Unconstrained and constrained minimization, localization, and the Grassmann manifold: Theory and application to electronic structure

David Raczkowski and C. Y. Fong
Department of Physics, University of California, Davis, California 95616-8677

Peter A. Schultz and R. A. Lippert
Sandia National Laboratories, Albuquerque, New Mexico 87185-1413

E. B. Stechel
Ford Motor Co. SRL MD 3083, Dearborn, Michigan 48121-2053

(Received 2 February 2001; published 20 September 2001)

An unconstrained minimization algorithm for electronic structure calculations using density functional for systems with a gap is developed to solve for nonorthogonal Wannier-like orbitals in the spirit of E. B. Stechel, A. R. Williams, and P. J. Feibelman [Phys. Rev. B **49**, 10 008 (1994)]. The search for the occupied subspace is a Grassmann conjugate gradient algorithm generalized from the algorithm of A. Edelman, T. A. Arias, and S. T. Smith [SIAM J. Matrix Anal. Appl. **20**, 303 (1998)]. The gradient takes into account the nonorthogonality of a local atom-centered basis, Gaussian in our implementation. With a localization constraint on the Wannier-like orbitals, well-constructed sparse matrix multiplies lead to $O(N)$ scaling of the computationally intensive parts of the algorithm. Using silicon carbide as a test system, the accuracy, convergence, and implementation of this algorithm as a quantitative alternative to diagonalization are investigated. Results up to 1458 atoms on a single processor are presented.

DOI: 10.1103/PhysRevB.64.155203 PACS number(s): 71.15.−m, 71.20.Nr, 31.15.Ar, 31.15.Pf

I. INTRODUCTION

In the past decade, numerous minimization techniques have appeared in the condensed matter literature that solve for the ground state in electronic structure calculations.[1–9] A common bottleneck in Hartree-Fock (HF), density functional theory (DFT), and tight-binding (TB) methods has been the $O(N^3)$ scaling in the computational effort required to generate the ground state solution, where N is proportional to the number of particles in the system. For systems with an energy gap, e.g., insulators, semiconductors, or molecules, minimization techniques offer methods to calculate the charge densities and total energies with computational effort that scales linearly $O(N)$ with the size of the system.

Minimization techniques achieve linear scaling by taking advantage of well-known chemical intuition recently summarized in the ''near-sightedness'' principle.[19] Two sufficient separated regions of a molecule, bulk system, surface, etc., should not interact strongly. Implicit in this principle is a localization condition: what is a range beyond which interactions can be neglected? This is not a deterministic question, but is a balance between accuracy and computational efficiency. Smaller localization regions (see Sec. V) imply longer range interactions are truncated, leading to reduced accuracy, but lead to computational savings as fewer matrix elements need be computed. $O(N)$ minimization techniques have mostly been implemented in TB methods, where the accuracy is not as great (see most HF and DFT formulations) and thus smaller localization regions are more tolerable. The resulting matrices are sparser, leading to an earlier crossover point to the regime where linear scaling minimization techniques are more efficient, with acceptable accuracy, than diagonalization to obtain the electronic ground state.

The basis sets of DFT and HF are larger and longer range than in TB, leading to Hamiltonian (or Fock) matrices that are larger and much less sparse. The longer range, and more computationally demanding, interactions of HF and DFT postpone the crossover point of $O(N)$ minimization compared to diagonalization, limiting the benefits of linear scaling algorithms for finding the ground state solution. The crossover point is achievable at smaller system sizes through tighter localization regions but only with a loss in accuracy. Hence, whether it is advantageous to substitute a linear scaling minimization for explicit diagonalization depends not only on system size, but also on the degree of accuracy desired in the solution. In this paper, we will examine the question of how desired accuracy affects the crossover point at which the linear scaling algorithms become more efficient than diagonalization in DFT calculations.

We present the implementation of a Grassmann conjugate gradient (GCG) minimization method into a Gaussian-based DFT code. We use nonorthogonal orbitals to span the occupied space, and discuss the technical issues related to the practical implementation of the minimization algorithm with localization. Using this implementation, we investigate the natural length scales of localization regions. The tradeoff between accuracy and computational efficiency and scaling behavior is discussed in detail. The method is demonstrated in full basis calculations of SiC bulk systems containing up to as many as 1458 atoms and 18 954 basis functions.

Section II gives some mathematical background comparing and contrasting the usual eigenvalue problem with the minimization algorithms. Section III introduces the common geometric framework of the minimization techniques: the Grassmann manifold.[20,21] The aim of this section is to present this recent mathematical insight in a concise, acces-

 ©2001 The American Physical Society

sible manner and to add new perspectives that pertain to linear scaling. Comments on the density matrix formalism are inserted for additional insight and completeness. Section IV presents our minimization algorithm, which has application to any problem that requires the sum of any number of the lowest eigenvalues of a standard or generalized eigenvalue problem and the subspace spanned by the corresponding eigenvectors. Effort is especially made to ease conceptual understanding while maintaining mathematical rigor. The description of vector spaces is general and deals with possible nonorthogonality at every level. Section V relates how localization of the physical system at different levels translates into linear scaling of the minimization algorithm. We also present information specific to the linear scaling implementation of the minimization algorithm. Section VI gives results and numerical analysis for our test system, silicon carbide (SiC). Our intent of this section is to evaluate the ability of an implementation of the algorithm to obtain quantitatively usable relative energies. We investigate this ability by mapping out the accuracy of different localization regions compared to diagonalization.

II. MATHEMATICAL BACKGROUND

The ground-state total energy E_T and charge density $n(\mathbf{r})$ of a molecule or condensed matter are fundamental quantities of a system.[22] Within an independent electron picture, DFT is a common method to calculate these quantities. There are two main variational formulations for these calculations—density matrix[12–19] and orbital[1–11] approaches. Both give the same total energy and charge density but obtain them in slightly different ways. Concerning notation, boldface will be used for vectors and matrices.

The standard orbital formulation for systems with a gap solves a generalized matrix eigenvalue problem

$$\mathbf{H}\Psi = \mathbf{S}\Psi\mathbf{E}. \tag{1}$$

For a given representation (basis set of M functions), $\mathbf{H}$ is the $M\times M$ Hamiltonian matrix. $\mathbf{S}$ is the overlap matrix, and equals $\mathbf{I}$, the unit matrix, in an orthonormal basis. Ψ is the $M\times N$ matrix comprised of the expansion coefficients of the M basis functions for the N lowest normalized eigenvectors of $\mathbf{H}$. Ψ diagonalizes $\mathbf{H}$, creating the diagonal $N\times N$ matrix $\mathbf{E}$, and defines the *occupied* vector space. In general, the eigenvectors are delocalized, making Ψ a dense matrix.

When calculating E_T and $n(\mathbf{r})$ of systems with a gap, all eigenvectors are weighted equally (two in a spin unpolarized calculation) allowing only the calculation of the collective properties of the eigenvectors. We define the band energy

$$E_B = 2\ \mathrm{Tr}[\mathbf{E}]. \tag{2}$$

We define $\Psi(\mathbf{r})$ to be the projection of the *occupied* space onto real space; therefore, if we approximate real space by a mesh of 100 points, $\Psi(\mathbf{r})$ would be a matrix of $100\times N$ yielding

$$n(\mathbf{r}) = \text{diagonal of } 2[\Psi(\mathbf{r})\Psi^{\dagger}(\mathbf{r}')], \text{ thus } \mathbf{r}=\mathbf{r}'. \tag{3}$$

E_T is obtained from E_B and $n(\mathbf{r})$ by Eq. (2.25) of Ref. 23. Since E_B depends upon the contribution $n(\mathbf{r})$ makes to $\mathbf{H}$, E_T is obtained only after a sufficient number of self-consistent cycles involving a diagonalization of $\mathbf{H}$ at each cycle.

This formulation alleviates the burden of orthogonalization of the wave functions. For transformations $\Psi\mathbf{A}=\Phi$, with $\mathbf{A}$ being any $N\times N$ nonsingular matrix, Φ spans the same space as Ψ. E_B and $n(\mathbf{r})$ are invariant as long as the equations are generalized to handle nonorthogonality and lack of explicit normalization. The essential information, E_B and $n(\mathbf{r})$, traditionally obtained by diagonalization, are now obtained in a different manner that allows for $O(N)$ scaling for the solution of the ground state.

For the N lowest eigenfunctions, the matrix eigenvalue problem is formally equivalent (i.e., $\Phi=\Psi$) to minimizing the trace of a matrix Rayleigh quotient equation[1,24]

$$E_B(\Phi) = 2\ \mathrm{Tr}[(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\Phi^{\dagger}\mathbf{H}\Phi] \tag{4}$$

with respect to Φ, under the constraints that $\Phi^{\dagger}\mathbf{S}\Phi=\mathbf{I}$, and $\Phi^{\dagger}\mathbf{H}\Phi=\mathbf{E}$ is a diagonal matrix.[7,8] As all occupations are equal; removing these constraints leaves E_B unaltered as long as $\Phi^{\dagger}\mathbf{S}\Phi$ is not a singular matrix. Φ is nonorthogonal, and most importantly for our purposes, is made local (see Sec. V). The charge density $n(\mathbf{r})$ is preserved using

$$n(\mathbf{r}) = \text{diagonal of } 2[\Phi(\mathbf{r})(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\Phi^{\dagger}(\mathbf{r}')], \tag{5}$$

thus making E_T equal to the traditional method.[25]

Another alternative to the eigenvalue equation is the density matrix formulation, which is similar to Eq. (4) also obtains information only about the entire system and not individual states. The band energy is minimized with respect to the density matrix $\mathbf{P}$, a Hermitian $M\times M$ matrix

$$E_B(\mathbf{P}) = \mathrm{Tr}[\mathbf{PH}]. \tag{6}$$

As $\mathbf{P}=\Phi(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\Phi^{\dagger}$, and from the relation $\mathrm{Tr}[\mathbf{AB}]=\mathrm{Tr}[\mathbf{BA}]$, we can see that Eq. (4) and Eq. (6) are equivalent. The matrix $\mathbf{P}$ is obtained directly without having to calculate Φ. However the idempotent constraint $\mathbf{PSP}=\mathbf{P}$, automatic when $(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}$ is calculated, must be enforced for the ground state solution. This is usually done via (see Sec. III) the McWeeny purification[26]

$$3\mathbf{PSP}-2\mathbf{PSPSP}\rightarrow\mathbf{P}. \tag{7}$$

III. GEOMETRY OF THE VECTOR SPACE: THE GRASSMANN MANIFOLD

The Grassmann manifold[20,21] of rank N is the set of all subspaces of rank N in some ambient (*primitive*) M-dimensional space. This manifold is comprised of orthogonal (or nonorthogonal, but linearly independent) wave functions or idempotent density matrices. In the density matrix representation, each Hermitian $M\times M$ matrix $\mathbf{P}$ defines the electronic occupation. Most matrices have occupation magnitudes (eigenvalues of $\mathbf{P}$) that violate the requirements of an electronic ground state. The occupation magnitudes

must be 0 or 1 and thus create an idempotent surface,[15] the Grassmann manifold.

In the nonorthogonal orbital representation, we rearrange Eq. (4) by defining

$$\bar{\Phi}^{\dagger}=(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\Phi^{\dagger}, \tag{8}$$

yielding

$$E_B(\bar{\Phi}^{\dagger},\Phi)=\mathrm{Tr}[\bar{\Phi}^{\dagger}H\Phi] \tag{9}$$

as the functional to be minimized. We choose $\bar{\mathbf{\Phi}}^{\dagger}$ as our variable instead of $\bar{\mathbf{\Phi}}$ as only the transpose appears in our equations. Equation (9) is cast in a dual basis[1,27–29] form where $\bar{\Phi}^{\dagger}$ is the covariant matrix, one form[30] or linear form,[31] of the matrix Φ. Biorthogonality, with orthogonality as a special case, is automatically satisfied if the inverse is calculated in Eq. (8)

$$\bar{\mathbf{\Phi}}^{\dagger}\mathbf{S}\Phi=(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\Phi^{\dagger}\mathbf{S}\Phi=\mathbf{I}. \tag{10}$$

The points in space are still defined by our occupied space, but the occupied space is now defined by $(\bar{\mathbf{\Phi}}^{\dagger},\Phi)$ and is related to $\mathbf{P}$ through[25]

$$\mathbf{P}=\Phi\bar{\mathbf{\Phi}}^{\dagger} \tag{11}$$

If $\bar{\mathbf{\Phi}}^{\dagger}$ and Φ are biorthogonal, then $\mathbf{P}$ is idempotent; therefore, the point [$\mathbf{P}$ or $(\bar{\mathbf{\Phi}}^{\dagger},\Phi)$] resides on the Grassmann manifold. Given a nonsingular $N\times N$ matrix $\mathbf{A}$ and $\bar{\mathbf{\Phi}}_A^{\dagger}$, the biorthogonal complement of $\Phi\mathbf{A}$, $(\bar{\mathbf{\Phi}}^{\dagger},\Phi)$ and $(\bar{\mathbf{\Phi}}_A^{\dagger},\Phi\mathbf{A})$ through Eq. (11) create the same $\mathbf{P}$ and thus define the same point on the Grassmann manifold.

An equivalent perspective defines each point by Φ and a covariant metric of the occupied subspace creating $\bar{\mathbf{\Phi}}^{\dagger}$ from Φ. If the correct metric is used, the point lies on the Grassmann manifold. It is in this space Φ and a metric, that we minimize $E_B(\Phi)$, with respect to Φ, under the constraint that the minimum resides on the Grassmann manifold. The reason for introducing a constraint manifold into a previously unconstrained problem is that, in an asymptotically linear scaling algorithm, the exact covariant metric of the occupied subspace $(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}$, is not calculated. Algorithms must address the possible departure and return to the manifold.

If a point $\mathbf{P}$ resides off of the Grassmann manifold, iterations of Eq. (7) define a set of translations, a McWeeny path,[15] resulting in idempotency. Using sparse multiplications, this is an $O(N)$ process enforcing idempotency after every update of $\mathbf{P}$. Alternatively the polynomial in Eq. (7) replaces $\mathbf{P}$ in Eq. (6) creating a new functional. In this method, the minimization path ideally moves in close proximity to the Grassmann manifold with the minimum satisfying idempotency.

In the orbital formulation, adhering to the Grassmann manifold when $\Phi^{\dagger}\mathbf{S}\Phi$ is not appreciably sparse (see Sec. V) is straightforward. For these system sizes, it is most efficient to calculate the dense $(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}$. Eventually the asymptotic $O(N^3)$ scaling of the dense inverse will dominate. Methods for satisfying the constraint with linear scaling effort are conceptually very similar to density matrix methods. Some orbital methods[2–5] also create a new functional with an implicit penalty that keeps the minimization path close to the Grassmann manifold. Other proposals use transformation iterations[21] on Φ or a direct calculation[1] of an approximate $(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}$ for remaining on the Grassmann manifold after each update of Φ.

It is not plainly evident which method is superior. A nonorthogonal orbital method might be preferable over orthogonal methods given that nonorthogonal orbitals are typically more localized[32,33] due to a prefactor.[34] Algorithms that are allowed to wander off of the Grassmann manifold have shown slower convergence rates than ones that enforce biorthogonality at every step.[35] This behavior was only studied for dense linear algebra, but the properties should also apply to sparse linear algebra.

IV. GRASSMANN CONJUGATE GRADIENT ALGORITHM

Within the orbital formulation, a Grassmann conjugate gradient algorithm[20,21] is used to minimize Eq. (4). Initially, we introduce the algorithm without the complexity of localization, i.e., no sparse matrices. The essential formulas in the algorithm appear in Ref. 21 with the exception of the parallel translation of the gradient and the Polak-Ribière formulation for creating conjugate directions.

A. Tangency

We need a gradient $\mathbf{G}=\nabla_{\Phi}E_B(\Phi)$ to produce new conjugate search directions for minimizing $E_B(\Phi)$. The gradient must be in the direction of greatest change of the functional. Even an infinitesimal movement in the direction of the gradient must cause E_B to change in value; therefore, the gradient must have the property that it has no component along Φ because such a direction would not change the value of E_B. This is a property of the functional and does not depend upon the representation. For this to be true, the gradient must lie in the tangent plane[20] of Φ, a constraint given by

$$\Phi^{\dagger}\mathbf{S}\mathbf{G}=0. \tag{12}$$

Since one desires the opportunity to change $E_B(\Phi)$ for any update, all search directions should also have this property.

B. Inner product

First, we must define an inner product. For an inner product $\langle\cdot\rangle$ and an infinitesimal steps in the direction of a displacement $\delta\mathbf{V}$ in the tangent plane, the inner product of the gradient with $\delta\mathbf{V}$ must equal the directional derivative in $\delta\mathbf{V}$, given by

$$\langle\mathbf{G}\cdot\delta\mathbf{V}\rangle=\delta E_B=d/ds\;\; E_B(\Phi+s\,\delta\mathbf{V})\big|_{s=0}. \tag{13}$$

The only form that keeps the inner product invariant between arbitrarily complete representations, for vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ defined for Φ at the origin, is

$$\langle\mathbf{X}_1\cdot\mathbf{X}_2\rangle=\mathrm{Real}\{\mathrm{Tr}[(\Phi^{\dagger}\mathbf{S}\Phi)^{-1}\mathbf{X}_1^{\dagger}\mathbf{S}\mathbf{X}_2]\}. \tag{14}$$

This gives $\langle \Phi \cdot \Phi \rangle = N$, formally the same constraint as normalization.

C. Gradient, search direction, and line minimization

Note that the gradient of a functional should not be confused with its differential.[21] The differential of the functional[35]

$$dE_B/d\Phi = 2(\mathbf{I} - \mathbf{S}\Phi\bar{\Phi}^\dagger)\mathbf{H}\Phi(\Phi^\dagger\mathbf{S}\Phi)^{-1} \tag{15}$$

does not satisfy Eq. (12). The gradient

$$\mathbf{G} = 2(\mathbf{S}^{-1} - \Phi\bar{\Phi}^\dagger)\mathbf{H}\Phi \tag{16}$$

does satisfy Eq. (12). The differential, Eq. (15) lies on a cone around the gradient. This can be graphically seen in Fig. 3 of White *et al.*[36] It may provide search directions that are suitable for minimization, but convergence will be degraded, as the gradient is not used. For efficiency, the differential may be preferable if incorporating $\mathbf{S}^{-1}$ is too expensive. This may be the case for finite elements[37,38] or Mehrstellen[39] finite difference representations, where the dimension of $\mathbf{S}$ is typically larger than for an atomic orbital basis.

The new search direction $\mathbf{Z}_{I+1}$ is updated by the Polak-Ribière (PR) formula[20]

$$\mathbf{Z}_{I+1} = \mathbf{G}_{I+1}[\langle(\mathbf{G}_{I+1} - \mathbf{G}_I)\cdot\mathbf{G}_{I+1}\rangle / \langle \mathbf{G}_I \cdot \mathbf{G}_I \rangle]\mathbf{Z}_I. \tag{17}$$

This formula gives slightly better convergence than the Fletcher-Reeves (FR) form. A step size λ in the search direction $\mathbf{Z}$ is chosen to minimize the functional in that direction. A quadratic approximation is used around the current Φ for the step size of λ in the direction $\mathbf{Z}$.

$$E_B(\Phi + \lambda\mathbf{Z}) = E_B(\Phi) + \lambda\mathbf{Z}\cdot dE_B/d\Phi + 1/2 \quad \lambda\mathbf{Z}\cdot\mathbf{H}\cdot\lambda\mathbf{Z}. \tag{18}$$

The second term is just the normal dot product of $\mathbf{Z}$ with the differential. The last term is the matrix element $H(\mathbf{Z},\mathbf{Z})$ of the Grassmann Hessian. The general formula of the matrix element of the Grassmann Hessian, given two tangent vectors $\mathbf{V}_1$ and $\mathbf{V}_2$, is[33]

$$H(\mathbf{V}_1, \mathbf{V}_2) = 2\,\mathrm{Tr}[\bar{\mathbf{V}}_1^\dagger\mathbf{H}\mathbf{V}_2 - \bar{\mathbf{V}}_1^\dagger\mathbf{S}\mathbf{V}_2\bar{\Phi}^\dagger\mathbf{H}\Phi]. \tag{19}$$

The overall cost of the calculation of $H(\mathbf{Z},\mathbf{Z})$ is equivalent to the calculation of the gradient because the expensive calculations of $\mathbf{HZ}$ and $\mathbf{SZ}$ are already necessary for the next functional evaluation.

To get λ, set

$$dE_B(\Phi + \lambda\mathbf{Z})/d\lambda = 0. \tag{20}$$

The step size is now easily evaluated by using Eq. (18), obtaining

$$\lambda = (\mathbf{Z}\cdot dE_B/d\Phi)/H(\mathbf{Z},\mathbf{Z}), \tag{21}$$

and Φ is updated according to

$$\Phi_{\mathrm{new}} = \Phi - \lambda\mathbf{Z}. \tag{22}$$

D. Parallel transport

For the search direction to be tangent to Φ, the current $\mathbf{Z}$ must be orthogonal to Φ_{new}, so that when Eq. (17) is used in the $I+1$ GCG iteration, $\mathbf{Z}_{I+1}$ will be orthogonal to Φ_{new}. In this manner, Hessian information is properly communicated from one iteration to the next. This is accomplished to all orders by a parallel transport of $\mathbf{Z}$ corresponding to the update in Φ. $\mathbf{G}_I$ is also translated. In Eq. (17), $\mathbf{Z}_I^{\mathrm{new}}$ and $\mathbf{G}_I^{\mathrm{new}}$ replace $\mathbf{Z}_I$ and $\mathbf{G}_I$ respectively.

$$\mathbf{Z}_I^{\mathrm{new}} = \mathbf{Z}_I + \lambda\Phi\bar{\mathbf{Z}}_I^\dagger\mathbf{S}\mathbf{Z}_I \tag{23}$$

$$\mathbf{G}_I^{\mathrm{new}} = \mathbf{G}_I + \lambda\Phi\bar{\mathbf{Z}}_I^\dagger\mathbf{S}\mathbf{G}_I. \tag{24}$$

E. Termination and convergence

The algorithm can be terminated when the change in $E_B(\Phi)$ or the norm of the gradient, $\mathrm{Re}\{\mathrm{Tr}[(\Phi^\dagger\mathbf{S}\Phi)^{-1}\mathbf{G}^\dagger\mathbf{S}\mathbf{G}]\}$, becomes smaller than a prescribed threshold. We adopt the latter. Once the threshold is achieved or when a maximum number of iterations is reached, we update the Coulomb and the exchange-correlation potential using the charge density created from Φ. For dense matrices, a threshold of 10^{-10} and a maximum iteration number of 15 was sufficient to obtain results within a few μRy of diagonalization.

This orbital minimization method has utility in any eigenvalue problem that extensively uses iterative solvers. In electronic structure, this comprises representations having a much larger number of basis functions than occupied states, e.g., plane waves,[7,10,11] finite difference,[39–41] and finite elements.[37,38]

V. SPARSE IMPLEMENTATION OF THE MINIMIZATION ALGORITHM

We use a pseudopotential Gaussian-based atomic orbital method implemented in the serial code SEQQUEST as the framework to implement the Grassmann conjugate gradient (GCG) minimization algorithm.[42] Each basis function is a single Gaussian $e^{-\alpha r^2}$ or a contracted Gaussian (a linear combination of Gaussians) multiplied by a spherical harmonic (Y_L^m) for the angular dependence. We will refer to a shell as the set of basis functions with the same radial Gaussian term and L value for the spherical harmonic but differing by the m value. We use a split-valence double zeta with polarization (DZP) basis set.

The inner shells ($L=0,1$ for silicon) are comprised of contracted Gaussians with four Gaussians that are short-ranged and thus give variational flexibility near the atom. The outer shells ($L=0,1$ for silicon) have a single Gaussian with a smaller α making it longer ranged. The polarization shell describes the lowest L value of an unoccupied atomic shell ($L=2$ for silicon). The localized nonorthogonal orbitals, which define the occupied space, are linear combinations of contracted Gaussians and are referred to as orbitals. This basis set is optimized for accuracy without explicit consideration of sparsity in $\mathbf{H}$ and $\mathbf{S}$.[43]

A. Sparsity

The implementation of linear scaling requires using a basis set whose elements are strictly localized in real space, e.g., finite elements[37,38] or a tight-binding basis,[2,6] or pseudolocalized in real space, e.g., Gaussians made local by the use of cutoffs. Using a local basis, **H** and **S** become sparse once a ''sufficient'' separation distance between basis functions is reached. The specifics of the $O(N)$ evaluation of **H** have been discussed elsewhere[42,44–48] and will not be repeated here.

Sparse storage of $M\times M$ matrices **H**, and **S** is easily adapted from the linear scaling dense creation scheme.[42] The cutoff values that determine if elements are non-negligible give errors in the total energy less than 1 μRy/atom. The sparsity pattern, a list of the positions of non-negligible elements, of **S** is used for the storage of surviving elements of **H** and those needed from **P**. A benefit of this method is that we only require the elements of **P** (used in the calculation of the total energy and charge density) that correspond to the nonzero elements **S**. The $M\times N$ matrices Φ, **Z**, and **G**, have strict distance cutoffs input by the user. Only shells within the localization sphere, measured from the center of an orbital (atom-centered or bond-centered) contribute to that orbital. Since each shell has a different spatial extent, every localized orbital has a separate localization radius for each shell of every type of atom.

The sparsity pattern of the $M\times N$ matrices of the type $\mathbf{S}\Phi$ can be determined two ways. The first method is by use of cutoffs as for Φ. In the second method, an element is kept if its value is above an input threshold value. In the results presented in Sec. VI, the sparsity pattern was calculated and held fixed for each SCF cycle. Since the initial estimate of Φ was not sufficient using the second method, the initial SCF cycle used the first method and subsequent cycles used the second method.

A feature of the orbital formulation is that while Φ becomes sparse and local we still have the information contained within an extended **P** (i.e., our method has the benefit that our local Φ has the same information as an extended **P** of which we only require a local piece). This is the result of the extended nature of the biorthogonal complement $\bar{\Phi}^{\dagger}$ (which is never explicitly stored). The ability to separate the localization lengths' scales (via localized nonorthogonal orbitals) while maintaining a less localized even delocalized density matrix **P**, is not possible with either localized orthogonal orbitals or a density matrix formulation.

$\Phi^{\dagger}\mathbf{S}\Phi$ and $\Phi^{\dagger}\mathbf{H}\Phi$ become sparse next and can be quite dense due to the interaction of the occupied orbitals mediated by **H** and **S**. Since sparsity occurs at different stages it is advantageous for the algorithm to exploit this property. If the matrices are sparse, the matrix multiplications can be carried out in $O(N)$ steps by computing only those elements that are nonzero. A matrix needs to be significantly sparse, about 10% of elements being nonzero, before sparse routines become faster than machine-dependent optimized dense routines. For the system sizes studied, the sparsity of $\Phi^{\dagger}\mathbf{S}\Phi$ and $\Phi^{\dagger}\mathbf{H}\Phi$ did not warrant sparse multiplications, and thus were kept dense. A similar observation was noted in Ref. 9.

In our implementation of the sparse matrix multiplications, we found BLAS routines to be very beneficial. We use a multiplication list for the $\mathbf{S}\Phi$ multiplication that holds information necessary to access via BLAS1 contiguous blocks of data for the multiplication of a shell-blocked row and the orbitals in a localization region. We then use BLAS3 routines for the multiplication of these dense submatrices. The other multiplications are done a similar manner.

B. Action of S^{-1}

With localization, the definition of the gradient, Eq. (16), changes. If the full S^{-1} is calculated, then the result of Eq. (16) is truncated in order for **G** to lie in the same space as Φ. This method includes geometric coupling from basis functions outside of the localization region and thus corrupts the gradient by incorporating this information which is inconsistent with the localization. This straightforward approach is unnecessarily expensive, an $O(N^3)$ process, and performed poorly compared to the following approach in empirical tests we ran.

With no localization, $\mathbf{S}^{-1}$ accounts for the curvature of the entire primitive space (basis set) in order to align the differential in the direction of greatest increase. With localization, the orbitals are restricted to lie in a certain subspace of the primitive space. An orbital is no longer affected by curvature of the entire primitive space but only by the geometry of the vector space in its localization region. Therefore, only a square submatrix of **S**, and consequently $\mathbf{S}^{-1}$, is associated with an orbital. This submatrix of **S** is inverted in order to obtain the localized gradient for a given orbital. The action of $\mathbf{S}^{-1}$ becomes a matrix multiply of a column of Φ, a single localized orbital, by its corresponding overlap inverse. This computation scales linearly.

C. Problems with convergence

With our implementation of the GCG algorithm including the effects of localization, the norm of the gradient no longer converges to zero. We do not know if this effect is a natural consequence of localization or our lack of fully understanding the geometry of the new vector space of localized orbitals. However, as long as the convergence is consistent and to a small enough value, the algorithm is reliable. The consistency provides a cancellation of errors in the relative energies, and some suitably small value of absolute error exists for any problem.

The norm of the gradient cannot be used as the stopping criteria. The convergence is measured by

$$\gamma = \mathbf{G}\cdot dE_B/d\Phi. \tag{25}$$

Without localization, this is exactly the norm of the gradient.

Also arising from localization, the step size, λ, can give an increase in the energy. This occurs more often with small localization regions. A version of Brent's algorithm[49] has been implemented to ensure a decrease in E_B. For small localization regions, sometimes a decrease in E_B cannot be found in the search direction. The basis functions necessary for a decrease have been made inaccessible due to localiza-

TABLE I. Structural and energetic results with diagonalization respectively compared to experimental and theoretical values.

	A (Bohr)	c/a	$E_{2(4)H}-E_c$ (mRy/atom)	
Cubic	4.095			Present[d]
	4.054			PW1[c]
	4.091			LMTO[b]
	4.118			PW2[a]
	4.120			EXPT[a]
2H	5.781	1.642	0.251 39	Present
	5.728	1.637	0.066 17	PW1
		1.644	0.198 52	LMTO
	5.805	1.641		PW2
	5.813	1.641		EXPT
4H	5.785	3.275	−0.034 25	Present
	5.730	3.274	−0.139 71	PW1
			−0.088 23	LMTO
	5.800			PW2
	5.807	3.271		EXPT

[a]Reference 51.
[b]Reference 52.
[c]Reference 50.
[d]Present work (LCAO).

tion. The GCG algorithm is restarted as an attempt to alleviate this problem. Sometimes a decrease cannot be found, even when the gradient is used. The algorithm is then terminated.

VI. RESULTS FOR SILICON CARBIDE

We test the accuracy of the GCG algorithm to resolve relative energies between different systems with different localization regions. SiC, a wide gap semiconductor that can be operated at high temperature and high pressure, was chosen for its technological importance[50] and for its multiple crystal phases, providing a good test suite for the accuracy of the total energy calculations. Relative energy differences between the 3C (cubic) and the 2H and 4H (hcp) phases[51] are used for this purpose.

As a check of the basis set, Table I compares our converged calculations using diagonalization within the local density approximation (LDA) to structural values from experiment,[52] and recent relative energy calculations using plane waves[51,52] (LDA) and LMTO (Ref. 53) (GGA). We relaxed the internal positions for the hcp phases. Our lattice constants lie within 1% of experiment as do one of the plane wave and the presented LMTO results (nonpresented results were stated to lie within 1%). All c/a ratios of the hcp phases are in good agreement with experiment. We obtain the proper energetic ordering of the phases, with differences on the order of 0.1 mRy/atom.

For the two-atom 3C phase, we use up to a $12\times12\times12$ Monkhorst-Pack[54] mesh. For the four-atom 2H and eight-atom 4H we used up to a $9\times14\times14$ and a $3\times10\times10$ mesh, respectively. We include the Γ point, $\mathbf{k}=0$, for the **k**-point sampling, to obtain (without explicit calculation) the total energy for a Γ point sampling of the Brillouin zone of the large supercells by using equivalent k-point sampling[55] of the primitive unit cells. Table II gives the energy/atom of the different phases for different supercells.

TABLE II. Γ point energies of supercells derived from equivalent **k**-point calculations involving the primitive unit cells

	Ry/atom		
No of atoms	3C	2H	4H
64	9.698 64	9.707 85	9.686 01
128	9.698 76	9.700 90	9.700 86
200		9.702 68	9.702 66
250	9.702 19		
288		9.703 16	9.703 14
432	9.703 12		
686	9.703 39		
1024	9.703 48	9.703 26	
1176			9.703 49
1458	9.703 52		

Varying the localization region, we investigated the accuracy, convergence, and scaling of computational effort with system size for Γ point calculations of 64 to 1458 atoms. The initial occupied orbitals were formed from sp^3 hybridizing the inner shells. The determination of the required accuracy, i.e., the energy differences to be resolved, and the corresponding localization region is logically the first step in these calculations. Generally, if the localization region is increased, higher accuracy is obtained. The localization used also affects directly the convergence and computational effort.

We start with the 64-atom system. Energies were calculated with the relaxed geometry from the converged **k**-point calculations. $\mathbf{S}\Phi$ had full growth. Table III gives the energies for the different phases for a given localization including the diagonalization result for reference and the corresponding energy difference with the 2H phase. The localization is given in order as the cutoff radius for the inner, outer, and polarization shells. A b corresponds to an orbital whose localization region is centered between two atoms and a is for an atom-centered orbital. For example, a $\{5,4,5;b\}$ setting defines bond-centered orbitals with the localization radius of the inner and polarization shells longer than outer shells. Four atom-centered orbitals on carbon were used for the larger localization region as the code takes advantage of shared sparsity patterns among orbitals to use less memory and run faster. At a 7 Bohr radius cutoff, the code was more accurate and efficient for atom-centered orbitals than bond-centered orbitals.

After four SCF cycles, the diagonalization calculations were converged to within 0.1 mRy/atom. For an equitable comparison with diagonalization and varying localization, the energy values in Table III are given after four SCF cycles and a maximum of 15 GCG iterations with a stopping criterion of $\gamma=10^{-10}$. We concentrate on the effects of the localization radius for the inner and outer shells on the absolute and relative energies. From the bond-centered calculations,

TABLE III. Energy in units of Ry/atom, for varying localization regions in the 64 atom SiC system.

	Energy/atom				
	3C	ΔE(2H−3C)	2H	ΔE(2H−4H)	4H
Diagonalization	9.698 64	0.009 20	9.707 85	0.021 83	9.686 01
$\{5,4,5;b\}$	9.690 65	0.010 83	9.701 49	0.023 28	9.678 21
$\{4,5,5;b\}$	9.678 91	0.011 92	9.690 84	0.024 17	9.666 67
$\{5,5,5;b\}$	9.692 08	0.011 71	9.703 78	0.024 09	9.679 69
$\{7,5,7;b\}$	9.693 43	0.012 00	9.705 43	0.023 59	9.681 84
$\{5,7,7;b\}$	9.693 64	0.011 33	9.705 27	0.023 36	9.681 91
$\{7,7,7;b\}$	9.694 18	0.011 65	9.705 83	0.023 08	9.682 75
$\{7,7,7;a\}$	9.697 06	0.009 83	9.706 90	0.023 06	9.683 84
$\{9,7,9;a\}$	9.697 97	0.009 32	9.707 29	0.022 37	9.684 92
$\{7,9,9;a\}$	9.697 95	0.009 26	9.707 21	0.022 16	9.685 05
$\{9,9,9;a\}$	9.698 24	0.009 17	9.707 41	0.022 00	9.685 41
$\{9,11,9;a\}$	9.698 38	0.009 18	9.707 56	0.022 01	9.685 55
$\{11,9,9;a\}$	9.698 31	0.009 21	9.707 52	0.021 98	9.685 54
$\{9,12,9;a\}$	9.698 60	0.009 05	9.707 65	0.021 90	9.685 75
$\{12,9,9;a\}$	9.698 43	0.009 11	9.707 56	0.021 95	9.685 61
$\{11,11,11;a\}$	9.698 43	0.009 21	9.707 65	0.022 03	9.685 62
$\{12,12,12;a\}$	9.698 63	0.009 14	9.707 77	0.021 88	9.685 89

we find that near the center of the localized orbital, the inner shells have a larger impact on the accuracy of the total and relative energies than the more diffuse outer shells. The inner shells give more variational freedom where the orbital has the highest probability and the strongest interactions, i.e., near the center of the bond.

With the localization radius between 5 and 9 Bohr, a larger radius for the inner shells gave lower total energies for the 3C and 2H structures, but higher for 4H. In this range, a larger radius for outer shells gave better accuracy for all energy differences. Above 9 Bohr, a larger outer-shell radius gave lower total energies for all structures, but the relative accuracy was mixed. Accuracy of the relative energies of less than 0.2 mRy/atom, which is on the order of the energy differences for the converged results, was obtained for a radius of 9 Bohr for each shell. The 9 Bohr radius corresponds to the extent of the orbitals of silicon determined by the cutoffs already in SEQQUEST, which give an accuracy of 1 μRy/atom. A radius of 12 Bohr gave an accuracy of better than 0.1 mRy/atom (also obtained with a 9 bohr radius after 10 SCF cycles). The SCF and GCG convergence with a larger radius for the outer shell did generally better.

The convergence to the GCG algorithm impacts the efficiency and accuracy of a calculation. As a measure of convergence for the GCG algorithm, we plot, in Fig. 1, the absolute value of γ for the $\{5,4,5;b\}$ setting of the 64-atom system for the first SCF cycle. Figure 1 shows three distinct stages for the convergence of the GCG algorithm: (1) starts with linear convergence typical of a conjugate gradient algorithm, (2) hits a region of flat convergence, and (3) then ends with linear convergence. This behavior is exhibited for all sizes of localization regions and occurs similarly in subsequent SCF cycles. The duration of each stage and the corresponding transition points vary with localization radius. Stage (3) is generally accompanied by γ being negative at some steps. In Fig. 1, γ becomes and stays negative at iteration number 39. Note that for no localization, the GCG algorithm converges linearly as in stage (1).

Initially, the largest gradient values are for the shells near the center of the localized orbital. This corresponds to the good initial convergence that results from the orbitals being allowed to change in the direction (shells) of steepest descent. As the gradient gets small, the size of the gradient for the shells within the localization region becomes comparable to the size for shells outside of the localization region. The error in the gradient due to localization is now of the same order of magnitude as its length. Thus, the GCG convergence stalls, as orbitals are not allowed to move in directions necessary for a significant decrease in the energy. Localization has effectively cut off the bottom of the bowl that makes the minimum. Since the direction that would take Φ to the bottom is not available, the orbitals wander around the edge of this cut. At the end, the sharp drop is most likely due to the two vectors $\mathbf{G}$ and $dE_B/d\Phi$ becoming perpendicular. $\mathbf{Z}$ and

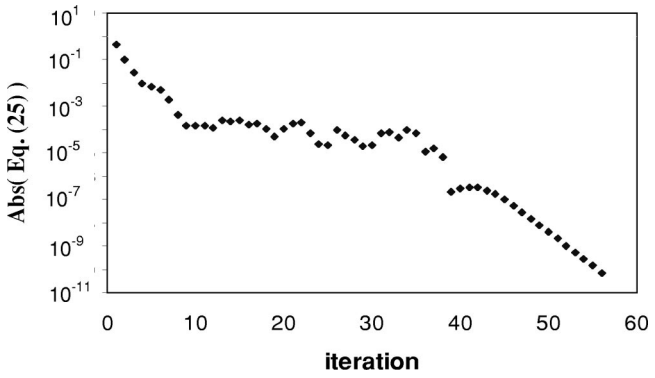

FIG. 1. $\mathbf{G}\cdot dE_B/d\Phi$ vs iteration number for the $\{5,4,5;b\}$ setting for the 64 atom unit cell of SiC.

TABLE IV. setting of $\{9,12,9;a\}$ with the growth, $\kappa=10^{-3}$.

$\{9,12,9;a\}$	Energy/atom		
	3C	2H	4H
128	9.698 50	9.700 53	9.700 48
200		9.702 22	9.702 18
250	9.701 77		
288		9.702 69	9.702 63

$dE_B/d\Phi$ also become perpendicular at this stage. This behavior can result in local minima being encountered, but depending on the localization region the difference in energy between the global and local minima may be within the desired accuracy.

As convergence with localization differs from the dense case and since accuracy depends on the convergence, convergence criteria require investigation. Because the stopping criterion of $\gamma=10^{-10}$ is rarely met in a reasonable number of iterations, the maximum number of iterations is more crucial. Generally, 15 iterations achieve the same accuracy as larger number of iterations. In some cases better accuracy is obtained with more iterations, but it is probably more efficient to use a larger localization region.

A suitable stopping criterion for SCF convergence also depends on the localization. A larger localization region will achieve SCF convergence comparable to diagonalization, and a smaller localization may never achieve the stopping criteria used for diagonalization calculations. Since less accuracy is expected from smaller localization regions, the SCF convergence can be stopped for a larger energy difference than for diagonalization and larger localization regions. For example, one can deem the $\{5,4,5;b\}$ converged once the energy change is less than 1 mRy/atom; higher accuracy is not expected. This is achieved in 4 SCF cycles. The $\{9,12,9;a\}$ converges to within 0.1 mRy/atom after 4 SCF cycles. If the SCF takes too many cycles for the desired convergence, then one needs a larger localization region.

We now concentrate on the growth parameter, κ. For the 250-atom cubic unit cell using the $\{5,4,5;b\}$ setting, a significant drop in total energy was seen above $\kappa=5\times10^{-3}$. For the accuracy expected of this system, growth cutoffs for the first SCF cycle were not crucial. For the $\{9,12,9;a\}$ localization, we looked at several system sizes to determine the growth value necessary to keep the same accuracy and the proper energetic ordering of the crystal phases. For the largest system we used, $\kappa=10^{-3}$ was sufficient. The growth for the first SCF cycle was (12,15,12) using the naming convention for the localization region. The results are presented in Table III. The accuracy for the $\{5,4,5;b\}$ localization is presented in Tables IV and V. This localization did not obtain the proper energetic ordering at every supercell size but did succeed with the proper ordering for the largest supercell done.

As a test of the scaling for the $\{5,4,5;b\}$ setting, we present in Fig. 2 the timing for 1 GCG step per orbital vs number of atoms. In contrast, perfect linear scaling gives a horizontal line. At smaller systems, linear scaling is not expected because the length scales are too small for the "nearsightedness" principle to take effect. The N^3 scaling of the inversion of the occupied space $\Phi^\dagger\mathbf{S}\Phi$ begins to be substantial around 1000 atoms. The timing for the 1458-atom system is smaller than expected. The impact of the N^3 parts should cause increasing deviation from a horizontal line. We attribute this discrepancy to fluctuations in timing due to running on a nondedicated machine or possibly cache effects. For the 1458 atom system, 11.6% of the elements of $\Phi^\dagger\mathbf{S}\Phi$ are nonzero so at this point an asymptotically linear scaling minimization might become more efficient.

TABLE V. Setting of $\{5,4,5;a\}$ with the growth $\kappa=5\times10^{-3}$.

$\{5,4,5;a\}$	Energy/atom		
	3C	2H	4H
128	9.690 78	9.690 89	9.691 10
200		9.691 62	9.691 83
250	9.692 96		
288		9.691 99	9.692 09
432	9.692 70		
686	9.692 44		
1024	9.693 49	9.691 59	
1176			9.692 67
1458	9.692 77		

In Fig. 3, we show a plot of time (s) vs diagonalization and optimization with two localization regions ($\{5,4,5;b\}$ and $\{9,12,9;a\}$). This plot displays the large disparity in the crossover point that occurs for different localization regions. For the $\{5,4,5;b\}$ setting, the crossover point is obtained early (less than 200 atoms), but the results are less accurate than diagonalization. For the $\{9,12,9;a\}$ setting, the results are as accurate as diagonalization, but the crossover point is obtained much later (approximately 700 atoms). These numbers are very important when deciding if a linear scaling algorithm is suitable for a certain problem. If one is looking at energy differences of $\sim$3 mRy/atom or only qualitative effects, the $\{5,4,5;b\}$ setting is sufficient. For very accurate calculations for systems up to at least 700 atoms, diagonalization is required. The Lapack routine DSYGVX, which cal-

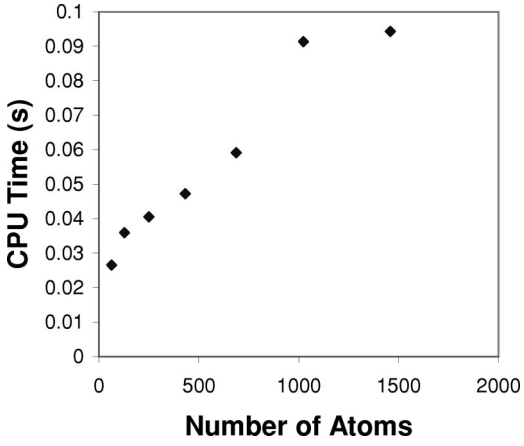

FIG. 2. CPU time for one GCG step per orbital for the $\{5,4,5;b\}$ setting vs No. of atoms.

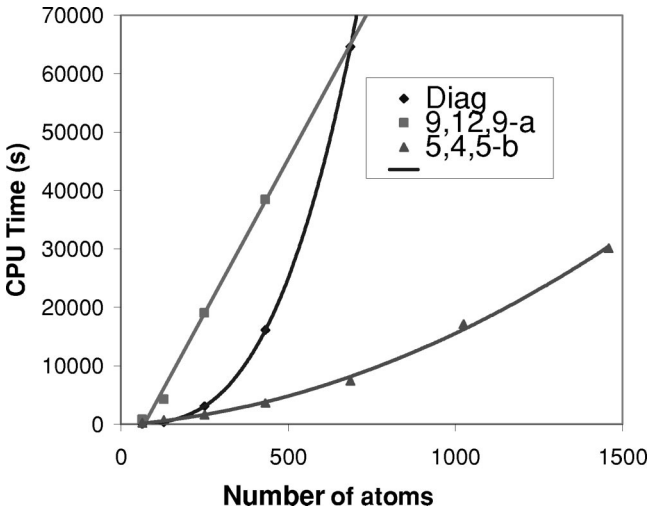

FIG. 3. Timing for 4 SCF cycles for diagonalization [9,12,9-a] and [5,4,5-b].

culates a given number of lowest eigenvalues and eigenvectors, from the optimized library DXML was used. All of the times are for a nondedicated serial 440 Mhz DEC workstation. The user time from the function ETIME was used.

VII. SUMMARY

We have introduced a method that solves for the electronic ground state in terms of localized nonorthogonal orbitals implemented in the Gaussian-based density functional code SEQQUEST. We have investigated the inherent length scales involved in the calculations and have tuned our algorithm to be as efficient as possible for systems in the range of 200–1500 atoms. We also have discussed the geometry of the involved vector spaces with reference to recent work concerning the Grassmann manifold.

The orbital solutions can be restricted in space, a localization region, with significant savings in memory requirements and computational effort. In our results, we have focused on accuracy as being the motivating factor in determining how and if the new minimization method code should be used instead of explicit diagonalization. The accuracy of relative energies of the cubic, 2H, and 4H phases of silicon carbide has been mapped out for different localization regions. We showed increasing levels of accuracy with increasing spatial extent of the localized orbitals. The crossover point with diagonalization for timings of the whole self-consistent cycle ranges from 200 to roughly 700 atoms depending on the accuracy desired. This establishes the method to be a promising quantitative tool for approaches utilizing Gaussians and other linear combination of atomic orbitals. Forces have been implemented and work utilizing them will be presented in a future paper.

ACKNOWLEDGMENTS

Partial support was provided by the Campus Laboratory collaboration of the University of California and Sandia National Laboratories. C.Y. Fong acknowledges support from a NSF Grant No. Int-9872053. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockhead Martin Company, for the United States Department of Energy, under Contract No. DEAC0494AL85000. D.B.R. acknowledges support and hospitality from Ford Motor Co. during the summer 1999.

[1] E. B. Stechel, A. R. Williams, and P. J. Feibelman, Phys. Rev. B **49**, 10 008 (1994).
[2] P. Ordejón, D. A. Diabold, R. H. Mortin, and M. P. Grumbach, Phys. Rev. B **51**, 1456 (1995).
[3] F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).
[4] W. Hierse and E. B. Stechel, Phys. Rev. B **50**, 17 811 (1994).
[5] J. Kim, F. Mauri, and G. Galli, Phys. Rev. B **52**, 1640 (1995).
[6] U. Stephan, D. A. Drabold, and R. M. Martin, Phys. Rev. B **58**, 13 472 (1998).
[7] M. P. Teter, M. C. Payne, and D. C. Allan, Phys. Rev. B **40**, 12 255 (1989).
[8] M. J. Gillan, J. Phys.: Condens. Matter **1**, 689 (1989).
[9] G. Galli and M. Parrinello, Phys. Rev. Lett. **69**, 3547 (1992).
[10] R. D. King-Smith and D. Vanderbilt, Phys. Rev. B **49**, 5828 (1994).
[11] I. Stich, R. Car, M. Parrinello, and S. Baroni, Phys. Rev. B **39**, 4997 (1989).
[12] X.-P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10 891 (1993).
[13] M. S. Daw, Phys. Rev. B **47**, 10 895 (1993).
[14] R. W. Nunes and D. Vanderbilt, Phys. Rev. B **50**, 17 611 (1994).
[15] D. R. Bowler and M. J. Gillan, Comput. Phys. Commun. **120**, 95 (1999).
[16] A. H. R. Palser and D. E. Manolopoulos, Phys. Rev. B **58**, 12 704 (1998).
[17] J. M. Millam and G. E. Scuseria, J. Chem. Phys. **106**, 5569 (1997).
[18] M. Challacombe, J. Chem. Phys. **110**, 2332 (1999).
[19] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).
[20] A. Edelman, T. A. Arias, and S. T. Smith, SIAM J. Matrix Anal. Appl. **20**, 303 (1998).
[21] R. A. Lippert and M. P. Sears, Report No. SAND99-2986, Sandia National Laboratories, Albuquerque, New Mexico, 1999.
[22] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[23] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[24] A. Edelman and S. T. Smith, BIT **36**, 494 (1996).
[25] P.-O. Löwdin, Int. J. Quantum Chem. **2**, 867 (1968).
[26] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).
[27] O. Sinanoglu, Theor. Chim. Acta. **65**, 233 (1984).
[28] E. Artacho and L. M. d. Bosch, Phys. Rev. A **43**, 5770 (1991).
[29] E. B. Stechel, in *Topics in Computational Materials Science*, edited by C. Y. Fong (World Scientific, Singapore, 1998), p. 1.
[30] B. F. Schutz, *Geometrical Methods of Mathematical Physics* (Cambridge University Press, Cambridge, 1980).
[31] R. W. R. Darling, *Differential Forms and Connections* (Cambridge University Press, Cambridge, 1994).

[32] S. Liu, J. M. Perez-Jorda, and W. Yang, J. Chem. Phys. **112**, 1634 (2000).
[33] P. W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).
[34] L. He and D. Vanderbilt, Phys. Rev. Lett. **86**, 5341 (2001).
[35] R. A. Lippert and M. P. Sears, Phys. Rev. B **61**, 12 772 (2000).
[36] C. A. White *et al.*, Chem. Phys. Lett. **276**, 133 (1997).
[37] J. Pask *et al.*, Phys. Rev. B **59**, 12 352 (1999).
[38] E. Tsuchida and M. Tsukada, J. Phys. Soc. Jpn. **67**, 3844 (1998).
[39] E. L. Briggs, D. J. Sullivan, and J. Bernholc, Phys. Rev. B **54**, 14 362 (1996).
[40] J. R. Chelikowsky, N. Troullier, and Y. Saad, Phys. Rev. Lett. **72**, 1240 (1994).
[41] N. A. Modine, G. Zumbach, and E. Kaxiras, Phys. Rev. B **55**, 10 289 (1997).
[42] P. Schultz and P. J. Feibelman (unpublished).
[43] P. A. Schultz (unpublished).
[44] C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
[45] K. N. Kudin and G. E. Scuseria, Chem. Phys. Lett. **289**, 611 (1998).
[46] E. Schwegler, M. Challacombe, and M. Head-Gordon, J. Chem. Phys. **106**, 9708 (1997).
[47] R. E. Stratmann, G. E. Scuseria, and M. J. Frisch, Chem. Phys. Lett. **257**, 213 (1996).
[48] J. M. Perez-Jorda and W. Yang, Chem. Phys. Lett. **241**, 469 (1995).
[49] J. J. More and D. J. Thuente (unpublished); R. P. Brent, *Algorithms for Minimization Without Derivatives* (Prentice-Hall, Englewood Cliffs, New Jersey, 1973), Chapter 5.
[50] V. A. Gubanov and C. Y. Fong, Appl. Phys. Lett. **75**, 88 (1999).
[51] P. Käckell, B. Wenzien, and F. Bechstedt, Phys. Rev. B **50**, 17 037 (1994).
[52] C. H. Park *et al.*, Phys. Rev. B **49**, 4485 (1994).
[53] S. Limpijumnong and W. R. L. Lambrecht, Phys. Rev. B **57**, 12 017 (1998).
[54] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[55] S. Froyen, Phys. Rev. B **39**, 3168 (1989).