Linear-scaling density-functional-theory technique: The density-matrix approach

E. Hernández and M. J. Gillan
Physics Department, Keele University, Keele, Staffordshire ST5 5BG, United Kingdom

C. M. Goringe
Materials Department, Oxford University, Oxford OX1 3PH, United Kingdom
(Received 15 August 1995; revised manuscript received 20 November 1995)

A recently proposed linear-scaling scheme for density-functional pseudopotential calculations is described in detail. The method is based on a formulation of density-functional theory in which the ground-state energy is determined by minimization with respect to the density matrix, subject to the condition that the eigenvalues of the latter lie in the range [0,1]. Linear-scaling behavior is achieved by requiring that the density matrix should vanish when the separation of its arguments exceeds a chosen cutoff. The limitation on the eigenvalue range is imposed by the method of Li, Nunes, and Vanderbilt. The scheme is implemented by calculating all terms in the energy on a uniform real-space grid, and minimization is performed using the conjugate-gradient method. Tests on a 512-atom Si system show that the total energy converges rapidly as the range of the density matrix is increased. A discussion of the relation between the present method and other linear-scaling methods is given, and some problems that still require solution are indicated.

I. INTRODUCTION

During the last decade, first-principles total-energy methods based on density-functional theory (DFT) combined with the pseudopotential method have become established as a major tool in the study of condensed matter.[1] The DFT pseudopotential approach is now widely used for both static and dynamic simulations on an enormous range of condensed-matter problems. However, these methods suffer from a severe drawback in that their computational cost generally increases as the cube of the number of atoms in the system. This unfavorable scaling limits the size of systems that can be studied with current methods and today's computers to a few hundred atoms at most. This $O(N^3)$ scaling appears in spite of the fact that the complexity of the problem increases only linearly with the system size. This observation suggests that the unfavorable scaling of current methods is a consequence of the way in which the electronic structure problem is being addressed. Conventional methods rely either on diagonalization of the Hamiltonian or orthonormalization of a set of occupied orbitals, both of which are intrinsically $O(N^3)$ operations. It is clear that more efficient methods in which the effort is proportional to the number of atoms must be possible, and in recent years a considerable effort has been devoted to finding such ''linear-scaling'' schemes.[2–20]

The earliest linear-scaling scheme appears to be the ''divide and conquer'' method of Yang.[2,3] This obtains the electronic density and hence the total energy by dividing the system into overlapping subsystems that can be treated independently. The density is calculated for each subsystem with conventional linear combination of atomic orbitals DFT. The Hamiltonian for each subsystem, which includes the potential due to the other subsystems, is diagonalized independently, thus avoiding the need to diagonalize the full Hamiltonian. This procedure is repeated until self-consistency is achieved. The divide-and-conquer strategy is being successfully applied to study the electronic structure of large molecular systems.[4] Baroni and Giannozzi[5] also proposed a scheme that directly determines the electron density. They do this by discretizing the Kohn-Sham Hamiltonian on a real-space grid, and then using the recursion method of Haydock, Heine and Kelly[21] to obtain the diagonal elements of the Green's function, from which the electron density can be computed by contour integration. In this case linear scaling results from the fact that the continued fraction used to evaluate a particular diagonal element of the Green's function can be truncated once a certain neighborhood of each point has been explored. This neighborhood is independent of the system size for sufficiently large systems.

More recently, several new schemes that resemble traditional first-principles methods have been reported. Galli and Parrinello[6] pointed out that some improvement could be achieved in the scaling of a conventional DFT calculation by requiring spatial localization of the electronic orbitals. This localization was achieved by adding certain nonlocal constraining terms to the Hamiltonian, or by using a filtering procedure. The total energy can then be obtained as a functional of the localized orbitals $|\phi_i\rangle$ and their *conjugate* orbitals $|\bar{\phi}_i\rangle = \Sigma_j S^{-1}_{ji} |\phi_j\rangle$, but in order to obtain these conjugate orbitals, the overlap matrix S has to be inverted. Since spatial localization implies sparsity of S, this can be achieved in $O(N^2)$ operations, so that some improvement with respect to $O(N^3)$ is obtained. A step further in this direction was made independently by Mauri, Galli, and Car[7,8] (hereafter referred to as MGC) and by Ordejón *et al.*[9,10] They introduced a new functional of the occupied orbitals that possesses the same ground state as the conventional energy functional, but with the added advantage of leading naturally to orthogonal orbitals when minimized. If this new functional is minimized with respect to orbitals that are constrained to remain localized in chosen regions of space, as suggested by Galli and Parrinello,[6] a linear scaling method results. In the original formulation, the number of orbitals entering the new func-

tional is equal to half the number of electrons in the system. This restriction seems to lead to very slow convergence, and to the appearance of spurious local minima in the functional. This problem has been recently overcome by Kim, Mauri, and Galli,[11] by generalizing the functional so that it depends on an arbitrary number of orbitals.

The linear-scaling scheme most relevant to the present work is that put forward by Li, Nunes, and Vanderbilt[12] (hereafter referred to as LNV) in the context of tight-binding semiempirical calculations. In this method, linear scaling is achieved by taking advantage of the real-space localization properties of the density matrix, $\rho(\mathbf{r},\mathbf{r}')$. By introducing a spatial cutoff R_c in ρ, such that $\rho(\mathbf{r},\mathbf{r}')$ is set to zero if $|\mathbf{r}-\mathbf{r}'|\geqslant R_c$, the number of nonzero elements in ρ increases only linearly with the system size. The electronic structure problem is then formulated as a minimization of the total energy with respect to the truncated density matrix, subject to the constraints of idempotency ($\rho^2=\rho$) and correct trace ($2\mathrm{Tr}\rho=N_e$, where N_e is the number of electrons). The scheme of LNV consists of an algorithm for imposing these constraints that at the same time fulfills the goal of linear scaling. The idempotency of ρ is the most difficult constraint to impose, and this scheme achieves it by expressing ρ in terms of an auxiliary matrix, which we denote in this paper by σ. This is subjected to a *purifying* transformation due to McWeeny.[22] If σ is a near-idempotent matrix, i.e., if its eigenvalues lie close to 0 or 1, this transformation will return ρ as a more nearly idempotent matrix, and thus it is possible to minimize the total energy with respect to σ while ensuring the near idempotency of ρ. By construction, the method is variational [i.e., $\min E(R_c)\geqslant\min E(\infty)$], and it has been shown that the convergence of calculated properties with the parameter R_c is fairly rapid.[12,13] It is now being widely used in tight-binding simulations of large systems.

Recently, the idea of working with the density matrix has been applied to DFT linear scaling schemes. This has been done independently by Hierse and Stechel[14] and by Hernández and Gillan.[15] In both cases, the density matrix is represented in real space as

$$\rho(\mathbf{r},\mathbf{r}')=\sum_{\alpha\beta}\phi_\alpha(\mathbf{r})K_{\alpha\beta}\phi_\beta^*(\mathbf{r}'), \tag{1}$$

where the ϕ_α are a set of localized functions, and $K_{\alpha\beta}$ is a symmetric matrix. The total energy is expressed in terms of $\rho(\mathbf{r},\mathbf{r}')$, and minimization is carried out with respect to both the ϕ_α and the $K_{\alpha\beta}$. Hierse and Stechel[14] use a number of functions ϕ_α equal to the number of occupied orbitals, but this restriction is not present in our scheme. The consequences of this and other differences between the two methods will be addressed later in this paper. Other methods have been proposed recently. Among them are the method due to Stechel, Williams, and Feibelman,[16] the method of Kohn,[17] the density-matrix method of Yang and Lee,[18] and the method due to Goedecker.[19,20]

Previously, only a brief description of our method has been published.[15] In this paper we give a detailed description of the method, together with some illustrations of its practical performance and a discussion of its relation to other methods. In Sec. II, the method is outlined and its theoretical foundations are discussed. The practical implementation of the method is then described in Sec. III. The tests we have performed to probe the practical usefulness of the scheme are presented in Sec. IV. In Sec. V, we assess what has been achieved and we discuss possible future developments, with particular attention to the problems that need to be overcome before the method can be generally applied. Some of the mathematical analysis is reported in an Appendix.

II. FORMULATION OF DFT IN TERMS OF THE DENSITY MATRIX

A. Density-functional theory

We need to recall briefly the principles of DFT.[23] The total energy E_{tot} of the system of valence electrons and atomic cores is expressed as

$$E_{\mathrm{tot}}=E_K+E_{\mathrm{ps}}+E_H+E_{\mathrm{xc}}+E_M, \tag{2}$$

where the terms on the right are the kinetic, pseudopotential, Hartree, and exchange-correlation energies of the electrons, and E_M is the Madelung energy of the cores. The first two energies are

$$E_K=2\sum_{i=1}^{N}\left\langle\psi_i\left|-\frac{\hbar^2}{2m}\nabla^2\right|\psi_i\right\rangle,$$

$$E_{\mathrm{ps}}=2\sum_{i=1}^{N}\langle\psi_i|\hat{V}_{\mathrm{ps}}|\psi_i\rangle, \tag{3}$$

where ψ_i are the Kohn-Sham (KS) orbitals, $\hat{V}_{\mathrm{ps}}$ is the total pseudopotential operator, and $N=\frac{1}{2}N_e$ is the number of occupied orbitals. The energies E_H and E_{xc} can be written in terms of the electron number density $n(\mathbf{r})$:

$$E_H=\frac{1}{2}e^2\int d\mathbf{r}d\mathbf{r}'n(\mathbf{r})n(\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|,$$

$$E_{\mathrm{xc}}=\int d\mathbf{r}\; n(\mathbf{r})\epsilon_{\mathrm{xc}}[n(\mathbf{r})], \tag{4}$$

where for simplicity we assume the local density approximation (LDA) for E_{xc}, with ϵ_{xc} the exchange-correlation energy per electron. The number density is

$$n(\mathbf{r})=2\sum_{i=1}^{N}|\psi_i(\mathbf{r})|^2. \tag{5}$$

The important principle for the present purposes is that the true ground-state energy and electron density are obtained by minimizing E_{tot} with respect to the KS orbitals, subject to the constraint that the latter are kept orthonormal.

In the standard formulation of DFT, which we have just summarized, all the occupied orbitals are fully occupied. However, it is frequently convenient, for physical, computational, or formal reasons, to generalize the theory so that orbitals can be partially occupied. Spatial orbital $\psi_i(\mathbf{r})$, rather than containing two electrons, may now contain $2f_i$ electrons, where the occupation number f_i lies in the range $0\leqslant f_i\leqslant 1$. The number density $n(\mathbf{r})$ now becomes

$$n(\mathbf{r})=2\sum_i f_i|\psi_i(\mathbf{r})|^2, \tag{6}$$

and the kinetic and pseudopotential energies are

$$E_K = 2\sum_i f_i \left\langle \psi_i \left| -\frac{\hbar^2}{2m}\nabla^2 \right| \psi_i \right\rangle ,$$

$$E_{\rm ps} = 2\sum_i f_i \langle \psi_i | \hat{V}_{\rm ps} | \psi_i \rangle . \tag{7}$$

The expressions for E_H and $E_{\rm xc}$ in terms of $n(\mathbf{r})$ are unchanged.

The usual physical reason for making this generalization is that one wishes to treat the electrons at a nonzero temperature, in which case the f_i are Fermi-Dirac occupation numbers;[24] computationally, the generalization is sometimes made in order to get rid of the troublesome discontinuity at the Fermi level in metallic systems.[25,26] Our reason for considering it here is that it will be relevant to the density matrix formulation. We shall assume that if $E_{\rm tot}$ is minimized both with respect to the ψ_i (subject to orthonormality) and with respect to the f_i (subject to the restriction $0 \leqslant f_i \leqslant 1$ and the condition that the sum f_i be equal to $\frac{1}{2}N_e$), then we arrive at exactly the ground state that is obtained by the more usual minimization with respect to fully occupied states ψ_i. Another way of putting this is that the energy cannot be reduced below the normal ground state by allowing partial occupation.

Now we turn to the density matrix, which is defined by

$$\rho(\mathbf{r},\mathbf{r}') = \sum_i f_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}') . \tag{8}$$

It follows from this definition that $\rho(\mathbf{r},\mathbf{r}')$ is a Hermitian operator whose eigenvalues are all in the interval $[0,1]$. The converse is also true: a Hermitian operator $\rho(\mathbf{r},\mathbf{r}')$ whose eigenvalues are f_i and whose eigenfunctions are $\psi_i(\mathbf{r})$ can be written as in Eq. (8). In terms of such an operator $\rho(\mathbf{r},\mathbf{r}')$, let the kinetic energy, pseudopotential energy, and number density be defined as

$$E_K = -\frac{\hbar^2}{m}\int d\mathbf{r} [\nabla_r^2 \rho(\mathbf{r},\mathbf{r}')]_{\mathbf{r}=\mathbf{r}'} ,$$

$$E_{\rm ps} = 2\int d\mathbf{r} d\mathbf{r}' V_{\rm ps}(\mathbf{r}',\mathbf{r}) \rho(\mathbf{r},\mathbf{r}') , \tag{9}$$

$$n(\mathbf{r}) = 2\rho(\mathbf{r},\mathbf{r}) ,$$

with E_H and $E_{\rm xc}$ expressed in the usual way in terms of $n(\mathbf{r})$. It follows from what we have said before that if $E_{\rm tot}$ is minimized with respect to $\rho(\mathbf{r},\mathbf{r}')$ subject to the condition that the eigenvalues of the latter are in the required interval and add up to $\frac{1}{2}N_e$, then we arrive at the usual ground state. This is the density-matrix formulation of DFT.

B. Localization of the density matrix

Since DFT is variational, any restriction placed on the class of density matrices $\rho(\mathbf{r},\mathbf{r}')$ that can be searched over has the effect of raising the minimum energy $E_{\rm min}$ above its true ground-state value E_0; progressive relaxation of such a restriction makes $E_{\rm min}$ tend to E_0. Now in general the density matrix in the true ground state tends to zero as the separation of its arguments $|\mathbf{r}-\mathbf{r}'|$ increases. This strongly suggests the usefulness of estimating E_0 by searching over $\rho(\mathbf{r},\mathbf{r}')$ with the following restriction:

$$\rho(\mathbf{r},\mathbf{r}') = 0, \quad |\mathbf{r}-\mathbf{r}'| > R_c , \tag{10}$$

where R_c is a chosen cutoff radius. The resulting estimate $E_{\rm min}(R_c)$ will tend to E_0 from above as $R_c \to \infty$. The manner in which $\rho(\mathbf{r},\mathbf{r}')$ goes to zero at large separations depends on the electronic structure of the system, and particularly on whether there is a gap between the highest occupied and lowest unoccupied states. It is rigourously established that in one-dimensional systems having a gap ρ decays exponentially with separation, while in gapless systems it decays only as an inverse power.[27] It is presumed that three-dimensional systems behave similarly. This suggests—though to our knowledge it is unproven—that $E_{\rm min}(R_c) \to E_0$ exponentially for insulators and algebraically for metals.

Clearly in practical calculations we cannot work directly with a six-dimensional function $\rho(\mathbf{r},\mathbf{r}')$, even if it vanishes beyond a chosen radius. It is essential that ρ be separable, i.e., representable in the form

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K_{\alpha\beta} \phi_\beta(\mathbf{r}') . \tag{11}$$

For practical purposes, there must be only a finite number of $\phi_\alpha(\mathbf{r})$ functions, which will be referred to as support functions. For ρ to be Hermitian, we must require that the matrix $K_{\alpha\beta}$ be Hermitian. The restriction to a finite number of support functions is equivalent to the condition that ρ have only this number of nonzero eigenvalues, and this is the essence of the separability requirement. With this, we now have two independent restrictions on ρ: localization and separability. The localization of ρ can be imposed by requiring that the support functions be nonzero only within chosen regions, which we call the support regions, and that the coefficients $K_{\alpha\beta}$ vanish if the separation of the support regions of ϕ_α and ϕ_β exceeds a chosen cutoff.

We now have a general framework for linear-scaling DFT schemes. In practical calculations, the ϕ_α functions will be represented either as a linear combination of basis functions, or simply by numerical values on a grid. Either way, the amount of information contained in a support function will be independent of the size of the system. The amount of information in the support functions will then scale linearly with the size of the system, and the number of $K_{\alpha\beta}$ coefficients will scale in the same way. This in turn implies that the electron density $n(\mathbf{r})$ and all the terms in the total energy can be calculated in a number of operations that scale linearly with system size.

C. Eigenvalue range of the density matrix

In this general scheme, the ground state is determined by searching over support functions and $K_{\alpha\beta}$ matrices. However, it is essential that this search be confined to those ϕ_α and $K_{\alpha\beta}$ for which the eigenvalues of $\rho(\mathbf{r},\mathbf{r}')$ lie in the interval $[0,1]$. This is a troublesome condition to impose, because we certainly do not wish to work directly with these

eigenvalues. We can achieve what we want by expressing ρ in a form that satisfies the condition automatically.

The scheme developed in this paper is the DFT analogue of the tight-binding scheme of LNV.[12] We write the density matrix as

$$\rho = 3\sigma^*\sigma - 2\sigma^*\sigma^*\sigma, \tag{12}$$

where $\sigma(\mathbf{r},\mathbf{r}')$ is an auxiliary function. The asterisk here indicates the continuum analogue of matrix multiplication. For arbitrary two-point functions $A(\mathbf{r},\mathbf{r}')$ and $B(\mathbf{r},\mathbf{r}')$, we use the notation $C=A^*B$ as a short-hand for the statement

$$C(\mathbf{r},\mathbf{r}') = \int d\mathbf{r}'' A(\mathbf{r},\mathbf{r}'')B(\mathbf{r}'',\mathbf{r}'). \tag{13}$$

The reason this works is that the eigenvalues λ_ρ of ρ automatically satisfy $0 \leqslant \lambda_\rho \leqslant 1$ provided the eigenvalues λ_σ of σ are in the range $-\frac{1}{2} \leqslant \lambda_\sigma \leqslant \frac{3}{2}$; in addition, λ_ρ has turning points at the values 0 and 1. Since the ground state is obtained when $\lambda_\rho = 0$ or 1, there is a natural mechanism whereby variation of σ drives ρ towards idempotency.

To obtain the separable form of ρ [Eq. (11)], we write

$$\sigma(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) L_{\alpha\beta} \phi_\beta(\mathbf{r}'), \tag{14}$$

which implies the matrix relation

$$K = 3LSL - 2LSLSL, \tag{15}$$

where $S_{\alpha\beta}$ is the overlap matrix of support functions:

$$S_{\alpha\beta} = \int d\mathbf{r} \phi_\alpha(\mathbf{r}) \phi_\beta(\mathbf{r}). \tag{16}$$

The ground state is now obtained by minimizing $E_{\rm tot}$ with respect to the ϕ_α and the $L_{\alpha\beta}$ matrix, with the $K_{\alpha\beta}$ matrix given by Eq. (15). In the practical calculations reported later, the ϕ_α are nonzero only inside spherical regions of radius $R_{\rm reg}$, and the $L_{\alpha\beta}$ are nonzero only if the centers of the regions α and β are separated by less than a cutoff distance R_L.

It will be useful for the purposes of later discussion to note how a closely related scheme leads back to the MGC method.[7] This scheme is obtained by writing

$$\rho = \sigma^*(2-\sigma), \tag{17}$$

where σ is required to be positive semidefinite. Since the eigenvalues λ_σ can be expressed as κ_σ^2 where κ_σ is real, the eigenvalues of ρ are given by

$$\lambda_\rho = \lambda_\sigma(2-\lambda_\sigma) = \kappa_\sigma^2(2-\kappa_\sigma^2). \tag{18}$$

This quartic function lies in the range [0,1] for $|\kappa_\sigma| \leqslant 2^{1/2}$ and has turning points when $\lambda_\rho = 0$ and 1. This gives an alternative mechanism for driving ρ towards idempotency. With σ given, as before, by Eq. (14), it is straightforward to show that σ is positive semidefinite if and only if the matrix $L_{\alpha\beta}$ is positive semidefinite, and this is equivalent to the condition that $L_{\alpha\beta}$ be expressible as

$$L_{\alpha\beta} = \sum_s b_\alpha^{(s)} b_\beta^{(s)}. \tag{19}$$

The result is that $\sigma(\mathbf{r},\mathbf{r}')$ must have the form

$$\sigma(\mathbf{r},\mathbf{r}') = \sum_s \chi^{(s)}(\mathbf{r}) \chi^{(s)}(\mathbf{r}'), \tag{20}$$

where

$$\chi^{(s)}(\mathbf{r}) = \sum_\alpha b_\alpha^{(s)} \phi_\alpha(\mathbf{r}). \tag{21}$$

Following arguments presented by Nunes and Vanderbilt,[28] it can now be shown that this scheme is exactly equivalent to the linear-scaling DFT scheme of MGC.

III. PRACTICAL IMPLEMENTATION OF THE METHOD

A. The real-space grid

We now give a prescription for the calculation of the energy functional, and of its derivatives with respect to the support functions ϕ_α and the $L_{\alpha\beta}$ parameters, and we describe how minimization of the energy can be carried out in practice. Central to our implementation of the method described in the previous section is the use of a regular cubic real-space grid, spanning the whole system under study. There have been a number of recent implementations of conventional DFT-pseudopotential calculations using real-space grids.[29–34]

The support functions are represented by their values at the grid points. Since these functions are required to be spatially localized, they have nonzero values only on the grid points inside the localization regions. In the present work, these regions are chosen to be spherical, and their centers are at the atomic positions. Real-space integration is replaced by summation over grid points, so that, e.g., the overlap matrix elements are calculated as

$$S_{\alpha\beta} \simeq \delta\omega \sum_{\mathbf{r}_\ell} \phi_\alpha(\mathbf{r}_\ell) \phi_\beta(\mathbf{r}_\ell), \tag{22}$$

where the sum goes over the set of grid points $\mathbf{r}_\ell$ common to the localization regions of both ϕ_α and ϕ_β, and $\delta\omega$ is the volume per grid point.

The action of the kinetic energy operator on the support functions is evaluated using a finite difference technique. To nth order in the grid spacing, h, we have that

$$\frac{\partial^2 \phi_\alpha}{\partial x^2}(n_x,n_y,n_z) \simeq \frac{1}{h^2} \sum_{m=-n}^{n} C_{|m|} \phi_\alpha(n_x+m,n_y,n_z), \tag{23}$$

where n_x, n_y, and n_z are integer indices labeling grid point $\mathbf{r}_\ell$, and the coefficients $C_{|m|}$ can be calculated beforehand. Equivalent expressions can be used for $\partial^2\phi_\alpha/\partial y^2$ and $\partial^2\phi_\alpha/\partial z^2$, and it is thus possible to evaluate $\nabla^2\phi_\alpha$ approximately at each grid point. From Eqs. (9) and (11), the kinetic energy is given by

$$E_K = 2\sum_{\alpha\beta} K_{\alpha\beta} T_{\beta\alpha}, \tag{24}$$

where

$$T_{\beta\alpha} = -\frac{\hbar^2}{2m}\int d\mathbf{r}\phi_\beta(\mathbf{r})\nabla_r^2\phi_\alpha(\mathbf{r}). \quad (25)$$

Once $\nabla^2\phi_\alpha(\mathbf{r})$ has been evaluated at each grid point using Eq. (23), the $T_{\alpha\beta}$ matrix elements are calculated by summing over grid points, just as for $S_{\alpha\beta}$ [see Eq. (22)]. It is worth noting that the use of a finite-difference approach can produce either positive or negative errors in the kinetic energy evaluation. Thus, strictly speaking, the variational characteristic of the method is lost with the introduction of this approximation. Nevertheless, the error incurred by such a finite-difference scheme will be small provided that the grid is sufficiently fine and the Laplacian approximation is of a sufficiently high order in the grid spacing h. Practical applications will be carried out using finite values of the parameters R_L and $R_{\rm reg}$, and since the method is still variational with respect to these, it is to be expected that the minimum energy provides an upper bound to the exact ground state.

In order to evaluate the exchange-correlation, Hartree, and pseudopotential contributions to the total energy, we first need to evaluate the electron density at each grid point. From Eqs. (9) and (11), the density at grid point $\mathbf{r}_{/\!/}$ is

$$n(\mathbf{r}_{/\!/}) = 2\sum_{\alpha\beta}\phi_\alpha(\mathbf{r}_{/\!/})K_{\alpha\beta}\phi_\beta(\mathbf{r}_{/\!/}). \quad (26)$$

From this, it is straightforward to evaluate the exchange-correlation energy by summing the quantity $n(\mathbf{r}_{/\!/})\epsilon_{\rm xc}[n(\mathbf{r}_{/\!/})]$ over grid points. The exchange-correlation potential $\mu_{\rm xc}$ can also be calculated at each point, and is given as

$$\mu_{\rm xc}(\mathbf{r}_{/\!/}) = \frac{d}{dn}\{n(\mathbf{r}_{/\!/})\epsilon_{\rm xc}[n(\mathbf{r}_{/\!/})]\}. \quad (27)$$

To obtain the Hartree energy and potential we use the fast Fourier transform (FFT) method to transform the calculated electronic density into reciprocal space, thus obtaining its Fourier components $\hat{n}_{\mathbf{G}}$. The Hartree energy is then given as

$$E_H = 2\pi\Omega e^2\sum_{\mathbf{G}\neq\mathbf{0}}|\hat{n}_{\mathbf{G}}|^2/G^2, \quad (28)$$

where Ω is the volume of the simulation cell. The Hartree potential in reciprocal space is

$$\hat{V}_H(\mathbf{G}) = 4\pi\Omega e^2\hat{n}_{\mathbf{G}}/G^2. \quad (29)$$

This can be constructed on the reciprocal-space grid, and transformed to obtain the Hartree potential in real space. FFT is, of course, an $O(N\log_2 N)$ operation rather than an $O(N)$ operation, but the difference is negligible for the present purposes.

We restrict ourselves here to local pseudopotentials, so that the value of the total pseudopotential $V_{\rm ps}(\mathbf{r}_{/\!/})$ at grid point $\mathbf{r}_{/\!/}$ is formally given by

$$V_{\rm ps}(\mathbf{r}_{/\!/}) = \sum_I v_{\rm ps}(|\mathbf{r}_{/\!/} - \mathbf{R}_I|), \quad (30)$$

where $v_{\rm ps}(r)$ is the ionic pseudopotential and $\mathbf{R}_I$ is the position of ion I. In practice, however, $V_{\rm ps}(\mathbf{r}_{/\!/})$ cannot be calculated like this, because $v_{\rm ps}(r)$ has a Coulomb tail $-Z|e|^2/r$ at large r, where Z is the core charge. In order to obtain a linear-scaling algorithm for $E_{\rm ps}$, we proceed as follows. The ionic pseudopotential is represented as the sum of the Coulomb potential due to a Gaussian charge distribution $\eta(r)$ and a short-range potential $v_{\rm ps}^0(r)$. The total charge in $\eta(r)$ is $Z|e|$, and the distribution is given by

$$\eta(r) = Z|e|(\alpha/\pi)^{3/2}\exp(-\alpha r^2), \quad (31)$$

where the parameter α governs the rate of decay of the Gaussian. We therefore have

$$v_{\rm ps}(r) = -\frac{Z|e|^2}{r}{\rm erf}(\alpha^{1/2}r) + v_{\rm ps}^0(r). \quad (32)$$

The part of $V_{\rm ps}$ coming from $v_{\rm ps}^0$ can now be calculated as a direct sum over ions, as in Eq. (30). Since $v_{\rm ps}^0$ can be neglected beyond a certain radius, this part of the calculation scales linearly. The part of $V_{\rm ps}$ coming from the array of Gaussians can be treated in exactly the same way as the Hartree potential. The pseudopotential energy is then calculated by summation over the real-space grid:

$$E_{\rm ps} = \delta\omega\sum_{/\!/} V_{\rm ps}(\mathbf{r}_{/\!/})n(\mathbf{r}_{/\!/}). \quad (33)$$

B. Derivatives and minimization

Once the contributions to the total energy have been obtained as outlined above, we need to vary both $L_{\alpha\beta}$ and ϕ_α in order to minimize it. The $L_{\alpha\beta}$ and ϕ_α are independent variables, and the problem breaks naturally into two separate minimizations that can be carried out in an alternating manner: one with respect to $L_{\alpha\beta}$ with fixed ϕ_α, and the other with respect to ϕ_α with fixed $L_{\alpha\beta}$. Indeed, the choice of object function can be different for the two types of variations, and when minimizing with respect to the $L_{\alpha\beta}$ we find it more convenient to take $\Omega = E_{\rm tot} - \mu N_e$ as our object function, where μ is the chemical potential and N_e is the electron number. We return to this point below.

Expressions for the derivatives with respect to $L_{\alpha\beta}$ and ϕ_α are obtained in the Appendix. The partial derivative of Ω with respect to $L_{\alpha\beta}$ is given by

$$\frac{\partial\Omega}{\partial L_{\alpha\beta}} = [6(SLH' + H'LS) - 4(SLSLH' + SLH'LS + H'LSLS)]_{\alpha\beta}, \quad (34)$$

where $H' = H - \mu S$, and H is the matrix representation of the KS Hamiltonian in the support function representation. It is worth noting that this expression is exactly the same as would be obtained in a nonorthogonal tight-binding formalism.[35] There is, however, one important difference: in self-consistent DFT calculations the Hamiltonian matrix elements depend on $L_{\alpha\beta}$ through the electronic density $n(\mathbf{r})$. The partial derivative of the total energy with respect to ϕ_α at grid point $\mathbf{r}_{/\!/}$ is given by

$$\frac{\partial E_{\rm tot}}{\partial\phi_\alpha(\mathbf{r}_{/\!/})} = 4\delta\omega\sum_\beta [K_{\alpha\beta}\hat{H} + 3(LHL)_{\alpha\beta} - 2(LSLHL + LHLSL)_{\alpha\beta}]\phi_\beta(\mathbf{r}_{/\!/}), \quad (35)$$

where $\hat{H}$ is the Kohn-Sham operator, which is made to act on support function ϕ_β.

It is important to notice that because of the spatial localization of the support functions, and the finite range of L, all the matrices involved in the calculation of these derivatives are sparse, when the system is large enough. Provided this sparsity is exploited in the computational scheme, the method scales linearly with the size of the system.

In the scheme of LNV,[12] it is proposed to work at constant chemical potential, rather than at constant electron number. We prefer to maintain the electron number constant. The variations with respect to $L_{\alpha\beta}$ and ϕ_α will in general cause the electron number to differ from the correct value, and it is therefore necessary to correct this effect as the minimization proceeds. We achieve this in the following manner: during the minimization with respect to L, the current search direction is projected so that it is tangential to the local surface of constant N_e, i.e., perpendicular to $\nabla_L N_e$ at the current position. This ensures that the minimization along this direction will cause only a small change in N_e, and it is expected that at the new minimum N_e will differ only slightly from the required value. In any case, it is possible to return to a position as close as desired to the constant N_e surface by following the local gradient $\nabla_L N_e$. If the value of the chemical potential μ is appropriately chosen, this correction step can be carried out without losing the reduction in Ω obtained by performing the line minimization, and this is why we prefer to take Ω as the object function instead of the total energy, when minimizing with respect to L. We find that this scheme is capable of maintaining the electron number close to its correct value throughout the minimization, and is also simple to implement. The gradient $\nabla_L N_e$ has elements

$$\frac{\partial N_e}{\partial L_{\alpha\beta}} = 12(SLS - SLSLS)_{\alpha\beta}, \tag{36}$$

which, as all other gradients discussed earlier, can be calculated in $O(N)$ operations. Minimization with respect to $\phi_\alpha(\mathbf{r}_\ell)$ will also have the effect of changing the electron number. However, given that the two types of variations are performed alternately, the correction during the L minimization is sufficient to counteract this effect.

Given that variation of $L_{\alpha\beta}$ causes the electronic density to change, and this in turn implies that the Hamiltonian matrix elements change, it would seem necessary to update the Hamiltonian at each step of the minimization with respect to L. However, we find that this can be avoided by considering H fixed during this part of the minimization. Strictly speaking, if H is held fixed while L is varied, we are not minimizing $\Omega = E_{\rm tot} - \mu N$ but rather $\Omega' = E' - \mu N$, where E' is given by

$$E' = \mathrm{Tr}[(6LSL - 4LSLSL)H]. \tag{37}$$

If this minimization were carried out through to convergence, this would be equivalent to diagonalizing H in the representation of the current support functions. At convergence, it will be found in general that L and H are not mutually consistent, and if consistency is required, one needs to update H and repeat the minimization, iterating this cycle until consistency was achieved. This is not necessary in practice, because H will be updated at the next variation with respect to the support functions. The minimization of Ω' has practical advantages in that it avoids the updating of the Hamiltonian at each step, and, because of its construction, it is a cubic polynomial in every possible search direction, so it is possible to find the exact location of line minima during its minimization.

The minimization with respect to $\phi_\alpha(\mathbf{r}_\ell)$ can be carried out by simply moving along the gradient $\partial E_{\rm tot}/\partial\phi_\alpha(\mathbf{r}_\ell)$ Eq. (35) (steepest descents) or by using this expression to construct mutually conjugate directions (conjugate gradients).

IV. TEST CALCULATIONS

In order to test our $O(N)$ DFT scheme, we have performed calculations on a system of 512 Si atoms treated using a local pseudopotential. The purpose of these tests is to find out how the total energy depends on the two spatial cutoff radii: the support-region radius $R_{\rm reg}$, and the L-matrix cutoff radius R_L. The practical usefulness of the scheme, and the size of system for which linear-scaling behavior is attained depend on the rate of convergence of $E_{\rm tot}$ to its exact value as $R_{\rm reg}$ and R_L are increased. Here, ''exact'' refers only to the absence of errors due to the truncation of $\rho(\mathbf{r},\mathbf{r}')$; other sources of inexactness, such as the use of a discrete grid and a local pseudopotential, are of no concern here.

The system treated is a periodically repeating cell containing 512 atoms of diamond-structure Si having the experimental lattice parameter (5.43 Å). The local pseudopotential is the one constructed by Appelbaum and Hamann,[36] which is known to give a satisfactory representation of the self-consistent band structure. The LDA exchange-correlation energy is calculated using the Ceperley-Alder formula.[37] We use a grid spacing of 0.34 Å, which is similar to the spacing typically used in pseudopotential plane-wave calculations on Si, and is sufficient to give reasonable accuracy. The second derivatives of the ϕ_α needed in the calculation of E_K are computed using the second-order formula given in Eq. (23).

A support region is centered on every atom, and each such region contains four support functions. One can imagine that these support functions correspond roughly to the single 3*s* function and the three 3*p* functions that would be used in a tight-binding description, but we stress that nothing obliges us to work with this number of support functions. In keeping with the tight-binding picture, the initial guess for the support functions is taken to be a Gaussian multiplied by a constant, x, y, or z, so that the functions have the symmetry of *s* and *p* states. As an initial guess for the L matrix, we take the quantity $2I - S$, where S is the overlap matrix calculated for the initial support functions. This guess for L, which represents the expansion of $S^{-1} \equiv [I-(I-S)]^{-1}$ to first order, is crude, and does not yield the correct value of $\mathrm{Tr}\rho$. This error is corrected by displacing L iteratively along the gradient $\nabla_L N_e$ until N_e is within a required tolerance of the correct value.

The initial guesses for the ϕ_α and the $L_{\alpha\beta}$ define the initial Hamiltonian and overlap matrices. From this starting point, we make a number of conjugate-gradient line searches to minimize Ω by varying L, with the Hamiltonian and overlap matrices held fixed. This is followed by a sequence of line searches in which the ϕ_α are varied. We refer to the sequence of L moves followed by a sequence of ϕ moves as

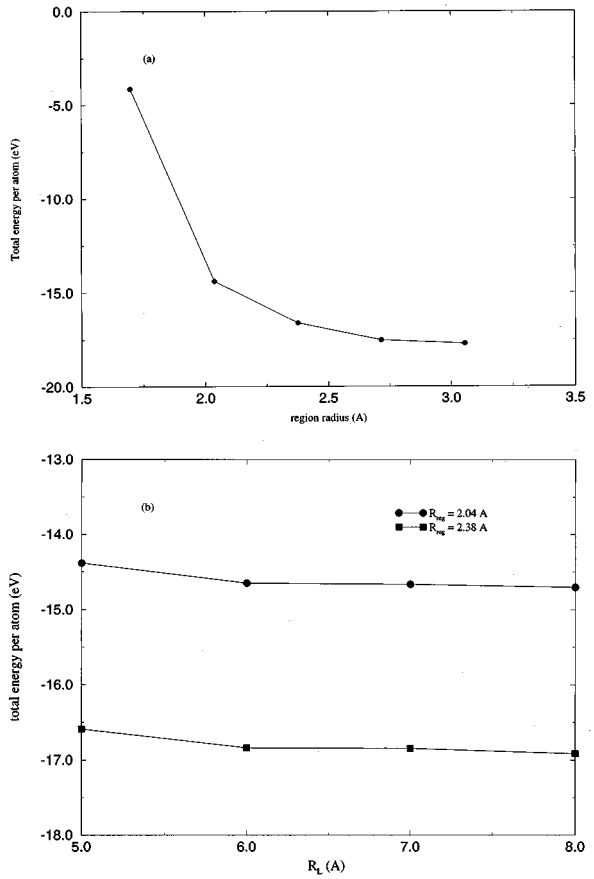

FIG. 1. (a) Total energy per atom as a function of the support region radius R_{reg} with $R_L = 5$ Å. (b) Total energy per atom as a function of the range of the L matrix, R_L, for two different support region radii, $R_{\mathrm{reg}} = 2.04$ and 2.38 Å.

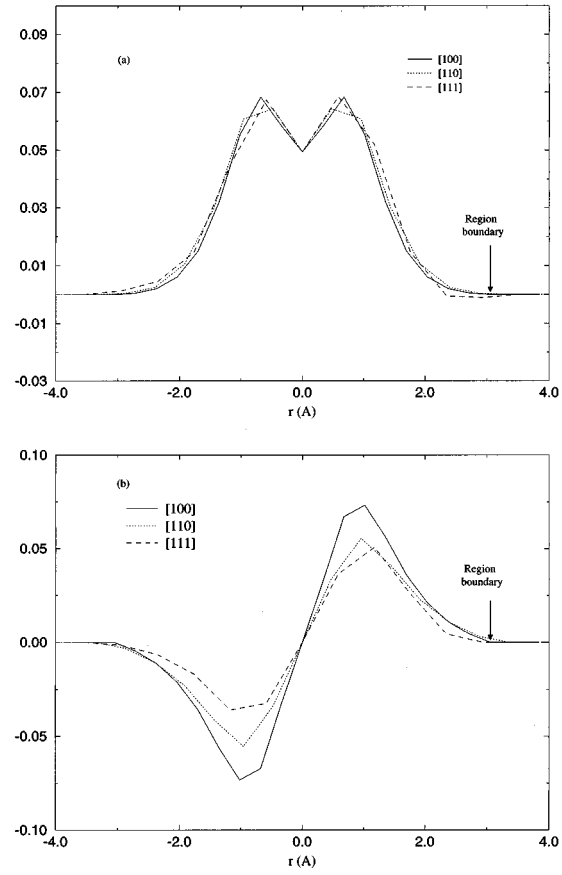

FIG. 2. Support functions after minimization of the total energy with $R_{\mathrm{reg}} = 3.05$ and $R_L = 5.0$ Å. (a) s-like support function, and (b) p_x-like support function.

a *cycle*. The entire energy minimization consists of a set of cycles. In practice, we have found that cycles consisting of five L moves and two ϕ moves work satisfactorily, and that E_{tot} is converged to within 10^{-4} eV/atom after typically 50–60 cycles. This would not be an efficient rate of convergence for routine applications, but is more than adequate for the present purposes.

Our test calculations confirm our earlier finding[15] that for the Si perfect crystal E_{tot} is already quite close to its exact value when $R_L = 5.0$ Å. We have therefore used this value of R_L to make calculations of E_{tot} as a function of R_{reg} [see Fig. 1(a)]. The results show that E_{tot} converges very quickly with increasing R_{reg}, and that it is within ~ 0.1 eV of its fully converged value for $R_{\mathrm{reg}} = 3.05$ Å. This would be a significant error on an absolute scale, but we would expect energy differences calculated with this technique to be much smaller. It is worth noting that errors of 0.1 eV/atom in the total energy are usually regarded as acceptable in conventional plane-wave calculations.

In order to show how E_{tot} depends on R_L, we present a series of results at the two region radii $R_{\mathrm{reg}} = 2.04$ and 2.38 Å [see Fig. 1(b)]. These results indicate that there is only a slow variation with R_L and that this variation is almost the same for different values of R_{reg}. This means that it is possible to converge the total energy to satisfactory accuracy with easily manageable spatial cutoffs.

It is interesting to know the form of the support functions for the self-consistent ground state. These are shown in Fig. 2 for the case $R_{\mathrm{reg}} = 3.05$, $R_L = 5$ Å. The support functions shown here are the first (initially s Gaussian) and second (p_x Gaussian). Profiles of the support functions along the [100], [110], and [111] directions are shown. The support functions are seen to be symmetric with respect to the center of the support region ($r = 0$) along the [100] and [110] directions. Along the [111] direction there is a slight asymmetry resulting from the presence of a nearest-neighbor ion, which lies at 2.35 Å from the origin in the positive direction. Remarkably, the s-like support function seems to be almost perfectly spherically symmetric, except near the peak at $r \approx 0.8$ Å. It is encouraging to see that the support functions go rather smoothly to zero at the region boundary, and this confirms that the boundary has little effect on the results.

V. DISCUSSION

We have tried to do three things in this work: to develop the basic formalism needed to underpin $O(N)$ DFT pseudopotential methods; to implement one such method and identify the main technical issues in doing so; and to present the results of tests on a simple but important system, which allow us to gauge the usefulness of the method. We have shown that a rather general class of $O(N)$ DFT pseudopotential methods can be based on a formulation of DFT in terms of the density matrix, and that this formulation is

equivalent to commonly used versions of DFT that operate with fractional occupation numbers. From this viewpoint, the key challenge is to ensure that the eigenvalues of the variable density matrix lie between 0 and 1, and we have seen that the method of LNV (Ref. 12) gives a way of doing this. The implementation of the basic ideas has been achieved by performing all calculations in real space, with the DFT integrals approximated by sums on a grid—except for the use of FFT to treat the Hartree term. An alternative here would be to work with atomiclike basis functions, but we note that the use of a grid preserves an important link with conventional plane-wave methods, as will be analyzed in more detail elsewhere. Our test results on perfect-crystal Si show that the total energy converges rapidly as the real-space cutoffs are increased, and that it is straightforward to achieve a precision comparable with that of normal plane-wave calculations.

An important question for any $O(N)$ method is the system size at which it starts to beat a standard $O(N^3)$ method—a plane-wave method in the present case. This will clearly depend strongly on the system, but even for Si it is too soon to answer it on the basis of practical calculations. The crossover point depends on the prefactor in the linear scaling, and this is strongly affected by the efficiency of the coding. All we have attempted to do here is to address the problem of achieving $O(N)$ behavior. The question of the prefactor is a separate matter, which will need separate investigation.

It should be clear that there is much more to do before the present method can be routinely applied to real problems. We have deliberately not discussed in detail the problems of doing calculations on a real-space grid. Such problems have been discussed outside the linear-scaling context in several recent papers,[29,30] and it should be possible to apply the advances reported there to $O(N)$ DFT calculations. In particular, curvilinear grids[31–34] for the treatment of strongly attractive pseudopotentials are likely to be very important for $O(N)$ calculations. We have also not discussed here the calculation of forces on the atoms, the problems that may arise when the boundaries of support regions cross grid points, and the general question of translational invariance within grid-based techniques.

In our current implementation, the rate of the convergence of the total energy during the minimization is somewhat slow for routine applications. In order to improve this convergence rate, a possible course of action would be to use *multigrid* techniques.[38,39] It is well known that standard relaxation techniques for the solution of partial differential equations (such as conjugate gradients) are efficient in reducing the Fourier components of the error of a trial solution with wavelengths in the range of the grid spacing (so-called oscillatory components). However, components of the error with longer wavelengths (smooth components) are left almost unchanged. Thus, the rate of convergence of relaxation methods is poor. By bringing into play multiple grids with different degrees of coarseness, it is possible to improve the rate of convergence, because smooth components on a fine grid will become oscillatory on a coarser grid. Multigrid methods are being recognized as a useful tool in electronic structure calculations.[30,40,41]

We have noted already that our method is related to other recently proposed methods. It was first pointed out by Vanderbilt[42] that Eq. (14) can be regarded as a general ansatz encompassing several linear-scaling methodologies. In both the MGC and the Hierse-Stechel methods, the number of functions employed is equal to the number of occupied orbitals. If this is the case, it is not necessary to use the transformation Eq. (12) to impose the approximate idempotency of the density operator, because it is only required that all eigenvalues of ρ be equal (or close to) unity. To achieve this, it is sufficient to use Eq. (17) as the purifying transformation. The scheme discussed in this paper and introduced in Ref. 15, as well as the original tight-binding density matrix formalism of LNV,[12] allows for the use of a number of support functions ϕ_α greater than the number of occupied orbitals. In this case it is necessary to use the transformation given by Eq. (12) to ensure the near idempotency of ρ. At first sight it would seem that the use of a higher number of support functions than occupied orbitals [and thus the need to use Eq. (12) rather than Eq. (17)] is unnecessary and wasteful. However, there is some indication that restricting the number of support functions to the number of occupied orbitals can result in slow convergence of the minimization and multiple-minima problems.[8] Our experience is that relaxing this constraint eliminates the multiple-minima problem. It is worth pointing out that Kim, Mauri, and Galli[11] found it necessary to increase the number of orbitals above the number of occupied orbitals in their generalization of the MGC scheme in order to avoid this problem.

Finally, we note that our linear-scaling scheme is intended for calculations on very large systems, and this means that parallel implementation will play a key role. The test calculations we have presented were, in fact, performed on a massively parallel machine, and the parallel-coding techniques we have developed will be described in a separate paper.

ACKNOWLEDGMENTS

The work of C.M.G. was supported by the High Performance Computing Initiative (HPCI) under Grant No. GR/K41649, and the work of E.H. by EPSRC Grant No. GR/J01967. The major calculations were done on the Cray T3D at Edinburgh Parallel Computing Centre using an allocation of time from the HPCI. Code development and subsidiary analysis were made using local hardware funded by EPSRC Grant No. GR/J36266. We gratefully acknowledge useful discussions with D. Vanderbilt.

APPENDIX: DERIVATIVES OF THE TOTAL ENERGY

We derive here expressions for the derivatives $\partial E_{\rm tot}/\partial L_{\alpha\beta}$ and $\delta E_{\rm tot}/\delta\phi_\alpha(\mathbf{r})$.

1. Derivative with respect to $L_{\alpha\beta}$

In DFT, the total energy Eq. (2) has two types of contributions: those that can be written as the trace of some operator acting on the density matrix, as is the case for the kinetic and pseudopotential energies [see Eq. (9)], and those that depend only on the diagonal elements of ρ, i.e., the electron density, namely, the Hartree and exchange-correlation energies. The Madelung term E_M does not depend on either $L_{\alpha\beta}$ or ϕ_α, so it will make no contribution to the variation in total energy as these are changed. Denoting by E_c the kinetic or pseudopotential contribution to the energy, we have

$$E_c = 2\sum_{\gamma\delta} [3(LSL)_{\gamma\delta} - 2(LSLSL)_{\gamma\delta}]C_{\delta\gamma}, \quad \text{(A1)}$$

where

$$C_{\delta\gamma} = \int d\mathbf{r}\ \phi_\delta(\mathbf{r})\left(-\frac{\hbar^2}{2m}\nabla^2\right)\phi_\gamma(\mathbf{r}) \quad \text{(A2)}$$

for the kinetic energy, and for the pseudopotential

$$C_{\delta\gamma} = \int d\mathbf{r}d\mathbf{r}'\,\phi_\delta(\mathbf{r}')V_{ps}(\mathbf{r},\mathbf{r}')\phi_\gamma(\mathbf{r}), \quad \text{(A3)}$$

where V_{ps} is in general a nonlocal pseudopotential operator. Clearly, $C_{\gamma\delta}$ does not depend on $L_{\alpha\beta}$ for either operator, so this term does not change as $L_{\alpha\beta}$ is varied. It is thus easy to see that

$$\frac{\partial E_c}{\partial L_{\alpha\beta}} = [6(SLC+CLS)_{\beta\alpha} - 4(SLSLC+SLCLS+CLSLS)_{\beta\alpha}], \quad \text{(A4)}$$

where C is the matrix representation of the corresponding operator (kinetic energy or pseudopotential) in the basis of the support functions.

For the Hartree and exchange-correlation contributions, denoted by E_v, we have that

$$\frac{\partial E_v}{\partial L_{\alpha\beta}} = \int d\mathbf{r}\,\frac{\delta E_v}{\delta n(\mathbf{r})}\frac{\partial n(\mathbf{r})}{\partial L_{\alpha\beta}}. \quad \text{(A5)}$$

The electron density $n(\mathbf{r})$ is simply $2\rho(\mathbf{r},\mathbf{r})$, so that

$$\frac{\partial n(\mathbf{r})}{\partial L_{\alpha\beta}} = 2\sum_{\gamma\delta}\phi_\gamma(\mathbf{r})\frac{\partial}{\partial L_{\alpha\beta}}[3(LSL)_{\gamma\delta} - 2(LSLSL)_{\gamma\delta}]\phi_\delta(\mathbf{r}). \quad \text{(A6)}$$

In the case of the Hartree contribution, we have

$$\frac{\delta E_H}{\delta n(\mathbf{r})} = e^2\int d\mathbf{r}'\,\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} = \Phi(\mathbf{r}), \quad \text{(A7)}$$

where $\Phi(\mathbf{r})$ is the Hartree potential, while in the case of the exchange-correlation contribution we have

$$\frac{\delta E_{\rm xc}}{\delta n(\mathbf{r})} = \frac{d}{dn}[n\epsilon_{\rm xc}(n)] = \mu_{\rm xc}(\mathbf{r}), \quad \text{(A8)}$$

where $\mu_{\rm xc}$ is the exchange-correlation potential. If we take $V(\mathbf{r})$ to represent either $\Phi(\mathbf{r})$ or $\mu_{\rm xc}(\mathbf{r})$ as the case may be, we see that expression (42) reduces to

$$\frac{\partial E_v}{\partial L_{\alpha\beta}} = [6(SLV+VLS)_{\beta\alpha} - 4(SLSLV+SLVLS+VLSLS)_{\beta\alpha}], \quad \text{(A9)}$$

where

$$V_{\alpha\beta} = \int d\mathbf{r}\phi_\alpha(\mathbf{r})V(\mathbf{r})\phi_\beta(\mathbf{r}). \quad \text{(A10)}$$

By comparing this expression with Eq. (A4), it is easy to see that the partial derivative of the total energy with respect to $L_{\alpha\beta}$ can be written more compactly as

$$\frac{\partial E_{\rm tot}}{\partial L_{\alpha\beta}} = [6(SLH+HLS)_{\beta\alpha} - 4(SLSLH+SLHLS+HLSLS)_{\beta\alpha}], \quad \text{(A11)}$$

where $H_{\alpha\beta}$ is the sum of the corresponding matrix elements of the kinetic, pseudopotential, Hartree, and exchange-correlation operators, i.e., the matrix representation of the Kohn-Sham Hamiltonian in the basis of the support functions. Recall that in practice, we do not vary $E_{\rm tot}$ but rather vary $\Omega = E_{\rm tot} - \mu N$ with respect to $L_{\alpha\beta}$. However, it is trivial to obtain $\partial\Omega/\partial L_{\alpha\beta}$ from Eq. (48) by simply substituting the matrix elements of H by those of $H-\mu S$. Once this is done, Eq. (A11) corresponds to Eq. (34).

2. Functional derivative of $E_{\rm tot}$ with respect to ϕ_α

According to Eq. (11), the density matrix, and hence the total energy, can be regarded as a function of the quantities ϕ_α and $K_{\alpha\beta}$. When ϕ_α is varied, $E_{\rm tot}$ therefore varies firstly because of its direct dependence on ϕ_α, and secondly because of the implicit dependence of the $K_{\alpha\beta}$ matrix on ϕ_α through its dependence on the overlap matrix elements $S_{\alpha\beta}$ [see Eq. (15)]; we call these two types of variations type 1 and type 2.

To see how variations of type 1 behave, consider first the kinetic and pseudopotential energies. The type 1 variation of either of these is given by

$$(\delta E_c)_1 = 2\sum_{\gamma\delta} K_{\delta\gamma}\delta\int d\mathbf{r}\phi_\gamma(\mathbf{r})\hat{C}\phi_\delta(\mathbf{r}), \quad \text{(A12)}$$

where $\hat{C}$ represents the kinetic energy or the pseudopotential operator. The variation of the integral gives

$$(\delta E_c)_1 = 2\sum_{\gamma\delta} K_{\delta\gamma}\int d\mathbf{r}(\delta\phi_\gamma\hat{C}\phi_\delta + \phi_\gamma\hat{C}\delta\phi_\delta) = 2\sum_{\gamma\delta} K_{\delta\gamma}\int d\mathbf{r}(\delta\phi_\gamma\hat{C}\phi_\delta + \delta\phi_\delta\hat{C}\phi_\gamma). \quad \text{(A13)}$$

The last equality follows from the fact that $\hat{C}$ is a Hermitian operator. The type 1 variation of E_c can therefore be expressed as

$$\left(\frac{\delta E_c}{\delta\phi_\alpha(\mathbf{r})}\right)_1 = 4\sum_\beta K_{\alpha\beta}(\hat{C}\phi_\beta)(\mathbf{r}), \quad \text{(A14)}$$

where $(\hat{C}\phi_\beta)(\mathbf{r})$ represents the action of the operator $\hat{C}$ on ϕ_β evaluated at the point $\mathbf{r}$.

Now consider the type 2 variation of E_c due to the variation of the overlap matrix elements. The variation of $S_{\alpha\beta}$ is

$$\delta S_{\alpha\beta} = \int d\mathbf{r}(\phi_\alpha \delta\phi_\beta + \delta\phi_\alpha \phi_\beta), \tag{A15}$$

and the type 2 variation of E_c is then obtained by applying this expression to Eq. (A1). After a little manipulation, one obtains

$$\left(\frac{\delta E_c}{\delta\phi_\alpha(\mathbf{r})}\right)_2 = 12\sum_\beta (LCL)_{\alpha\beta}\phi_\beta(\mathbf{r}) - 8\sum_\beta (LSLCL + LCLSL)_{\alpha\beta}\phi_\beta(\mathbf{r}). \tag{A16}$$

Here, C is the matrix whose elements are

$$C_{\alpha\beta} = \int d\mathbf{r}\phi_\alpha \hat{C}\phi_\beta. \tag{A17}$$

Combining Eqs. (A14) and (A16), we obtain the following expression for the total variations of the kinetic and pseudopotential energies:

$$\frac{\delta E_c}{\delta\phi_\alpha(\mathbf{r})} = 4\sum_\beta [K_{\alpha\beta}\hat{C} + 3(LCL)_{\alpha\beta} - 2(LSLCL + LCLSL)_{\alpha\beta}]\phi_\beta(\mathbf{r}). \tag{A18}$$

For the remaining terms (Hartree and exchange-correlation), variation in the energy results from variation in the electron density. Thus we need to calculate

$$\frac{\delta n(\mathbf{r}')}{\delta\phi_\alpha(\mathbf{r})} = \frac{\delta}{\delta\phi_\alpha(\mathbf{r})} 2\sum_{\beta\gamma} \phi_\beta(\mathbf{r}') \times [3(LSL)_{\beta\gamma} - 2(LSLSL)_{\beta\gamma}]\phi_\gamma(\mathbf{r}'). \tag{A19}$$

Again, we will have variations coming directly from the change in $\phi_\alpha(\mathbf{r})$ and variations coming indirectly from changes in the overlap matrix elements. The total variation of $n(\mathbf{r}')$ will be

$$\frac{\delta n(\mathbf{r}')}{\delta\phi_\alpha(\mathbf{r})} = 2\,\delta(\mathbf{r}-\mathbf{r}')\sum_\beta [\phi_\beta(\mathbf{r})K_{\beta\alpha} + K_{\alpha\beta}\phi_\beta(\mathbf{r})] + 2\sum_{\beta\gamma} \phi_\beta(\mathbf{r}')\phi_\gamma(\mathbf{r}')\frac{\delta}{\delta\phi_\alpha(\mathbf{r})} \times [3(LSL)_{\beta\gamma} - 2(LSLSL)_{\beta\gamma}]. \tag{A20}$$

Substituting this expression into

$$\frac{\delta E_v}{\delta\phi_\alpha(\mathbf{r})} = \int d\mathbf{r}' V(\mathbf{r}')\frac{\delta n(\mathbf{r}')}{\delta\phi_\alpha(\mathbf{r})}, \tag{A21}$$

where the quantity $V(\mathbf{r}) \equiv \delta E_v / \delta n(\mathbf{r}')$ represents the Hartree or exchange-correlation potential, we find, after some manipulation:

$$\frac{\delta E_v}{\delta\phi_\alpha(\mathbf{r})} = 4\sum_\beta [K_{\alpha\beta}V(\mathbf{r}) + 3(LVL)_{\alpha\beta} - 2(LSLVL + LVLSL)_{\alpha\beta}]\phi_\beta(\mathbf{r}). \tag{A22}$$

Combining this expression for the Hartree and exchange-correlation derivatives with Eq. (A18) for the kinetic and pseudopotential derivatives, we find

$$\frac{\delta E_{\rm tot}}{\delta\phi_\alpha(\mathbf{r})} = 4\sum_\beta [K_{\alpha\beta}\hat{H} + 3(LHL)_{\alpha\beta} - 2(LSLHL + LHLSL)_{\alpha\beta}]\phi_\beta(\mathbf{r}), \tag{A23}$$

where $\hat{H}$ is the Kohn-Sham operator, and H is its matrix representation in the basis of support functions.

Note that in the practical grid-based calculations, the derivative we actually want is $\partial E_{\rm tot}/\partial\phi_\alpha(\mathbf{r}_\not)$, which describes the variation of $E_{\rm tot}$ with respect to change of ϕ_α at the grid point $\mathbf{r}_\not$. The formula for $\partial E_{\rm tot}/\partial\phi_\alpha(\mathbf{r}_\not)$ is identical to Eq. (A23) except that we need to multiply by the volume per grid point $\delta\omega$.

[1] See, e.g., M.C. Payne, M.P. Teter, D.C. Allan, T.A. Arias, and J.D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1993); M.J. Gillan, in *Computer Simulation in Materials Science*, edited by M. Meyer and V. Pontikis (Kluwer, Dordrecht, 1991); G. Galli and A. Pasquarello, in *Computer Simulation in Chemical Physics*, edited by M.P. Allen and D.J. Tildesley (Kluwer, Dordrecht, 1993).

[2] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).

[3] W. Yang, J. Mol. Str. (Theochem) **255**, 461 (1992).

[4] Examples are given in C. Lee and W. Yang, J. Chem. Phys. **96**, 2408 (1992); J.-P. Lu and W. Yang, Phys. Rev. B **49**, 11 421 (1994).

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

[6] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).

[7] F. Mauri, G. Galli, and R. Car, Phys. Rev. B **47**, 9973 (1993).

[8] F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).

[9] P. Ordejón, D.A. Drabold, M.P. Grumbach, and R.M. Martin, Phys. Rev. B **48**, 14 646 (1993).

[10] P. Ordejón, D.A. Drabold, R.M. Martin, and M.P. Grumbach, Phys. Rev. B **51**, 1456 (1995).

[11] J. Kim, F. Mauri, and G. Galli, Phys. Rev. B **52**, 1640 (1995).

[12] X.P. Li, R.W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993).

[13] S.Y. Qiu, C.Z. Wang, K.M. Ho, and C.T. Chan, J. Phys. Condens. Matter **6**, 9153 (1994).

[14] W. Hierse and E.B. Stechel, Phys. Rev. B **50**, 17 811 (1994).

[15] E. Hernández and M.J. Gillan, Phys. Rev. B **51**, 10 157 (1995).

[16] E.B. Stechel, A.R. Williams, and P.F. Feibelman, Phys. Rev. B **49**, 10 088 (1994).

[17] W. Kohn, Int. J. Quantum Chem. **56**, 229 (1995).

[18] W. Yang and T.-S. Lee, J, Chem. Phys. **103**, 5674 (1995).

[19] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).

[20] S. Goedecker and M. Teter, Phys. Rev. B **51**, 9455 (1995).

[21] R. Haydock, V. Heine, and M.J. Kelly, J. Phys. C **5**, 2845 (1972).
[22] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).
[23] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L.J. Sham, *ibid.* **140**, A1133 (1965); for a review, see R.O. Jones and O. Gunnarsson, Rev. Mod. Phys. **61**, 689 (1989).
[24] N.D. Mermin, Phys. Rev. **137**, A1441 (1965)
[25] M.J. Gillan, J. Phys. Condens. Matter **1**, 689 (1989)
[26] M.P. Grumbach, D. Hohl, R.M. Martin, and R. Car, J. Phys. Condens. Matter **6**, 1999 (1994)
[27] W. Kohn, Phys. Rev. **115**, 809 (1959)
[28] R.W. Nunes and D. Vanderbilt, Phys. Rev. Lett. **73**, 712 (1994)
[29] J.R. Chelikowsky, N. Troullier, and Y. Saad, Phys. Rev. Lett. **72**, 1240 (1994).
[30] E.L. Briggs, D.J. Sullivan, and J. Bernholc, Phys. Rev. B **52**, R5471 (1995).
[31] F. Gygi, Europhys. Lett. **19**, 617 (1992); Phys. Rev. B **48**, 11 692 (1993); **51**, 11190 (1995).
[32] G. Zumbach, N.A. Modine, and E. Kaxiras (unpublished).
[33] D.R. Hamann, Phys. Rev. B **51**, 7337 (1995).
[34] D.R. Hamann, Phys. Rev. B **51**, 9508 (1995).
[35] R.W. Nunes and D. Vanderbilt, Phys. Rev. B **50**, 17 611 (1994).
[36] J.A. Appelbaum and D.R. Hamann, Phys. Rev. B **8**, 1777 (1973).
[37] D.M. Ceperley and B.J. Alder, Phys. Rev. Lett. **45**, 566 (1980); J. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[38] A. Brandt, Math. Comput. **31**, 333 (1977).
[39] W.L. Briggs, *A Multigrid Tutorial* (SIAM Books, Philadelphia, 1987).
[40] S. Costiner and S. Ta'asan, Phys. Rev. E **51**, 3704 (1995).
[41] S. Costiner and S. Ta'asan, Phys. Rev. E **52**, 1181 (1995).
[42] D. Vanderbilt (private communication).