ISAS - INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

Conjugate Gradient Approach to Energy Minimization in DFT

Thesis submitted for the degree of Magister Philosophiæ

Condensed Matter Sector

Candidate:
Riccardo Valente

Supervisor:
Prof. Stefano Baroni

October 1992

SISSA - SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
TRIESTE
Strada Costiera 11

TRIESTE

Conjugate Gradient Approach to Energy Minimization in DFT

Thesis submitted for the degree of
Magister Philosophiæ

Condensed Matter Sector

Candidate:
Riccardo Valente

Supervisor:
Prof. Stefano Baroni

October 1992

Contents

1

Introduction

There is great interest in large-scale electronic structure calculations, based on Density-Functional Theory (DFT) in the Local Density Approximation (LDA). Within DFT, the electronic ground state is given by the minimum of the energy functional, which is usually obtained by solving the Kohn-Sham equations. This problem is in turn converted into self-consistent matrix eigenvalue equations. If M is the size of the basis set employed in the expansion of the electronic wavefunctions, this approach requires $\mathcal{O}(M^3)$ floating-point operations for each diagonalization; M may be very large, especially in plane-wave pseudopotential calculations. Moreover, in order to achieve self-consistency, the diagonalization must be iterated I times; typically, I is of the order of ten.

In order to reduce the size of the calculation, one tries to exploit the fact that only the lowest N eigenstates are needed to compute the ground-state energy within DFT; usually N is much less than M.

One of the possible approaches is to replace the full matrix diagonalization with a partial one, which yields only the lowest N eigenstates (e. g. Davidson block iterative technique [1, 2]). In this way the size of the calculation reduces to $\mathcal{O}(N^2M)$ in plane-wave pseudopotential calculations.

A different approach, which has been proposed in the context of *ab initio* molecular-dynamics simulations [3], regards the minimization of the energy functional as an optimization problem: this allows one to obtain simultaneously self-consistency and diagonalization. The size of calculations performed in this scheme scales again as N^2M.

In this preliminary work, we implemented two minimization techniques, based

on steepest descent (SD) and conjugate gradient (CG) methods [4]; since both allow only for downhill moves, they are suitable in situations where a single minimum is present in the functional, which is usually the case in total energy calculations performed at fixed ions. Our attention is mainly devoted to CG approach, which is generally more efficient than SD, in particular when dealing with low-symmetry systems. A further advantage of conjugate gradient method is that no parameters are needed to control convergence, such as the time step in SD or the potential mixing parameter in Davidson scheme.

2

Minimization Techniques

2.1 Steepest Descent (SD)

The simplest and most intuitive procedure to find the minimum of a function of several variables is the SD method [4], which can be cast in terms of the following:

$$\frac{\partial \psi_i(\mathbf{r},t)}{\partial t} = -\frac{\delta E}{\delta \psi_i^*(\mathbf{r},t)} - \sum_j \Lambda_{ij}\psi_j(\mathbf{r},t) \tag{2.1}$$

$$= -H\psi_i(\mathbf{r},t) - \sum_j \Lambda_{ij}\psi_j(\mathbf{r},t) \tag{2.2}$$

where t is a fictitious time variable and Λ_{ij} are Lagrange multipliers introduced in order to satisfy the orthonormalization constraints. The SD step consists in the following:

$$\psi_i(t+\Delta t) = \varphi_i(t) - \Delta t\left[H\varphi_i(t) + \sum_j \Lambda_{ij}\varphi_j(t)\right] \tag{2.3}$$

$$= \overline{\varphi}_i(t+\Delta t) - \Delta t\sum_j \Lambda_{ij}\varphi_j(t) \tag{2.4}$$

$\overline{\psi}_i(t+\Delta t)$ is the unconstrained evolved of $\varphi_i(t)$. Setting $\mathbf{X} = \Delta t\Lambda^*$, the constraint equation is

$$\mathbf{X}^2 + \mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{B} + \mathbf{A} - \mathbf{1} = 0 \tag{2.5}$$

where

$$A_{ij} = \langle\overline{\psi}_j(t)|\overline{\psi}_i(t)\rangle \tag{2.6}$$

$$B_{ij} = \langle\overline{\psi}_j(t+\Delta t)|\psi_i(t)\rangle; \tag{2.7}$$

the wavefunctions at time t are assumed to be orthonormal. Since $A = 1 + \mathcal{O}(t^2)$ and $B = 1 + \mathcal{O}(t)$, Eq. (2.5) is usually solved iteratively [5, 6]:

$$\mathbf{X}^{(0)} = \frac{1}{2}(\mathbf{1} - \mathbf{A}) \tag{2.8}$$

$$\mathbf{X}^{(n)} = \frac{1}{2}\left[\mathbf{1} - \mathbf{A} + \mathbf{X}^{(n-1)}(\mathbf{1} - \mathbf{B}) + (\mathbf{1} - \mathbf{B})\mathbf{X}^{(n-1)} - \mathbf{X}^{(n-1)^2}\right] \tag{2.9}$$

We actually solved Eq. (2.5) in a different way; when written by components, it is a quadratic equation, whose solution is given by the usual formula

$$\mathbf{X} = -\mathbf{B} \pm \sqrt{\mathbf{B}^2 - \mathbf{A} + \mathbf{1}} \tag{2.10}$$

2.2 Conjugate Gradient (CG)

Suppose that the function f to be minimized (energy) is approximated by a quadratic form

$$f(\mathbf{x}) \simeq c - \mathbf{b}\cdot\mathbf{x} + \frac{1}{2}\mathbf{x}\cdot\mathbf{A}\cdot\mathbf{x} \tag{2.11}$$

where

$$\mathbf{x} = (x_1, \ldots, x_N) \tag{2.12}$$

$$c = f(0) \tag{2.13}$$

$$\mathbf{b} = -\nabla f|_0 \tag{2.14}$$

$$A_{ij} = \left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_0 \tag{2.15}$$

One then proceeds in an iterative fashion, defining a sequence

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \lambda_i \mathbf{h}_i \tag{2.16}$$

where λ_i are scalars and $\mathbf{h}_i$ a set of directions in the N-dimensional space. As we have seen, in SD one simply sets

$$\mathbf{h}_i = -\nabla f(\mathbf{x}_i) \tag{2.17}$$

and λ_i can be either a small scalar (the time step Δt, in the previous notation) or such as to minimize f along $\mathbf{h}_i$.

The former implementation of SD may require too many function and gradient evaluations to converge; furthermore, it requires an empirical choice for the convergence parameter Δt.

On the other hand, the latter is not at all optimal either, even with quadratic forms, because two consecutive directions resulting from Eq. (2.16) are usually nearly orthogonal, possibly causing the method to perform many tiny steps down to the minimum.

This can be avoided exploiting the knowledge of the hessian matrix $\mathbf{A}$. In the quadratic approximation, the gradient ∇f at $\mathbf{x}$ is $\mathbf{A} \cdot \mathbf{x} - \mathbf{b}$; if we move along some direction we have

$$\delta(\nabla f) = \mathbf{A} \cdot (\delta \mathbf{x}) \tag{2.18}$$

Suppose that we have moved along some direction $\mathbf{u}$ to a (line) minimum and now set a new direction $\mathbf{v}$. In order not to spoil the minimization along $\mathbf{u}$, the gradient has to stay perpendicular to $\mathbf{u}$, that is

$$0 = \mathbf{u} \cdot \delta(\nabla f) = \mathbf{u} \cdot \mathbf{A} \cdot \mathbf{v} \tag{2.19}$$

When this property holds, $\mathbf{u}$ and $\mathbf{v}$ are said to be conjugate. When doing successive line minimizations along a conjugate set of directions, there is no need to reconsider any of the previous ones, and each step is virtually an improvement over all the

preceding ones.

In the CG method [4], $\mathbf{A}$ is used to define an optimal set of directions $\mathbf{h}_i$ along which to line-minimize. Let $\mathbf{g}_0$ be an arbitrary vector and $\mathbf{g}_0 = \mathbf{h}_0$. For $i = 0, 1, 2, \ldots$ let

$$\mathbf{g}_{i+1} = \mathbf{g}_i - \lambda_i \mathbf{A} \cdot \mathbf{h}_i \tag{2.20}$$

$$\mathbf{h}_{i+1} = \mathbf{g}_{i+1} + \gamma_i \mathbf{h}_i \tag{2.21}$$

where λ_i and γ_i are such that $\mathbf{g}_{i+1} \cdot \mathbf{g}_i = 0$ and $\mathbf{h}_{i+1} \cdot \mathbf{A} \cdot \mathbf{h}_i = 0$, that is

$$\lambda_i = \frac{\mathbf{g}_i \cdot \mathbf{g}_i}{\mathbf{g}_i \cdot \mathbf{A} \cdot \mathbf{h}_i} \tag{2.22}$$

$$\gamma_i = \frac{\mathbf{g}_{i+1} \cdot \mathbf{A} \cdot \mathbf{h}_i}{\mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_i} \tag{2.23}$$

Then it can be shown by induction that for all $i \neq j$

$$\mathbf{g}_i \cdot \mathbf{g}_j = 0 \quad \text{and} \quad \mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_j = 0 \tag{2.24}$$

The important fact is that the hessian matrix is not actually necessary. Let $\mathbf{g}_i = -\nabla f(\mathbf{x}_i)$ at some point $\mathbf{x}_i$; suppose to move from $\mathbf{x}_i$ along $\mathbf{h}_i$ to a new point $\mathbf{x}_{i+1}$, and set $\mathbf{g}_{i+1} = -\nabla f(\mathbf{x}_{i+1})$; then $\mathbf{g}_{i+1}$ can be shown to be the same vector which would result from Eq. (2.20), but has been constructed without explicitly using $\mathbf{A}$.

Going back to our physical problem, we also have to take into account the orthonormalization constraint involving the wavefunctions. Since using Lagrange multipliers, as in the previous section, would imply performing line minimizations on curved trajectories, in order to keep things simple we give up the ideas exposed in the previous section, recasting the minimum problem in terms of non-orthonormal wavefunctions. For the purpose, we introduce a new energy functional which is the composition of the initial one with a transformation which maps a set of non-orthonormal orbitals $\{\phi_i\}$ onto an orthonormal one, $\{\psi_i\}$ [2, 7]. This scheme is feasible when both sets span the same subspace. The latter function can be ex-

pressed in terms of the overlap matrix $S_{ij} = \langle\phi_j|\phi_i\rangle$:

$$\psi_i = \sum_j S_{ij}^{-1/2}\phi_j \tag{2.25}$$

The energy functional and the gradient are then

$$E[\{\phi\}] = \sum_{ij}\langle\phi_i|H|\phi_j\rangle S_{ij}^{-1} \tag{2.26}$$

$$\frac{\delta E}{\delta\phi_i^*} = \sum_j H\phi_j S_{ij}^{-1} \tag{2.27}$$

In order for the overlap matrix to remain non-singular, the $\{\psi_i\}$ set is orthonormalized at every CG step; this is demanded by numerical convenience only, since the value of E is invariant under such kind of transformation.

3

Numerical Tests

A simple test for the conjugate gradient method was to find the ground state energy and the first excitation energies of a non self-consistent hamiltonian. In particular, we considered solid GaAs in the zincblend structure, by means of the empirical pseudopotentials by Cohen and Bergstresser [8]; our results reproduced those obtained with full matrix diagonalization.

A second important test is to check how the CPU time needed for convergence scales with the size M of the hamiltonian.

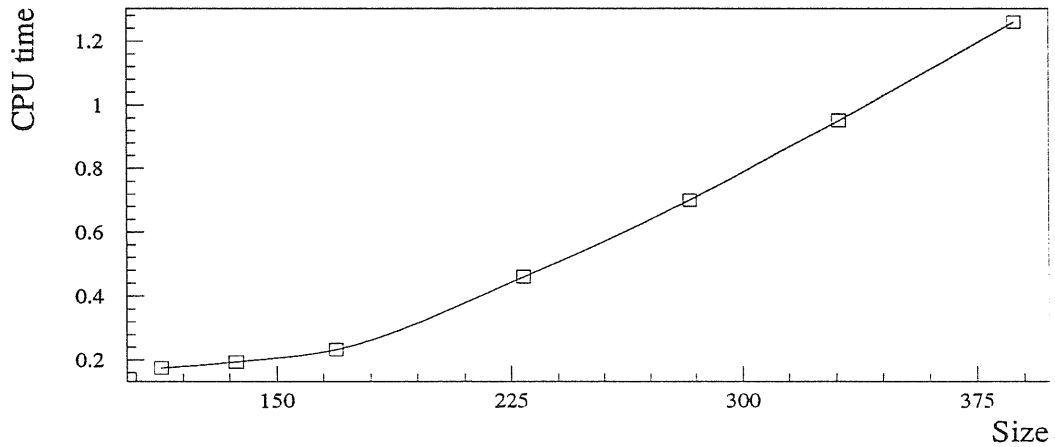

Figure 3.1: CPU time required for convergence as a function of the size of the hamiltonian

The computation of $\hat{E}(\lambda)$ (see previous section) requires the evaluation of the matrices $\langle\phi|H|\phi\rangle$, $\langle\phi|H|h\rangle$, $\langle o|h\rangle$ and $\langle h|h\rangle$. When using local pseudopotentials,

$\mathcal{O}(NM' \log_2 M')$ [1] operations are needed for computing $H\phi$ and Hh, $\mathcal{O}(N^2M)$ for each of the above matrices, $\mathcal{O}(N^3)$ for the inversion of S and $\mathcal{O}(N^2M)$ to orthonormalize the orbitals at each CG step. On the whole, the number of floating-point operations required is asymptotically $\alpha NM' \log_2 M' + \beta N^2 M$; thus the big dimension of the system enters only linearly in the count of the operations.

Fig. 3.1 summarizes several calculations performed at different cutoffs; it clearly shows that the asymptotic behaviour is substantially linear, confirming our expectations.

[1] $M' = xM$ is the number of points used in FFT operations

4

Conclusions

The main characteristics of the CG method as applied to total energy minimization can be summarized as follows:

- The cost of a CG step has the correct scaling behaviour with the size M of the hamiltonian, i. e., is linear with M;
- The rate of convergence is considerably better than SD when low-symmetry systems are considered;
- The cost of a (partial) matrix diagonalization is comparable in CG and Davidson methods; but in SCF calculations the former achieves simultaneously diagonalization and self-consistency, whereas in the latter the diagonalization has to be repeated I times to obtain self-consistency;
- There are no convergence-controlling parameters. such as the time step in SD or the potential mixing parameter in the traditional method.

Bibliography

[1] E. R. Davidson, in *Methods in Computational Molecular Physics*, edited by G. H. F. Diercksen and S. Wilson Vol. 113 of *NATO Advanced Study Institute, Series C* (Plenum, New York, 1983)

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

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

[4] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling. *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press. Cambridge, England, 1989)

[5] R. Car, M. Parrinello, in *Simple Molecular Systems at Very High Density* (Plenum, New York, 1989)

[6] J. P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977)

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

[8] M. L. Cohen. T. K. Bergstresser. Phys. Rev. **141**. 789 (1966)