Computer Physics Communications 185 (2014) 873–883

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

The orbital minimization method for electronic structure calculations with finite-range atomic basis sets

Fabiano Corsetti *

CIC nanoGUNE, 20018 Donostia-San Sebastián, Spain

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form
27 November 2013
Accepted 4 December 2013
Available online 12 December 2013

Keywords:
Finite-range numerical atomic orbitals
Density-functional theory
Unconstrained minimization
Preconditioning
Parallel scaling

ABSTRACT

The implementation of the orbital minimization method (OMM) for solving the self-consistent Kohn–Sham (KS) problem for electronic structure calculations in a basis of non-orthogonal numerical atomic orbitals of finite-range is reported. We explore the possibilities for using the OMM as an exact cubic-scaling solver for the KS problem, and compare its performance with that of explicit diagonalization in realistic systems. We analyze the efficiency of the method depending on the choice of line search algorithm and on two free parameters, the scale of the kinetic energy preconditioning and the eigenspectrum shift. The results of several timing tests are then discussed, showing that the OMM can achieve a noticeable speedup with respect to diagonalization even for minimal basis sets for which the number of occupied eigenstates represents a significant fraction of the total basis size (>15%). We investigate the hard and soft parallel scaling of the method on multiple cores, finding a performance equal to or better than diagonalization depending on the details of the OMM implementation. Finally, we discuss the possibility of making use of the natural sparsity of the operator matrices for this type of basis, leading to a method that scales linearly with basis size.

1. Introduction

Over the last two decades, density-functional theory [1,2] (DFT) has become a ubiquitous tool for studying molecular and condensed matter systems at the atomic level, with applications ranging from the Earth sciences to nanotechnology [3,4]. This is due in no small part to the proliferation of fast, accurate, easily available and user friendly software packages for performing DFT calculations (see, e.g., Refs. [5–12]). Much work has gone into developing the methods used by such codes, and ongoing optimization of the underlying algorithms is essential to keep up with the possibilities offered by new computer architectures [13–16].

Within the standard Kohn–Sham (KS) approach [2], the many-electron problem is reduced to a self-consistent eigenvalue problem with an effective Hamiltonian. When solving the problem in a basis, either repeated explicit diagonalizations (for small enough bases), or one of a number of iterative minimization algorithms [5,17–19] can be used. The latter proceed either by minimizing the KS total energy functional directly, or as an alternative to diagonalization for a fixed Hamiltonian, within an outer self-consistency cycle for the electronic density. The use of an iterative algorithm of either form is essential for plane-wave methods, as the large number of basis functions/atom makes diagonalization prohibitively expensive for all but the smallest systems. One of the main advantages of using such algorithms is that only the occupied subspace needs to be computed, and for a plane-wave basis this typically corresponds to a very small fraction (<1%) of the total number of eigenstates. Methods based on localized atomic-like orbitals [6,7,11], on the other hand, employ a much smaller number of basis functions/atom; in such cases, therefore, diagonalization is feasible even for large systems. Furthermore, the fraction of occupied eigenstates is much larger than for plane waves (10%–20%), making diagonalization not just competitive but indeed more efficient than most iterative algorithms.

In this paper, we present our implementation of an iterative minimization algorithm, which we refer to as the orbital minimization method (OMM) following Refs. [15,20], as an alternative to explicit diagonalization in the SIESTA [7] code. The OMM works by finding the $N/2$ Wannier functions (WFs) describing the occupied subspace of an N-electron system by direct *unconstrained* minimization of an appropriately-constructed energy functional. This functional was originally proposed independently by Mauri, Galli and Car [21,22], and Ordejón et al. [23,24], in the context of linear-scaling DFT methods [15] with spatially confined WFs. In fact, a serial implementation of the linear-scaling OMM was an integral part of the original version of SIESTA; nowadays, however, most applications of the code employ diagonalization with LAPACK [25]/ScaLAPACK [26]. It is important to note that our new implementation is completely separate from this old one, and is *not* a linear-scaling solver; our aim, instead, is to explore the

* Tel.: +34 943574063.
E-mail addresses: f.corsetti@nanogune.eu, fabiano.corsetti@gmail.com.

0010-4655/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2013.12.008

potential of the OMM as a conventional cubic-scaling iterative algorithm, which solves the KS problem exactly without introducing spatial truncations. Although previous investigations of cubic-scaling iterative algorithms in SIESTA (both Jacobi–Davidson [27] and Lanczos-like [28]) have found that a very small fraction of occupied eigenstates is needed to rival the efficiency of diagonalization [29], we show that this is indeed possible with the OMM even for a standard double-ζ basis with a single polarization shell (dζ + p), for which the fraction of occupied eigenstates is significant, almost 20%.

The use of the OMM as a cubic-scaling DFT solver has previously been described for a plane-wave basis by Pfrommer, Demmel and Simon [19] (alongside closely related methods), and implemented in the PARATEC [30] plane-wave code. In contrast, SIESTA makes use of a minimal basis of numerical atomic orbitals (NAOs) of finite range, leading to formally sparse Hamiltonian and overlap matrices. Our implementation is therefore somewhat different, in particular in the choice of preconditioner. Furthermore, the natural sparsity of the operators in our case can be used to eliminate the most expensive type of matrix–matrix multiplication present in the algorithm, leading to qualitatively different scaling behavior with basis size.

The rest of the paper is organized as follows: in Section 2, we describe in detail the implementation of the OMM and investigate its convergence properties. We give a theoretical overview of the method (Sections 2.1 and 2.2), describe how the number of matrix operations can be minimized by careful consideration of the line search algorithm (Section 2.3), discuss the issue of preconditioning (Section 2.4) and empirically assess its efficiency (Section 2.5), detail the use of sparse–dense matrix operations (Section 2.6), and comment on the issue of fractional occupancies (Section 2.7). In Section 3, we present scaling tests performed in serial and in parallel, and then investigate the hard and soft parallel scaling of the method (Section 3.1), and the scaling with basis size (Section 3.2). Finally, in Section 4, we give a summary of our main conclusions.

2. Formalism and algorithmic considerations

2.1. OMM overview

We work in a basis of m finite-range NAOs $\{\phi_\mu(\mathbf{r})\}$, typically (but not necessarily) atom-centered, and want to solve the generalized eigenvalue problem

$$\mathbf{H}\mathbf{c}_\mu = \varepsilon_\mu \mathbf{S}\mathbf{c}_\mu, \tag{1}$$

where

$$H_{\mu\nu} = \left\langle \phi_\mu \right| \hat{H}^{\mathbf{KS}}[\rho] \left| \phi_\nu \right\rangle, \tag{2}$$

and

$$S_{\mu\nu} = \left\langle \phi_\mu | \phi_\nu \right\rangle. \tag{3}$$

Within a single inner self-consistency (SCF) cycle, $\mathbf{H}$ depends on the fixed electronic density $\rho(\mathbf{r})$, and within an outer molecular dynamics (MD) step, both $\mathbf{H}$ and $\mathbf{S}$ depend on the atomic positions $\{\mathbf{R}_I\}$.

Explicit diagonalization computes the KS eigenenergies $\{\varepsilon_\mu\}$ and the matrix of KS eigenvectors $\mathbf{c}_\mu$, from which the full density matrix can be obtained. Instead, in the OMM we define a set of $n = N/2$ non-orthogonal WFs $\{\chi_i(\mathbf{r})\}$:

$$|\chi_i\rangle = \sum_{\mu=1}^{m} C_i^\mu \left|\phi_\mu\right\rangle. \tag{4}$$

The reduced subspace operators defined by the WFs are then, in matrix form:

$$\mathbf{H}_\mathrm{W} = \mathbf{C}^\mathrm{H}\mathbf{H}\mathbf{C} \tag{5}$$

and

$$\mathbf{S}_\mathrm{W} = \mathbf{C}^\mathrm{H}\mathbf{S}\mathbf{C}. \tag{6}$$

We can also define a subspace energy E, which, when minimized with respect to the coefficients $\{C_i^\mu\}$, will (by the variational principle) give the sum of the lowest n eigenvalues of the original problem [19]:

$$E[\mathbf{C}] = 2\mathrm{Tr}\left\{\mathbf{S}_\mathrm{W}^{-1}\mathbf{H}_\mathrm{W}\right\} \tag{7}$$

(we include a factor of two for spin degeneracy).

The minimization of the functional given in Eq. (7) has the advantage of being unconstrained, since no orthonormalization is required. Nevertheless, performing such a minimization is computationally demanding due to the presence of the inverse overlap $\mathbf{S}_\mathrm{W}^{-1}$, which needs to be recomputed at every trial step; furthermore, this means that the line search in a steepest descent (SD) or conjugate gradient (CG) algorithm has to be solved numerically.

The OMM substitutes Eq. (7) with a different functional, one that does not contain the inverse operation:

$$\tilde{E}[\mathbf{C}] = 2\mathrm{Tr}\left\{\left[\mathbf{I}_n + (\mathbf{I}_n - \mathbf{S}_\mathrm{W})\right]\mathbf{H}_\mathrm{W}\right\} = 4\mathrm{Tr}\left\{\mathbf{H}_\mathrm{W}\right\} - 2\mathrm{Tr}\left\{\mathbf{S}_\mathrm{W}\mathbf{H}_\mathrm{W}\right\}. \tag{8}$$

It can be shown that this new functional drives the WFs towards orthonormality as it is minimized [21–24]; at the minimum, therefore, $\mathbf{S}_\mathrm{W} = \mathbf{I}_n$, and $\tilde{E}[\mathbf{C}_0] = E[\mathbf{C}_0] = E_0$, where $\mathbf{C}_0$ describes the occupied subspace, and E_0 is the ground-state KS (band) energy. Eq. (8) can be derived either by replacing the inverse overlap matrix in Eq. (7) with a first-order Taylor expansion [21], or by using a Lagrange multiplier approach to enforce the desired orthonormality requirement on Tr $\{\mathbf{H}_\mathrm{W}\}$ at the solution [23].

The OMM functional, therefore, allows for unconstrained minimization without requiring any matrix inversion; this is particularly suitable for developing linear-scaling methods, since the inverse of a formally sparse $\mathbf{S}_\mathrm{W}$ matrix (obtained by constraining the radii of the WFs) will not itself be sparse. However, the OMM in its original form was quickly abandoned by the linear-scaling community, as the localization of the WFs introduces many local minima that were found to lead to serious difficulties in obtaining the true ground state [22,24,31]. Ultimately, this led to the development and implementation of various generalized OMMs to overcome the convergence problem [20,31–33]. For non-linear-scaling implementations of the OMM, however, this problem does not present itself, as the WFs are not spatially constrained; the simplicity of the original functional is therefore ideal in our case for developing an efficient algorithm.

2.2. Eigenspectrum shift

As shown by Ordejón et al. [24] and Kim, Mauri and Galli [31], a stationary point of the OMM functional is obtained when all the WFs are either eigenvectors of Eq. (1) or zero, subject to any arbitrary unitary transformation between them. If all the corresponding eigenvalues are negative, this point will be a minimum; however, if any of them are positive, it will become a saddle point. Therefore, the stationary point at $\mathbf{C}_0$ will be a minimum provided that all the occupied eigenvalues are negative.[1]

Because of this, it is necessary to shift the eigenspectrum by $\eta > \varepsilon_n$ for the minimization procedure to be able to find the correct ground state. This is achieved by the transformation

$$\mathbf{H} \to \mathbf{H} - \eta\mathbf{S}, \tag{9}$$

[1] Furthermore, it can be shown that it will only be a global minimum if the entire eigenspectrum of $\hat{H}$ is negative; otherwise, it will be a local minimum, and the functional will have no lower bound.