Large Scale Electronic Structure Calculations

Giulia Galli[1],[2] and Michele Parrinello[2]

[1] *Institut Romand de Recherche Numérique en Physique des Matériaux (IRRMA), PHB Ecublens, 1015 Lausanne, Switzerland*[a]

[2] *IBM Research Division, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland*

We formulate the Kohn-Sham density functional theory in terms of nonorthogonal, localized orbitals. Within this formulation we introduce a simple and effective method to localize the orbitals. Our approach leads to a plane-wave-based algorithm for total energy calculations whose computational complexity is of $O(N)$, where N is the number of electrons. This opens the way to calculations of unprecedented scale. Our method appears to be of general character and applicable in other contexts such as quantum chemical or projected quantum Monte Carlo calculations.

PACS numbers: 71.10.+x, 71.25.Cx.

Recent years have seen considerable progress in the scale and scope of electronic structure computations based on density functional (DF) theory [1]. Total energy calculations and *ab initio* molecular dynamics (MD) simulations [2] involving a few hundreds of atoms are now performed in many laboratories.

Much of this progress is due to the use of pseudopotentials and plane-wave (PW) basis sets, in conjunction with iterative techniques for the solution of the Schrödinger equation [3]. Very recently the use of parallel algorithms has added further power to this approach [4, 5]. The choice of PW expansions for the single particle orbitals has several advantages. It permits the use of fast Fourier transforms (FFT's) which are computationally very efficient; PW's do not depend on atomic positions and therefore forces acting on atoms can be easily computed via a straightforward application of the Hellman-Feynman theorem. PW's are free of basis set superposition errors and allow one to compute the total energy of different atomic arrangements with the same accuracy. This is particularly relevant in MD simulations, where rather different atomic configurations are explored. Finally the convergence of PW calculations can be controlled in a very simple manner.

The disadvantage of PW expansions is the very large number (M) of basis functions needed to represent the electronic orbitals; M is proportional to the number of electronic states N. In large scale applications, M is so large that straightforward diagonalization is not possible and use of iterative techniques to obtain the lowest N eigenvalues is necessary.

Two are the basic steps of iterative diagonalizations: The evaluation of $H\psi_i$, namely, the application of the Hamiltonian H to the N orbitals ψ_i, and the orthogonalization of the ψ_i's. If a split-operator technique is used, and the properties of convolutions [3] and of the localized nature of the pseudopotential operators [6] are taken into account, the computational cost of $H\psi_i$ can be reduced to $O(NM\log(M))$. The dominant term is the orthogonalization step which scales as $O(N^2M)$ and therefore

constitutes the real bottleneck of these calculations [3]. In principle, use of localized basis set might alleviate the orthogonalization step, at the expense however of losing the advantages of PW expansions.

In this Letter we propose a novel formulation of the Kohn-Sham (KS) equations [1] in terms of nonorthogonal localized orbitals, which are expanded in a PW basis set. The orbitals are localized with an efficient and simple method. We demonstrate that a scaling of order $NM\log(M)$ can be obtained with straightforward modifications of existing iterative procedure: No orthogonalization is needed and advantage is taken of the rapid decay of the overlap and Hamiltonian matrix elements as a function of distance, due to the localized character of the orbitals. We then show how the goal of a PW-based method of order N can be achieved by using only the volume where the localized wave functions are different from zero to expand the KS orbitals and to express the total energy and potential.

In conventional approaches to DF–local-density calculations, the density matrix ρ is expanded in terms of single particle orthonormal orbitals ψ_i:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_i \psi_i^*(\mathbf{r}')\psi_i(\mathbf{r}) . \tag{1}$$

This leads to the well known KS equations $H\psi_i = \epsilon_i\psi_i$, where the Hamiltonian $H = -\frac{1}{2}\nabla^2 + V_{\text{SCF}}(\mathbf{r})$, and $V_{\text{SCF}}(\mathbf{r})$ is the sum of the external, Hartree, and local exchange and correlation potentials. Rather than representing $\rho(\mathbf{r})$ as in Eq.(1), we introduce a set of nonorthogonal orbitals ϕ_i and write ρ as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i,j} \phi_i^*(\mathbf{r}')S_{ij}^{-1}\phi_j(\mathbf{r}) , \tag{2}$$

where $S_{ij} = \langle\phi_i|\phi_j\rangle$. This is a legitimate representation of the density matrix since $\rho^2 = \rho$ and $\int \rho(\mathbf{r}, \mathbf{r})d\mathbf{r} = N$. The choice of the overlap matrix S is of course not unique and needs to be specified. It is convenient to introduce, together with the ϕ_i, their conjugate orbitals

$|\overline{\phi_i}\rangle = \sum_j S_{ji}^{-1}|\phi_j\rangle$ which satisfy the relation $\langle\overline{\phi_i}|\phi_j\rangle = \delta_{ij}$. The total energy E of interacting electrons in the field of a fully separable pseudopotential can be written in such a form that all the relevant quantities depend upon the product $\overline{\phi_i}^*\phi_i$. In particular, the kinetic energy E_{kin} is given by

$$E_{\text{kin}} = \sum_i \langle\overline{\phi_i}| - \tfrac{1}{2}\nabla^2|\phi_i\rangle \qquad (3)$$

and

$$\rho(\mathbf{r}) = \sum_i \overline{\phi_i}^*(\mathbf{r})\phi_i(\mathbf{r}). \qquad (4)$$

The Euler equations associated with the minimization of E are obtained from $\delta E/\delta\overline{\phi_i}^* = 0$:

$$-H|\phi_i\rangle + \sum_l \langle\phi_i|H|\phi_l\rangle|\overline{\phi_l}\rangle = 0. \qquad (5)$$

Equation (5) can be recast into the standard KS equations by applying the transformation $|\psi_i\rangle = \sum_j S_{ij}^{-1/2}|\phi_j\rangle$. A practical way of finding the minimum of E is, e.g., to introduce a fictitious dynamics of the type:

$$\dot{\phi_i}(\mathbf{r}) = -\frac{\delta E}{\delta\overline{\phi_i}^*(\mathbf{r})}. \qquad (6)$$

From Eqs. (5) and (6) it is easily seen that $\dot{S} = 0$. This is true in principle, i.e., if Eq. (5) is integrated with infinite precision. However, in practice a finite precision has to be adopted to solve Eq. (5) and therefore numerical instabilities may arise, leading to an ill-conditioned overlap matrix as the iterative procedure is carried on.

This instability can be avoided by requiring that different orbitals ϕ_i be localized in different regions of space. As discussed, e.g., in Ref. [7], nonorthogonal functions can be better localized than Wannier functions. Localization can be obtained in different ways; in order to be computationally efficient it should not imply the evaluation of the entire S matrix. In our calculations, we have kept the orbitals localized in given regions of space by adding to the Hamiltonian H the nonlocal potential $\mathcal{V} = \sum_i V_I^i|\phi_i\rangle\langle\overline{\phi_i}|$, where

$$V_I^i = A[1 - \theta(|\mathbf{R}_I - \mathbf{r}| - \sigma)]. \qquad (7)$$

A is the strength of the potential, $\mathbf{R}_I$ the center of the localization region (LR) of radius σ associated to the state i, and θ is a step function [8]. We note that $\mathcal{V}|\phi_i\rangle = V_I^i|\phi_i\rangle$, i.e., V_I^i acts only on the state $|\phi_i\rangle$. The presence of $\mathcal{V}$ brings in Eq. (5) additional terms, which are, however, easily implemented and evaluated. The role of V_I^i is to break the symmetry among the infinite number of S matrices which give rise to the same ρ, and to *choose* the one built from maximally localized ϕ_i. An alternative way of localizing orbitals (which we call a filtering procedure) is that of multiplying the ϕ_i by $\theta(|\mathbf{R}_I - \mathbf{r}| - \sigma)$ at each step, after Eq. (5) has been numerically integrated.

Several wave functions can be associated to the same LR, but in this case orbitals belonging to the same region have to be kept mutually orthogonal. Clearly if A is too large or σ too small the introduction of V_I^i does not amount to a simple unitary transformation and leads to unphysical results. However the V_I^i can be engineered so that they lead to total energies which are practically the same as those obtained by conventional solutions of the KS equations.

We have checked this to be the case for a system of 64 Si atoms in the diamond structure [9]. In our test calculations we find a total energy which is the same as that obtained in conventional computations within 10^{-5} eV/atom. The same accuracy is achieved when a filtering procedure is used to impose localization constraints. Two different sets of $\mathbf{R}_I$ have been tried. In the first one the $\mathbf{R}_I$ coincide with the atomic positions, and there are two occupied states per LR. In the second one the $\mathbf{R}_I$ form a uniformly spaced grid of points arranged in a bcc lattice, with each LR being singly occupied. The first choice is physically more appealing, but introduces a dependence of ϕ_i on the atomic positions which would lead to Pulay-like forces in MD simulations. Such forces are absent in the second case [10].

In the formulation presented so far, there are some advantages such as the removal of the orthogonalization step, which allows the use of unconstrained minimization methods for the calculation of $\delta E/\delta\phi_i = 0$, and the implementation of parallel codes, since the system is naturally divided into subunits. However, no gain in the scaling of the CPU time with N has yet been obtained.

This can be achieved by taking advantage of the localized nature of ϕ_i. Indeed the S_{ij} and H_{ij} matrix elements decay rapidly to zero as a function of $(\mathbf{R}_I - \mathbf{R}_J)$, and need to be computed only for $|\mathbf{R}_I - \mathbf{R}_J| \leq n\sigma$ ($n \simeq 4$-5). In such a manner their evaluation scales as NM. $|\overline{\phi_i}\rangle$ can be obtained in NM operations as a solution of the sparse linear system

$$\sum_j S_{ij}|\overline{\phi_j}\rangle = |\phi_i\rangle \qquad (8)$$

which is solved either with conjugate gradient techniques or iteratively as $(\overline{\phi_i})^n = (1 - S)(\overline{\phi_i})^{n-1} + \phi_i$ [11]. Similar considerations may be adopted for the calculation of the nonlocal pseudopotential energy E_{NL}. Alternatively E_{NL} can be evaluated as suggested in Ref. [6]. The dominant part of the whole minimization procedure is thus the evaluation of $H|\phi_i\rangle$, which scales as $NM\log(M)$. We note that the localization region of $\overline{\phi_i}$ is larger than that of ϕ_i; however, $\overline{\phi_i}$ needs to be evaluated only at the points where ϕ_i is different from zero, since all the relevant quantities depend upon the product $\overline{\phi_i}\phi_i$.

We have tested numerically the validity of our formulation by performing a calculation on the same 64 Si atom system [9] as before, but this time using for S_{ij} and H_{ij} the cutoff distances $d_{\text{cut}} = 16$ and $d'_{\text{cut}} = 20$ a.u., re-

spectively. The optimized energy differs once more from that obtained in conventional calculations by a negligible amount (10^{-5} eV/atom).

We have also tested the scaling of the algorithm in practice by performing a calculation on a 216 Si atom system [9]. Its performance on vectorial machines is even better than the theoretical one, due to improved performances on longer arrays. The break-even point between ours and a conventional approach is at about 64 atoms, for Si. A further important facet of our scheme, which becomes crucial for very large systems, is that it offers considerable storage savings since ϕ_i need to be stored only at the mesh points where they are significantly different from zero.

In the scheme presented so far, we have not fully exploited the localized character of ϕ_i and unnecessarily used the whole volume (Ω) to expand the KS orbitals. We are presently exploring various strategies to take full advantage of the localization of ϕ_i. (Details of our formulations and an account of their relative merits will be presented elsewhere.) One can for instance form superfunctions which are obtained from the union of nonoverlapping localized orbitals. The number of such superfunctions $n_S \sim N/[\Omega/(d'_{cut})^3]$ does not vary as the size of the system is increased. The cost of the FFT part is therefore of $O(n_S M \log M)$. The individual wave functions, when needed, can be obtained in real space and matrix elements evaluated in real space. Since ϕ_i are localized, this is an $O(N)$ operation.

Alternatively, one can define a region of space of volume Ω_s slightly larger then $(d'_{cut})^3$, centered on the LR's, which contains all the points where ϕ_i differs significantly from zero. This region (small mesh) is much smaller than the cell volume and does not scale as a function of the

system size. Most operations can be performed on the small mesh, and by an appropriate mapping ϕ_i can be projected into the full mesh. The full mesh needs to be used only when calculating the total charge density and the potential. We give here the expressions for the calculation of the key quantities in the case where the LR length and the edge of the cubic total volume are L and $2L$, respectively .

The overlap, kinetic energy $E_{kin} = \sum_{ij} S_{ij}^{-1} \sum_{\eta=1}^{3} K_{ij}^{\eta}$, and density $\rho = \sum_{ij} S_{ij}^{-1} \rho_{ij}$ matrix elements are given by

$$S_{ij} = \frac{1}{\Omega_s} \int_{\Omega_s} \phi_i^*(\mathbf{r})\Gamma_{IJ}(\mathbf{r})\phi_j(\mathbf{r})\, d\mathbf{r}, \tag{9}$$

$$K_{ij}^{\eta} = \frac{1}{\Omega_s} \int_{\Omega_s} \frac{\partial \phi_i^*(\mathbf{r})}{\partial r_\eta} \Lambda_{IJ}^{\eta}(\mathbf{r}) \frac{\partial \phi_j(\mathbf{r})}{\partial r_\eta}\, d\mathbf{r}, \tag{10}$$

$$\rho_{ij} = \phi_i^*(\mathbf{r} - \alpha\mathbf{L})\, [\Gamma_{IJ}(\mathbf{r} - \alpha\mathbf{L})\Delta_J(\mathbf{r})] \\ \times \phi_j(\mathbf{r} - \alpha\mathbf{L}). \tag{11}$$

$\alpha = 0$ if $\mathbf{r}$ belongs to Ω_s and 1 otherwise. If we define $\tau_I^{\eta} = R_I^{\eta} - r^{\eta} + L/2$, we have

$$\Gamma_{IJ}(\mathbf{r}) = \prod_{\eta}[1 + \gamma(\tau_I^{\eta})\gamma(\tau_J^{\eta})], \tag{12}$$

$$\Lambda_{IJ}^{\eta}(\mathbf{r}) = \prod_{\xi \neq \eta}[1 + \gamma(\tau_I^{\xi})\gamma(\tau_J^{\xi})], \tag{13}$$

$$\Delta_J(\mathbf{r}) = \prod_{\eta}[1 + (-1)^{\alpha}\gamma((1-\alpha)L - \tau_J^{\eta})], \tag{14}$$

where the products run over the Cartesian components and

$$\gamma(R_I^{\eta} - r^{\eta} + L/2) = \begin{cases} 1 \text{ if } R_I^{\eta} - L/2 - m \le r^{\eta} < R_I^{\eta} + L/2 - m \\ -1 \text{ otherwise}. \end{cases} \tag{15}$$

Here m is an arbitrary integer . We note that Γ and Λ are different from zero only for overlapping sites. The role of Γ_{IJ} is to map onto Ω_s the points of overlapping regions and to appropriately discard those of nonoverlapping sites. Similar *mapping* and *selection* roles are played by the functions Λ and Δ. We also note that the generalized forces of Eq. (5) need to be computed only on the small mesh points.

An analysis of the scaling of the method sketched above shows it has a computational complexity of $O(N)$. In principle, the evaluation of S^{-1} requires $O(N^2)$ operations (S is a banded matrix); however, the number of operations can be reduced to $O(N)$, since for the construction of ρ and E_{kin} only the S^{-1} matrix elements between overlapping orbitals are needed [12]. For problems where the self-consistent calculation of $\rho(\mathbf{r})$ is not needed the same $O(N)$ scheme can also be used. This

is the case, for instance, in the solution of Schrödinger equations in an external potential or in calculations that use the Harris functional [13] or atomic-orbital-based *ab initio* MD [14].

Recently, two other schemes for total energy calculations which scale linearly with N have been proposed in the literature [15, 16]. The approach of Ref. [15] is based on the division of a given system into independent subsystems connected through a common chemical potential, and introduces certain approximations for the evaluation of the charge density. The scheme proposed by Baroni and Giannozzi [16] is instead an elegant implementation of DFT without orbitals. This scheme becomes favorable compared to conventional approaches for N of the order of 1000. Both schemes, however, strongly depart from conventional PW algorithms and their accuracy and ro-

bustness especially for MD calculations is still to be established. This has to be contrasted with the present approach which mantains the advantage of conventional PW calculations.

We believe that the algorithm described here has other virtues besides its favorable scaling. Several different applications can be envisaged. For instance, divide-and-conquer strategies are now possible and different parts of a given system can be treated differently, if needed. In a longer perspective the application of ideas from renormalization group to electronic structure calculations is probably easier in the present framework. DFT calculations are not the only instances in which an automatic generation of localized orbitals is desirable. The application of our localization technique in multireference self-consistent-field quantum chemical calculations can lead to substantial reduction in the number of integrals that need to be evaluated and to faster algorithms. Another field of interesting applications is projected quantum Monte Carlo, which could be stabilized by suitable extensions of our localization approach rather than by more costly orthogonalization procedures.

(a) Present address.

[1] See, e.g., R.M. Dreizler and E.K.U. Gross, *Density Functional Theory* (Springer-Verlag, Berlin, 1990).

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

[3] See, e.g., G. Galli and M. Parrinello, in *Computer Simulation in Material Science,* edited by M. Meyer and V. Pontikis (Kluwer, Dordrecht, 1991).

[4] I. Stich *et al.*, Phys. Rev. Lett. **68**, 1351 (1992).

[5] K.D. Brommer *et al.*, Phys. Rev. Lett. **68**, 1355 (1992).

[6] R.D. King-Smith, M.C. Payne, and J-S. Lin, Phys. Rev. B **44**, 13 063 (1991).

[7] P.W. Anderson, Phys. Rev. Lett. **21**, 13 (1968).

[8] In practice the θ function is smoothed for numerical convenience.

[9] We have used a PW cutoff of 12 Ry, separable nonlocal pseudopotentials, and a spherical LR of radius $\sigma = 3.5$ a.u. A was varied between 10^{-2} and 10^{-3} a.u. during the iterative procedure.

[10] Another possible choice, not yet explored, would be to optimize also the positions of the localization centers, following a method similar to that introduced by M. Sprik and M.L. Klein [J. Chem. Phys. **89**, 1592 (1989)] in the context of floating Gaussian calculations.

[11] See, e.g., *Numerical Recipes* (Cambridge Univ. Press, Cambridge, 1987), pp. 306-307. The number of iterations needed for the solution of Eq. (8) can be made as small as 1 to 3 if $\overline{\phi}_i$ is stored at each step.

[12] The methods for inverting banded matrices are so efficient that for realistic sizes one can afford the evaluation of all S_{ij}^{-1}: This amounts to a tiny fraction of the global CPU time, and the practical scaling of the algorithm is still $O(N)$.

[13] J. Harris, Phys. Rev. B **31**, 1770 (1985).

[14] O.F. Sankey and D.J. Niklewski, Phys. Rev. B **40**, 3979 (1989).

[15] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).

[16] S. Baroni and P. Giannozzi, Europhys. Lett. **17**, 547 (1992).