lengths of the lattice vectors to the diagonal components of the stress tensor integrated over the unit cell and relate the accelerations of the angles between the lattice vectors to the off-diagonal components of the stress tensor integrated over the unit cell.

The equations of motion for the degrees of freedom associated with the dynamics of the ions and of the unit cell can be integrated at the same time as the equations of motion for the electronic states and, as will be shown below, provide a method for performing *ab initio* dynamical simulations of the ionic system. However, a relaxation of the ionic system can be performed using these equations of motion simply by removing kinetic energy from the electronic system, the ionic system, and the motion of the unit cell. In this case the system will evolve until the total energy of the system is minimized with respect to all of these degrees of freedom, and the ionic configuration will have reached a local energy minimum. However, integration of the equations of motion for the ions and for the unit cell is not as straightforward as it first appears. This is because *physical* ground-state forces on the ions and integrated stresses on the unit cell cannot be calculated for arbitrary electronic configurations, as shown in the following section.

C. The Hellmann-Feynman theorem

The force on ion I, $\mathbf{f}_I$, is minus the derivative of the total energy of the system with respect to the position of the ion,

$$\mathbf{f}_I=-\frac{dE}{d\mathbf{R}_I}\ . \qquad (7.4)$$

As an ion moves from one position to another, the wave functions must change to the self-consistent Kohn-Sham eigenstates corresponding to the new position of the ion if the value of the Kohn-Sham energy functional is to remain physically meaningful. The changes in the electronic wave functions contribute to the force on the ion, as can be clearly seen by expanding the total derivative in (7.4),

$$\mathbf{f}_I=-\frac{\partial E}{\partial \mathbf{R}_I}-\sum_i\frac{\partial E}{\partial\psi_i}\frac{d\psi_i}{d\mathbf{R}_I}-\sum_i\frac{\partial E}{\partial\psi_i^*}\frac{d\psi_i^*}{d\mathbf{R}_I}\ . \qquad (7.5)$$

Equation (7.5) should be compared with the Lagrange equation of motion for the ion (7.2). It can be seen that the "force" in Eq. (7.2) is only the partial derivative of the Kohn-Sham energy functional with respect to the position of the ion. In the Lagrange equations of motion for the ion, the force on the ion is not a physical force. It is the force that the ion would experience from a particular electronic configuration. However, it is easy to show that when each electronic wave function is an eigenstate of the Hamiltonian the final two terms in Eq. (7.5) sum to zero. Since $\partial E/\partial\psi_i^*$ is just $H\psi_i$, these two terms can be rewritten

$$\sum_i\left\langle\frac{\partial\psi_i}{\partial\mathbf{R}_I}\right|H\psi_i\Big\rangle+\sum_i\Big\langle\psi_iH\left|\frac{\partial\psi_i}{\partial\mathbf{R}_I}\right\rangle\ . \qquad (7.6)$$

However, if each ψ_i is an eigenstate of the Hamiltonian,

$$H\psi_i=\lambda_i\psi_i\ , \qquad (7.7)$$

so Eq. (7.6) is equal to

$$\sum_i\left\langle\frac{\partial\psi_i}{\partial\mathbf{R}_I}\right|\lambda_i\psi_i\Big\rangle+\sum_i\Big\langle\psi_i\lambda_i\left|\frac{\partial\psi_i}{\partial\mathbf{R}_I}\right\rangle =\sum_i\lambda_i\frac{\partial}{\partial\mathbf{R}_I}\langle\psi_i|\psi_i\rangle=0\ , \qquad (7.8)$$

since $\langle\psi_i|\psi_i\rangle$ is a constant by normalization.

This shows that when each ψ_i is an eigenstate of the Hamiltonian the partial derivative of the Kohn-Sham energy with respect to the position of an ion gives the real *physical* force on the ion. This result is usually referred to as the *Hellmann-Feynman theorem* (Hellmann, 1937; Feynman, 1939). The Hellmann-Feynman theorem holds for any derivative of the total energy. Hence, when each ψ_i is an eigenstate of the Hamiltonian, only the explicit dependence of the energy on the size and the shape of the unit cell has to be calculated to determine the integrated stresses.

1. Errors in Hellmann-Feynman forces

Forces calculated using the Hellmann-Feynman theorem are very sensitive to errors in the wave functions ψ_i. The error in the force is first order with respect to errors in the wave functions. Therefore accurate forces can only be calculated when the wave functions are very close to exact eigenstates. The error in the Kohn-Sham energy is second order with respect to errors in the wave function, so that it is significantly easier to calculate an accurate total energy than to calculate an accurate force.

2. Consequences of the Hellmann-Feynman theorem

The Hellmann-Feynman theorem simplifies the calculation of the physical forces on the ions and the integrated stresses on the unit cell. However, the electronic wave functions must be eigenstates of the Kohn-Sham Hamiltonian for the Hellmann-Feynman theorem to be applicable. Therefore the forces on the ions and the integrated stresses on the unit cell should not be calculated until the electronic configuration is near its ground state. Once the forces and stresses have been calculated, the positions of the ions and the size and shape of the unit cell may be changed. Each time that the positions of the ions or the size and shape of the unit cell are changed, the electrons must be brought close to the ground state of the new ionic configuration in order to calculate forces and stresses for the new ionic configuration.

When the ionic configuration is relaxed to a local energy minimum, the relaxation of the ionic configuration can be partially overlapped with the initial relaxation of the electronic configuration. Provided that the magnitudes of the Hellmann-Feynman forces are larger than the errors in the forces, moving each ion in the direction of the calculated force will lower the total energy of the system and move the ionic configuration towards the local energy minimum. However, if the Hellmann-Feynman forces are smaller than the errors in the forces, displacement of the ions in the directions of the forces may not decrease the total energy and could take the ionic configuration away from the global energy minimum. In this case, overlapping the ionic relaxation with the electronic relaxation will increase the total number of iterations needed to relax the system to the global energy minimum.

It might be argued that, as long as kinetic energy is continuously removed from all the degrees of freedom in the system, the total energy in the system must continuously decrease, so that the ionic configuration must relax to a local energy minimum. However, this is only true if the time steps are made sufficiently short. Moving the ions a finite distance can add energy to the electronic system. If the energy added to the electronic system each time step becomes too large, the electronic system will never relax to its ground state, and the ionic system will never reach a local energy minimum. Therefore some caution has to be exercised when one overlaps ionic relaxation with the electronic relaxation, to ensure that the ionic system reaches the local energy minimum in the shortest possible time.

D. Pulay forces

In principle, there should be an additional term in Eq. (7.5) to represent the derivative of the basis set with respect to the position of the ion. This contribution to the force on the ion is called the Pulay force (Pulay, 1969). If the value of the Pulay force is not calculated, there is a further error in the value of the Hellmann-Feynman force. It can be shown that the Pulay force vanishes if the derivatives of all the basis states $\delta\phi/\delta\lambda$ are spanned by the basis set $\{\phi\}$ (Scheffler *et al.*, 1985). For a plane-wave basis set, the derivatives of each basis state with respect to the position of an ion are zero and the Pulay force is zero. The calculated Hellmann-Feynman force then will be exactly equal to the derivative of the total energy with respect to the position of the ion, provided that the electronic wave functions are Kohn-Sham eigenstates. This is one of the great advantages of using a plane-wave basis. If the Pulay force does not vanish and if it is not calculated, the computed Hellmann-Feynman force will not be equal to the derivative of the total energy with respect to the position of the ion. This error is independent of how close the electronic configuration is to its ground state. In this case, moving an ion in the direction of the calculated force may increase the total energy. When the Pulay force is nonzero, a local energy minimum of the ionic system cannot be located by calculating the Hellmann-Feynman forces on the ions. The only ways of finding a local energy minimum are by trial and error or by calculating the actual force on each ion by calculating the change in the total energy on displacing each ion in turn. This calculation on N_I ions requires $3N_I$ total-energy calculations. Therefore the number of total-energy calculations that are required to take the system to the local energy minimum using this method is $3N_I$ times as many as would be required if the forces on all the ions could be calculated from a single total-energy calculation. This represents a hidden increase in computational time with the size of the system, which could completely negate the apparent efficiency of a computational method.

E. Pulay stresses

If a plane-wave basis set is used in a total-energy calculation, the Pulay forces on the ions will be zero. However, the Pulay stresses on the unit cell may be nonzero with a plane-wave basis set. If the number of plane-wave basis states remains constant, changing the size of the unit cell changes the cutoff energy for the basis set. Increasing the number of plane-wave basis states by increasing the cutoff energy for the basis set will usually reduce the total energy of the system. Only if the cutoff energy is large enough to achieve absolute convergence will the change in the total energy be zero. Most total-energy pseudopotential calculations are performed with a cutoff energy at which energy differences have converged but at which the total energies have not converged. In this case the diagonal components of the Pulay stresses on the unit cell will be nonzero.

It may be surprising that increasing the number of basis states can change the total energy of a system but not change the differences in energy between a number of systems. However, the energy differences between systems arise mainly from the differences in bonding in each system, so the energy differences are dominated by the regions outside the ion cores. Provided that the additional basis states introduced by increasing the cutoff energy for the basis set do not change the charge density in the bonding regions, energy differences between systems will not change. The additional basis states introduced by increasing the cutoff energy for the basis set merely provide a more accurate description of the wave functions inside the core regions. The additional basis states will change the total energy per atom of each system by a constant amount and so leave energy differences unaltered.

The magnitude of the Pulay stress in a pseudopotential calculation can be determined by calculating the change in the total energy per atom as the cutoff energy for the basis set varies (Froyen and Cohen, 1986; Gomes Dacosta *et al.*, 1986). The crucial significance of Pulay stress correction for surface stress calculations has been emphasized by Vanderbilt (1987). The change in the total energy per atom will be independent of the details of the ionic configuration provided that the cutoff energy for the basis set is large enough for energy differences to have converged. Hence the Pulay stress due to the plane-wave basis set can be calculated once and for all from the change in the total energy of a small unit cell as the cutoff energy for the plane-wave basis set varies.

F. Local energy minimization

The simplest use of Hellmann-Feynman forces is to locate the position of a local energy minimum of the ionic system. The ions are moved along the directions of the Hellmann-Feynman forces until the residual forces on all the atoms are smaller than a given value. In such a calculation the errors in the Hellmann-Feynman forces due to the deviation of the electronic configuration from the ground state can be regarded as a source of thermal noise. These forces will cause the ions to fluctuate around their equilibrium positions, and the magnitudes of the residual forces on the ions will never reach zero. The magnitudes of the errors in the Hellmann-Feynman forces must be reduced as the system approaches the local energy minimum if the system is to continue approaching that minimum. Therefore the electronic configuration must be relaxed closer and closer to the instantaneous ground state as the ionic configuration approaches the local energy minimum.

1. Local energy minimization with the molecular-dynamics method

In molecular-dynamics methods it is sensible to treat the electronic and ionic systems independently when relaxing the ions to their equilibrium positions and to use different time steps for the two systems. The time step for the ionic system should be progressively reduced as the ionic configuration approaches the local energy minimum. This allows the electronic configuration to relax closer to its instantaneous ground-state configuration as the ions approach their equilibrium positions, to ensure that the errors in the Hellmann-Feynman forces are always smaller than the actual forces on the ions.

2. Local energy minimization with the conjugate-gradients method

The conjugate-gradients method converges the electronic configuration to its ground state in far fewer iterations than molecular-dynamics methods. In this case, moving the ions small distances along the directions of the Hellmann-Feynman forces at each iteration will be an inefficient method for performing a local energy minimization. Many more iterations will be required to reach the energy minimum than would be required to converge the electronic configuration to its ground state. In this case it is sensible to use a more sophisticated scheme for relaxing the ionic configuration, one which can locate the equilibrium positions of the ions in the minimum number of iterations. Ideally the number of iterations required to locate a local energy minimum of the ionic system should be of the same order as the number of iterations required to relax the electronic configuration to its ground state.

G. Global energy minimization

Accurate forces on the ions can be calculated relatively quickly when conjugate-gradients or molecular-dynamics methods are used to perform a total-energy pseudopotential calculation. It has been shown how these forces can be used to relax a system of ions to a local energy minimum. The technique of moving the ions in the directions of the Hellmann-Feynman forces until the forces on the ions become zero is basically a zero-temperature quench, because the ions do not acquire any kinetic energy during the relaxation. At the end of this process, the system will be in a local energy minimum. By performing zero-temperature quenches from a variety of initial configurations of the ionic system, one can obtain information about the local energy minima of the system. However, there is no guarantee that this method will locate the *global* energy minimum of the system. In theory, a very-low-energy minimum can only be found if a simulated annealing process is carried out (Kirkpatrick *et al.*, 1983). But even with a simulated annealing procedure, there is no guarantee that the *global* energy minimum will be located.

1. Simulated annealing

The success of any simulated annealing technique is very sensitive to the structure of the phase space being explored. The purpose of performing a simulated anneal is to determine the lowest-energy configuration of the ionic system. For a system that contains many ions there will be a large number of ionic configurations that are local energy minima. The simulated annealing procedure has to explore the phase space of the system to locate the lowest-energy local minimum. The phase space for a particularly simple system is shown schematically in Fig. 22. The diagram shows two local energy minima, separated by energy ΔE. If the position of the lowest-energy minimum is to be located using the technique of simulated annealing, the thermal energy kT must be smaller than ΔE. If the thermal energy is larger than this, the energies of the two minima cannot be distinguished within the thermal smearing. The diagram shows an energy barrier of height E_B separating the local energy minima. The ionic configuration can only move between the two local energy minima by gaining at least E_B in energy through a thermal fluctuation. The time spent waiting for a thermal fluctuation of this magnitude is $(1/\nu)\exp(E_B/kT)$, where ν is the attempt frequency. In a typical simulation with the molecular-dynamics method, the total time of the simulation is of the order of $10-100$ times $1/\nu$. Therefore the probability that the ionic configuration will traverse an energy barrier during the simulation is dominated by the exponential factor. When the temperature is low enough to distinguish between the energies of the local energy minima ($kT \sim \Delta E$), the time taken for the system to move between the energy minima is proportional to $\exp(E_B/\Delta E)$. The system must move between the minima at least once to locate the lower-energy minimum. If $E_B/\Delta E$ is large, it is extremely unlikely that the simulation will locate the global energy minimum, but if $E_B/\Delta E$ is small, the global energy minimum can be located easily.

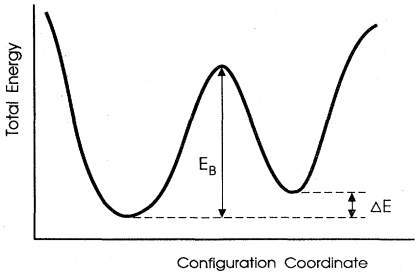

FIG. 22. Representation of two energy minima differing by ΔE, separated by a barrier E_B.

Simulated annealing is ideal for escaping from local energy minima separated from the global energy minimum by small energy barriers. However, the method is very inefficient when the energy barriers separating the energy minima are large. Unless the structure of the phase space of the system is very favorable, any simulated annealing process will leave the system in a local energy minimum rather than at the global energy minimum.

The success of simulated annealing techniques also depends on the number of local energy minima that have energies close to the energy of the global energy minimum. In principle, all of these local energy minima have to be visited during the simulated annealing process in order to determine which is the true global energy minimum. As the number of local energy minima increases exponentially with the number of atoms in the system, simulated annealing is only likely to be successful for small systems.

2. Monte Carlo methods

The difficulty of locating the position of the global energy minimum using the technique of simulated annealing is due to the difficulty of traversing the energy barriers that separate the local energy minima. If the energy barriers are large, the system only occasionally traverses an energy barrier. Between traversals of the energy barriers, the system explores the region of phase space around a single local energy minimum. However, only the value of the energy at the local minimum has any relevance to the process of locating the global energy minimum. The time spent exploring the phase space around each local energy minimum serves no useful purpose in the simulated annealing process, although it does provide information about the free energy of the system. It is easy to locate the local energy minimum in each region of phase space, but it is difficult to move between regions of phase space that have low-energy local minima.

The process of locating the global energy minimum is sometimes attempted by using the method of steepest descents or simple molecular dynamics to sample the region of phase space around each local energy minimum, followed by a discontinuous jump through phase space into the region around a different local energy minimum.

However, there is no point in exploring regions of phase space that have very high energies, and so a sampling criterion should to be applied to determine whether to explore the region of phase space reached by the discontinuous jump. The sampling criterion generally adopted compares the energies of the new and the old ionic configurations. The new configuration is accepted or rejected according to a Monte Carlo algorithm (Metropolis *et al.*, 1953): if the new configuration is of lower energy than the old, it is accepted; if the new configuration is ΔE higher in energy than the old configuration, it is accepted with a probability $\exp(-\Delta E/kT)$. This method allows the system to cross energy barriers without waiting for a thermal fluctuation large enough to traverse the barrier.

The Monte Carlo technique described above is computationally expensive to implement with molecular-dynamics schemes for relaxing the electronic configuration to its ground state. When the ionic configuration makes a discontinuous jump through phase space, the electrons will not be close to the ground state of the new ionic configuration. Each change in the ionic configuration must be followed by a complete relaxation of the electronic configuration to the new ground state before the energies of the initial and final configurations can be compared. In contrast, the Monte Carlo technique could be efficiently implemented with the conjugate-gradients method because the energy of the new ionic configuration can be calculated rapidly.

3. Location of low-energy configurations

Location of global energy minima is a complex problem. No scheme can be guaranteed to find the global energy minimum in a single calculation. The only way of being reasonably confident that the global energy minimum has been located is to find the same lowest-energy configuration in a number of different calculations. In practice a number of low-energy configurations will be located by successive calculations. When subsequent calculations do not locate any new low-energy configurations and the ionic configuration always reaches one of the low-energy configurations found previously or a configuration of significantly higher energy, then there is a very high probability that all the low-energy configurations of the system have been located and, hence, that the global energy minimum has been located.

VIII. DYNAMICAL SIMULATIONS

There is an enormous literature associated with studies of the dynamical behavior of systems. The book by Allen and Tildesley (1987) provides an excellent introduction to the subject. Obvious areas of interest include diffusion, melting, and the calculation of free energies. These studies are generally carried out using empirical potentials (i.e., some model of the interaction between the atoms in the system parametrized according to experimental data). Empirical potentials have the drawback that it is impossible to know their region of validity. The potentials are often parametrized using data for the perfect crystal or data describing only small perturbations from the perfect crystal. Even if these potentials do work perfectly for the crystal, there is no reason why they should be capable of describing diffusion in the solid, which can involve configurations very different from those found in the crystal, let alone a liquid, whose structure may bear no relation whatsoever to the parent crystal. The problem of determining an accurate interatomic potential is particularly acute in the case of silicon, for which many years of effort have yet to produce a general-purpose potential. In contrast, the total-energy pseudopotential method has been shown to be applicable in a much larger region of phase space than any empirical potential. Hence a dynamical simulation performed using these forces should accurately describe a real system, irrespective of the region of phase space that is explored under the dynamical evolution of the system.

A. Limitations to dynamical simulations

If simulations are performed using a finite supercell, as they must be when plane-wave basis sets are used, the systems cannot undergo true phase transitions, and the range of correlations in the system will be limited by the size of the supercell. It should also be appreciated that the electron temperature will, in general, be zero in such a simulation. Thermal excitation of the electronic system can be described in density-functional theory (Mermin, 1965); however, there are fundamental problems with density-functional theory which make it difficult to describe a system at finite temperature without performing an extremely time-consuming calculation for the excited states of the system (Hybertson and Louie, 1985; Godby *et al.*, 1986). Provided that the thermal energy is much smaller than the smallest excitation energy of the electronic system, the effects of a finite electron temperature should be small. If this is the case, the error introduced by using zero-temperature density-functional theory in a dynamical simulation should not be significant. The effect of setting the electronic temperature to zero recently has been shown to be negligible in a study of the structural phase transition of GeTe (Rabe and Joannopoulos, 1987).

B. Accuracy of dynamical trajectories

In Sec. VII it was pointed out that the calculated forces on the ions are only the true physical forces when the electronic system is in its exact ground-state configuration. Therefore, to generate correct dynamical trajectories for the ions, the electrons must be relaxed to the ground state at each ionic time step. Although any of the methods described in Secs. III, IV, and V can be used to relax the electronic configuration to its ground state, most of these prove to be extremely expensive computationally for performing dynamical simulations of the ionic system. The most efficient of these, the conjugate-gradients method, is fast enough to allow dynamical simulations, but even in the case of this technique it is important to generate good sets of initial wave functions according to the technique outlined in Sec. VIII.E below. However, there is an alternative approach to performing dynamical simulations, which forms the basis of the Car-Parrinello method. Rather than insisting that the electronic configuration be in the exact ground-state configuration at each ionic time step, one may be able to perform dynamical simulations even if the electronic configuration is only close to the exact ground state. Although this implies that there are errors in the Hellmann-Feynman forces at each time step, dynamical simulations will be successful provided that the errors in the forces remain small *and* that the effect of these errors remains bounded in time. The Car-Parrinello method can fulfill both of these criteria (Remler and Madden, 1990; Pastore *et al.*, 1991). It is this latter point about the boundedness of the errors which provides the distinction between the Car-Parrinello method and the "improved" methods outlined in Secs. IV and V. While these improved methods will for a fixed computational effort move the electronic system closer to the ground-state configuration than the simple molecular-dynamics method, the errors introduced by these improved methods, although smaller than the error in the simple molecular-dynamics method, does not remain bounded in time. The boundedness of the error in the Car-Parrinello method results from an "error cancellation" that occurs when the Car-Parrinello Lagrangian is used to generate the dynamics of the electronic and ionic system. This effect is most easily demonstrated by the simple example in the following section, which clarifies this point by comparing second-order and first-order equations of motion.

C. Error cancellation in dynamical simulations

The origin of the cancellation of the errors in the Hellmann-Feynman forces under the equations of motion generated by the Car-Parrinello Lagrangian can be illustrated by considering a system that contains a single atom, which has a single occupied electronic orbital, as shown in Fig. 23. The molecular-dynamics equation of motion for the evolution of the electronic wave function is

$$\mu\ddot{\psi} = -[H-\lambda]\psi \ . \tag{8.1}$$

If the atom is at rest and the electronic wave function is the ground-state wave function, then $[H-\lambda]\psi=0$, and the wave function will be stationary. If the orbital is displaced away from the ion, the magnitude of the acceleration of the wave function will increase roughly linearly with the magnitude of the displacement. If the ion is moving at constant velocity and the orbital begins to lag behind the ion, the acceleration of the orbital will increase. The velocity of the orbital will increase until the orbital overtakes the ion. As the orbital overtakes the ion, the acceleration of the wave function will change sign and the orbital will begin to slow down. The orbital continues to slow down until the ion overtakes it, at which point the whole process starts again. Hence, if the ion were to move at constant velocity, the electronic orbital would oscillate around the instantaneous position of the ion. The value of the Hellmann-Feynman force exerted on the ion by the orbital will oscillate around the correct value, so that the error in the Hellmann-Feynman force will tend to cancel when averaged over a number of wave-function oscillations. The oscillation of the error in the Hellmann-Feynman force will prevent a continuous transfer of energy between the ionic and the electronic degrees of freedom, as long as the fictitious oscillations occur over time scales much shorter than the physical ionic time scales. This is a reflection of the fact that, given a sufficiently large mass ratio, there is an adiabatic isolation of the (much) "lighter" electronic coordinates from the "heavier" ionic degrees of freedom.

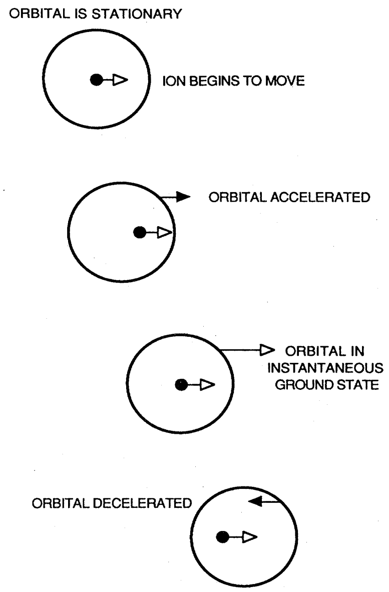

FIG. 23. Schematic illustration of how an orbital will oscillate around a moving ion during a simulation with $\mu\ddot{\psi}=-[H-\lambda]\psi$, as discussed in the text. Velocities and accelerations are designed as open and filled arrows, respectively.

A first-order equation of motion, on contrast, gives the following expression for the evolution of the electronic orbital:

$$\dot{\psi} = -[H-\lambda]\psi \ . \tag{8.2}$$

With this equation of motion the velocity of the orbital increases roughly linearly with the displacement of the orbital from the ion. Once the ion has begun to move,

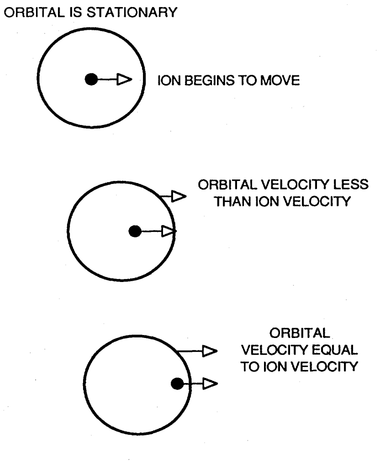
ORBITAL IS STATIONARY
ION BEGINS TO MOVE
ORBITAL VELOCITY LESS THAN ION VELOCITY
ORBITAL VELOCITY EQUAL TO ION VELOCITY

FIG. 24. Schematic illustration of how an orbital will eventually lag behind a moving ion during a simulation with $\mu\dot{\psi}=-[H-\lambda]\psi$, as discussed in the text. Convention the same as in Fig. 23.

the orbital falls further behind the ion until its velocity is equal to the velocity of the ion, and the orbital then remains a fixed distance behind the instantaneous position of the ion. This process is illustrated in Fig. 24. The orbital then exerts a constant damping force on the ion due to the systematic error in the value of the Hellmann-Feynman force, which is proportional to the velocity of the ion. Hence using the first-order equation of motion to evolve the electronic wave function for a dynamical simulation results in a viscous drag on the ions. This simple model suggests that a second-order equation of motion for the electronic degrees of freedom should give a more accurate account of the dynamics of the ionic system than a first-order equation of motion, a conclusion supported by detailed analysis of the evolution of the electronic wave function (Payne, 1989; Car *et al.*, 1991) and by simulations using first- and second-order equations of motion (Remler and Madden, 1990). The success of the Car-Parrinello method comes from this error cancellation, which turns the first-order error in the Hellmann-Feynman forces into a second-order error when integrated along the ionic trajectories.

D. Car-Parrinello dynamics; constraints

The successful implementation of Car-Parrinello dynamics relies on a number of features. The error cancellation occurs only if the time scales in the electronic system are all shorter than the shortest time period in the ionic system. From considerations similar to those in Sec. III.D.2, it can be seen that the longest time period in the electronic system is related to the difference in energy between the highest occupied state and the lowest unoccupied state. The actual magnitude of this time period can be adjusted by changing the value of the fictitious mass, so that even in systems where this energy gap is arbitrarily small the fictitious mass can be chosen sufficiently small to ensure the nonoverlap of time periods in the electronic and ionic systems. However, as the value of the fictitious mass is decreased, the highest frequency in the electronic system increases, requiring a shorter time step to integrate the equations of motion stably for the electrons, thus increasing the computational effort required for a given simulation. Therefore the energy gap can become so small that it becomes impractical to carry out a simulation.

The Car-Parrinello Lagrangian is invariant in time, and hence the total energy in the electronic and ionic systems will be constant provided that no damping is applied to any of the kinetic degrees of freedom and that the "forces" required to impose the constraints of orthonormality of the wave functions are conservative. The energy constant of motion in a classical molecular-dynamics simulation is made up only of the kinetic energy of the ions and the potential energy, which is the Kohn-Sham energy in the *ab initio* simulation. However in a Car-Parrinello simulation the energy constant of motion includes the fictitious kinetic energy of the electrons. Even in a "perfect" Car-Parrinello calculation, in which the electrons moved at exactly the correct velocity to remain in the instantaneous ground state of the ionic configuration, the electronic wave functions would speed up and slow down as the ions moved along their trajectories, and the kinetic energy of the electrons would vary in time in direct proportion to the ionic kinetic energy, but smaller by a factor of the ratio of the fictitious electronic and physical ionic masses, typically less than 0.01. Because of this fictitious electronic kinetic energy, the sum of the kinetic energy of the ions and the potential energy is not a constant, even in a "perfect" Car-Parrinello simulation. In this situation the fictitious temperature of the electronic degrees of freedom is at least two orders of magnitude smaller than the temperature of the ionic system, so that this situation is thermodynamically unstable.

In addition to the kinetic energy required for the electrons to follow the ions exactly, there are fluctuations of energy between the ionic and electronic systems. The deviation of the electronic configuration from its ground state is related to the magnitude of these energy fluctuations. If the longest time period in the electronic system is not significantly shorter than the shortest time period in the ionic system, energy will be continuously transferred between the electronic and ionic systems until the fictitious temperature of the electronic degrees of freedom and the temperature of the ionic degrees of freedom are equal. In such a situation there are large amounts of energy in the electronic degrees of freedom, and the electronic configuration is far from its ground state. When a significant transfer of energy from the ions to the electrons occurs in a Car-Parrinello calculation, the simulation must be stopped periodically and the electrons returned to their ground-state configuration before restarting the simulation (Zhang *et al.*, 1990). In the process of returning the electrons to their ground state, the fictitious kinetic and potential energies in the electronic degrees of freedom are removed from the system so there is no longer a conserved energy associated with the Car-Parrinello dynamics. If no further action were taken, the temperature of the ionic system would decrease continuously during the simulation. To compensate for this irreversible heat flow from the ionic system, it is usual to attach a Nosé thermostat to the ionic system (Nosé, 1984). This Nosé thermostat has the primary role of supplying energy to the ionic degrees of freedom, to compensate for the loss of energy from the ions to the electronic degrees of freedom. A systematic study of the range and validity of the Nosé thermostat is given by Cho and Joannopoulos (1992). It has been demonstrated, however, that in simulations where the total energy tends to drift, the Nosé thermostat breaks down and fails to produce a correct canonical thermal distribution (Toxvaerd, 1991). As yet no one has attempted to analyze the errors in the ionic trajectories that arise when the time periods in the electronic and ionic degrees of freedom begin to overlap. An alternative method has recently been proposed to control the buildup of energy in the electronic degrees of freedom by attaching a separate Nosé thermostat to the electronic degrees of freedom, set at much lower temperature than the ionic thermostat (Bloechl and Parrinello, 1991). Once again no attempt has yet been made to quantify the errors introduced by this method.

The correct application of the constraints of orthogonality and normalization is essential for performing a successful Car-Parrinello dynamical simulation. This is relatively easily understood from the following considerations. Consider two wave functions that are not orthogonal. There are an infinite number of pairs of orthogonal wave functions that can be formed from these two wave functions, and each of these possible choices will have a different "velocity" associated with each of the wave functions. However, only one of these choices is consistent with a "conservative" constraint force acting on the wave functions that does not change the kinetic energy of the electronic system. The simplest method for understanding which form of application of constraints is correct is to appreciate that the constraint forces must not change if the labeling of the wave functions is changed—the constraint forces should be invariant under rotations within the subspace of the occupied electronic states. It is clear, then, that the Gram-Schmidt orthogonalization technique cannot be applied in a dynamical simulation, because the forces change according to the labeling of the states—for instance, whichever wave function is labeled 1 is not changed by the orthogonalization procedure. Car and Parinello apply the constraints in two steps (Car and Parrinello, 1989). The first is an application of constraints directly in the equations of motion, using the Lagrange multipliers. These constraints ensure that if the accelerations of the wave functions were all zero the wave functions would remain orthonormal. This constraint is required because, despite their orthonormality at the last and present time steps, the wave functions would become nonorthonormal if they continued to move with constant velocity. The Lagrange multipliers that ensure this orthonormality (to order dt^4) are

$$\Lambda_{ij} = \mu\langle \dot{\psi}_j | \dot{\psi}_i \rangle \ . \tag{8.3}$$

Although application of these Lagrange multipliers alone would be sufficient to ensure orthonormality of the wave functions, to the same accuracy as the error in Verlet algorithm in the absence of any accelerations, this is no longer true if accelerations are present. To ensure the orthonormality of the wave functions at the end of the time step, one can either modify the above Lagrangian multipliers to take account of the accelerations of the wave functions, or one can retain the Lagrange multipliers given by Eq. (8.3) and Car and Parrinello's iterative method (3.23), or one can employ a similar rotationally invariant method, such as determining the similarity transform required to diagonalize the overlap matrix.

E. Conjugate-gradients dynamics

It has been pointed out that the Car-Parrinello algorithm permits accurate dynamical simulations of ionic systems to be performed, providing the time scales in the ionic and electronic systems are decoupled. Although there are many systems in which this is the case, this decoupling of the time scales is generally difficult to obtain in the case of metallic systems, where the gap vanishes (unless the "simulation" is so artificial that the system used in the simulation is actually an insulator as a result of limited **k**-point sampling). In such cases, where the long-term stability of the Car-Parrinello dynamics is in doubt, there is considerable motivation for seeking an alternative technique for performing dynamical simulations. It has already been pointed out that using one of the alternative techniques to relax the electrons to the ground state requires much more computational effort to achieve the same accuracy in the evolution of the ionic system, and so, at first sight, it is simply too expensive computationally to perform dynamical simulations on systems for which the Car-Parrinello algorithm fails. However, it is obvious that, if the initial electronic configuration can be moved closer to the correct instantaneous ground-state configuration, less computational effort is required to converge it to its exact ground state, and hence a faster simulation is possible.

A simple method has been developed that allows an accurate prediction to be made for the initial electronic configuration at each ionic time step by extrapolating forward from electronic configurations at previous time steps. Typically, this method of extrapolation is found to bring the initial wave functions two orders of magnitude closer to the minimizing energy functional than simply using the wave functions from the previous time step. This typically reduces by a factor of two the computational effort required to bring the electronic system to within a few micro-eV per ion of its ground state. When this extrapolation technique is combined with the conjugate-gradients method, the resulting computational scheme is sufficient to make the entire dynamical simulation comparable in speed to a Car-Parrinello simulation. However, the technique has the advantage that it can be applied to a broader class of systems. The details of the scheme can be found in Arias *et al.* (1991); it will only be described briefly below.

The basis for the trial wave-function scheme is the first-order extrapolation

$$\Psi'_{nk}(\{r(t_{i+1}\})\equiv\Psi_{nk}(\{r(t_i\}) \ , \\ +\alpha[\Psi_{nk}(\{r(t_i\})-\Psi_{nk}(\{r(t_{i-1}\})] \ , \tag{8.4}$$

where $\{r(t_i)\}$ are the ionic coordinates at time t_i with i the ionic iteration number, α is a fitted parameter, and the prime indicates a trial wave function, as opposed to the fully converged $\Psi_{nk}(\{r(t_i)\})$. This scheme produces trial wave functions correct to first order in dr (and, by the $2N+1$ theorem, energies correct to third order in the time step) when the ionic coordinates are

$$\{r'(t_{i+1})\}=\{r(t_i)+\alpha[r(t_i)-r(t_{i-1})]\} \ . \tag{8.5}$$

To ensure that the resulting wave functions are in as close correspondence as possible with the actual ionic locations, $\{r(t_{i+1})\},\alpha$ is taken to minimize the discrepancy

$$|r(t_{i+1})-r'(t_{i+1})| \\ =|r(t_{i+1})-(1+\alpha)r(t_i)+\alpha r(t_{i-1})| \ . \tag{8.6}$$

One may imagine generalizing this scheme to higher orders, employing more of the preceding wave functions and producing ever smaller errors in the extrapolated wave functions. However, higher-order schemes suffer from an instability that pushes the wave-function errors into regions of phase space where convergence is so difficult that the net effect is to slow the simulation.

As in the Car-Parrinello scheme, orthonormality of the wave functions must be maintained; however, Eq. (8.4) yields wave functions that are not properly orthonormal. In the present case, one can simply Gram-Schmidt orthonormalize the resulting wave functions, because there is no longer any concern for maintaining a proper electron dynamic and because this procedure will not disturb the correctness to first order of the wave functions, a consequence of the fact that

$$\langle\Psi'_{nk}(\{r(t_{i+1})\})|\Psi'_{mk}(\{r(t_{i+1})\})\rangle=\delta_{n,m}+\mathrm{O}(dr^2) \ . \tag{8.7}$$

Once the wave functions extrapolated according to Eq. (8.4) have been Grahm-Schmidt orthonormalized, they are then relaxed to within a set tolerance of the Born-Oppenheimer surface by the conjugate-gradient procedure; this completes one cycle of iteration of the ionic motion.

F. Comparison of Car-Parrinello and conjugate-gradient dynamics

The Car-Parrinello and conjugate-gradients schemes for performing dynamical simulations are very different, and it is important to understand these differences in order to apply either technique successfully. The most important point is the difference between the time steps used in the two methods. In this respect conjugate-gradients dynamics is closer to conventional dynamical simulations, in which the time step is chosen to ensure an accurate integration of the ionic equations of motion. In simulations employing empirical potentials and those using the conjugate-gradients scheme, the forces on the ions are, to high precision, true derivatives of the total potential energy of the ions. In the case of empirical potentials, the only differences between the computed forces and the derivatives of the total ionic energy are rounding errors due to finite machine accuracy, but in the case of the conjugate-gradients simulation, the differences also include contributions due to the failure of the Hellmann-Feynman theorem because the electronic system is not exactly converged to its ground state. In the Car-Parrinello simulation, at each time step there are significantly larger errors in the Hellmann-Feynman forces, because the electronic configuration is not maintained so close to its exact ground-state configuration. The time step used in a Car-Parrinello simulation has to be much shorter than the one used for an equivalent conjugate-gradients simulation to integrate the electronic equations of motion stably. Additionally, the longest time period in the electronic system must be less than the shortest ionic time period, to ensure that the errors in the Hellmann-Feynman forces average to zero along the ionic trajectories.

At first sight the Car-Parrinello method and the wave-function extrapolation combined with conjugate-gradient relaxation are rather similar, in that each essentially performs an integration of the wave functions forward in time. However, the spirit of each technique and the behavior of the wave-function coefficients in the two cases are very different. In the case of the conjugate-gradients dynamics, the wave functions are propagated as close as possible to the instantaneous ground state, in order to reduce the effort required to fully relax them to the ground state. In the Car-Parrinello method, the motion of the electronic degrees of freedom preserves a delicate adiabatic separation between the electronic and ionic degrees of freedom. The electronic coefficients oscillate artificially about their ground-state values, which leads to a cancellation of the errors in the ionic forces.

G. Phonon frequencies

The phonon frequencies of a system can be obtained by performing a dynamical simulation and then Fourier-transforming either the velocity or the position autocorrelation functions. However, for this procedure to give phonon frequencies to high accuracy, the original ionic trajectories must be extremely accurate, since any noise in the trajectories will broaden the phonon frequencies. The conjugate-gradients dynamics scheme generates extremely accurate ionic trajectories, in which the noise can be reduced to an arbitrarily low level, and thus provides an excellent set of input data with which to determine phonon frequencies. Figure 25 shows the transform of the longitudinally polarized autocorrelation, as determined by the maximum-entropy method (Press *et al.*, 1989), of 40 silicon atoms in a periodic system over 50 Å long in the [100] direction. Each peak represents a natural frequency in the system. Neither the heights of the peaks nor this integrated intensities are meaningful, in that the system has not yet reached thermal equilibrium. Note that the primary caveat when working with the maximum-entropy method is that it produces spurious peaks when working with noisy data. No such peaks are obtained, indicating very clean data. The frequencies of the peak values of these spectra, as well as their transverse counterparts, are then compared with the experimentally measured phonon frequencies (Dolling, 1963; Nilsson and Nelin, 1972) in Fig. 26. As can be seen, there is good agreement between the results of the calculation and experiment, particularly in resolving the delicate splitting of the optic bands along Δ, which beat against each other with periods on the order of one picosecond. This technique for obtaining phonon frequencies requires no information about the displacements associated with each phonon mode and is particularly attractive for complex systems in which the phonon displacements are not known and for which it would be extremely expensive to compute the full matrix of second derivatives of the ionic potential energy—a calculation normally required to calculate phonon frequencies and eigenvectors.

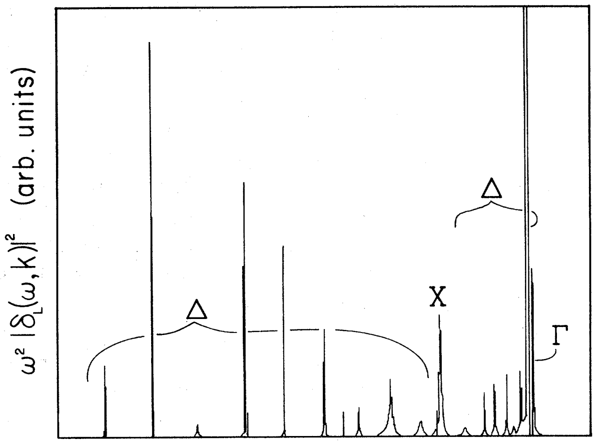

FIG. 25. Superposition of maximum-entropy-method spectral fits for each class of allowed phonon k state. For clarity, only the longitudinal spectra are displayed. The frequencies have been scaled so that the optic phonon frequency ω_0, as calculated from a frozen phonon calculation at the same cutoff in the same supercell, is normalized to unity.

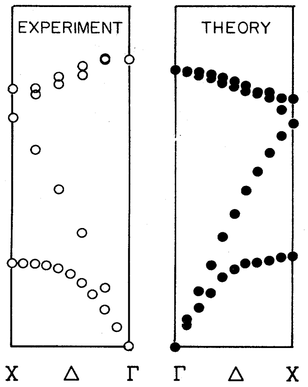

FIG. 26. Phonon spectrum as determined from maximum peak values of maximum-entropy-method fits. These values are completely *ab initio*, with no free parameters. Empty circles represent experimental data (Dolling, 1963; Nilsson and Nelin, 1972), and filled circles represent results of an extrapolated conjugate-gradient dynamics simulation.

IX. NONLOCAL PSEUDOPOTENTIALS

The computational speeds of the molecular dynamics and conjugate gradients techniques are significantly enhanced by using local pseudopotentials rather than nonlocal pseudopotentials. This allows the number of operations required to multiply each of the wave functions by the Hamiltonian to be reduced from N_{PW}^2 to $16N_{PW}\ln(N_{PW})$, where N_{PW} is the number of plane wave basis states. However, it is not possible to produce accurate local pseudopotentials for all atoms. To apply molecular-dynamics and conjugate-gradients methods to systems containing atoms that can only be represented by nonlocal pseudopotentials, it is necessary to use an efficient scheme for dealing with the nonlocality of the pseudopotential. Nonlocal pseudopotentials generally require fewer plane-wave basis states than do local pseudopotentials to expand the electronic wave functions. Therefore, although it will require additional computational effort to use nonlocal pseudopotentials in molecular-dynamics and conjugate-gradients calculations, some of the loss in computational speed will be recouped because fewer plane-wave basis states are required. However, it is essential to find an efficient method for using the nonlocal pseudopotentials. The methods that have been used employ only a partial projection of the nonlocal components of the wave functions. Examples of such methods are described in the following two sections. It has long been appreciated that all of these partial-projection methods could be applied in either real space or reciprocal space. The computational cost scales as the cube of the system size using a reciprocal-space method but only as the square of the system size with the real-space method. As will be seen in Sec. IX.C.1, there are difficulties associated with the real-space projection method that have delayed its implementation. These problems have now been overcome, and Sec. IX.D describes a successful real-space projection method for nonlocal pseudopotentials.

A. Kleinman-Bylander potentials

The most general form for a nonlocal pseudopotential is

$$V_{\text{ion}}=\sum_{lm}|Y_{lm}\rangle V_l\langle Y_{lm}| \,, \tag{9.1}$$

where Y_{lm} are spherical harmonics and V_l is the pseudopotential acting on the component of the wave function that has angular momentum l. Outside the core radius the potentials V_l are identical for all the angular momentum components of the wave function. To implement this form for the nonlocal pseudopotential, one needs a complete projection of the angular momentum components of the wave functions. In contrast, the Kleinman-Bylander pseudopotential (Kleinman and Bylander, 1982; Allan and Teter, 1987) is a norm-conserving pseudopotential that uses a single basis state for each angular momentum component of the wave function. The Kleinman-Bylander pseudopotential has the form

$$V_{\text{ion}}=V_{\text{LOC}}+\sum_{lm}\frac{|\phi^0_m\delta V_l\rangle\langle\delta V_l\phi^0_{lm}|}{\langle\phi^0_{lm}|\delta V_l|\phi^0_{lm}\rangle} \,, \tag{9.2}$$

where V_{LOC} is a local potential, ϕ^0_{lm} are the wave functions of the pseudoatom, and δV_l is

$$\delta V_l=V_{l,\text{NL}}-V_{\text{LOC}} \,. \tag{9.3}$$

Here $V_{l,\text{NL}}$ is the l angular momentum component of any nonlocal pseudopotential. Kleinman and Bylander suggested using the arbitrariness of V_{LOC} to produce an accurate and transferable pseudopotential.

The Kleinman-Bylander pseudopotential projects each spherical harmonic component of the wave function onto a single basis state. When applied to the pseudoatom, the potential gives identical results to the nonlocal pseudopotential it was derived from, independent of the choice for the local potential V_{LOC}. However, the potential does not produce identical results when applied in another environment, because the wave function is not projected onto a radially complete set of spherical harmonics. Some of the difficulties that can be encountered with this approach have been discussed recently by Gonze *et al.* (1990). All of the known problems can be overcome by the proper choice of local potential, a simple reduction in the core radius, or the application of the ideas of extended norm conservation (Shirley *et al.*, 1989). It may be necessary, however, to include pseudo core states to achieve a high degree of transferability for certain transition-metal atoms. These improvements all typically require a larger number of plane waves in the basis set.

1. Enhanced projections

The Kleinman-Bylander form of the pseudopotential can be systematically improved by adding more basis functions for the projection of the spherical harmonics (Bloechl, 1990). This allows the accuracy of the nonlocal potential to be checked by plotting the total energy as a function of the number of basis states used for the projection of the spherical harmonics of the wave functions.

2. Computational cost

The contribution to the product of the Hamiltonian and the wave function ψ_i at wave vector $\mathbf{k}+\mathbf{G}$ for the Kleinman-Bylander pseudopotential is given by

$$\sum_{lm}\left[\chi_{lm,\mathbf{k}+\mathbf{G}}\left[\sum_{\mathbf{G}'}\chi_{lm,\mathbf{k}+\mathbf{G}'}c_{i,\mathbf{k}+\mathbf{G}'}\right]\right] \,, \tag{9.4}$$

where

$$\chi_{lm,\mathbf{k}+\mathbf{G}}=\frac{\int r^2dr\, j_1(|\mathbf{k}+\mathbf{G}|r)\delta V_l(r)\phi^0_{lm}(r)}{[\langle\phi^0_{lm}|\delta V_l|\phi^0_{lm}\rangle]^{1/2}} \tag{9.5}$$

and j_l is the spherical Bessel function of order l. The spherical Bessel function $j_l(|\mathbf{k}+\mathbf{G}|r)$ gives the amplitude of the l angular momentum component of the plane wave $\exp[i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}]$ at distance r from the origin.

The contributions to the product of the Hamiltonian with each and every electronic wave function from the nonlocal pseudopotential can be calculated in $3N_BN_{PW}N_IN_L$ operations per $\mathbf{k}$ point, using the Kleinman-Bylander pseudopotential, where N_L is the number of nonlocal spherical harmonic components in the pseudopotential. $3N_BN_{PW}N_IN_L$ operations are required to calculate the contributions to the forces on all the ions from the nonlocal pseudopotential using the Kleinman-Bylander scheme in reciprocal space, and $6N_BN_{PW}N_IN_L$ operations are required to compute the diagonal and off-diagonal stresses on the unit cell. Hence the computational cost of all these operations scale as the third power of the number of ions in the unit cell, since N_B and N_{PW} are proportional to N_I. This computational cost is usually significantly larger than the cost of orthogonalizing the wave functions $(N_B^2N_{PW})$. Therefore the application of the nonlocal potential in reciprocal space will dominate the computational cost for large systems.

B. Vanderbilt potentials

A rather more radical approach to modifying pseudopotentials for use in plane-wave calculations has been suggested by Vanderbilt (1990). The basic aim with these potentials, in common with the other schemes described in Sec. II.D.1.d, is to allow calculations to be performed with as low a cutoff energy for the plane-wave basis set as possible. The rationale behind the Vanderbilt potential is that in most cases a high cutoff energy is required for the plane-wave basis set only when there are tightly bound orbitals that have a substantial fraction of their weight inside the core region of the atom. In this case the cutoff energy for the plane-wave basis set cannot be substantially reduced, because there must be plane-wave components up to a large enough wave vector to allow the majority of the weight of the wave function to be kept within the core. However, if the norm conservation rule is relaxed, then the resulting wave function can be expanded using a much smaller plane-wave basis set, as shown in Fig. 27. All that is required is that the wave function be projected back to the correct pseudovalence wave function before normalization. Unfortunately, the procedure is rather more complex, because the relaxation of the norm conservation condition from the pseudopotential *also* causes the correct first-order change of the phase shift with energy to be lost. Therefore this scheme also requires an energy-dependent potential to ensure that the correct phase shift is generated over the range of energies of the electrons in the system. Fortunately, this modification can be included at a relatively modest computational cost in any iterative method, although it would be disastrous in a conventional matrix diagonalization method, since each matrix diagonalization would yield only a single band. Details of the implementation of Vanderbilt potentials can be found in Vanderbilt (1990) and Laasonen *et al* (1991).

Although Vanderbilt potentials require lower cutoff energies for the plane-wave basis set than even optimized pseudopotentials, they require more operations to compute the nonlocal components of the pseudopotential at each iteration. It is also possible that iterating to energy self-consistency in the potential may increase the total time required for an energy minimization calculation or require an even shorter time step in a Car-Parrinello dynamical calculation. If the operations for the nonlocal pseudopotential are carried out in reciprocal space in a large system, where these operations dominate the computational cost, it is not clear that a calculation using Vanderbilt potentials will be any cheaper than a calculation using Kleinman-Bylander pseudopotentials. If the real-space projection technique described in Secs. IX.C and IX.D below is used, so that operations required to implement the nonlocal pseudopotential no longer dominate the computational cost, then a Vanderbilt potential will be more efficient.

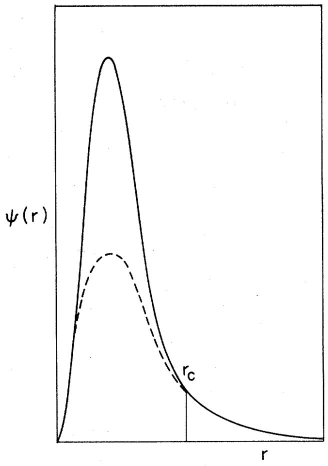

FIG. 27. Illustration of a pseudo wave function that is strongly peaked inside the core and the modified wave function in Vanderbilt's scheme.

C. Real-space projection; nonuniqueness of spherical harmonic projection

The nonlocality of the pseudopotential extends only over the region occupied by the core of the atom. As the core region is relatively small, it should be possible to deal efficiently with the nonlocality of the pseudopotential by working in real space, since only a small number of operations should be required to project the angular momentum components of each wave function in the core of each atom. Furthermore, the number of operations needed to project the angular momentum components of a single wave function around a single atom in real space will be independent of the size of the system, thus leading to a more efficient scaling than the reciprocal-space projection. If N_P is the number of points in the core of each atom used to project each angular momentum component of a single wave function, then the number of operations required to incorporate a nonlocal pseudopotential using a real-space method is $N_B N_I N_P N_L$ per **k** point. The electronic wave functions are routinely transformed to real space in the molecular-dynamics and conjugate-gradients methods. No further operations besides those described above are required to implement a real-space projection of the angular momentum components provided that the product of the wave function and the nonlocal potential is computed at the point in the calculation where the product of the wave function and the local potential is computed. The number of operations required to calculate the forces on the ions is $3N_B N_I N_P N_L$ per **k** point using the real-space projection method; $6N_B N_I N_P N_L$ operations are required to compute the diagonal and off-diagonal stresses on the unit cell. However, it may also be necessary to perform an additional FFT to generate the wave functions in real space before performing these operations.

The cost of computing the product of the Hamiltonian and the wave functions, the forces on the atoms, and the stresses on the unit cell using the real-space projection method scales only as the second power of the number of ions in the system. This is in contrast to cubic scaling for the reciprocal-space projection methods. The reason is that, in a reciprocal-space formulation, computation of the force on an ion requires a sum over all reciprocal-lattice vectors. In contrast, calculation of the force on an ion using the real-space method requires only operations involving the wave function in the immediate vicinity of the atom.

As we shall see shortly, the real-space projection method for nonlocal pseudopotentials is not simple to implement using a representation of the wave function on the existing Fourier transform grid because of its coarseness. Calculation of the wave functions at a specialized dense set of points near each atom (ruling out the FFT) results in the same scaling as the reciprocal-space methods. Alternatively, a fast Fourier transform of the wave functions onto a denser real-space *grid* preserves the favorable scaling of the real-space methods. However, a scheme that allowed computation on the original grid would be most efficient. Such a scheme exists and will be described below.

If fast Fourier transform techniques are used to transform the electronic wave functions between real and reciprocal space, the values of the electronic wave functions in real space are known only on a grid of points. The incomplete knowledge of the wave function in real space introduces an ambiguity between the angular and radial dependences of the wave function. This can be illustrated with a simple example as shown in Fig. 28. Consider a real-space Fourier transform grid that has equal distances between grid points in the x and y directions but a distance between grid points in the z direction that is 50% greater. Suppose an atom is positioned at one of the grid points and the value of the wave function is 0 at the position of the atom, 1.0 at the adjacent grid points in the $\pm x$ and $\pm y$ directions, and 2.25 at the adjacent grid points in the $\pm z$ directions. The wave function at these points could be an "s" wave function equal to $1.0(r^2/a^2)$, where r is the radial coordinate and a is the distance between grid points in the x direction. However, the values of the wave function at the points can also be fit by the sum of an "s" wave function equal to $\frac{7}{6}(r/a)$ and a "d_{zz}" wave function equal to $\frac{1}{6}(r/a)[3\cos^2\theta-1]$, where θ is the angle between the radius vector and the z axis. The distinction between these two wave function can only be made by considering the value of the wave function at points lying between the Fourier transform grid points. However, it is computationally expensive to calculate the value of the wave function at arbitrary points of a Fourier transform grid. If the real-space projections cannot be performed from the Fourier transform grid, then the real-space method will be too costly to implement directly.

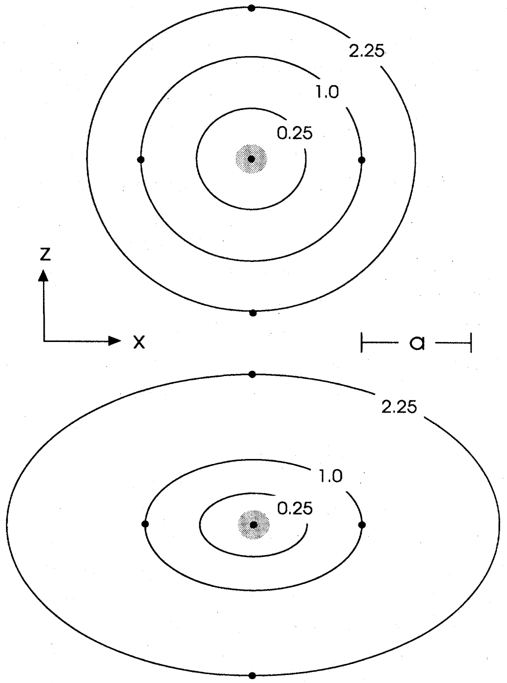

FIG. 28. Illustration of nonuniqueness of spherical harmonic projection, as discussed in the text. Top, an s state equal to $1.0(r^2/a^2)$. Bottom, a mixed state equal to $\frac{7}{6}(r/a)+\frac{1}{6}(r/a)(3\cos^2\theta-1)$. The grey discs represent the ion positions.

The difficulty highlighted above rises because the discrete Fourier transform grid is poorly suited to projecting the angular momentum components of the wave functions. The problem of ambiguity between the radial and angular dependences of the wave functions becomes even more difficult to resolve if the atom is not positioned at a grid point or halfway between grid points, if the distances between the grid points is different along all three of the cell axes, or if the unit cell axes are not perpendicular to each other.

D. Potentials for real-space projection

Some of the problems outlined above become somewhat less severe if, instead of insisting on a complete projection of the spherical harmonic components of the wave functions around each ion, one performs only a partial projection as in the Kleinman-Bylander scheme. However, the scheme outlined below can be applied to any number of projections and hence allows the spherical harmonic components of the wave functions to be projected to any degree of accuracy (as in the enhanced projection schemes mentioned above). Let us consider Kleinman-Bylander projectors for the present.

An alternative way to understand the difficulties associated with the real-space projection technique and thus identify a possible solution is by considering the properties of the Fourier transform grid. If a double-density Fourier transform grid (see Sec. III.K.1) is used in the calculation, the reciprocal-space Fourier transform grid is of length $4G_{\max}$ along each reciprocal-lattice vector, where $G_{\max}$ is the cutoff wave vector for the electronic wave functions. On the corresponding real-space Fourier transform grid, the phase of a plane wave of wave vector $4G_{\max}$ changes by a factor of 2π between each grid point, and so the values of this plane wave on the grid points are indistinguishable from those generated by a plane wave with wave vector $G=0$ or any other integer multiple of $4G_{\max}$. This is a general result; any function defined only on the real-space grid points is unchanged if any of its Fourier components are changed by any multiple of the wave vector $4G_{\max}$. This is the origin of the so-called "wraparound error" of discrete representations of Fourier transforms. The use of a double-density Fourier transform grid eliminates the wraparound error in all the parts of a pseudopotential calculation considered so far. Unfortunately, this is not true in the case of a real-space projection of the nonlocal Kleinman-Bylander projectors because these projectors must have Fourier components at all wave vectors if they are to be strictly localized in the core of the atom. If the Fourier transform grid moves with respect to a fixed ionic and electronic configuration, there will be interference between the components of the Kleinman-Bylander projectors at wave vectors $G+n4G_{\max}$ (n integer), and the value of the Kleinman-Bylander matrix elements will vary. For a fixed ionic and electronic configuration, these matrix elements must be independent of the origin of the Fourier transform grid for their values to be physically meaningful, and without a solution of this problem the real-space projection technique cannot be applied.

It should be remembered that the local part of the pseudopotential will also have components at large wave vectors. However, the real-space representation of this potential includes only wave vectors up to $2G_{\max}$, since this representation is generated by Fourier-transforming the potential from reciprocal space, where only components of the potential up to wave vector $2G_{\max}$ are represented on the double-density Fourier transform grid. This analogy provides an immediate solution to the problem associated with the real-space Kleinman-Bylander projectors. Rather than using the actual real-space Kleinman-Bylander projectors, we should use projectors that have been Fourier filtered so that they do not contain any components at wave vectors larger than a wave vector G_P, which must be less than $4G_{\max}$. If this is the case, these modified projectors will not be subject to any wraparound error.

The cost of forcing the high Fourier components of the modified Kleinman-Bylander projectors to be strictly zero at wave vectors above G_P is that in real space the projectors are no longer localized in the core, but are nonzero over the whole of the real-space grid. Obviously, if the projectors extend over the whole of the real-space grid, there is nothing to be gained from the real-space projection technique. The trick is to use the arbitrariness of the modified Kleinman-Bylander projectors over the range of wave vectors $G_{\max}<G<G_P$ to ensure that the magnitude of the modified projectors is negligible at distances greater than roughly $2r_c$ from the ion core, where r_c is the core radius of the pseudopotential. Of course, the Fourier components of the projectors at wave vectors less than $G_{\max}$ must not be changed, or the real- and reciprocal-space Kleinman-Bylander projections will not be identical. The Kleinman-Bylander projections are carried out only over the grid points where the projectors are non-negligible, thus restoring the favorable scaling of the real-space projection technique. The procedure for generating the modified projectors is detailed in King-Smith *et al.* (1991). The test of the generation technique is simple. The modified real-space projectors must yield identical results to the reciprocal-space Kleinman-Bylander projectors, irrespective of the relative positions of the ion and the Fourier transform grid, a condition that the old, unmodified real-space Kleinman-Bylander projectors failed to meet.

X. PARALLEL COMPUTERS

The series of algorithmic improvements described in this article yield a method for performing total-energy pseudopotential calculations whose computational time requirements scale essentially as a constant times the theoretical minimum number of operations required to update all the wave functions in a single iteration. (This can *never* be reduced below the cost of orthogonalizing the wave functions without introducing an error.) The value of this constant depends on the system but always lies between several tens and several hundreds. Although there may be some possibility of reducing this constant, it is clear that, since this constant can never be less than one and is more likely to be of the order of ten, there are ultimate limits to the gains to be achieved by improvements in the numerical methods. There is certainly no longer the possibility of increases in computational speed by many orders of magnitude, of the sort that have been gained in the last few years, without fundamentally changing the essential features of the total-energy pseudopotential method. This does not, of course, exclude the possibility of a completely different method proving to be more efficient.

The developments in the total-energy pseudopotential method allow calculations to be performed on any reasonably powerful computer (anything from a modern workstation to a conventional supercomputer) for quite large systems containing up to 150 atoms in the unit cell, provided that the pseudopotentials are moderately weak. However, to allow studies of significantly larger systems, much more powerful computers are required, which combine both faster processing speeds with extremely large amounts of memory. Without a fundamental change in computer technology (the speed of each processor of a conventional supercomputer has changed by significantly less than one order of magnitude in the last decade), these requirements can only be fulfilled by combining a number of processors into a "parallel" computer. In principle, by combining enough compute nodes, parallel computers can be constructed that achieve arbitrarily large numbers of operations per second. No parallel computer will achieve 100% efficiency on a real computation, since there are overheads associated with communication between the processors, and achievement of a significant fraction of full efficiency is quite difficult on any but the most trivial of tasks. A relatively low efficiency is not too great a cause for concern, as long as it is maintained as the number of compute nodes is increased. In this case, any required computational speed is achievable with a large enough number of processors—even if this computational speed is significantly less than the theoretical speed of that number of processors. The real problem, however, is that in all applications the efficiency falls steadily above a critical number of processors. This effect is associated with the parts of the code that cannot be run in parallel. At the critical number of processors this part of the calculation starts to dominate the computational time, and no further reduction in computational time is possible by increasing the number of processors. In this regime the efficiency varies as the inverse of the number of compute nodes! It is clear that, for any calculation to be run efficiently and "scaleably" (i.e., so that computational time decreases with number of processors), on a parallel machine containing n processors not more than $1/n$ of the entire calculation may run sequentially.

There are many possible architectures for parallel machines. Each tends to have its own strengths and weaknesses and, more importantly, its own suitability for any particular computation. Interestingly total-energy pseudopotential calculations have been successfully implemented on two very different classes of parallel machine. One, the Connection Machine, consists of an extremely large number of relatively modest-performance compute nodes. The other class of machine consists of a smaller number of extremely powerful compute nodes; examples of this latter class of machine are those manufactured by Intel, Meiko, and N-Cube. Although the strategies for implementing the codes is the same on both classes of machine, the detailed methods required to implement the codes are rather different. The Connection Machine is programmed using a standard high-level computer language, Fortran 90. When the total-energy pseudopotential calculation is expressed using vector-oriented Fortran 90 statements, parallel execution is implemented by the compiler. The programmer is not required explicitly to implement communications among the thousands of processors constituting the fine-grained, massively parallel architecture. To further accelerate processing, the vendor supplies a library of hand-microcoded FFT subroutines, which are directly callable from Fortran 90 programs. The combination of programming in Fortran 90 and the use of the distributed machine FFT makes it possible for most of the total-energy calculation to be performed in parallel.

In the case of the other class of machine, there is normally no fully distributed three-dimensional FFT, and a major task in implementing total-energy pseudopotential codes on these machines is the implementation of the communications required to perform a distributed FFT. A full description of the implementation of a set of pseudopotential codes on this class of machine can be found in Clarke *et al.* (1992).

The potential for pseudopotential calculations on parallel computers has been demonstrated by the successful calculations of the surface energy and relaxed structure of the 7×7 Takayanagi reconstruction (Takayanagi *et al.*, 1985) of the (111) surface of silicon on a 64-node Meiko machine (Stich *et al.*, 1992) and on a Connection Machine (Brommer *et al.*, 1992). The supercells used for these calculations contained 400 atoms and had a volume of 784 times the atomic volume. Basis sets of up to 35 000 plane waves were used to expand the electronic wave functions, and the electronic minimization involved up to 2.8×10^8 degrees of freedom.

XI. CONCLUDING REMARKS

Car and Parrinello's molecular-dynamics method stands as a landmark in the field of total-energy pseudopotential calculations. Iterative matrix diagonalization techniques were in use before the molecular-dynamics method was developed, but these schemes only partially exploited the benefits to be gained by the use of iterative techniques. Car and Parrinello's technique exploited the advantages of overlapping the processes of calculating eigenstates and iterating to self-consistency and the use of fast Fourier transform techniques. Efficient schemes for using nonlocal pseudopotentials were only implemented as a result of the molecular-dynamics method. The combination of all of these features rather than just the replacement of a conventional matrix diagonalization scheme by an iterative scheme is responsible for the significant increase in the power of the total-energy pseudopotential technique brought about by the molecular-dynamics method. Any iterative matrix diagonalization technique must exploit all of these features to be able to challenge the molecular-dynamics method. Car and Parrinello's method did not change the basic total-energy pseudopotential technique but offered an enormous increase in computational efficiency, so that much larger and more complex systems became accessible to the technique. It also allowed the first *ab initio* dynamical simulations to be performed. Only with the reduction in the scaling of computational time with system size that comes from Car and Parrinello's molecular-dynamics method and from conjugate-gradients techniques is it worthwhile performing calculations for large systems on parallel computers.

As mentioned previously, there is now only limited scope for improvements in the algorithms used to perform total-energy pseudopotential calculations. However, there are a number of areas where progress can still be made. At present a definitive scheme for generating pseudopotentials that are fully transferable and computationally efficient is still lacking. There is considerable scope for improved density functionals. Perhaps the most ambitious objective is to couple quantum-mechanical modeling of small, critical regions of a system (such as a dislocation core) with a less rigorous modeling of the noncritical regions (which could be modeled using classical elasticity theory).

Even if all else is forgotten, the authors would like the reader to retain just one idea. This is that *ab initio* quantum-mechanical modeling using the total-energy pseudopotential technique is now capable of addressing an extremely large range of problems in a wide range of scientific disciplines.

ACKNOWLEDGMENTS

The authors wish to thank Y. Bar-Yam, E. Kaxiras, M. Needels, K. M. Rabe, E. Tarnow, and D. H. Vanderbilt for help and advice in developing, applying, and understanding the molecular-dynamics and conjugate-gradients methods. MCP wishes to thank Volker Heine for many insightful discussions on the pseudopotential approach and R. J. Needs, R. W. Godby, and C. M. M. Nex for helpful discussions on iterative matrix diagonalization techniques.

REFERENCES

Alan, D. C., and M. P. Teter, 1987, Phys. Rev. Lett. **59**, 1136.

Allen, M. P., and D. J. Tildesley, 1987, *Computer Simulation of Liquids* (Clarendon, Oxford).

Arias, T., J. D. Joannopoulos, and M. C. Payne, 1991, Phys. Rev. B, in press.

Ashcroft, N. W., and N. D. Mermin, 1976, *Solid State Physics* (Holt Saunders, Philadelphia), p. 113.

Bachelet, G. B., D. R. Hamann, and M. Schlüter, 1982, Phys. Rev. B **26**, 4199.

Baraff, G. A., and M. A. Schluter, 1979, Phys. Rev. B **19**, 4965.

Bar-Yam, Y., S. T. Pantelides, and J. D. Joannopoulos, 1989, Phys. Rev. B **39**, 3396.

Benedek, R., L. H. Yang, C. Woodward, and B. I. Min, 1991, unpublished.

Bloechl, P. E., 1990, Phys. Rev. B **41**, 5414.

Bloechl, P. E., and M. Parrinello, 1991, unpublished.

Brommer, K., M. Needels, B. Larson, and J. D. Joannopoulos, 1992, Phys. Rev. Lett. **68**, 1355.

Broughton, J., and F. Khan, 1989, Phys. Rev. B **40**, 12 098.

Car, R., and M. Parrinello, 1985, Phys. Rev. Lett. **55**, 2471.

Car, R., and M. Parrinello, 1989, in *Simple Molecular Systems at Very High Density,* edited by A. Polian, P. Lebouyre, and N. Boccara (Plenum, New York), p. 455.

Car, R., M. Parrinello, and M. C. Payne, 1991, J. Phys.: Condens. Matter **3**, 9539.

Chadi, D. J., and M. L. Cohen, 1973, Phys. Rev. B **8**, 5747.

Cho, K., and J. D. Joannopoulos, 1992, Phys. Rev. A **45**, 7089.

Clarke, L. J., I. Stich, and M. C. Payne, 1992, unpublished.

Cohen, M. L., 1984, Phys. Rep. **110**, 293.

Cohen, M. L., and V. Heine, 1970, Solid State Physics Vol. 24, p. 37.

Denteneer, P., and W. van Haeringen, 1985, J. Phys. C **18**, 4127.

Dolling, G., 1963, in *Inelastic Scattering of Neutrons in Solids and Liquids Vol. II* (International Atomic Energy Agency, Vienna), p. 37.

Drabold, D. A., P. A. Fedders, Otto F. Sankey, and John D. Dow, 1990, Phys. Rev. B **42**, 5135.

Dreizler, R. M., and J. da Providencia, 1985, *Density Functional Methods in Physics* (Plenum, New York).

Evarestov, R. A., and V. P. Smirnov, 1983, Phys. Status Solidi **119**, 9.

Ewald, P. P., 1917a, Ann. Phys. (Leipzig) **54**, 519.

Ewald, P. P., 1917b, Ann. Phys. (Leipzig) **54**, 557.

Ewald, P. P., 1921, Ann. Phys. (Leipzig) **64**, 253.

Fahy, S., X. W. Wang, and S. G. Louie, 1988, Phys. Rev. Lett. **61**, 1631.

Fernando, G. W., Guo-Xin Qian, M. Weinert, and J. W. Davenport, 1989, Phys. Rev. B **40**, 7985.

Fetter, A. L., and J. D. Walecka, 1971, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York), p. 29.

Feynman, R. P., 1939, Phys. Rev. **56**, 340.

Francis, G. P., and M. C. Payne, 1990, J. Phys.: Condens. Matter **17**, 1643.

Froyen, S., and M. L. Cohen, 1986, J. Phys. C **19**, 2623.

Gill, P. E., W. Murray, and M. H. Wright, 1981, *Practical Optimization* (Academic, London).

Gillan, M. J., 1989, J. Phys.: Condens. Matter **1**, 689.

Godby, R. W., M. Schluter, and L. J. Sham, 1986, Phys. Rev. Lett. **56**, 2415.

Gomes Dacosta, P., O. H. NIelsen, and K. Kunc, 1986, J. Phys. C **19**, 3163.

Gonze, X., P. Kackell, and M. Scheffler, 1990, Phys. Rev. B **42**, 12 264.

Gunnarsson, O., and B. I. Lundqvist, 1976, Phys. Rev. B **13**, 4174.

Hamann, D. R., 1989, Phys. Rev. B **40**, 2980.

Hamann, D. R., M. Schlüter, and C. Chiang, 1979, Phys. Rev. Lett. **43**, 1494.

Harris, J., and R. O. Jones, 1974, J. Phys. F **4**, 1170.

Hedin, L., and B. Lundqvist, 1971, J. Phys. C **4**, 2064.

Hellmann, H., 1937, *Einfuhrung in die Quantumchemie* (Deuticke, Leipzig).

Hohenberg, P., and W. Kohn, 1964, Phys. Rev. **136**, 864B.

Hybertson, M. S., and S. G. Louie, 1985, Phys. Rev. Lett. **55**, 1418.

Ihm, J., A. Zunger, and M. L. Cohen, 1979, J. Phys. C **12**, 4409.

Janak, J. F., 1978, Phys. Rev. B **18**, 7165.

Joannopoulos, J. D., 1985, in *Physics of Disordered Materials,* edited by D. Adler, H. Fritzsche, and S. R. Ovshinsky (Plenum, New York), p. 19.

Joannopoulos, J. D., P. Bash, and A. Rappe, 1991, Chemical Design Automation News **6**, No. 8.

Joannopoulos, J. D., and M. L. Cohen, 1973, J. Phys. C **6**, 1572.

Joannopoulos, J. D., T. Starkloff, and M. A. Kastner, 1977, Phys. Rev. Lett. **38**, 660.

Jones, 1991, Phys. Rev. Lett. **67**, 224.

Jones, R. O., and O. Gunnarsson, 1989, Rev. Mod. Phys. **61**, 689.

Kerker, G., 1980, J. Phys. C **13**, L189.

King-Smith, R. D., M. C. Payne, and J-S. Lin, 1991, Phys. Rev. B **44**, 13 063.

Kirkpatrick, S., C. Gelatt, Jr., and M. Vecchi, 1983, Science **220**, 671.

Kleinman, L., and D. M. Bylander, 1982, Phys. Rev. Lett. **4**, 1425.

Kohn, W., and L. J. Sham, 1965, Phys. Rev. **140**, 1133A.

Kryachko, E. S., and E. V. Ludena, 1990, *Energy Density Functional Theory of Many Electron Systems* (Kluwer Academic, Boston).

Laasonen, K., R. Car, C. Lee, and D. Vanderbilt, 1991, Phys.

Rev. B **43**, 6796.

Langreth, D. C., and M. J. Mehl, 1981, Phys. Rev. Lett. **47**, 446.

Langreth, D. C., and M. J. Mehl, 1983, Phys. Rev. B **28**, 1809.

Langreth, D. C., and J. P. Perdew, 1977, Phys. Rev. B **15**, 2884.

Li, X. P., D. M. Ceperley, and Richard M. Martin, 1991, Phys. Rev. B **44**, 10929.

Lin, C. C., and L. A. Segel, 1974, *Mathematics Applied to Deterministic Problems in the Natural Sciences* (Macmillan, New York), p. 161.

Mathews, J., and R. L. Walker, 1970, *Mathematical Methods of Physics* (Benjamin, New York), p. 355.

Mermin, N. D., 1965, Phys. Rev. **137**, 1441A.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953, J. Chem. Phys. **21**, 1087.

Monkhorst, H. J., and J. D. Pack, 1976, Phys. Rev. B **13**, 5188.

Nilsson, G., and G. Nelin, 1972, Phys. Rev. B **6**, 3777.

Nosé, S., 1984, J. Chem. Phys. **81**, 511.

Pastore, G., E. Smargiassi, and F. Buda, 1991, Phys. Rev. A **44**, 6334.

Payne, M. C., 1989, J. Phys.: Condens. Matter **1**, 2199.

Payne, M. C., J. D. Joannopoulos, D. C. Allan, M. P. Teter, and D. H. Vanderbilt, 1986, Phys. Rev. Lett. **56**, 2656.

Payne, M. C., M. Needels, and J. D. Joannopoulos, 1988, Phys. Rev. B **37**, 8138.

Pederson, M. R., and K. A. Jackson, 1991, Phys. Rev. B **43**, 7312.

Perdew, J. P., and A. Zunger, 1981, Phys. Rev. B **23**, 5048.

Perdew, J. P., Robert G. Parr, Mel Levy, and Jose L. Balduz, Jr., 1982, Phys. Rev. Lett. **49**, 1691.

Phillips, J. C., 1958, Phys. Rev. **112**, 685.

Pickett, W., 1989, Comput. Phys. Rep. **9**, 115.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterlin, 1989, in *Numerical Recipes: The Art of Scientific Computing* (Cambridge University, Cambridge, England), p. 431.

Pulay, P., 1969, Mol. Phys. **17**, 197.

Qian, Guo-Xin, M. Weinert, G. W. Fernando, and J. W. Davenport, 1990, Phys. Rev. Lett. **64**, 1146.

Rabe, K., and J. D. Joannopoulos, 1987, Phys. Rev. B **36**, 6631.

Rappe, A., and J. D. Joannopoulos, 1991, in *Computer Simulation in Materials Science,* edited by M. Meyer and V. Pontikis, NATO ASI Vol. 205, p. 409.

Rappe, A., K. Rabe, E. Kaxiras, and J. D. Joannopoulos, 1990, Phys. Rev. B **41**, 1227.

Redondo, A., W. A. Goddard III, and T. C. McGill, 1977, Phys. Rev. B **15**, 5038.

Remler, D. K., and P. A. Madden, 1990, Mol. Phys. **70**, 921.

Robertson, I. J., and M. C. Payne, 1990, J. Phys.: Condens. Matter **2**, 9837.

Robertson, I. J., and M. C. Payne, 1991, J. Phys.: Condens. Matter **3**, 8841.

Scheffler, M., J. P. Vigneron, and G. B. Bachelet, 1985, Phys. Rev. B **31**, 6541.

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

Starkloff, T., and J. D. Joannopoulos, 1977, Phys. Rev. B **16**, 5212.

Stich, I., R. Car, M. Parrinello, and S. Baroni, 1989, Phys. Rev. B **39**, 4997.

Stich, I., M. C. Payne, R. D. King-Smith, J-S. Lin, and L. J. Clarke, 1992, Phys. Rev. Lett. **68**, 1351.

Takayanagi, K., Y. Tanashiro, S. Takahashi, and M. Takahashi, 1985, Surf. Sci. **164**, 367.

Teter, M. P., M. C. Payne, and D. C. Allan, 1989, Phys. Rev. B **40**, 12255.

Toxvaerd, Soren, 1991, Mol. Phys. **72**, 159.

Trouillier, N., and J. L. Martins, 1991, Phys. Rev. B **43**, 1993.

Vanderbilt, D., 1987, Phys. Rev. Lett. **59**, 1456.

Vanderbilt, D., 1990, Phys. Rev. B **41**, 7892.

Verlet, L., 1967, Phys. Rev. **159**, 98.

von Barth, U., 1984, in *Many Body Phenomena at Surfaces,* edited by D. Langreth and H. Suhl (Academic, New York), p. 3.

Vosko, S. H., L. Wilk, and M. Nusair, 1980, Can. J. Phys. **58**, 1200.

Wigner, E. P., 1938, Trans. Faraday Soc. **34**, 678.

Williams, A., and J. Soler, 1987, Bull. Am. Phys. Soc. **32**, 562.

Woodward, C., B. I. Min, R. Benedek, and J. Garner, 1989, Phys. Rev. B **39**, 4853.

Yin, M. T., and M. L. Cohen, 1982a, Phys. Rev. B **25**, 7403.

Yin, M. T., and M. L. Cohen, 1982b, Phys. Rev. B **26**, 5668.

Zhang, Q.-M., G. Chiarotti, A. Selloni, R. Car, and M. Parrinello, 1990, Phys. Rev. B **42**, 5071.

Zunger, A., and M. L. Cohen, 1979, Phys. Rev. B **20**, 4082.