LBNL-41735

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Unconstrained Energy Functionals for Electronic Structure Calculations

Bernd G. Pfrommer, James Demmel,
and Horst Simon

Computing Sciences Directorate
National Energy Research
Scientific Computing Division

April 1998

REFERENCE COPY
Does Not
Circulate
Bldg. 50 Library - Ref.
Lawrence Berkeley National Laboratory

LBNL-41735
Copy 1

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBNL-41735

Unconstrained Energy Functionals for Electronic Structure Calculations

Bernd G. Pfrommer

Department of Physics
University of California, Berkeley
Berkeley, California 94720

James Demmel

Department of Electrical Engineering and Computer Science
Computer Science Division
University of California, Berkeley
Berkeley, California 94720

and

Horst Simon

Computing Sciences Directorate
National Energy Research Scientific Computing Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

April 1998

This work was supported by the Director, Office of Energy Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Unconstrained Energy Functionals for Electronic Structure Calculations

Bernd G. Pfrommer

Department of Physics, University of California at Berkeley, CA.

James Demmel

Department of Electrical Engineering and Computer Science,

Computer Science Division, University of California at Berkeley

Horst Simon

NERSC Division,

Lawrence Berkeley National Laboratory, Berkeley, CA.

April 4, 1998

subject classification: 65B99, 65C20, 65F10, 65F15, 65N25, 81V55

keywords: electronic structure, density functional theory, unconstrained, energy functional, conjugate gradients, convergence.

Running head: Unconstrained Energy Functionals...

Send proofs to:

Bernd Pfrommer

Department of Physics

University of California at Berkeley

Berkeley, CA 94720

E-mail: pfrommer@physics.berkeley.edu

Telephone: (650) 786 7523

Abstract

The performance of conjugate gradient schemes for minimizing unconstrained energy functionals in the context of electronic structure 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 confirms the analysis.

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 many-electron Schrödinger Equation is well known, and describes the behavior of non-relativistic electrons correctly. It can be solved analytically for some important special cases like a uniform potential, the harmonic oscillator, or the hydrogen atom. For real materials such as molecules or solids, where the potential is complicated, and several or even a large number of electrons are present, analytic solutions are not known. The numerical solution of the many-electron Schroedinger equation in some external potential $V_{ext}(\mathbf{r})$

$$\left(\sum_{i=1}^{N_{el}} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N_{el}} V_{ext}(\mathbf{r}_i) + \sum_{i=1,j>i}^{N_{el}} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - E\right) \Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_{el}}) = 0 \tag{1}$$

becomes very demanding as the number of electrons N_{el} grows. Since the many-body wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_{N_{el}})$ is represented in the *product* space of the single-electron positions $\mathbf{r}_i$, the number of degrees of freedom grows *exponentially* with N_{el}. A brute force approach is not feasible.

Two different, but similar approximations to the many-particle Schrödinger equation have enjoyed great success during the last three decades: The Hartree-Fock (HF) approach and Density Functional Theory (DFT) in the Local Density Approximation (LDA) . The Hartree-Fock equations were discovered in 1951 [1], and were readily embraced by the quantum chemistry community because they describe the chemical bond of molecules reasonably well, and also reproduce the experimentally known binding energies of many molecules better than DFT/LDA. Density Functional Theory was founded in 1964 [2, 3], and is in principle an exact approach. However, it buries all

the difficult many-body effects inside an "exchange-correlation" energy term E_{xc}, which is proven to exist, but no simple and exact expression is known for it. In the Local Density Approximation, 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) [4] improve upon the LDA by including the gradient of the electron charge density into E_{xc}. The resulting computational procedure is not substantially different from the LDA, but the results are in general more accurate, e.g. binding energies are comparable or better than those from HF calculations.

Both HF and DFT/LDA reduce (1) to a single-particle problem, such that the individual particles are decoupled, and interact with each other only through an average effective potential. This simplifies the problem substantially, and renders calculations on real materials feasible. The DFT/LDA equations are somewhat simpler than Hartree-Fock, and allow for larger systems. These days, up to several hundred atoms can be treated within DFT/LDA[5]. Many algorithms have been proposed to solve the DFT/LDA equations (see, e.g. [6, 7, 8, 9, 10]), but the search for more efficient schemes is still an active field[11].

2 Formalism

From a computational point of view, the DFT/LDAT electronic structure problem is simply a minimization of a function in a large parameter space. This section will motivate the objective function, and give a brief introduction to the subject.

The fundamental theorems of DFT are that a) the ground state energy 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[2]. Although in principle the ground state energy E of an electron system

is a functional of $\rho(\mathbf{r})$ only, in practice a "Kohn-Sham" expression[3] 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 $m = N_{el}/2$ wave functions $\{|\psi\rangle\}$ correspond to orbitals occupied by electrons. The Kohn-Sham functional then reads

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

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

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{m} |\langle\psi_i|\mathbf{r}\rangle|^2 \; . \tag{3}$$

The functional $F[\rho]$ contains the ionic, exchange-correlation, and Hartree energy of the Kohn-Sham functional [3]. Minimizing (2) seems straight forward, but is impeded by the orthonormality constraints $\langle\psi_i|\psi_j\rangle = \delta_{ij}$.

Fortunately, the first derivative of E with respect to the parameters $|\psi_i\rangle$ is available:

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

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

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

This derivative *does not* take the orthonormality constraints 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 (2) becomes a non-linear eigenvalue problem

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

where the operator $\hat{H}[\rho]$ depends on the solutions $\{|\psi\rangle\}$ through (3), (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 [7, 6, 8, 9] to directly minimize (2).

A different, but simpler functional than (2) is the "non-selfconsistent" functional

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

in which the operator $\hat{H}_{fixed}$ *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[12] or Lanczos[13] 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-selfconsistent functional (8). This simplifies the presentation substantially. At first it seems like $E_{non-scf}$ is a rather different problem than the original one (2). 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 (2) 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:

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

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 [14] expose this concept explicitly. A straight forward use of e.g. the conjugate gradient algorithm is not possible, because the columns of Y have to be kept orthonormal during the iteration[8]. 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[15]. 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 if 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 [16], 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 Stich et al [9], 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[9, 10], especially for order-N schemes[17, 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

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

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:

$$\min_t E_{S^{-1}}[X(t)] \ = \ \min_t 2 \ \mathrm{tr}(S^{-1}(t)X(t)^T HX(t)) \tag{15}$$

$$
\begin{aligned}
X(t) &= X + tD \\
S(t) &= X(t)^T X(t) \; .
\end{aligned}
$$

At this point, one should use the true energy functional (2) – suitably generalized to nonorthonormal wave functions X – to do the line minimization. However, it is more convenient and faster to minimize the non-selfconsistent 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-selfconsistent 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}, $S^{-1}(t_{trial})$ 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 Ref.[18]. Since the eigenvectors $\mathbf{y}_i^{(0)}$ corresponding to the smallest eigenvalues $\epsilon_i, i = 1, \ldots m$ 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 $c_l^{(i)}$:

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

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[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[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-selfconsistent 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

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

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. Eq. (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:

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

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-selfconsistent 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[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[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 $S = X^T X = I$, and for $\kappa > 0$ will drive the X to become orthonormal. At first it seems from the proof in Ref.[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:

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

Now the only second order term leading to orthonormality are due to the second term 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) without the second term 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) such as to minimize the condition number of $\mathcal{H}_{3I-3S+S^2}$:

$$\epsilon_{m+1} - \epsilon_m \leq 4\kappa \leq \epsilon_N - \epsilon_1 . \quad (26)$$

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.* It is especially under preconditioning where the differences between E_{2I-S} and $E_{3I-3S+S^2}$ 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[6] 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:

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

By construction, $V_{\mathbf{GG}'}$ decays for large $|\mathbf{G}|$ or $|\mathbf{G}'|$, so for large $\mathbf{G}$, $\mathbf{G}'$, the "kinetic energy" term $\frac{1}{2}|\mathbf{G}|^2\delta_{\mathbf{GG}'}$ dominates, and H is almost diagonal. This is exploited to construct an approximate inverse K of H which is essentially the one from Ref. [6]:

$$K_{\mathbf{GG}'} \quad = \quad \delta_{\mathbf{GG}'} \frac{27 + 18x + 12x^2 + 8x^3}{27 + 18x + 12x^2 + 8x^3 + 16x^4} \tag{29}$$

$$x = |\mathbf{G}|^2/T \, .$$

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" $\frac{1}{2}\sum_{\mathbf{G}} |\mathbf{G}|^2(x^{(i)}(G))^2$ 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 to use MATLAB and an explicit representation of H for numerical experimentation.

A direct diagonalization of the full matrix is first performed to get the spectrum shown in the inset of figure 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 eigen pairs. 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 $\kappa = 0.4 \ldots 1.0$, $E_{3I-3S+S^2}$ shows the same performance as $E_{S^{-1}}$.

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[18], is found to be sensitive to the choice of its free parameter η, and, under 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 $E_{S^{-1}}$.

10 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 supported by

the Director, Office of Energy Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

References

[1] C. Roothaan, Mod. Phys. **23**, 69 (1951).

[2] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[3] W. Kohn and L. Sham, Phys. Rev. **140**, A1133 (1965).

[4] J. Perdew, in *Electronic Structure of Solids '91*, edited by P. Ziesche and H. Eschrig (Akademie Verlag, Berlin, 1991), pp. 11–20.

[5] K. Brommer, B. Larson, M. Needels, and J. Joannopoulos, Computers in Physics **7**, 350 (1993).

[6] M. Teter, M. Payne, and D. Allan, Phys. Rev. B **40**, 12255 (1989).

[7] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[8] M. Payne *et al.*, Rev. Mod. Phys. **64**, 1045 (1992).

[9] I. Štich, R. Car, M. Parrinello, and S. Baroni, Phys. Rev. B **39**, 4997 (1989).

[10] T. A. Arias, M. C. Payne, and J. D. Joannopoulos, Phys. Rev. Lett. **69**, 1077 (1992).

[11] N. Marzari and D. Vanderbilt, Phys. Rev. Lett. **79**, 1337 (1997).

[12] E. Davidson, J. Comput. Phys. **17**, 87 (1975).

[13] C. Lanczos, J. Res. Nat. Bur. Stand. **45**, 255 (1950).

[14] A. H. Sameh and J. A. Wisniewski, SIAM Journal on Numerical Analysis **19**, 1243 (1982).

[15] A. Edelman, T. Arias, and S. T. Smith, unpublished .

[16] A. Edelman and S. T. Smith, BIT **36**, 494 (1996).

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

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

[19] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes*, 2nd ed. (Cambridge University Press, New York, 1992).

[20] J. Stoer and R. Bulirsch, *Numerische Mathematik*, 3rd ed. (Springer-Verlag, New York, 1990), Vol. 2.

[21] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

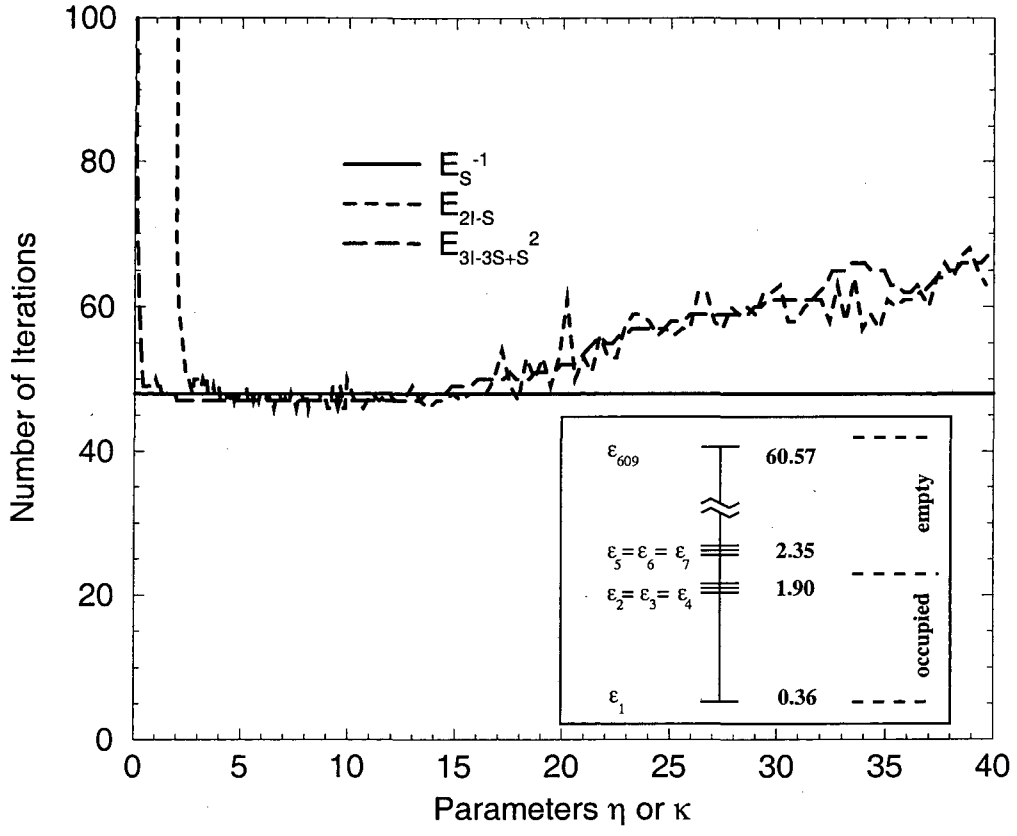
[22] L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).

Figure 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 $0.11 < \kappa < 15.05$.

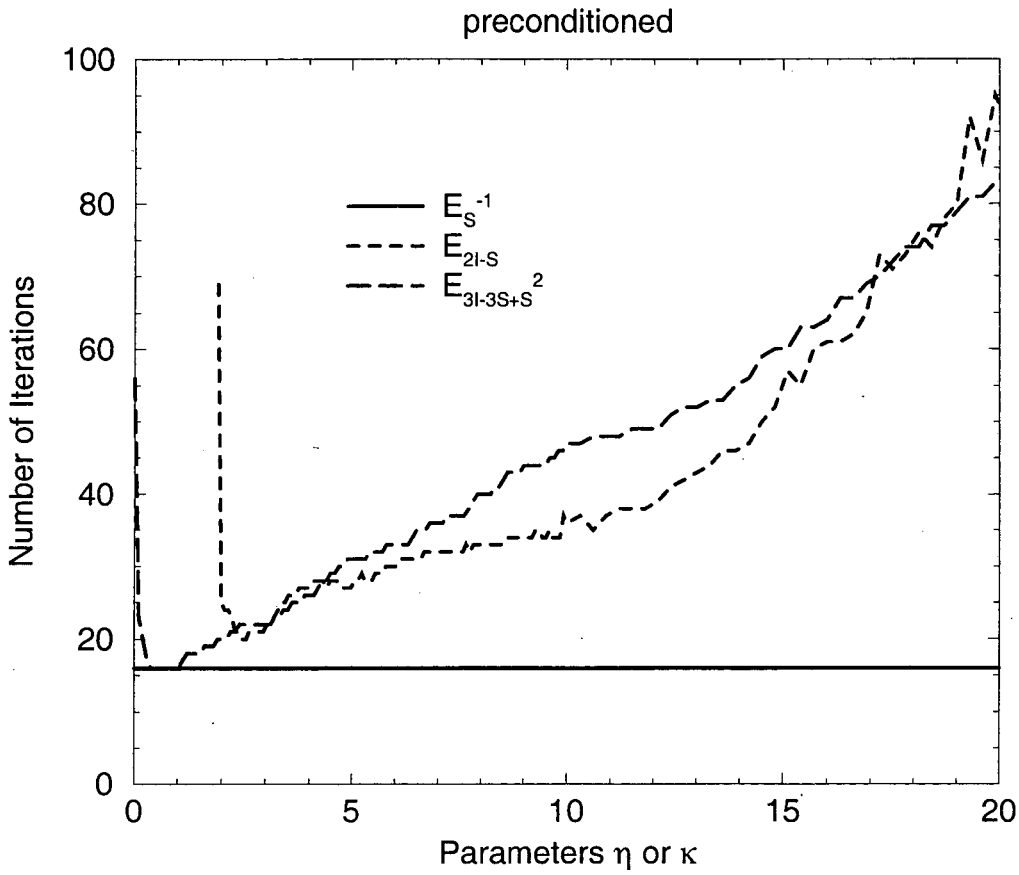
Figure 2: Number of iterations to reach an error of 10^{-13} in the objective functions. On the abscissa are the shift parameters η or κ for a conjugate gradient algorithm performed on the energy functionals $E_{S^{-1}}$, E_{2I-S}, and $E_{3I-3S+S^2}$. The preconditioning results in better performance,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

Convergence of unconstrained functionals

without preconditioning



Convergence of unconstrained functionals



Ernest Orlando Lawrence Berkeley National Laboratory
One Cyclotron Road | Berkeley, California 94720

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.