Journal of Computational Physics **150,** 287–298 (1999)
Article ID jcph.1998.6181, available online at http://www.idealibrary.com on IDEAL®

Unconstrained Energy Functionals for Electronic Structure Calculations

Bernd G. Pfrommer,* James Demmel,† and Horst Simon‡

*Department of Physics and †Department of Electrical Engineering and Computer Science, Computer Science Division, University of California, Berkeley, California; and ‡NERSC Division, Lawrence Berkeley National Laboratory, Berkeley, California

Received July 6, 1998; revised December 3, 1998

The performance of conjugate gradient schemes for minimizing unconstrained energy functionals in the context of condensed matter electronic structure density functional calculations is studied. The unconstrained functionals allow a straightforward application of conjugate gradients by removing the explicit orthonormality constraints on the quantum-mechanical wave functions. However, the removal of the constraints can lead to slow convergence, in particular when preconditioning is used. The convergence properties of two previously suggested energy functionals are analyzed, and a new functional is proposed, which unifies some of the advantages of the other functionals. A numerical example derived from a diamond crystal confirms the analysis.

Key Words: electronic structure; density functional theory; unconstrained; energy functional; conjugate gradients; convergence.

1. INTRODUCTION

There is little need to motivate the interest of science in electronic structure calculations. The description of the chemical bond is probably the most celebrated success. Many other important properties of matter, such as, for example, the response to electric and magnetic fields, are also determined by the electronic structure.

The behavior of non-relativistic electrons is described by the many-electron Schrödinger equation, which is too numerically demanding to solve for most real materials, since the effort grows exponentially with the number of electrons. In density functional theory (DFT) [1, 2], the task is reduced to dealing with an effective single-particle system with much more favorable scaling properties. The fundamental theorems of DFT are that (a) the ground state energy E of a quantum-mechanical system is a functional of the *electron number density* $\rho(\mathbf{r})$ only, and (b) the true ground state density minimizes this functional [1]. Although in principle E is a functional of $\rho(\mathbf{r})$ only, in practice a "Kohn–Sham" expression [2]

0021-9991/99 $30.00

is used for accuracy reasons, involving single-particle wave functions $|\psi_i\rangle$, $i = 1, \ldots, m$. Restricting the system to be a spin-compensated insulator with N_{el} electrons, the $m = N_{\text{el}}/2$ wave functions $\{|\psi\rangle\}$ correspond to orbitals occupied by electrons. With these definitions at hand, Kohn–Sham theory reduces to the optimization problem

$$E_0 = \min_{\{|\psi\rangle\}} E[\{|\psi\rangle\}] = \min_{\{|\psi\rangle\}} 2\sum_{i=1}^{m} \langle\psi_i| - \frac{1}{2}\nabla^2|\psi_i\rangle + F[\rho]. \tag{1}$$

The electron number density $\rho(\mathbf{r})$ is a scalar function of the spatial position $\mathbf{r}$ and depends on the wave functions as

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{m} |\langle\psi_i|\mathbf{r}\rangle|^2. \tag{2}$$

Notice that the wave functions $\{|\psi\rangle\}$ are subject to orthonormality constraints:

$$\langle\psi_i|\psi_j\rangle = \delta_{ij}. \tag{3}$$

The functional $F[\rho]$ contains the ionic, exchange-correlation, and Hartree energy of the Kohn–Sham functional [2]. The exchange-correlation energy captures the complicated many-body effects, and while it is proven to exist [2], no simple and exact expression is known for it. In the local density approximation (LDA), this exchange-correlation term is approximated by a simple functional form, which depends on the local electron density only. The recently developed generalized gradient approximations (GGA) [3] improve upon LDA by including the gradient of the electron charge density into the expression for the exchange-correlation term. The resulting computational procedure is not substantially different from LDA, but the results are generally more accurate.

These days, up to several hundred atoms can be treated within DFT/LDA [4]. Many algorithms have been proposed to solve the DFT/LDA equations (see, e.g., [5–9]), but the search for more efficient schemes is still an active field [10].

2. FORMALISM

From a computational point of view, the DFT/LDA electronic structure problem is simply a minimization of a function (cf. Eq. (1)) in a large parameter space. This section introduces the necessary notation and a model functional which will be analyzed subsequently.

For optimizing (1), it is useful to know the first derivative of E with respect to the parameters $|\psi_i\rangle$:

$$\frac{\partial E}{\partial\langle\psi_i|} = 2\hat{H}|\psi_i\rangle \tag{4}$$

$$\hat{H} = -\frac{1}{2}\nabla^2 + \hat{V} \tag{5}$$

$$\hat{V} = \int_{\mathbf{r}} d^3r \frac{\delta F}{\delta\rho(\mathbf{r})}|\mathbf{r}\rangle\langle\mathbf{r}|. \tag{6}$$

This derivative *does not* take the orthonormality constraints (3) into account. Both the Hamiltonian operator $\hat{H}$ and the potential operator $\hat{V}$ are in general Hermitian operators, but for simplicity will be assumed real and symmetric here.

The constraints can be treated by introducing a set of Lagrange multipliers ϵ_i, $i = 1, \ldots, m$ (also known as Kohn–Sham eigenvalues), such that (1) becomes a non-linear eigenvalue problem

$$(\hat{H}[\rho] - \epsilon_i)|\psi_i\rangle = 0, \qquad i = 1, \ldots, m, \tag{7}$$

where the operator $\hat{H}[\rho]$ depends on the solutions $\{|\psi\rangle\}$ through (2), (5), and (6). The standard procedure for many years has been to solve (7) with a fast, iterative eigensolver, then update ρ and $\hat{H}[\rho]$ by forming ρ from the m eigenvectors with the smallest eigenvalues ϵ, and solve again, until "self-consistency" is achieved. For a large number of electrons, this scheme becomes unstable, and it is more efficient [6–8] to directly minimize (1).

A functional different from but simpler than (1) is the "non-self-consistent" functional

$$E_{\text{non-scf}}[\{|\psi\rangle\}] = 2\sum_{i=1}^{m}\langle\psi_i|\hat{H}_{\text{fixed}}|\psi_i\rangle, \qquad \langle\psi_i|\psi_j\rangle = \delta_{ij}, \tag{8}$$

in which the operator $\hat{H}_{\text{fixed}}$ *does not* depend on ρ. This functional represents simply an eigenvalue problem and can be efficiently minimized by an iterative eigensolver, e.g., based on the Davidson [11] or Lanczos [12] schemes. However, these eigensolvers have not been designed to handle a matrix $\hat{H}$ that depends on the eigenvectors.

In the following sections, the unconstrained functionals will be developed based on the non-self-consistent functional (8). This simplifies the presentation substantially. At first it seems like $E_{\text{non-scf}}$ is a rather different problem from the original one (1). However, if just $H[\rho]$ is updated as the $\{|\psi\rangle\}$ converge (i.e., at any instance ρ is consistent with $\{|\psi\rangle\}$), it retains one essential feature of the original functional: it yields the same first derivative, provided that the dependence of $\hat{H}$ on ρ is ignored when the derivative is computed. This means that the algorithms presented below are easily generalized to the "self-consistent" case by keeping H and $\{|\psi\rangle\}$ consistent. Where the differences between (1) and (8) become important, special mention will be made.

An explicit representation of the wave functions $\{|\psi\rangle\}$ allows a compact matrix notation. Expanding in terms of a finite set of N orthonormal basis functions $\{|\varphi\rangle\}$,

$$|\psi_i\rangle = \sum_{l=1}^{N} Y_{li}|\varphi_l\rangle, \tag{9}$$

the orthonormality constraint can be expressed as

$$Y^T Y = I_m \qquad (I_m \text{ is the } m \times m \text{ identity}), \tag{10}$$

since column i of Y contains the expansion coefficients of $|\psi_i\rangle$. For simplicity, Y is assumed to be real. N varies depending on the basis set and the system under study, but for the popular plane-wave basis used in the subsequent test calculations, N typically ranges from 20 to 1000 times m. Thus Y is a $(N \times m)$ tall and skinny matrix. With the expansion (9) the operator $\hat{H}$ turns into a matrix H, and the objective function (8) becomes

$$E_{\perp}[Y] = 2\,\text{tr}(Y^T H Y), \qquad Y^T Y = I_m, \tag{11}$$

where the subscript $\perp$ denotes that the Y are subject to orthonormality constraints.

3. MINIMIZING THE CONSTRAINED FUNCTIONAL

All eigensolvers minimize (11) when they compute the smallest eigenvalues and corresponding eigenvectors. In particular the trace minimization algorithms [13] expose this concept explicitly. A straightforward use of, e.g., the conjugate gradient algorithm is not possible, because the columns of Y have to be kept orthonormal during the iteration [7]. The inclusion of the constraint is not trivial, and many algorithms proposed in the literature do not exhibit some of the desirable properties of true conjugate gradients, such as quadratic convergence near the minimum [14]. Admittedly, the regime of quadratic convergence is never reached in practice, since the dimensionality of the search space (up to several millions) is orders of magnitude larger than the number of iterations (a few hundred at the most). However, since most of the proposed algorithms cannot claim to progress in conjugate directions, it is questionable whether the rate of convergence in the linear convergence regime is as good as conjugate gradients. This has been pointed out in a recent paper by Edelman *et al.* [15], who present a "correct" conjugate gradient algorithm with superlinear speedup near the minimum.

The present work will not discuss the constrained minimization, but follow the lines of Štich *et al.* [8] and eliminate the constraints by rewriting the objective function (11).

4. UNCONSTRAINED FUNCTIONAL WITH OVERLAP MATRIX INVERSION

The constraints in (11) can be removed by transforming to a set of vectors X spanning the same subspace,

$$Y = XS^{-1/2}, \qquad S = X^T X, \tag{12}$$

but not necessarily being orthonormal. The overlap matrix S is a measure of the non-orthonormality of X. This approach has been used for electronic structure calculations before [8, 9], especially for order-N schemes [16–18]. In terms of X the energy functional reads

$$E_{S^{-1}}[X] = 2\,\mathrm{tr}(S^{-1}X^T HX), \tag{13}$$

but now there are no constraints, and a standard optimization technique can be used to minimize $E_{S^{-1}}[X]$, which is a function of Nm variables. Since Nm can easily grow to several millions, conjugate gradients is the method of choice.

Conjugate gradients needs two basic ingredients: the gradient of the objective function, and a rule how to do the line search. For $E_{S^{-1}}[X]$, the gradient is

$$\frac{\partial E}{\partial X_{ij}} = 4[HXS^{-1} - XS^{-1}(X^T HX)S^{-1}]_{ij}. \tag{14}$$

From the gradient, a search direction D (a $N \times m$ matrix) is computed according to, e.g., the Polak–Ribière prescription [19]. Once D is picked, a line minimization is performed along D:

$$\begin{aligned} \min_t E_{S^{-1}}[X(t)] &= \min_t 2\,\mathrm{tr}(S^{-1}(t)X(t)^T HX(t)) \\ X(t) &= X + tD \\ S(t) &= X(t)^T X(t). \end{aligned} \tag{15}$$

At this point, one should use the true energy functional (1)—suitably generalized to non-orthonormal wave functions X—to do the line minimization. However, it is more convenient and faster to minimize the non-self-consistent functional $E_{S^{-1}}[X(t)]$ instead. Then, the line minimization becomes an inexact one. Our experience, however, shows that the inexact line search degrades the rate of convergence of the algorithm only negligibly.

Even using the simpler non-self-consistent functional, the line search is cumbersome, because one has to find the minimum of (15) by numerical methods, and for each trial step length t_{trial}, $S^{-1}(t_{\text{trial}})$ has to be computed. This is one of the main motivations for the approximate functionals presented later.

In order to compare $E_{S^{-1}}[X]$ with the other functionals discussed below, it is useful to understand the rate of convergence with which a conjugate gradient scheme will minimize (13). For *quadratic forms*, one can find rigorous upper bounds on the convergence rate of the conjugate gradient algorithm in the regime of linear convergence [20]. Linear convergence is observed when the eigenvalues are sufficiently spread out, and the number of iterations is much smaller than the number of distinct eigenvalues. Then, the error ρ_k in the objective function at iteration step k is bounded by

$$\rho_k \leq 2\left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^k \rho_0. \tag{16}$$

Here, c is the condition number of the Hessian matrix $\mathcal{H}$ associated with (13). When the eigenvalues are clustered, then the conjugate gradient algorithm may converge much faster than the above bound indicates. Indeed, in the absence of roundoff error, the algorithm will converge in k steps on a matrix with only k distinct eigenvalues. To get insight into the expected rate of convergence near the minimum, we compute the eigenvalues of $\mathcal{H}$ following Refs. [17, 18]. Since the eigenvectors $\mathbf{y}_i^{(0)}$ corresponding to the smallest eigenvalues ϵ_i, $i = 1, \ldots, m$ are known to minimize (13), one can choose them as the origin,

$$\mathbf{x}_i = \mathbf{y}_i^{(0)} + \sum_{l=1}^{N} c_l^{(i)} \mathbf{y}_l^{(0)}, \tag{17}$$

and express the deviation in terms of the *full spectrum* of the N eigenvectors of H. Inserting (17) into (13) yields to second order in the expansion coefficients $c_l^{(i)}$:

$$E_{S^{-1}} - E_0 = 2\sum_{i=1}^{m}\sum_{k=m+1}^{N} (\epsilon_k - \epsilon_i)\left(c_k^{(i)}\right)^2. \tag{18}$$

Notice that the sum over k covers the full spectrum beyond m, but the sum over i is just over the m eigenvectors with smallest eigenvalues. Since the ϵ are labeled in ascending order, we can immediately read off the smallest eigenvalue of $\mathcal{H}$ as $2(\epsilon_{m+1} - \epsilon_m)$ and the largest as $2(\epsilon_N - \epsilon_1)$. Hence the condition number c of $\mathcal{H}$ is determined by the ratio of H's spread and "gap":

$$c = \frac{\epsilon_N - \epsilon_1}{\epsilon_{m+1} - \epsilon_m}. \tag{19}$$

For fast convergence, a large gap and a small spread are necessary. Because $(\epsilon_N - \epsilon_1) \geq (\epsilon_{m+1} - \epsilon_m)$, of course, $c \geq 1$.

5. UNCONSTRAINED FUNCTIONAL WITH APPROXIMATE OVERLAP MATRIX INVERSION

As has been pointed out in Section 4, the inverse of S in the functional $E_{S^{-1}}[X]$ is undesirable. Assuming for the moment that the columns of X are almost orthonormal, S^{-1} is to first order in $(S - I)$:

$$S^{-1} \approx (2I - S). \tag{20}$$

After shifting H by η to be *negative definite*, one can show [17, 18] that the resulting functional

$$E_{2I-S}[X] = 2\,\mathrm{tr}((2I - S)X^T(H - \eta)X) \tag{21}$$

still has the "right" minimum. This means that the X minimizing $E_{2I-S}[X]$ span the same subspace as the X minimizing $E_{S^{-1}}[X]$ or the Y obtained by minimizing $E_{\perp}[Y]$. In fact, at the minimum (21) automatically yields [17, 18] a set of orthonormal X. With a proper choice of η (potentially a larger value) this holds also for the self-consistent functional, not just for the non-self-consistent functional in (21). The intuitive reason for the automatic orthonormality of X at the minimum is that $E_{2I-S}[X]$ has built-in "forces" driving the X to become orthonormal, which in turn justifies the expansion (20).

The aforementioned "forces" become evident when an expansion (17) of $E_{2I-S}[X]$ around the minimum is carried out as in Section 4. To second order one obtains

$$\begin{aligned} E_{2I-S} - E_0 = {} & 2\sum_{i=1}^{m}\sum_{k=m+1}^{N}(\epsilon_k - \epsilon_i)\left(c_k^{(i)}\right)^2 + \sum_{i=1}^{m} 8(\eta - \epsilon_i)\left(c_i^{(i)}\right)^2 \\ & + \sum_{i,j=1,j>i}^{m} 8\left(\eta - \frac{\epsilon_i + \epsilon_j}{2}\right)\left(\frac{c_i^{(j)} + c_j^{(i)}}{\sqrt{2}}\right)^2. \end{aligned} \tag{22}$$

In addition to the first term (also present in (18)), there is the second term which drives the X to be of unit length, and the third term leading to orthogonality. Equation (22) shows that the shift η should be at least $\eta > \epsilon_m$ to make all eigenvalues of the Hessian $\mathcal{H}_{2I-S}$ positive. For $X^{(0)}$ to be a *global* minimum of (21), η must be greater than the largest eigenvalue ϵ_N.

To get fast convergence, η should be chosen such that the condition number of $\mathcal{H}_{2I-S}$ is as small as possible. In other words, the eigenvalues of $\mathcal{H}_{2I-S}$ from the second and third term should fall within the range of eigenvalues generated by the first term. The proper choice of η is

$$\frac{\epsilon_{m+1} - \epsilon_m}{4} + \epsilon_m \le \eta \le \frac{\epsilon_N - \epsilon_1}{4} + \epsilon_1. \tag{23}$$

In case such an η exists, the condition numbers of $\mathcal{H}_{2I-S}$ and $\mathcal{H}_{S^{-1}}$ are identical, and therefore the conjugate gradient algorithm converges at the same rate. A numerical example of this will be shown in Section 8.

The main advantage of E_{2I-S} over $E_{S^{-1}}$ is the simplicity of the line minimization, which now does not involve an explicit inverse of S. Rather, the line minimization can be done exactly by finding the minimum of a fourth order polynomial (this is only valid for the non-self-consistent functional). The order-N schemes prefer E_{2I-S} because it does not involve a poorly scaling explicit matrix inverse.

6. IMPROVED UNCONSTRAINED FUNCTIONAL WITH APPROXIMATE OVERLAP MATRIX INVERSION

As shown in Section 5, the expansion (20) of the matrix S^{-1} to first order simplifies the line minimization, and automatically [17, 18] leads to orthonormal vectors X. However, the Hessian matrix is altered, which could increase the condition number. The functional presented in this section maintains the simplicity of E_{2I-S} but reduces the potentially adverse effects on the Hessian matrix.

It has been proven [17, 18] that the expansion of S^{-1} in (13) to *even orders* in $(S-I)$ also yields a functional which has orthonormal $X^{(0)}$ at the minimum, but now H has to be shifted to be *positive definite*. Furthermore, the $X^{(0)}$ at the minimum span the subspace in which $E_{S^{-1}}$ is minimal. Expanding S^{-1} to second order in $(S-I)$ yields the first term of the functional

$$E_{3I-3S+S^2} = 2\,\mathrm{tr}((3I-3S+S^2)X^T(H+\eta')X) + 2\kappa\,\mathrm{tr}((S-I)^2). \tag{24}$$

Here, η' should be chosen to make $H+\eta'$ *positive definite*, and a second term with κ in front *has been introduced to facilitate the minimization*. Obviously, this new term will vanish at the minimum when $S = X^TX = I$, and for $\kappa > 0$ will drive the X to become orthonormal. At first it seems from the proof in Ref. [17, 18] that there is no need for the second term in (24), since the X should become automatically orthonormal. Its need will become clear when the Hessian matrix $\mathcal{H}_{3I-3S+S^2}$ of (24) is discussed in the following paragraph.

Using the expansion (17) of $E_{3I-3S+S^2}$ around the minimum as in Section 4 yields

$$E_{3I-3S+S^2} - E_0 = 2\sum_{i=1}^{m}\sum_{k=m+1}^{N}(\epsilon_k-\epsilon_i)\left(c_k^{(i)}\right)^2 + 8\kappa\left(\sum_{i=1}^{m}\left(c_i^{(i)}\right)^2 + \sum_{i=1,\,j>i}^{m}\left(\frac{c_i^{(j)}+c_j^{(i)}}{\sqrt{2}}\right)^2\right). \tag{25}$$

Now, the only second-order terms leading to orthonormality are due to the second expression in (24). Without it, a conjugate gradient scheme cannot be used to minimize $E_{3I-3S+S^2}$, since there would be special directions in parameter space along which the objective function has vanishing first and second derivatives, but is not completely flat (as it is in the case of $E_{S^{-1}}$). As numerical experiments show, an attempted conjugate gradient minimization of (24) with $\kappa = 0$ stagnates at a finite error.

The line minimization for $E_{3I-3S+S^2}$ is only slightly more effort than for E_{2I-S}. Instead of a fourth-order polynomial, now a sixth-order polynomial needs to be minimized. To get fast convergence, κ should be picked analogously to η in (23) so as to minimize the condition number of $\mathcal{H}_{3I-3S+S^2}$:

$$\epsilon_{m+1} - \epsilon_m \le 4\kappa \le \epsilon_N - \epsilon_1. \tag{26}$$

In contrast to E_{2I-S}, the shift η' of H can be picked without impact on the Hessian matrix near the minimum. Furthermore, there always exists a κ for which (26) is satisfied. The same need not be true for η in (23). Notice that only a single eigenvalue of 8κ is introduced to $\mathcal{H}_{3I-3S+S^2}$ by the second term in (24), whereas in (22), there is a range of eigenvalues due to the orthonormality terms.

In case a proper shift η exists for E_{2I-S}, and κ in $E_{3I-3S+S^2}$ satisfies (26), the two functionals should show the same rate of convergence. In practice, this is often the case *if no preconditioning is used*. Under preconditioning, the differences between E_{2I-S} and $E_{3I-3S+S^2}$ do become important (Section 8).

7. PRECONDITIONING

Preconditioning [20] accelerates the convergence of the conjugate gradient scheme by using a $(Nm \times Nm)$ matrix $\mathcal{K}$ which, when applied from the left to the Hessian matrix $\mathcal{H}$, brings the condition number of $\mathcal{K}\mathcal{H}$ as close as possible to one. Preferably, the application of $\mathcal{K}$ should not increase the operation count significantly. A simple and effective diagonal preconditioner [5] is known for the case when a Fourier basis is used in (9) to represent the wave functions. First, an approximate inverse K of H is constructed, and then an approximate inverse $\mathcal{K}$ of $\mathcal{H}$ is deduced.

When Fourier expanding the (not necessarily orthonormal) wave functions $\{|\phi\rangle\}$ corresponding to X,

$$\langle \mathbf{r}|\phi_i\rangle = \sum_{\mathbf{G}} x^{(i)}(G) e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{27}$$

the vector indices are ordered ascending with $|\mathbf{G}|$, and the expansion is truncated at a suitably large $|\mathbf{G}| = G_{\max}$. The Hamiltonian operator $\hat{H} = -\frac{1}{2}\nabla^2 + \hat{V}$ turns into a matrix:

$$H_{\mathbf{G}\mathbf{G}'} = \frac{1}{2}|\mathbf{G}|^2 \delta_{\mathbf{G}\mathbf{G}'} + V_{\mathbf{G}\mathbf{G}'}. \tag{28}$$

By construction, $V_{\mathbf{G}\mathbf{G}'}$ decays for large $|\mathbf{G}|$ or $|\mathbf{G}'|$, so for large $\mathbf{G}$, $\mathbf{G}'$, the "kinetic energy" term $\frac{1}{2}|\mathbf{G}|^2\delta_{\mathbf{G}\mathbf{G}'}$ dominates, and H is almost diagonal. This is exploited to construct an approximate inverse K of H which is essentially the one from Ref. [5]:

$$\begin{aligned} K_{\mathbf{G}\mathbf{G}'} &= \delta_{\mathbf{G}\mathbf{G}'} \frac{27 + 18s + 12s^2 + 8s^3}{27 + 18s + 12s^2 + 8s^3 + 16s^4} \\ s &= |\mathbf{G}|^2/T. \end{aligned} \tag{29}$$

The parameter T determines the value of $|\mathbf{G}|$ for which the preconditioner K starts to become $\propto 1/|\mathbf{G}|^2\delta_{\mathbf{G}\mathbf{G}'}$. For $|\mathbf{G}|^2 < T$, the preconditioner in (29) approaches the identity, since the assumption of H being diagonal is not valid here, and it is better not to precondition. In practice, T is chosen to be the maximum "kinetic energy" $T = \max_i \frac{1}{2}\sum_{\mathbf{G}}|\mathbf{G}|^2(x^{(i)}(G))^2$ of all columns $\mathbf{x}^{(i)}$ $i = 1, \ldots, m$. This turns out to give a good estimate for the regime $|\mathbf{G}|^2 > T$ where the diagonal terms start dominating $H_{\mathbf{G}\mathbf{G}'}$. In principle, K must be kept fixed during the course of the minimization to get truly conjugate directions. Numerical experiments show that T changes only little as the $\mathbf{x}^{(i)}$ converge, and sacrificing exact conjugacy by adjusting K does not change the rate of convergence.

With K as an approximate inverse of H at hand, the preconditioner $\mathcal{K}$ is constructed by replicating K onto the diagonal of $\mathcal{K}$. This preconditioner reduces the condition number of $\mathcal{H}$ by compressing the spectrum of H. As a consequence, it becomes more difficult or even impossible to find a proper choice of η in $E_{S^{-1}}$ to satisfy the condition (23). At that point, the more liberal condition (26) gives the functional $E_{3I-3S+S^2}$ an advantage over $E_{S^{-1}}$. The numerical example in Section 8 will illustrate this.

8. NUMERICAL EXAMPLE

It is instructive to look at a simple, but relevant example for testing the statements of the preceding sections. Here, the performance of the conjugate gradient algorithm is studied for a diamond crystal. Only the valence electrons are treated, assuming the core electrons do not participate in the chemical bond. The ionic cores are represented by norm-conserving pseudopotentials [21] in a separable Kleinman–Bylander form [22]. The pseudopotentials are designed to give the same energy E as the real potential, but with a much smaller Fourier basis set. Since there are two atoms in the unit cell with two valence electrons per spin for each atom, one needs to compute $m = 4$ wave functions. In the plane-wave representation, the matrix H has a size of $N = 609$. This is much smaller than typical problem sizes studied today, but it allows us to use MATLAB and an explicit representation of H for numerical experimentation. For larger matrix sizes, a straightforward parallelization is possible [23].

A direct diagonalization of the full matrix is first performed to get the spectrum shown in the inset of Fig. 1. The smallest four "occupied" eigenvalues are grouped into a smaller single eigenvalue and a triplet. They are well separated from the larger, "unoccupied" eigenvalues. This gap is critical for achieving fast convergence, since it affects the condition number of the Hessian according to (19).

The starting guess for the conjugate gradient procedure is generated by diagonalizing a 27 by 27 submatrix from the upper left corner of H, and selecting the smallest four eigenpairs.

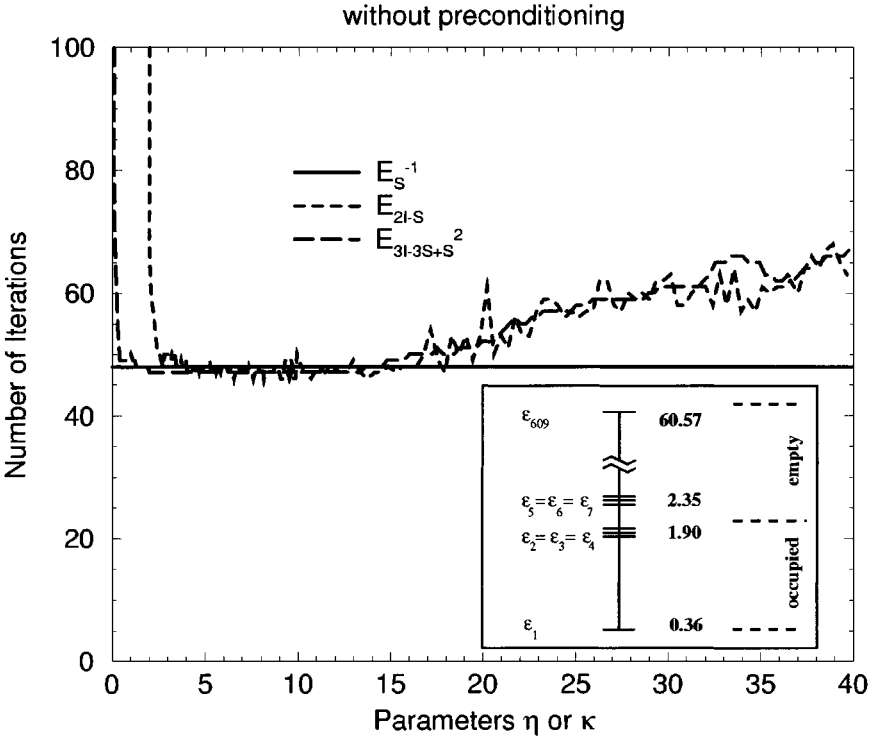

FIG. 1. Number of iterations to reach an error of 10^{-13} in the objective functions. On the abscissa are the shift parameters η or κ for a conjugate gradient algorithm performed on the energy functionals $E_{S^{-1}}$, E_{2I-S}, and $E_{3I-3S+S^2}$. No preconditioning is performed. The inset shows the spectrum of the matrix H. According to (23) and (26), the rate of convergence should be the same for all functionals if $2.01 < \eta < 15.41$ and $0.11 < \kappa < 15.05$.

The other (609-27) components of the start vectors are filled up with 0.001* rand () to ensure that the full spectrum is represented in the starting guess. The resulting vectors are orthonormalized with the MATLAB orth () command.

Without preconditioning, all three functionals $E_{S^{-1}}$, E_{2I-S}, and $E_{3I-3S+S^2}$ should exhibit similar convergence rates when minimized with a conjugate gradient algorithm. According to Eq. (23), the functional E_{2S-I} should perform best for $2.01 \leq \eta \leq 15.41$. Likewise, from (26), $E_{3I-3S+S^2}$ should give best performance for $0.11 \leq \kappa \leq 15.05$. Figure 1 shows the number of iterations to reach an error of 10^{-13} as a function of η (for E_{2S-I}) and κ (for $E_{3I-3S+S^2}$). Since $E_{S^{-1}}$ has no free parameters, it is represented by a horizontal line corresponding to 48 iterations.

As is obvious from Fig. 1, as long as the parameters η and κ are chosen within the intervals given by (23) or (26), all three functionals lead to the same rate of convergence. Once η or κ are outside these intervals, the condition numbers of the Hessian matrices for E_{2S-I} and $E_{3I-3S+S^2}$ increase, and the convergence slows down.

Under preconditioning, convergence is more rapid ($E_{S^{-1}}$ converges in 16 instead of 48 iterations), but the functionals E_{2S-I} and $E_{3I-3S+S^2}$ now show more sensitivity to the choice of η and κ (Fig. 2). The parameter T for the preconditioner (29) has been set to $T = 4$ (the physical units are Rydbergs) in order to be sure the same, fixed preconditioner is used for all functionals. No shift η exists for which E_{2S-I} converges as fast as $E_{S^{-1}}$. In contrast, for $0.4 \leq \kappa \leq 1.0$, $E_{3I-3S+S^2}$ shows the same performance as $E_{S^{-1}}$.

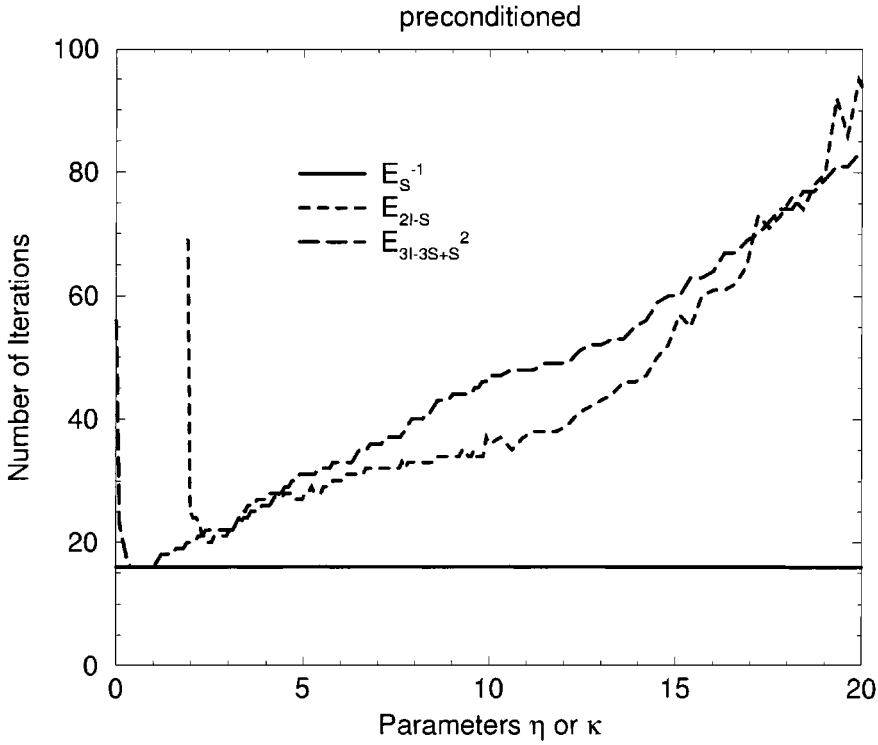

FIG. 2. Number of iterations to reach an error of 10^{-13} in the objective functions. On the abscissa are the shift parameters η or κ for a conjugate gradient algorithm performed on the energy functionals $E_{S^{-1}}$, E_{2I-S}, and $E_{3I-3S+S^2}$. The preconditioning results in better performance, but also in increased sensitivity to the choice of the parameters η and κ for E_{2I-S} and $E_{3I-3S+S^2}$.

We currently only have a parallel implementation of E_{2S-I}, which we have used for many systems, some as large as $m = 288$, and $N = 367{,}672$ (this corresponds to an optimization in a parameter space of dimension 106×10^6!). We did not encounter any numerical instabilities related to, e.g., the approximate inversion of S, but we did encounter convergence problems for systems with a small gap, consistent with our convergence analysis.

9. CONCLUSION

Three different variants of unconstrained energy functionals, $E_{S^{-1}}$, E_{2S-I}, and $E_{3I-3S+S^2}$, for electronic structure calculations have been studied comparatively. The rate of convergence for a conjugate gradient minimization of those functionals is discussed. While $E_{S^{-1}}$ does not require any shift parameters and performs best under preconditioning, it has the disadvantages of a tedious line minimization and an explicit inversion of a (small) matrix. The functional E_{2S-I}, which has been previously used for order-N calculations [17, 18], is found to be sensitive to the choice of its free parameter η and, in certain circumstances, does not achieve optimal performance under preconditioning. A new functional $E_{3I-3S+S^2}$ is proposed which is less sensitive to its shift parameter κ, while avoiding the complicated line minimization of $E_{S^{-1}}$.

ACKNOWLEDGMENTS

B.G.P. acknowledges useful discussions with S. G. Louie, and particularly with F. Mauri. A. Canning is thanked for his critical reading of the manuscript. This work was carried out at the National Energy Research Scientific Computing Center (NERSC) and is based in part upon work supported by the Advanced Research Projects Agency Contract DAAH04-95-1-0077 (via Subcontract ORA4466.02 with the University of Tennessee), the Department of Energy Grants DE-FG03-94ER25219 and DE-AC03-76SF00098, and Contract W-31-109-Eng-38 (via Subcontracts 20552402 and 941322401 with Argonne National Laboratory), the National Science Foundation Grants ASC-9313958, ASC-9005933, and CCR-9196022, and NSF Infrastructure Grants CDA-8722788 and CDA-9401156.

REFERENCES

1. P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* **136**, B864 (1964).
2. W. Kohn and L. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.* **140**, A1133 (1965).
3. J. Perdew, in *Electronic Structure of Solids '91*, edited by P. Ziesche and H. Eschrig (Akademie Verlag, Berlin, 1991), p. 11.
4. K. Brommer, B. Larson, M. Needels, and J. Joannopoulos, Implementation of the Car–Parrinello algorithm for ab initio total energy calculations on a massively parallel computer, *Comput. Phys.* **7**, 350 (1993).
5. M. Teter, M. Payne, and D. Allan, Solution of Schrödinger's equation for large systems, *Phys. Rev. B* **40**, 12255 (1989).
6. R. Car and M. Parrinello, Structural, dynamical, and electronic properties of amorphous silicon: an ab initio molecular-dynamics study, *Phys. Rev. Lett.* **55**, 2471 (1985).
7. M. Payne *et al.*, Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients, *Rev. Mod. Phys.* **64**, 1045 (1992).
8. I. Štich, R. Car, M. Parrinello, and S. Baroni, Conjugate gradient minimization of the energy functional: A new method for electronic structure calculation, *Phys. Rev. B* **39**, 4997 (1989).
9. T. A. Arias, M. C. Payne, and J. D. Joannopoulos, Ab initio molecular dynamics: Analytically continued energy functionals and insights into iterative solutions, *Phys. Rev. Lett.* **69**, 1077 (1992).

10. N. Marzari and D. Vanderbilt, Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators, *Phys. Rev. Lett.* **79**, 1337 (1997).
11. E. Davidson, The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices, *J. Comput. Phys.* **17**, 87 (1975).
12. B. N. Parlett, *The Symmetric Eigenvalue Problem* (Prentice Hall, Englewood Cliffs, NJ, 1980).
13. A. H. Sameh and J. A. Wisniewski, A trace minimization algorithm for the generalized eigenvalue problem, *SIAM J. Numer. Anal.* **19**, 1243 (1982).
14. A. Edelman, T. Arias, and S. T. Smith, Conjugate gradients on the Stiefel and Grassman manifolds, unpublished manuscript.
15. A. Edelman and S. T. Smith, On conjugate gradient-like methods for eigen-like problems, *BIT* **36**, 494 (1996).
16. G. Galli and M. Parrinello, Large scale electronic structure calculations, *Phys. Rev. Lett.* **69**, 3547 (1992).
17. F. Mauri, G. Galli, and R. Car, Orbital formulation for electronic-structure calculations with linear system-size scaling, *Phys. Rev. B* **47**, 9973 (1993).
18. F. Mauri and G. Galli, Electronic-structure calculations and molecular-dynamics simulations with linear system-size scaling, *Phys. Rev. B* **50**, 4316 (1994).
19. E. Polak, *Computational Methods in Optimization* (Academic Press, New York, 1971).
20. J. Stoer and R. Bulirsch, *Numerische Mathematik*, 3rd ed. (Springer-Verlag, New York, 1990), Vol. 2.
21. N. Troullier and J. L. Martins, Efficient pseudopotentials for plane-wave calculations, *Phys. Rev. B* **43**, 1993 (1991).
22. L. Kleinman and D. M. Bylander, Efficacious form for model pseudopotentials, *Phys. Rev. Lett.* **48**, 1425 (1982).
23. B. G. Pfrommer, Master's thesis, University of California at Berkeley, 1997.