ELSEVIER

Computational Materials Science 6 (1996) 15–50

COMPUTATIONAL
MATERIALS
SCIENCE

Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set

G. Kresse [a,*], J. Furthmüller [b]

[a] *Institut für Theoretische Physik, Technische Universität Wien, Wiedner Hauptstraße S-10, A-1040 Wien, Austria*

[b] *Institut für Festkörpertheorie und Theoretische Optik, Friedrich-Schiller-Universität Jena, Max-Wien-Platz 1, D-07743 Jena, Germany*

Received 4 February 1996; accepted 15 March 1996

Abstract

We present a detailed description and comparison of algorithms for performing ab-initio quantum-mechanical calculations using pseudopotentials and a plane-wave basis set. We will discuss: (a) partial occupancies within the framework of the linear tetrahedron method and the finite temperature density-functional theory, (b) iterative methods for the diagonalization of the Kohn–Sham Hamiltonian and a discussion of an efficient iterative method based on the ideas of Pulay's residual minimization, which is close to an order N_{atoms}^2 scaling even for relatively large systems, (c) efficient Broyden-like and Pulay-like mixing methods for the charge density including a new special 'preconditioning' optimized for a plane-wave basis set, (d) conjugate gradient methods for minimizing the electronic free energy with respect to all degrees of freedom simultaneously. We have implemented these algorithms within a powerful package called VAMP (Vienna ab-initio molecular-dynamics package). The program and the techniques have been used successfully for a large number of different systems (liquid and amorphous semiconductors, liquid simple and transition metals, metallic and semi-conducting surfaces, phonons in simple metals, transition metals and semiconductors) and turned out to be very reliable.

1. Introduction

1.1. General

In recent years, ab-initio calculations have made a profound impact on the investigation of material properties. The main reason for the enormous success of ab-initio methods lies in the fact that they are parameter-free and require no other input than the atomic number. In addition, improvements in computer performance and algorithms allow to apply these methods to a steadily increasing number of physical and chemical phenomena. Probably, the most successful method currently tractable is the local density functional (LDF) theory proposed by Kohn and Sham [1]. In conjunction with the Hellmann–Feynman theorem [2] forces can be evaluated easily, allowing the *simultaneous* investigation of structural, electronic and dynamic properties. The first successful ab initio calculation in this context goes back to a seminal paper written by Car and Parrinello (CP) [3]. In their work Car and Parrinello proposed a simulated annealing approach in which electrons and ions are treated on the same footing via

[*] Corresponding author. E-mail: kresse@tph20.tuwien.ac.at.

PII S0927-0256(96)00008-0

a Quasi-Newton equation of motion. This approach allows for an efficient simultaneous update of electrons and ions, but also possesses some serious restrictions: The time step for the CP technique is limited by the requirement that the electrons are always close to the exact electronic groundstate. Indeed, it can be shown that this is only the case if the typical excitation frequencies of the electronic subsystem are much higher than that of the ionic system [4] (in this case electrons and ions decouple adiabatically, and the electrons oscillate around the real electronic groundstate). This also implies that the time step in a CP simulation is determined by the electronic degrees of freedom, and usually the time step is an order of magnitude smaller than that necessary to simulate the ionic subsystem.

A straightforward alternative to the simultaneous update of electrons and ions is the exact calculation of the electronic groundstate after each ionic move. This is possible if the algorithms for calculating the electronic groundstate are sufficiently efficient. Recently several approaches have been proposed and most of these methods differ significantly from the standard original CP implementation, except for one aspect: For a plane-wave basis set, CP introduced an efficient way to calculate the action of the Hamiltonian onto the electronic wavefunctions. They used the fact that the local potential part of the Hamiltonian is diagonal in real space and that the kinetic energy part of the Hamiltonian is diagonal in reciprocal space. Therefore, the evaluation of the action of the Hamiltonian is very fast if the wavefunctions are transformed from reciprocal to real space and backwards using fast Fourier transformations. In addition, it is easy to evaluate the nonlocal part of the Hamiltonian using separable factorized pseudopotentials [5]. These features make all 'iterative' algorithms for calculating the electronic groundstate tremendously more efficient than the previously used schemes based on an exact diagonalization of the Kohn-Sham (KS) Hamiltonian. Here, the term 'iterative' refers to any technique requiring the repeated evaluation of the action of the Hamiltonian onto the wavefunctions as a key step. In general two different techniques can be distinguished:

(i) Methods for determining the minimum of the KS energy functional directly (in the future simply called direct methods); and

(ii) iterative methods for the diagonalization of the KS-Hamiltonian in conjunction with an iterative improvement (i.e. mixing) of the charge density (we will refer to these methods as selfconsistency cycle (SC) methods).

The direct methods (i) have been pioneered by CP. They are based on the fact that the Kohn-Sham energy functional is minimal at the electronic groundstate. Therefore, minimization of the functional with respect to the variational degrees of freedom leads to a convenient scheme for calculating the electronic groundstate. The only problem to be solved is the inclusion of the orthonormality constraint on the wavefunctions, which is done with a Lagrange formalism in the original method of CP. Generally the standard CP algorithm is relatively slow if it is applied to the electrons only. Small improvements might be obtained by integrating the equations of motions analytically [6], or by introducing an improved pre-conditioning for the gradient [7]. In addition it is possible to replace the second order CP equations by a first-order steepest descent [8,9] equation. Nevertheless, recently Tassone, Mauri and Car [7] showed that a preconditioned damped second order equation of motion for the electrons is generally more efficient than this first-order steepest descent equation.

Even more promising than CP like techniques are conjugate gradient (CG) schemes. Within these schemes it is necessary to minimize the KS functional along a given search direction exactly (which is usually not done within the CP like techniques), and in successive steps the new search direction is conjugated to previous directions. The main problem within the CG methods is that the orthonormality constraint is not easy to incorporate. For semiconductors and insulators Teter, Payne and Allan proposed a reliable algorithm which optimizes the electronic energy in a band-by-band fashion [10]. In their algorithm the total energy is minimized for a single orbital within the sub-space orthonormal to the current set of trial wavefunctions. Despite the advantage of small storage requirements, the algorithm is relatively slow because only a limited number of CG steps per orbital can be done, and because the charge density and the potential must be recalculated after each single update of each orbital. Therefore, algorithms which update all orbitals simultaneously should be superior. These algorithms were pioneered independently by Stich, Car, Parrinello and Baroni [11] and by Gillan [12]. The most systematic and elegant way to incorporate the orthogonality constraint in this case is to generalize the KS functional for nonorthogonal orbitals [13].

The results of this paper indicate that the direct methods (i) discussed up to now are in general not as efficient as the traditional SC-methods (ii) which are based on the repeated diagonalization of the KS-Hamiltonian and a charge density mixing. This is especially true for metallic systems. At first sight this is a clear contradiction to the mathematical crudeness of the SC-methods, considering that the selfconsistent minimization of the KS functional is replaced by an independent improvement of the eigenfunctions and the charge density. But the reason for this behavior might lie in the following points: First, iterative methods for the diagonalization of the KS-Hamiltonian are easier to implement and more mature than methods for minimizing the total energy selfconsistently. Second, and more important: methods for an iterative improvement (i.e. mixing) of the charge density can retain information from all previous mixing steps. This is an important difference to all direct schemes which take into account only information from one or two previous steps. In principle CG methods should overcome this shortcoming by creating a set of conjugated directions, but the speed of a CG method is always limited by the accuracy of the minimization into the search direction, which becomes especially cumbersome for metallic systems, and slows down the net convergence.

We have applied the SC-technique successfully to several different systems including liquid simple metals (Na, Ge) [14], liquid transition metals (V, Cu) [15,16], the transition from a liquid metal to an amorphous semiconductor by the rapid quenching of Ge [17,16], and the metal non-metal transition in l-Hg [18]. One advantage of an efficient electronic minimization is that the relaxation of the ions to their instantaneous equilibrium position is much faster. Successful calculations for clean and hydrogenated C(100) surfaces [19], the site-selective adsorption of C-atoms on Al(111) surfaces [20] and Rh surface properties [21] demonstrate the feasibility of our method in this respect. Finally, we have also performed calculations of bulk-phonons in insulators and metals (cubic diamond and graphite see Ref. [22]), indicating that forces can be evaluated efficiently and accurately within the SC methods.

1.2. Outline of the paper

In this paper we will mainly concentrate on methods based on the repeated diagonalization of the KS-Hamiltonian and a charge density mixing (SC-methods). After a general introduction of the Kohn–Sham energy functional (Section 2.1) the impact of partial occupancies on the Kohn–Sham functional will be explained (Section 2.2). The discussion includes newest improvements of the tetrahedron method as well as a comparison of the tetrahedron method with finite temperature methods. Section 2.3 contains a broad outline of the steps involved in methods relying on the selfconsistency cycle, followed by a brief explanation of the Hellmann–Feynman force theorem (Section 2.4). Some important technical details for the calculation of forces are pointed out.

An in-depth discussion and comparison of several iterative matrix diagonalization schemes is given in Section 3. The ideas discussed in Section 3 have partly been published in different papers by different authors – but to our knowledge this is the first consistent review. In addition technical aspects which are important for an actual implementation will be explained. We also review an efficient iterative matrix diagonalization scheme based on the ideas of residual vector minimization (direct inversion of iterative subspace). This scheme will outperform any other iterative matrix diagonalization scheme for very large matrices.

The second main ingredient of our scheme is the charge density mixing discussed in Section 4. We will concentrate on Broyden mixing [23] (especially Broyden's second method or inverse Jacobian update) and a mixing method proposed by Pulay [24]. A close relationship between both methods will be pointed out, and a special metric optimized for a plane-wave basis set will be introduced.

Finally, we have also included a section, which discusses methods to determine the minimum of the KS energy functional directly (Section 5). Special attention will be given to the conjugate gradient method.

In Section 6 a comparison between different methods is given. We have divided this section into a part which compares the non-selfconsistent case (i.e. iterative matrix diagonalization only, Section 6.1) and a part which concentrates on mixing and on the direct methods (Section 6.2).

2. The Kohn–Sham energy functional for partial occupancies

2.1. The Kohn–Sham energy functional

In general the Kohn–Sham energy functional for an ultrasoft (US) Vanderbilt pseudopotential (PP) can be written as [25–27]

$$\begin{aligned} E_{\mathrm{KS}}[\{\phi\},\{\boldsymbol{R}\}] &= \sum_n f_n \langle \phi_n | T + V_{\mathrm{NL}}^{\mathrm{ion}} | \phi_n \rangle + E^H[\rho] + E^{\mathrm{xc}}[\rho] \\ &\quad + \int \mathrm{d}^3 r V_{\mathrm{loc}}^{\mathrm{ion}}(\boldsymbol{r}) \rho(\boldsymbol{r}) + \gamma_{\mathrm{Ewald}}(\{\boldsymbol{R}\}), \end{aligned} \tag{1}$$

with $f_n = 1$ for occupied bands and $f_n = 0$ for unoccupied bands. To simplify the notation we have dropped the k-point index. The first sum runs over all bands N_{b} included in the calculation. The KS functional depends on the positions $\boldsymbol{R}_N$ of the ions and the electronic wavefunctions ϕ_n only. E^{H} is the Hartree-energy, E^{xc} the exchange correlation energy functional, $V_{\mathrm{loc}}^{\mathrm{ion}}$ the local ionic pseudopotential, $T = -\hbar^2/2m_{\mathrm{e}}\nabla^2$ the kinetic energy operator and γ_{Ewald} the Madelung energy of the ions. For US PP the nonlocal part of the PP can be written as

$$V_{\mathrm{NL}}^{\mathrm{ion}} = \sum_{ij} D_{ij}^{\mathrm{ion}} | \beta_j \rangle \langle \beta_i | \tag{2}$$

and the charge density $\rho(\boldsymbol{r})$ is given by

$$\begin{aligned} \rho(\boldsymbol{r}) &= \sum_n f_n |\phi_n(\boldsymbol{r})|^2 \\ &\quad + \sum_{n,ij} f_n \langle \phi_n | \beta_j \rangle \langle \beta_i | \phi_n \rangle Q_{ij}(\boldsymbol{r}), \end{aligned} \tag{3}$$

where β_i are localized projection states, and $Q_{ij}(\boldsymbol{r})$ localized augmentation functions. The total energy has to be minimized subject to the constraint of orthonormalization

$$\langle \phi_n | S | \phi_m \rangle = \delta_{nm} \tag{4}$$

where S is defined as

$$S = 1 + \sum_{ij} q_{ij} | \beta_j \rangle \langle \beta_i |, \tag{5}$$

with

$$q_{ij} = \int Q_{ij}(\boldsymbol{r}) \, \mathrm{d}^3 \boldsymbol{r}. \tag{6}$$

Ultrasoft pseudopotentials are discussed in detail in Ref. [25–28]. Their general advantage is that they reduce the necessary energy cutoff for transition metals and first row elements by a factor of 2–4. The resulting basis sets are comparable in size with the basis sets for typical ‘pseudopotential’ elements like Na, Al, Si and Ge.

The most important property of the KS functional is to be extremal in the ground-state with respect to arbitrary variations of the wavefunctions. Under the constraint of orthonormalization, variation with respect to the wavefunctions leads to the well known KS eigenvalue equations

$$\boldsymbol{H} | \phi_n \rangle = \epsilon_n S | \phi_n \rangle, \tag{7}$$

where $\boldsymbol{H}$ is the Kohn–Sham Hamiltonian

$$\boldsymbol{H} = T + V_{\mathrm{loc}}^{\mathrm{sc}} + V_{\mathrm{NL}}^{\mathrm{sc}}, \tag{8}$$

with

$$V_{\mathrm{loc}}^{\mathrm{sc}} = V_{\mathrm{loc}}^{\mathrm{ion}} + V^{\mathrm{H}}[\rho] + V^{\mathrm{xc}}[\rho], \tag{9}$$

where $V^{\mathrm{H}}[\rho]$ is the Hartree potential and $V^{\mathrm{xc}}[\rho]$ the exchange-correlation potential. For ultrasoft pseudopotentials the nonlocal part of the pseudopotential $V_{\mathrm{NL}}^{\mathrm{sc}}$ depends also on the total local potential and must be calculated accordingly via (compare Eq. (2))

$$D_{ij}^{\mathrm{sc}} = D_{ij}^{\mathrm{ion}} + \int Q_{ij}(\boldsymbol{r}) V_{\mathrm{loc}}^{\mathrm{sc}} \, \mathrm{d}^3 \boldsymbol{r}. \tag{10}$$

From inspection it is clear that only occupied orbitals contribute to the total energy, and it can be shown that the total energy is invariant under an unitary transformation of the wavefunctions ϕ_n if only occupied bands are taken into account (compare with Section 5). In this case it is sufficient to calculate a set of wavefunctions which fulfill the less stringent equation

$$\boldsymbol{H} | \phi_n \rangle = \sum_m \gamma_{nm} S | \phi_m \rangle, \tag{11}$$

where γ_{nm} is an Hermitian matrix. If partial occupancies are included, i.e. if the f_n are treated as additional variational degrees of freedom, it is necessary to calculate the KS orbitals exactly (Eq. (7)), making the calculation of the electronic groundstate more complex for metallic systems, where partial occupancies should be used.

2.2. Metallic systems and partial occupancies

At this point it is necessary to review the impact of partial occupancies on the local density functional (LDF). There are two different approaches to the introduction of partial occupancies to the Kohn–Sham functional:

First, Mermin [29] extended the LDF to finite temperatures. This approach becomes physically significant if the temperature of the system is comparable to characteristic excitation energies. In this case it is also important to use the finite-temperature exchange correlation functional $E_T^{xc}[\rho]$ [30]. Considering the limited accuracy of the LDA it seems to be reasonable to replace $E_T^{xc}[\rho]$ by $E_{T=0}^{xc}[\rho]$. For the finite temperature LDF, the impact of partial occupancies on the forces has probably been first discussed independently by Weinert and Davenport [31] and by Wentzcovitch, Martins and Allen [32].

The second approach concentrates on the evaluation of the energy at zero temperature: In this case partial occupancies are introduced as a tool for reducing the number of k-points in the Brillouin zone which are necessary to evaluate the band structure energy. At zero temperature, the band-structure energy is defined as

$$\sum_n \frac{1}{\Omega_{\mathrm{BZ}}} \int_{\Omega_{\mathrm{BZ}}} \epsilon_{nk} \Theta(\epsilon_{nk} - \mu) \, \mathrm{d}^3 \boldsymbol{k}, \qquad (12)$$

where $\Theta(x)$ is the Dirac step function, and μ the Fermi-energy. This integral has to be evaluated using a discrete set of k-points

$$\frac{1}{\Omega_{\mathrm{BZ}}} \int_{\Omega_{\mathrm{BZ}}} \mathrm{d}^3 \boldsymbol{k} \rightarrow \sum_m w_k. \qquad (13)$$

For completely filled bands (i.e. semiconductors and insulators) no discontinuity exists, and the integral can be calculated accurately using a set of Monkhorst Pack special k-points (see Ref. [34]). But for metals the sum converges exceedingly slowly with the number of k-points included, because the occupancies jump discontinuously from 1 to 0 at the Fermi level. The convergence with respect to the number of k points can be improved by replacing the step function $\Theta(\epsilon_{nk} - \mu)$ by a smoother function $f(\{\epsilon_{nk}\})$. Several functional forms have been proposed for $f(\{\epsilon_{nk}\})$, among these the linear tetrahedron method is probably the most unambiguous approach.

2.2.1. Linear tetrahedron method

Within the linear tetrahedron (LT) method, the one-electron energies ϵ_{nk} are interpolated linearly between the k-points, and the integral for the band-structure energy is performed analytically within each tetrahedron [35]. Blöchl [36] has recently revised the tetrahedron method to give effective weights $f(\{\epsilon_{nk}\})$ for each band and k-point. This new formulation gives strictly the same results as the conventional tetrahedron method but is easier to implement in most existing codes. In a second step, Blöchl was able to derive a correction formula which removes the quadratic error inherent in the LT method by going beyond the linear approximation and by including the effects of the curvature of the bands at the Fermi surface (we will refer to this method as LT-C, whereas LT is the standard linear tetrahedron method). The LT-C method converges very fast with the number of k-points, and we consider this method to be the most accurate and most unambiguous method for calculating the total energy of bulk materials containing a small number of atoms.

Nevertheless the method is not applicable to large supercells, because usually only a very small number of k-points is used in this case. In addition we have shown in Appendix B that the LT-C method makes the calculation of forces at least inconvenient (see also Ref. [36]), whereas the calculation of forces is straightforward for the conventional LT method: Blöchl points out that the total energy is variational with respect to the partial occupancies ('the traditional tetrahedron method is variational with respect to a change in the Fermi surface'), and therefore it is not necessary to recalculate the occupancies to get first order energy changes or forces. This behavior is clear, considering the basic foundations of the LT method. Within the conventional tetrahedron method the energy is linearly interpolated between a set of k-points resulting in a band structure $\epsilon_n(\boldsymbol{k})$ and the occupancies are set according to the step function $f_n(\boldsymbol{k}) = \Theta(\epsilon_n(\boldsymbol{k}) - \mu)$. Because these occupancies minimize the zero temperature KS functional, it is possible to evaluate first order energy changes without recalculating the 'Fermi surface', i.e. keeping the occupancies fixed. This property is still valid for the revised linear tetrahedron method with effective occupancies $f(\{\epsilon_{nk}\})$.

But if the additional correction formula given by Blöchl (LT-C) is used the variational property with respect to the occupancies is destroyed [36] (see Appendix B), and additional terms have to be included for an exact evaluation of the forces. These additional terms arise from the derivatives of the partial occupancies with respect to the ionic positions. For US PP the corresponding terms can not be evaluated easily, making the LT-C method an inconvenient tool if exact forces are required. Therefore it is necessary to resort to different methods like the smearing methods, explained in the next section.

2.2.2. *Finite-temperature approaches – 'smearing methods'*

We have already pointed out that finite-temperature LDF methods were first introduced by Mermin [29]. In the context of ab-initio calculations it is possible to use these methods as a tool for the reduction of the necessary number of k-points to calculate the total energy of a metallic system. In this case, the term 'smearing methods' is probably more appropriate. Within these methods the step function is simply replaced by a smoothly varying function, for example the Fermi–Dirac function

$$f\left(\frac{\epsilon-\mu}{\sigma}\right) = \frac{1}{\exp((\epsilon-\mu)/\sigma)+1} \tag{14}$$

or the integral over a Gaussian

$$f\left(\frac{\epsilon-\mu}{\sigma}\right) = \frac{1}{2}\left(1-\operatorname{erf}\left[\frac{\epsilon-\mu}{\sigma}\right]\right). \tag{15}$$

The Gaussian has been used first by Fu and Ho [33] in the context of plane wave pseudopotential calculations. It turns out that the total energy is no longer minimal with respect to variations of f_n at the electronic groundstate, and to obtain a variational functional it is necessary to replace the total energy E by a generalized free energy F [31,32]

$$F = E - \sum_n \sigma S(f_n) \tag{16}$$

with a correct form for the entropy term $S(f_n)$. For the Fermi–Dirac function S is given by

$$S(f) = -[f\ln f + (1-f)\ln(1-f)]. \tag{17}$$

If the constraint on the number of electrons is taken into account it is easy to show that the variation of Eq. (16) with respect to f_n is zero if the partial occupancies are set according to Eq. (14). For the Gaussian smearing the 'entropy' is defined as [37]

$$S\left(\frac{\epsilon_n-\mu}{\sigma}\right) = \frac{1}{2\sqrt{\pi}}\exp\left(-\left(\frac{\epsilon_n-\mu}{\sigma}\right)^2\right). \tag{18}$$

Formally it is necessary to express S as a function of f (see Eq. (16)), but this is not possible for Gaussian smearing because no analytical inversion of the error function exists. During an actual calculation based on the SC-methods (see Section 2.3) it is sufficient to evaluate the entropy term from Eq. (18), because the eigenvalues ϵ_n are always available.

In conjunction with Fermi–Dirac statistics the free energy might be interpreted as the free energy of the electrons at some finite-temperature $\sigma = k_B T$ [29], but the physical significance of the free energy remains undefined for Gaussian smearing. For a continuous density of states at the Fermi-level it might be shown that the free energy deviates quadratically with σ from $E_{\sigma=0}$ [38]

$$F(\sigma) \approx E_{\sigma=0} + \gamma\sigma^2. \tag{19}$$

Using $S(\sigma) = -\mathrm{d}F(\sigma)/\mathrm{d}\sigma$ it is possible to obtain for the energy E the equation

$$E(\sigma) = F(\sigma) + \sigma S(\sigma) \approx E_{\sigma=0} - \gamma\sigma^2. \tag{20}$$

It is now easy to see that it is possible to obtain an accurate extrapolation for $\sigma \to 0$ from results at finite σ using the formula

$$E_{\sigma=0} \approx \tilde{E}(\sigma) = \tfrac{1}{2}(F(\sigma) + E(\sigma)). \tag{21}$$

This way the leading quadratic error in σ is removed from $F(\sigma)$, and a functional $\tilde{E}(\sigma)$ which deviates only slowly from $E_{\sigma=0}$ might be obtained.

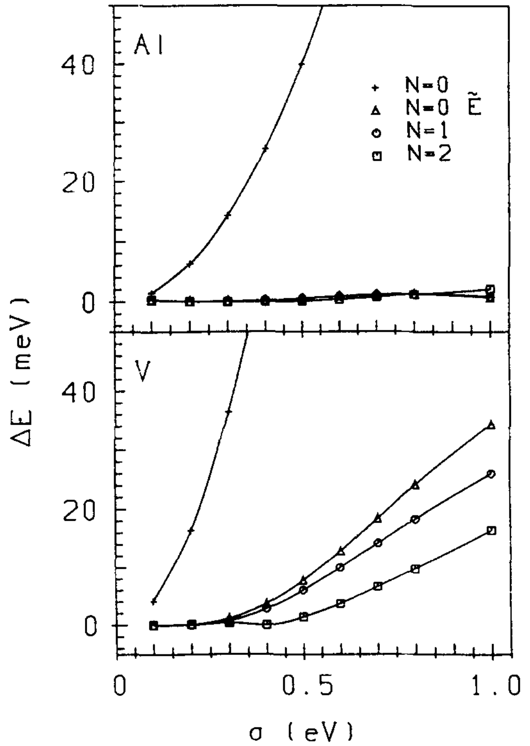

Fig. 1. Convergence of the free energy for Gaussian smearing ($N = 0$) and the MP-method ($N = 1$, $N = 2$) with respect to σ. Also shown is the convergence of the quantity $\tilde{E}$ (Eq. (21)) for $N = 0$ i.e. Gaussian smearing. A k-point grid consisting of 35×35 ×35 k-points (i.e. 1140 k-points in the irreducible wedge of the Brillouin zone has been used). The energy zero was calculated with the LT-C method.

As an illustration the convergence with respect to σ for $F(\sigma)$ and $\tilde{E}(\sigma)$ is shown for Aluminium and Vanadium in Fig. 1. In the context of ab-initio calculations Eq. (21) was first used by Gillan [12] and later generalized by De Vita and Gillan [37] and it allows an easy and accurate evaluation of zero temperature properties. Even for a relatively large σ $\tilde{E}(\sigma)$ converges to the exact zero temperature total energy $E_{\sigma=0}$, and the evolution of $\tilde{E}(\sigma)$ requires a substantially smaller number of k-points than a calculation with $\sigma = 0$. Typical settings for σ for different materials are compiled in Table 1.

2.2.3. Improved functional form for $f(\epsilon)$ – Method of Methfessel and Paxton

The method described in the last section has two distinct shortcomings:

- Forces are usually defined as the derivatives of the variational quantity, i.e. the free electronic energy F (see Section 2.4). Therefore the forces *cannot* be used to obtain the equilibrium zero 'temperature' groundstate or exact phonon frequencies which corresponds to an energy-minimum of $\tilde{E}(\sigma)$.
- The parameter σ has to be chosen with great care. If σ is too large the energy $\tilde{E}(\sigma)$ will converge to the wrong value even for an infinite k-point mesh, and if σ is too small the convergence with the number of k-points will deteriorate. An optimum choice for σ can be found only by comparing results for different k-point meshes and different values of σ.

These problems can be solved by adopting a slightly different functional form for $f(\{\epsilon\})$, which was first proposed by Methfessel and Paxton (MP) [38]. They expanded the step function in a complete orthonormal set of functions. Within this approach the integral of the Gaussian is only the first approximation ($N = 0$) of the step function, further successive approximations ($N = 1, 2, \ldots$) can be obtained easily. In analogy to the Gaussian method, the total energy is no longer variational with respect to the partial occupancies and has to be replaced by a generalized free energy functional (one feature missing in the original work of Methfessel and Paxton). The variational quantity is defined by

$$F = E - \sum_n \sigma S_N\left(\frac{\epsilon_n - \mu}{\sigma}\right), \tag{22}$$

Table 1
Convenient settings for the smearing parameter σ for different metals

	σ (eV)
Aluminium	1.0
Lithium	0.4
sc-Tellurium	0.8
Copper	0.4
Vanadium	0.3
Rhodium	0.3

The smearing parameter σ was determined so that the entropy term $\Sigma_n \sigma S_N(f_n)$ was less than 1 meV/atom in the method of Methfessel and Paxton with $N = 1$. Aluminium, Lithium and Tellurium show a fairly simple structure of the DOS at the Fermi level, therefore σ might be large. For Copper σ is restricted by the fact that the d-band lies approximately 0.5 eV beneath the Fermi level. Due to the complicated structure of the DOS at the Fermi level σ must be small for most transition metals like Vanadium.

where S_N is given by

$$S_N(x) = \tfrac{1}{2}A_N H_{2N}(x)\,\mathrm{e}^{-x^2} \tag{23}$$

and the partial occupancies are given by

$$f_0(x) = \tfrac{1}{2}(1 - \mathrm{erf}(x)),$$

$$f_N(x) = f_0(x) + \sum_{m=1}^{N} A_m H_{2m-1}(x)\mathrm{e}^{-x^2}. \tag{24}$$

with

$$x = \frac{\epsilon - \mu}{\sigma}. \tag{25}$$

H_m is the Hermite polynomial of degree m, and explicit formulas for A_m can be found in Ref. [38].

In contrast to the Gaussian method the entropy term $\Sigma_n \sigma S_N((\epsilon_n - \mu)/\sigma)$ will be very small for a reasonable choice of σ, and the deviations from $E_{\sigma=0}$ are only of the order $N+2$ in σ (see also Fig. 1)

$$F(\sigma) = E_{\sigma=0} + \mathrm{O}(\sigma^{2+N}). \tag{26}$$

Extrapolation to zero σ is usually not necessary, but in principle it might be done using

$$E_{\sigma=0} \approx \tilde{E}(\sigma) = \frac{1}{N+2}((N+1)F(\sigma) + E(\sigma)). \tag{27}$$

The values given in Table 1 will result in an entropy which is less than 1 meV per atom and in a very accurate description of the lattice constant and bulk moduli. We found that the 1 meV threshold is sufficient for most calculations of elastic properties and phonon frequencies.

To summarize: For the MP-method the entropy term is a simple error estimation for the difference between the free energy F and the 'physically' correct energy $E_{\sigma=0}$. σ can be increased until this error estimation gets larger than an allowed threshold (usually 1 meV). Because the free energy and the 'physical' energy $E_{\sigma=0}$ are the same except for this small error the forces which are calculated as a derivative of the free energy are also correct and can be used to determine the zero 'temperature' groundstate. Especially the last property makes the method of Methfessel and Paxton very appealing for situations where the k mesh is not sufficient for the application of the tetrahedron method, or applications where accurate forces are required (see Section 2.2.5 and Table 2).

2.2.4. *Convergence of the total energy with the number of k-points*

We want to illustrate the convergence with respect to the number k-points for the LT-C and the MP-method for two simple examples — bulk Aluminium and bulk Vanadium. In Vanadium the convergence is especially cumbersome, because s and d like bands exist close to the Fermi surface.

Fig. 1 shows the convergence of different functionals with respect to σ for Al and V. For the conventional Gaussian method the free energy $F(\sigma)$ deviates even for small σ rapidly from $E_{\sigma=0}$. But the functional $\tilde{E}$ (Eq. (21)) and the free energy functionals for the MP-methods with $N \geq 1$ converge rapidly to the correct energy, allowing a much larger σ.

In Fig. 2 the convergence of energy for the LT-C and of the free energy for Gaussian smearing ($N = 0$) and the MP-method ($N = 1$) with respect to the k-point mesh is shown. For each calculation σ was chosen so that the error in the k-point converged energy was less than 1 meV. It might be seen that

Table 2
Phonon frequencies for Rh at the K-point (i.e. in (111) direction, Brillouin zone boundary) calculated using a frozen-phonon approach

f (THz)	Energy MP	Force MP	Energy LT-C	Force LT-C	Energy LT	Force LT
111 trans.	4.29	4.28	4.30	4.10	3.91	3.93
111 long	7.95	7.96	7.93	7.68	7.46	7.49

A Monkhorst Pack grid consisting of 9×9×3 k-points was used corresponding to 70 k-points in the irreducible wedge for the transversal branch and 31 k-points for the longitudinal branch. MP is the method of Methfessel Paxton for $N = 2$ and $\sigma = 0.4$, LT-C is the tetrahedron method including the correction terms proposed by Blöchl and LT the linear tetrahedron method without corrections. The k-point mesh is not sufficient for an accurate calculation of phonon frequencies with the LT method, but results are converged for the MP and LT-C method. For the LT-C method the weights f were naively kept fixed, therefore the forces are not consistent with the energy.

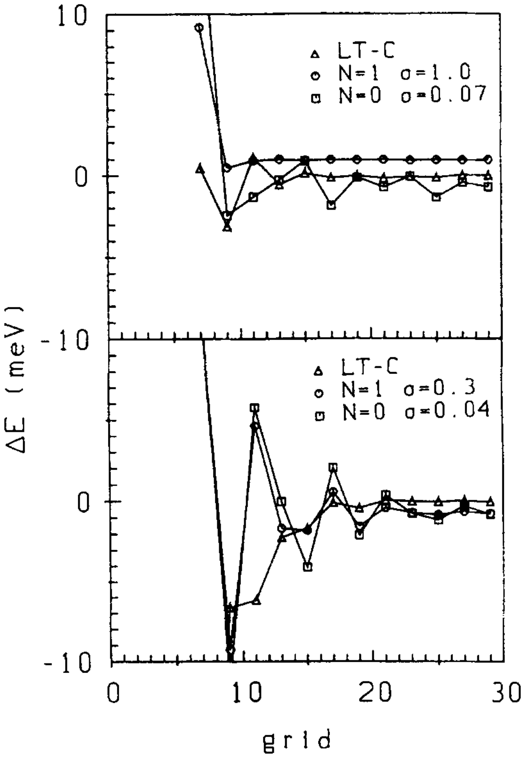

Fig. 2. Convergence of energy for the LT-C and of the free energy for Gaussian smearing ($N = 0$) and the MP-method ($N = 1$) with respect to the k-point mesh. The k-point meshes consisted of grid×grid×grid k-points. For the smearing methods *sigma* σ was chosen so that the error in the k-point converged energy was less than 1 meV with respect to the LT-C method.

the LT-C method converges fastest for Al and V and in addition the LT-C does not require to find an optimal σ. Nevertheless the convergence of the MP-method ($N = 1$) is also quite good, mainly because σ can be relatively large. Gaussian smearing without the extrapolation $\tilde{E}$ (Eq. (21)), requires a small σ resulting in the slowest convergence.

2.2.5. Calculation of phonon frequencies for metals

To illustrate that the LT-C method might be problematic for the calculation of phonon frequencies based on the forces we show results for the phonon frequencies of Rh for a high symmetry point in Table 2. The phonon frequencies were calculated from the change of the free energy induced by a small displacement and from the forces for a displaced structure. Different methods for calculating the partial occupancies have been used: For the MP-method and the standard linear tetrahedron (LT) method the phonon frequencies derived from the forces and the energy are equivalent, but because the k-point mesh is not sufficient for a good convergence of the LT method errors in the LT method are up to 10%.

Phonon frequencies derived from the free energy change are almost indistinguishable for the MP and the linear tetrahedron method with Blöchl corrections (LT-C) approach indicating that the k-point mesh is sufficient for both methods. But for the LT-C method, the partial occupancies were kept fixed for the evaluation of the forces, resulting in a considerable error of the forces and a serious error of the corresponding phonon frequencies (≈ 5%). This indicates that the MP-method is the best choice for calculations where accurate forces are required. Especially phonon frequencies can be evaluated accurately and easily within this approach. We have used the MP-method recently with good success for the calculation of phonon frequencies in bulk Li, Na, K, Rh and Mo and for the calculation of properties of metallic Rh surfaces (including surface phonons) [21].

2.3. Selfconsistency loop and iterative methods

We have pointed out in the introduction that iterative methods for the diagonalization of the KS-Hamiltonian seem to be the most efficient schemes for calculating the finite temperature KS groundstate. These methods must be used in conjunction with a charge density mixing to get a reliable scheme; Fig. 3 shows a typical flowchart for this situation: At the beginning an appropriate set of trial wavefunction $\{\phi_n,\ n = 1,\ldots,N_b\}$ and a reasonable input charge density ρ_{in} is chosen. Usually, the start charge density corresponds to the superposition of the atomic pseudo charge densities of the constituents. From the input charge density the local potential

$$V_{\text{loc}} = V_{\text{loc}}^{\text{ion}} + V^{\text{H}}[\rho_{\text{in}}] + V^{\text{xc}}[\rho_{\text{in}}] \tag{28}$$

and the corresponding double counting corrections

$$E_{\text{d.c.}}[\rho_{\text{in}}] = -\tfrac{1}{2}E^{\text{H}}[\rho_{\text{in}}] + E^{\text{xc}}[\rho_{\text{in}}] - \int \mathrm{d}^3 r\, V^{\text{xc}}(r)\,\rho_{\text{in}}(r) \tag{29}$$

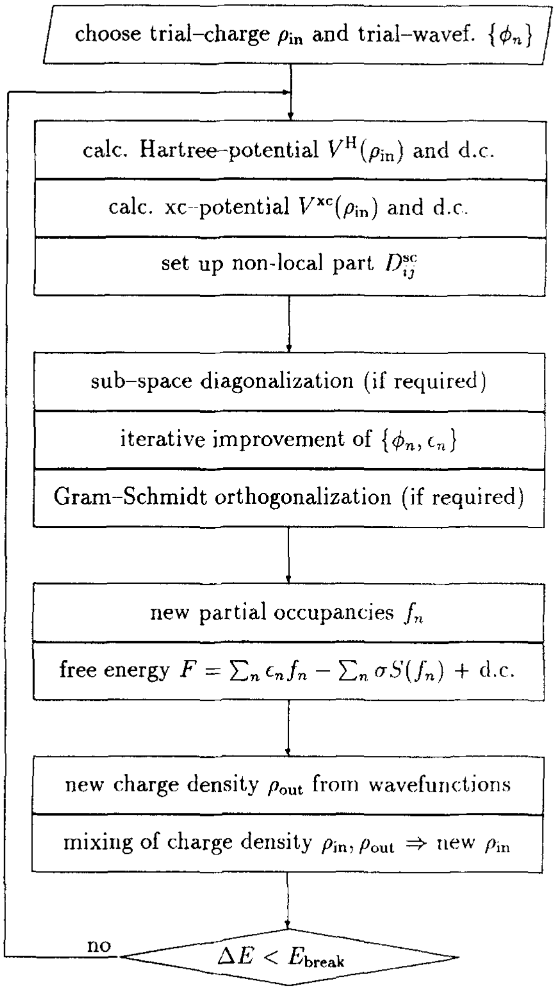
choose trial–charge $\rho_{\rm in}$ and trial–wavef. $\{\phi_n\}$
calc. Hartree–potential $V^{\rm H}(\rho_{\rm in})$ and d.c.
calc. xc–potential $V^{\rm xc}(\rho_{\rm in})$ and d.c.
set up non-local part $D_{ij}^{\rm sc}$
sub–space diagonalization (if required)
iterative improvement of $\{\phi_n, \epsilon_n\}$
Gram–Schmidt orthogonalization (if required)
new partial occupancies f_n
free energy $F = \sum_n \epsilon_n f_n - \sum_n \sigma S(f_n)$ + d.c.
new charge density $\rho_{\rm out}$ from wavefunctions
mixing of charge density $\rho_{\rm in}, \rho_{\rm out} \Rightarrow$ new $\rho_{\rm in}$
no
$\Delta E < E_{\rm break}$

Fig. 3. Flow chart for iterative methods for the diagonalization of the KS-Hamiltonian in conjunction with an iterative improvement (i.e. mixing) of the charge density.

are evaluated. For ultrasoft pseudopotentials the non-local part of the pseudopotential depends also on the local potential and must be calculated accordingly (Eq. (10)). In the next step the N_b trial wavefunctions are improved using an iterative method, and the new eigenenergies are used to calculate a new Fermi energy and new partial occupancies. The total free energy for the current iteration is calculated as the sum of the band structure energy plus the entropy term plus double counting corrections

$$F = \sum_n f_n \epsilon_n^{\rm app} - \sum_n \sigma S\left(\frac{\epsilon_n - \mu}{\sigma}\right) + E_{\rm d.c.}[\rho_{\rm in}]. \tag{30}$$

Conceptually the calculated energy corresponds to the energy evaluated from the Harris–Foulkes (HF) functional [39–41], which is non-selfconsistent — in contrast to the KS functional: the HF functional (defined in Eq. (30)) requires the calculation of the band structure energy for a fixed charge density $\rho_{\rm in}$. With our code it is easy to evaluate this energy keeping the initial charge density fixed (for instance to the superposition of atomic pseudo charge densities) and iterating the eigenvectors only until they are converged.

To get the exact KS-groundstate-energy selfconsistency with respect to the input charge density requires that the charge density residual vector $R[\rho_{\rm in}]$

$$R[\rho_{\rm in}] = \rho_{\rm out} - \rho_{\rm in} \tag{31}$$

is zero, where the output charge density $\rho_{\rm out}$ is calculated from the wavefunctions using Eq. (3). The residual vector $R[\rho_{\rm in}]$ – and possibly information from previous mixing steps – allows to calculate a new charge density $\rho_{\rm in}$ for the next selfconsistency loop. In principle it is necessary to evaluate the eigenfunctions ϕ_n exactly for each new input charge density making $\rho_{\rm out}$ and the residual vector R functionals of the input charge density $\rho_{\rm in}$ only. Nevertheless, even in conjunction with complex Broyden like mixing techniques, it turns out that this is not necessary if the final wavefunctions of the previous mixing iteration are used as new initial trial wavefunctions. In this case a few steps in the iterative matrix diagonalization are sufficient to get a reliable result for the charge density residual vector R. In Section 3 we will concentrate on different iterative methods for the diagonalization of the KS-Hamiltonian, Section 4 will discuss algorithms for the charge density mixing.

2.4. Forces

Forces for the finite-temperature KS functional can be obtained easily, but the calculation is somewhat complicated by the US PP. To obtain the forces it is convenient to use a basis set oriented notation. In our case the wavefunctions are written as a sum of a finite set of plane waves $|q\rangle$, i.e.

$$|\phi_n\rangle = \sum_q C_{nq}|q\rangle, \tag{32}$$

(once again the k-point index $\boldsymbol{k}$ has been omitted) and the KS energy functional E, respectively the free energy F becomes a function of the expansion coefficients C_{nq} the partial occupancies f_n and the ionic positions $\boldsymbol{R}_N$,

$$F \rightarrow F[\{C\},\{f\},\{\boldsymbol{R}\}]. \tag{33}$$

To incorporate the orthonormality constraint and the conservation of the number of electrons it is convenient to use the Lagrange formalism and to introduce the function

$$\bar{F}[\{C\},\{\gamma\},\{f\},\mu,\{\boldsymbol{R}\}] = F - \sum_{nn'qq'} \gamma_{nn'} C^{*}_{nq'} S_{q'q} C_{nq} - \mu\left(\sum_{n} f_n - N_{\mathrm{el}}\right), \tag{34}$$

where $\boldsymbol{A}_{q'q}$ is defined as

$$\boldsymbol{A}_{q'q} = \langle \boldsymbol{q}' | \boldsymbol{A} | \boldsymbol{q} \rangle. \tag{35}$$

At the KS-groundstate the Lagrange multipliers are given by $\gamma_{nn'} = \delta_{nn'} \epsilon_n f_n$, where ϵ_n are the exact KS-eigenvalues (compare with Eq. (117)), and $\bar{F}$ is minimal with respect to arbitrary variations of C_{nq}, $\gamma_{nn'}$, f_n and μ. The change in the free energy up to first-order is exactly given by (see Appendix A)

$$\mathrm{d}F = \sum_{N} \frac{\partial \bar{F}[\{C\},\{\gamma\},\{f\},\mu,\{\boldsymbol{R}\}]}{\partial \boldsymbol{R}_N} \mathrm{d}\boldsymbol{R}_N, \tag{36}$$

and it is convenient to define the forces $\boldsymbol{F}_N$ as

$$\boldsymbol{F}_N = \frac{\partial \bar{F}}{\partial \boldsymbol{R}_N}. \tag{37}$$

This formula is exact and contains Hellmann–Feynman [2] as well as Pulay contributions [42] (for the pseudopotential approach, no Pulay contributions exist, but Eq. (37) is also exact for other basis sets). A similar formula also holds for the stress tensor, derivatives with respect to the basis set are implicitly contained in this definition. It is now easy to show that the forces can be rewritten as (for the selfconsistent case this equation was first derived in Ref. [43])

$$\boldsymbol{F}_N = \sum_{nqq'} f_n C^{*}_{nq'} \frac{\partial(\boldsymbol{H}[\rho,\{\boldsymbol{R}\}] - \epsilon_n S[\{\boldsymbol{R}\}])_{q'q}}{\partial \boldsymbol{R}_N} C_{nq}, \tag{38}$$

where changes of the Hamiltonian $\boldsymbol{H}$ due to changes in the selfconsistent charge density ρ must *not* be calculated. For further details we refer to Ref. [26,27].

It is also possible to obtain a correct formula for the forces if the Harris–Foulkes functional instead of the Kohn–Sham functional is used. If the input charge density ρ_{in} for the Harris–Foulkes functional is calculated from the atomic charge density of the constituents, only one additional term arises which is due to the fact that the input charge density depends on the atomic coordinates. In this case $\boldsymbol{H}$ in Eq. (38) has to be replaced by the Hamiltonian calculated from the atomic charge density $\boldsymbol{H}[\rho_{\mathrm{atom}},\{\boldsymbol{R}\}]$, and the term

$$\int \mathrm{d}^3\boldsymbol{r}\left(\frac{\partial V^{\mathrm{H}}(\rho_{\mathrm{atom}}) + V^{\mathrm{xc}}(\rho_{\mathrm{atom}})}{\partial \boldsymbol{R}_N} \times(\boldsymbol{r})(\rho_{\mathrm{out}}(\boldsymbol{r}) - \rho_{\mathrm{atom}}(\boldsymbol{r}))\right). \tag{39}$$

has to be added to the forces. In Eq. (38) changes of the Hamiltonian $\boldsymbol{H}$ due to changes in the input charge density ρ_{atom} have to be omitted, as in the selfconsistent case.

We have found that the similar correction formula

$$\int \mathrm{d}^3\boldsymbol{r}\left(\frac{\partial V^{\mathrm{H}}(\rho_{\mathrm{atom}}) + V^{\mathrm{xc}}(\rho_{\mathrm{atom}})}{\partial \boldsymbol{R}_N}(\boldsymbol{r}) \times(\rho_{\mathrm{out}}(\boldsymbol{r}) - (\rho_{\mathrm{in}}(\boldsymbol{r})))\right). \tag{40}$$

also improves the convergence of the forces during a selfconsistent calculation. In Eq. (38) $\boldsymbol{H}$ has to be replaced by $\boldsymbol{H}[\rho_{\mathrm{in}},\{\boldsymbol{R}\}]$, where ρ_{in} is the charge density obtained in the previous iteration. In principle it is necessary to evaluate the change of ρ_{in} if the ions move (i.e. the first term in Eq. (40) should be replaced by $\partial(V^{\mathrm{H}} + V^{\mathrm{xc}})(\rho_{\mathrm{in}})/\partial \boldsymbol{R}_N$), which is not possible, but Eq. (40) seems to be an excellent approximation. This correction formula improves the precision of the forces by almost two orders of magnitude, and allows to stop the selfconsistency cycle much earlier.

This is demonstrated in Fig. 4, where the convergence for the forces is compared for different algorithms for a Pd(111) surface with a mono-layer hydrogen (see Section 6.1.2). It can be seen that the

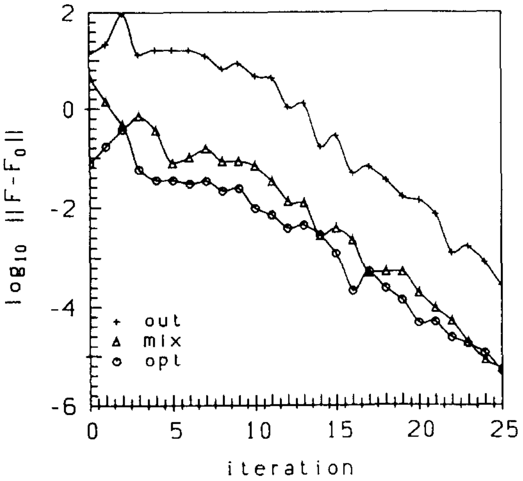

Fig. 4. Convergence of the forces (in eV/Å/atom) for different algorithms for the Pd (111) surface with a monolayer hydrogen located in the hollows between three Pd atoms, for a selfconsistent calculation. 'out' – output charge density was used for the calculation; 'mix' – mixed charge density was used; 'opt' – is the optimized scheme explained in the text.

optimized scheme (opt) explained here results in the best overall performance. A similar convergence rate might be obtained by using the mixed charge density (mix) (Section 4) for the calculation of the local contribution to the forces, i.e.

$$\sum_{nqq'} f_n C^*_{nq'} \frac{\partial V^{\text{ion}}_{\text{loc}\, q'q}}{\partial \boldsymbol{R}_N} C_{nq} \to \int d^3r \frac{\partial V^{\text{ion}}_{\text{loc}}}{\partial \boldsymbol{R}_N}(\boldsymbol{r})\, \rho_{\text{mixed}}(\boldsymbol{r}). \tag{41}$$

This part of the forces is very sensible to changes in the charge density. The use of the output charge density (out) – i.e. the left side of Eq. (41) – without the correction term (Eq. (40)) makes the forces worse by a factor 100 (see Fig. 4).

3. Iterative methods for the diagonalization of the KS-Hamiltonian

In this section we will discuss several iterative approaches for the diagonalization of the KS Hamiltonian, including the residual minimization method direct inversion in the iterative subspace (RMM-DIIS or simply RMM) proposed by Pulay [24] and Wood and Zunger [44], which is superior for very large matrices. For ultrasoft pseudopotentials we are confronted with a generalized eigenvalue problem (see Eq. (7))

$$H|\phi_n\rangle = \epsilon_n S|\phi_n\rangle, \tag{42}$$

which has to be solved. For a small basis set, this eigenvalue problem is usually solved by straightforward diagonalization of the Hamiltonian (for instance using the Choleski–Householder procedure). Nevertheless, this procedure becomes intractable for large matrices, because it scales as N^3_{plw}, N_{plw} being the number of plane-waves included in the basis set. For three reasons iterative methods are an order of magnitude faster for the calculation of the eigenfunctions: (i) only $N_{\text{b}} \ll N_{\text{plw}}$ occupied orbitals must be calculated, (ii) the calculation of $H|\phi_n\rangle$ is especially convenient for a plane-wave basis set (see introduction), and (iii) iterative methods are efficient in conjunction with a selfconsistent calculation, because optimization of the charge density and wavefunctions can be done almost simultaneously.

A good review of different iterative methods can be found in [44] and we will try to adopt the notation of this paper. Nevertheless we also want to point out that the examples discussed in Ref. [44] are only small to medium sized from today's point of view, and not all of the algorithms of Ref. [44] work reliably for very large systems. In addition new iterative methods (like the band-by-band conjugate gradient algorithm [10,45]) have been proposed recently, making a new comparison between different methods necessary.

As pointed out in [44] most iterative algorithms construct an expansion set $\{|b_i\rangle, i = 1, \ldots, N_{\text{a}}\}$ from which the best approximation to the exact eigenvalues and eigenvectors is calculated. This expansion set is much smaller than the number of plane waves $N_{\text{a}} \ll N_{\text{plw}}$ and depending on the algorithm it might be smaller or larger than the number of bands N_{b} to be calculated. In each iteration new vectors are added to the expansion set. It is possible to differentiate between blocked and unblocked methods. Unblocked methods (or sequential band by band methods) optimize one band at a time and the expansion set usually starts with only one vector $|b_0\rangle$, which should be a reasonable approximation to the eigenvector $|\phi_n\rangle$. In each iteration i a single correction vector $|b_i\rangle$ is added to the expansion set. Blocked methods on the other hand optimize all orbitals or a set of orbitals simultaneously, increasing the size of the expansion set at each step by N new vectors.

In most blocked and unblocked methods it is necessary to obtain a new best approximation of the exact eigenvalues and eigenvectors from the expansion set at each iteration. This is done via the Rayleigh–Ritz [44,46] scheme: In principle the Hamiltonian is diagonalized in the space spanned by the expansion set, i.e. the $N_a \times N_a$ eigenvalue problem

$$\sum_m \bar{H}_{nm} B_{mk} = \sum_m \epsilon_k^{\mathrm{app}} \bar{S}_{nm} B_{mk} \tag{43}$$

with

$$\bar{H}_{nm} = \langle b_n | H | b_m \rangle, \qquad \bar{S}_{nm} = \langle b_n | S | b_m \rangle, \tag{44}$$

must be solved. The m lowest eigenvalue/eigenvector pairs

$$\epsilon_k^{\mathrm{app}}, |\bar{b}_k\rangle = \sum_m B_{mk} | b_m \rangle \tag{45}$$

correspond to the best approximation of the exact lowest m eigenvalues and eigenvectors within the sub-space spanned by the expansion set.

3.1. Residual vector and preconditioning

A key step within all iterative methods is the calculation of a correction vector which is added to the expansion set, and a central quantity within most methods is the Rayleigh quotient, which is defined as

$$\epsilon_{\mathrm{app}} = \frac{\langle \phi_n | H | \phi_n \rangle}{\langle \phi_n | S | \phi_n \rangle}. \tag{46}$$

This quantity possesses a saddle point at the exact eigenvector, and variation of the Rayleigh quotient with respect to $\langle \phi_n |$ leads to the residual vector defined as

$$|R(\phi_n)\rangle = (H - \epsilon_{\mathrm{app}} S) | \phi_n \rangle \tag{47}$$

if $\langle \phi_n | S | \phi_n \rangle = 1$. The norm of the residual vector $\langle R | R \rangle$ is an accepted measure for the error in the eigenvector. Formally a good approximation to the difference between the exact and the approximate eigenvector $|\phi_n\rangle$ is obtained from the residual vector using the equation

$$|\delta\phi_n\rangle = -\frac{1}{H - \epsilon_{\mathrm{app}} S} |R\rangle. \tag{48}$$

It is easy to show that $|\bar{\phi}_n\rangle = |\phi_n\rangle + |\delta\phi_n\rangle$ results in a minimum residual vector, which satisfies

$$0 = (H - \epsilon_{\mathrm{app}} S) | \bar{\phi}_n \rangle. \tag{49}$$

Nevertheless, the inversion of $H - \epsilon_{\mathrm{app}} S$ is not easier than the diagonalization of the matrix H and definitely intractable for large matrices. This makes a more approximate treatment necessary. In the following, the step which calculates the approximative error $|\delta\phi_n\rangle$ from the residual vector will be called preconditioning, and the matrix K which is multiplied with the residual vector to obtain $|\delta\phi_n\rangle$

$$|\delta\phi_n\rangle = K |R\rangle \tag{50}$$

will be called preconditioning matrix.

Frequently only the diagonal elements of the exact matrix in Eq. (49) are used, i.e.

$$K = -\sum_q \frac{|q\rangle\langle q|}{\langle q | H - \epsilon_{\mathrm{app}} S | q \rangle}, \tag{51}$$

where q runs over all plane-waves included in the basis set. Instead of $|q\rangle$ it is possible to chose a different complete set of vectors in Eq. (51). Note that Eq. (51) is equivalent to Eq. (48) if $|q\rangle$ is replaced by the exact eigenvectors of the Hamiltonian H. This inspired Wood and Zunger [44] to use the eigenvectors $|a_i^0\rangle$ of a small approximate Hamiltonian H_0 plus a set of plane-waves to make a complete set. The Hamiltonian H_0 might be calculated for a plane-wave basis set consisting of N_0 plane-waves; this results in a preconditioning matrix

$$K = -\sum_{i=1}^{N_0} \frac{|a_i^0\rangle\langle a_i^0|}{\langle a_i^0 | H - \epsilon_{\mathrm{app}} S | a_i^0 \rangle} - \sum_q{}' \frac{|q\rangle\langle q|}{\langle q | H - \epsilon_{\mathrm{app}} S | q \rangle}, \tag{52}$$

where the prime in the second sum means exclusion of all plane waves included in the initial Hamiltonian H^0. We have tested this preconditioning to some extent and it works reasonably well for small to medium sized problems ($N_{\mathrm{plw}} < 1000$), but for large basis sets ($N_{\mathrm{plw}} > 1000$) the size N_0 of the initial matrix soon becomes the limiting factor. If the initial size is equal to the number of bands $N_0 = N_b$ the improvement over the diagonal approximation Eq. (51) is small. If N_0 is considerably larger than N_b

($N_0 > 3N_b$) the calculation of the first sum in Eq. (52) becomes the most expensive part of the calculation. Therefore, we actually adopted the preconditioning function proposed by Teter et al. [10]

$$K = -\sum_q \frac{2|q\rangle\langle q|}{\frac{3}{2}E^{\mathrm{kin}}(R)} \times \frac{27 + 18x + 12x^2 + 8x^3}{27 + 18x + 12x^2 + 8x^3 + 16x^4}$$

$$\text{with } x = \frac{\hbar^2}{2m_e} \frac{q^2}{\frac{3}{2}E^{\mathrm{kin}}(R)}, \qquad (53)$$

and $E^{\mathrm{kin}}(R)$ being the kinetic energy of the residual vector. There are only two changes with respect to Ref. [10]: First, we use $\frac{3}{2}E^{\mathrm{kin}}(R)$ instead of $E^{\mathrm{kin}}(R)$ in the definition of x, resulting in a slightly improved convergence speed for most elements. Second, we multiply the preconditioning function by a constant factor $2/(\frac{3}{2}E^{\mathrm{kin}}(R))$. Using this factor the diagonal part of the preconditioning matrix K converges to

$$\frac{2m_e}{\hbar^2 q^2} \qquad (54)$$

for large q, resulting in a more convenient length of the correction vector $|\delta\phi_n\rangle$. Although the length of the correction vector does not enter in any of the schemes discussed in the next sections, it is important to have a reasonable length for the algorithm in which the KS functional is minimized directly (see Section 5).

Slightly different preconditioning functions were proposed by several other authors: Furthmüller [47] and recently Tassone et al. [7] used the matrix

$$K = -\sum_q |q\rangle\langle q| \times \begin{cases} 1, & q < q_{\mathrm{cut}}, \\ q_{\mathrm{cut}}^2/q^2, & q > q_{\mathrm{cut}}. \end{cases} \qquad (55)$$

The functional form of this matrix is very similar to Eq. (53) and requires the determination of an optimum cut off q_{cut}. Generally it does not improve the convergence speed in comparison to the 'smoother' preconditioning function (53).

3.2. Blocked Davidson scheme

This method was originally proposed by Davidson [48] and later modified by Liu for a simultaneous update of all bands [49]. The expansion set increases in each step M by N_b – where N_b is the number of bands included in the calculation $N_b \geq N_{\mathrm{elect}}/2$ – residual vectors

$$\{|b_i\rangle, \ldots, i = 1, N_b(M+1)\} = \{|\phi_i^0\rangle, i = 1, \ldots, N_b/|P_i^0\rangle, i = 1, \ldots, N_b/|P_i^1\rangle, i = 1, \ldots, N_b/\ldots\}. \qquad (56)$$

$|P_i^0\rangle = K|R(\phi_i^0)\rangle$ are the preconditioned residual vectors for the initial wavefunctions. In each iteration M the Rayleigh–Ritz scheme is used to obtain the lowest eigenvalue/eigenvector pairs $\epsilon_i^M, |\phi_i^M\rangle$. From these eigenvectors the new preconditioned residual vectors $|P_i^M\rangle = K|R(\phi_i^M)\rangle$ are calculated and added to the expansion set. For large problems the storage of all previous residual vectors and of the initial vectors $|\phi^0\rangle$ becomes a problem, therefore M must be restricted to a relatively small value. In our case we generally use $M = 1$: In each step the final expansion set consists of $2N_b$ vectors, the preconditioned residual vectors $\{|P_i^0\rangle\}$ and the initial set $\{|\phi_i^0\rangle\}$. Then a diagonalization is performed in this $2N_b$-dimensional sub-space, and the N_b lowest eigenvectors are calculated. In the next step these lowest eigenvectors and the new residual vectors form the new expansion set. For a selfconsistent calculation usually two steps are necessary between the charge density mixing. After the charge density mixing the final wavefunctions of the previous step are used as new initial trial vectors. We will refer to this algorithm as DAV2, for DAV2 one step always consist of 2 successive simple blocked Davidson steps.

3.3. Unblocked algorithms

Unblocked or sequential algorithms are generally considered to be 'slower' than blocked algorithms, nevertheless the blocked Davidson scheme requires the storage of at least $2N_b$ vectors, which is inconvenient for large systems. Schemes which optimize a single band at a time require less storage, and as we will show they are also more efficient for very large systems, because the number of iterations can be larger than in the blocked Davidson scheme. Generally it is favorable to restrict the search vector in the sequential methods to the sub-space orthonormal to the current wavefunctions. After a band-by-band sequential update of *all* wavefunctions it is necessary to determine the optimal wavefunctions in the sub-space spanned by the N_b final trial wavefunctions $\{|\phi_n\rangle, n = 1,\ldots, N_b\}$ using the Rayleight–Ritz scheme (Eqs. (43)–(45)). We will refer to this step as sub-space diagonalization or sub-space rotation. In any efficient sequential scheme sub-space rotation and sequential update should alternate.

3.4. Correction vector for sequential, band-by-band methods

Generally the correction vector must be chosen with a little bit more care in the sequential methods. Consider that one specific band m, has to be optimized: As already pointed out, it is convenient in the sequential methods to restrict the search direction for this band m to the sub-space orthogonal to the current trial set $\{|\phi_n\rangle, n = 1,\ldots, N_b\}$. This can be done easily using e.g. the Lagrange formalism: Minimization of the Rayleigh quotient with the additional orthonormality constraint

$$\sum_n \gamma_{mn}(\langle\phi_m|S|\phi_n\rangle - \delta_{mn}) = 0 \quad \forall n = 1,\ldots, N_b, \tag{57}$$

results in a gradient vector

$$|g(\phi_m)\rangle = |g_m\rangle = H|\phi_m\rangle - \sum_n S\gamma_{mn}|\phi_n\rangle. \tag{58}$$

which can be made orthonormal to the current trial set by requiring

$$\langle\phi_n|g_m\rangle = 0 \quad \forall n = 1,\ldots, N_b, \tag{59}$$

and after evaluating the Lagrange multiplicators explicitly we obtain

$$|g(\phi_m)\rangle = |g_m\rangle = \left(1 - \sum_n S|\phi_n\rangle\langle\phi_n|\right)H|\phi_m\rangle. \tag{60}$$

For a set of trial wavefunctions which diagonalizes the Hamiltonian, i.e. $\langle\phi_n|H|\phi_m\rangle = \delta_{mn}\epsilon_m^{\text{app}}$, Eq. (60) reduces simply to the residual vector (47). Mind that the gradient vector allows to calculate the first-order change of the Rayleigh quotient ϵ_m^{app} (Eq. (47)) using

$$d\epsilon_m^{\text{app}} = \langle\delta\phi_m|g_m\rangle + \text{c.c.}, \tag{61}$$

where c.c. is the complex-conjugate of the first expression. The choice of the constraint (59), which actually determines the Lagrange multipliers, is inspired by the requirement of getting correct first-order energy changes: For a change $|\delta\phi_m\rangle$ parallel to any vector contained in the set $\{\phi_n\}$ the energy change, given by Eq. (61) should be zero.

Because in our implementation the sequential band-by band algorithms generally alternate with the sub-space rotation, it is reasonable to replace the exact gradient (60) with its 'diagonal' approximation the residual vector given by Eq. (47). The residual is preconditioned and then explicitly S-orthogonalized to the set $\{\phi_n\}$, i.e.

$$|p(\phi_m)\rangle = |p_m\rangle = \left(1 - \sum_n |\phi_n\rangle\langle\phi_n|S\right) \times K(H - \epsilon_{\text{app}}S)|\phi_m\rangle. \tag{62}$$

This preconditioned 'search vector' fulfills the condition

$$\langle\phi_n|S|p_m\rangle = 0 \quad \forall n = 1,\ldots, N_b. \tag{63}$$

The sequential methods differ in the way this correction vector is added to the wavefunctions $\{\phi_n\}$.

3.5. Unblocked Davidson-like update

In the conventional unblocked Davidson method a single preconditioned correction vector $K|R(\phi_m)\rangle$ is added to the expansion set $\{b_i\}$ at each iteration. At startup the expansion set consists of the current set of trial wavefunctions, therefore the expansion set is given in each iteration by the set

$$\{|b_i\rangle, i = 1,\ldots, N_b + M\} = \{|\phi_n^0\rangle, n = 1,\ldots, N_b/|P_m^0\rangle/|P_m^1\rangle/\ldots\}. \tag{64}$$

If band m is optimized, then in the first iteration the preconditioned residual is evaluated from the initial trial vector $|P_m^0\rangle = K|R(\phi_m^0)\rangle$, added to the expansion set and a new optimal $|\phi_m^1\rangle$ is calculated applying the Rayleigh–Ritz scheme. In the next step the preconditioned residual $|P_m^1\rangle = K|R(\phi_m^1)\rangle$ is calculated from the new wavefunction $|\phi_m^1\rangle$ and once again added to the expansion set.

This scheme is relatively inconvenient and requires the diagonalization of a large matrix at each step. To simplify the scheme we first replace $|P_m^M\rangle$ by the preconditioned gradient $|p_m^M\rangle = |p(\phi_m^M)\rangle$ (Eq. (62)). By inspection we see that this does not change the result of the iteration scheme, but the overlap matrix $\bar{S}_{ij}$ (Eq. (44)) in the Rayleigh–Ritz scheme is now simply the unity matrix for i or $j \leq N_b$ and $i, j \neq m$. Second, in $\overline{H}_{ij}$ the terms (Eq. (44)) in the off-diagonal elements are neglected for i or $j \leq N_b$ and $i, j \neq m$. This leads to a simple 'decoupling' of the vector sets $\{|\phi_i^0\rangle, i = 1, \ldots, N_b; i \neq m\}$ and $\{|\phi_m^0\rangle/|p_m^0\rangle/|p_m^1\rangle/\ldots\}$, requiring only the diagonalization of a much smaller matrix.

To summarize: In this case the expansion set starts with the trial vector $|\phi_m^0\rangle$, and in successive iterations M the preconditioned and orthonormalized gradient $|p_m^M\rangle = |p(\phi_m^M)\rangle$ is added:

$$\{|b_i\rangle, i = 1, \ldots, M\} = \{|\phi_m^0\rangle/|p_m^0\rangle/|p_m^1\rangle/\cdots\} \tag{65}$$

In each iteration the optimal wave vector $|\phi_m^M\rangle$ is calculated from this expansion set using the Rayleigh–Ritz scheme. After updating one state several times a move to the next band is done, and at last a sub-space rotation for all final wavefunctions $\{|\phi_n^M\rangle, n = 1, \ldots, N_b\}$ is performed. The sub-space rotation at the end is strictly necessary to obtain the exact groundstate eigenvectors. Without the sub-space rotation this scheme would converge to an arbitrary linear combination of the exact lowest eigenvectors.

3.6. Conjugate gradient minimization

It is possible to reduce the number of numerical operations even further by applying the ideas of conjugated gradients (CG) [50,51]. In the context of a sequential energy minimization this was first done by Teter et al. [10], and the same algorithm was also used by Bylander, Kleinman and Lee [45] for the iterative diagonalization of the KS-Hamiltonian.

Instead of storing all previous preconditioned gradients it is possible to conjugate each new search direction to the previous directions applying a standard conjugate gradient scheme, i.e. the search direction $|f^M\rangle$ for iteration M is now given by

$$|f^M\rangle = |p_m^M\rangle + \frac{\langle p_m^M|g_m^M\rangle}{\langle p_m^{M-1}|g_m^{M-1}\rangle}|f^{M-1}\rangle. \tag{66}$$

In this equation $|g_m^M\rangle = |g(\phi_m^M)\rangle$ is the gradient vector defined in Eq. (60), and $|p_m^M\rangle = |p(\phi_m^M)\rangle$ is the preconditioned gradient defined in Eq. (62). In each iteration the optimum new wave vector $|\phi_m^{M+1}\rangle$ is determined from the set $\{|\phi_m^M\rangle/|f^M\rangle\}$ applying the Rayleigh–Ritz scheme. In Eq. (66) it is possible to replace the gradient vector $|g_m^M\rangle$ by the residual vector $|R_m^M\rangle$ because the preconditioned gradient is orthogonal to all wavefunctions and therefore $\langle p_m^M|g_m^M\rangle = \langle p_m^M|R_m^M\rangle$. Except for small (mainly numerical) rounding errors, the improvement per iteration is the same for the conjugate gradient minimization and for the scheme introduced in the previous section (Eq. (65)), and we will restrict the following discussions to the computationally slightly more efficient CG algorithm.

3.7. Residual minimization method – direct inversion in the iterative subspace (RMM-DIIS)

The sequential conjugate gradient scheme discussed so far is relatively fast and very stable. The only remaining drawback is the necessity for an explicit orthonormalization of the preconditioned residual vector $\boldsymbol{K}|R(\phi_m)\rangle$ to the current set of trial wavefunctions (Eq. (62)). Unfortunately avoiding the orthonormalization is not possible applying the algorithms discussed above. The Rayleigh–Ritz scheme tries to obtain the lowest possible eigenvalues in the sub-space spanned by the expansion set, actually it is easy to show that the algorithm minimizes the Rayleigh quotient for a given set of wavefunctions. The Rayleigh quotient is stationary at each eigenvector but it does *not* possess a minimum. Without explicit orthonormalization (62) the schemes investigated in Sections 3.5 and 3.6 will converge for any start vector to the lowest eigenvector of the Hamiltonian. In this case the algorithms are equivalent to a simple Lanczos [44,52] method, only the explicit orthonormalization makes it possible to converge to a selected eigenvalue efficiently.

Fortunately a solution to this problem is available and was first proposed in Ref. [44]. Minimizing *the norm of the residual vector* instead of the Rayleigh quotient makes the orthonormalization unnecessary, because the norm of the residual vector has an unconstrained minimum at each eigenvector.

In our implementation we follow the original work of Pulay [24] and not the variant proposed by Wood et al. [44]. This variant would require the additional calculation and storage of $S|\phi\rangle$, and is therefore slower than our algorithm. We start with an evaluation of the preconditioned residual vector $\boldsymbol{K}|R^0_m\rangle = \boldsymbol{K}|R(\phi^0_m)\rangle$ for a selected band m. Then a trial step into this direction is done

$$|\phi^1_m\rangle = |\phi^0_m\rangle + \lambda \boldsymbol{K}|R^0_m\rangle \tag{67}$$

and the new residual vector $|R^1_m\rangle = |R(\phi^1_m)\rangle$ is evaluated (mind that we update ϵ_{app} in the evaluation of $R(\phi^1_m)$, see Eq. (47)). Next a linear combination of the initial $|\phi^0_m\rangle$ and the trial wavefunction $|\phi^1_m\rangle$

$$|\bar{\phi}^M\rangle = \sum_{i=0}^{M} \alpha_i |\phi^i_m\rangle \quad \text{with } M = 1 \tag{68}$$

is found which minimizes the norm of the residual vector. Assuming linearity in the residual vector i.e.

$$|\bar{R}^M\rangle = |R(\bar{\phi}^M)\rangle = \sum_{i=0}^{M} \alpha_i |R^i_m\rangle \tag{69}$$

this requires the minimization of

$$\frac{\sum_{ij} \alpha_i^* \alpha_j \langle R^i_m | R^j_m \rangle}{\sum_{ij} \alpha_i^* \alpha_j \langle \phi^i_m | S | \phi^j_m \rangle}. \tag{70}$$

This step is usually called direct inversion in the iterative subspace (DIIS). The problem stated in Eq. (70) is equivalent to solving the Hermitian eigenvalue problem

$$\sum_{j=0}^{M} \langle R^i_m | R^j_m \rangle \alpha_j = \epsilon \sum_{j=0}^{M} \langle \phi^i_m | S | \phi^j_m \rangle \alpha_j. \tag{71}$$

The next trial step starts from $|\bar{\phi}^M\rangle$ along the direction $\boldsymbol{K}|\bar{R}^M\rangle$. In each iteration M a new wavefunction $|\phi^M_m\rangle = |\bar{\phi}^{M-1}\rangle + \lambda \boldsymbol{K}|\bar{R}^{M-1}\rangle$ and a new residual vector $|R(\phi^M_m)\rangle$ are added to the iterative sub-space. The size of the trial step λ is a critical value. We have found that a reasonable choice for the trial step can be obtained from the minimization of the Rayleigh quotient along the search direction in *the first step* (this is in the spirit of Section 3.5), this optimal λ is used until a move to the next band is performed. The line minimization can be done *without* additional computational requirements. Usually the obtained step size is between 0.3 and 1 for the preconditioning function given in Eq. (53). In rare cases – especially if the minimization of the Rayleigh quotient starts to go for the wrong band – the trial step might become very large. Therefore we restrict the size of the trial step to a value between 0.1 and 1. With this choice we have found that it is always possible to finish with the trial step. The trial step approaches already the exact position of the minimal residual vector.

The scheme explained in this section requires approximately the same number of iterations as the CG algorithm, but it avoids any explicit orthonormalization and is therefore much faster for very large problems where the orthonormalization is the leading factor. Even more important is the fact that the residual minimization is inherently local, and it is therefore very easy to implement the algorithm on a parallel machine. For instance each processor might handle a certain number of bands, information about other bands is not required (see also Section 3.9).

One drawback of the RMM method is that it always finds the vector which is closest to the initial trial vector. Therefore, initialization becomes a critical task and it might happen that in the final solution one vector is 'missing'. To avoid this the initialization must be done with great care: We usually start with a set of random trial vectors, and perform three sweeps over all bands. Each initial sweep consists of one sub-space rotation and two steepest descent steps into the direction of the preconditioned residual vectors (Eq. (67)) per band (see also Section 3.8). During this initial phase the Hamiltonian is also kept fixed, after this 'delay' we switch to the RMM scheme and start to update the potential.

As already explained sub-space rotation and sequential update of the bands alternate. In the residual minimization the final vectors are no longer orthogonal. Applying the Rayleigh–Ritz scheme the vectors are correctly orthonormalized. We want to emphasize, that in principle the RMM method would also converge without any explicit sub-space rotation or orthonormalization, but for current system sizes we have found that the sub-space rotation speeds up the calculations although it is an order $O(N^3)$ operation (see Section 3.1). The main problem is that the 'barrier' in the norm of the residual vector between two neighboring eigenvectors with eigenvalues ϵ and $\epsilon + \delta\epsilon$ is only of the order $\delta\epsilon$ [53]. Therefore two eigenvectors which are close in energy are lying in one long steep valley and only a shallow hill spearates them – a typical example of a badly conditioned minimization problem. The sub-space rotation solves this problem because after the rotation the residual vectors are orthonormal to the current trial set (see explanation following Eq. (60)), and search vectors parallel to the long valleys are effectively suppressed.

3.8. The complete algorithm

The complete selfconsistency loop consists of several steps (the section where the algorithm has been discussed is given in brackets, also see Fig. 3):

- sub-space rotation (3),
- DAV2 (3.2), CG (3.6) or RMM (3.7) algorithm,
- orthonormalization using Gram–Schmidt method (only required for the RMM scheme),
- update of partial occupancies and charge density for a selfconsistent calculation.

The initial trial set $\{|\phi_n\rangle, n=1,\ldots, N_b\}$ in each iteration is equivalent to the final set of the previous iteration, initialization is usually done with a random number generator. This loop is repeated until selfconsistency is reached, for a non-selfconsistent calculation no charge density update is done. The orthonormalization is only necessary in conjunction with the RMM, in addition the DAV2 method requires no sub-space rotation.

We have found that the sub-space rotation should be performed between the update of the charge density and the RMM or CG algorithm, especially at the beginning of a selfconsistent calculation. In this case the calculated residual vectors $|R(\phi_m)\rangle$ agree with the exact gradients $|g(\phi_m)\rangle$. For this reason and because the wavefunctions should be orthonormal for a recalculation of the charge density, it is necessary to separate the orthonormalization and the diagonalization of the sub-space Hamiltonian, which is done at once in the Rayleigh–Ritz scheme.

In addition it is necessary to find an optimal break condition for the sequential RMM and CG algorithms. A static criterion, for example 2 steps per band, is not a good choice, because lower bands converge much faster than higher bands. Therefore, we have adopted the following dynamic criterion (which is inspired by Ref. [54]): (i) Both algorithms are stopped if the change in the total eigenvalue becomes smaller than $E_{\text{accuracy}}/N_b/4$, where E_{accuracy} is the required accuracy of the calculation and N_b is the number of occupied bands. (ii) The RMM is stopped if the square of the norm of the residual vector gets smaller than 30% of its initial value, and the minimization always stops with the trial step. (iii) The CG is stopped if the change in the eigenvector gets smaller than 30% of the change in the first i.e. the steepest descent step. (iv) The maximum number of steps is always four. For the RMM the residual vector is minimized three times and at last a fourth trial step is performed. (v) Empty bands are optimized only twice.

By now, these criteria have been used for a large number of system and are very robust. In most cases two CG or two RMM steps are done per band, but problematic eigenvalue/eigenvector pairs are iterated more frequently. Usually more iterations are done for the higher bands, and the total speed of convergence for all bands is very good.

3.9. Computational considerations

To make a fair comparison of different techniques it is necessary to count the number of operations for each algorithm carefully. The CG minimization of the Rayleigh quotient requires always slightly less evaluations of the Hamiltonian multiplied with a wavefunction than the RMM, but for large systems the most expensive part is the orthonormalization of the wavefunctions. For our implementation the evaluation of $(\boldsymbol{H}-\epsilon_n\boldsymbol{S})|\phi_n\rangle$ is an order

$$T^{\mathrm{H}} = N_b N_{\mathrm{plw}} \ln N_{\mathrm{plw}} \propto N^2 \ln N \tag{72}$$

operation, where N qualifies the system size. The limiting factors are the fast Fourier transformations ($N_b N_{\mathrm{plw}} \ln N_{\mathrm{plw}} \propto N^2 \ln N$) and the evaluation of the nonlocal projection operators. For large systems we calculate the non-local projection operators in real space [55] and therefore the number of operations per band increases linearly with the system size (CN_{ions}), for all bands this is only an order N^2 operation. The Gram–Schmidt orthonormalization takes

$$T^{\mathrm{GS}} = N_b^2 \times N_{\mathrm{plw}} \propto N^3 \tag{73}$$

steps, whereas the explicit orthogonalization of the gradients of each band to all other bands in Eq. (62) takes twice as many steps

$$T^{\text{ort}} = 2N_{\text{b}}^2 \times N_{\text{plw}} \propto 2N^3. \quad (74)$$

But even worse, the explicit orthogonalization makes any efficient memory caching impossible. The CG algorithm is strictly sequential and at each iteration the new gradient must be orthogonalized to all other bands, requiring a large band width from the main memory. We found that this is a problem on some machines like the Silicon Graphics Power Challenge architecture where several processors share a large main memory (on vector processors this operation is generally reasonably fast). For the Gram–Schmidt orthonormalization a routine with good data locality which avoids this problem can be found easily and T^{ort} is therefore usually 3–10 times larger than T^{GS} on scalar machines. Efficient routines with good data locality can also be found for the sub-space rotation, and the number of operations is

$$T^{\text{diag}} = T^{\text{H}} + 2N_{\text{b}}^2 \times N_{\text{plw}}. \quad (75)$$

For the blocked Davidson scheme the number of operations is

$$T^{\text{dav}} = 2T^{\text{H}} + 5\tfrac{1}{2}N_{\text{b}}^2 \times N_{\text{plw}} \quad (76)$$

for the first iteration and

$$T^{\text{dav}} = 1T^{\text{H}} + 4N_{\text{b}}^2 \times N_{\text{plw}} \quad (77)$$

for all further iterations if the potential is fixed. As we will show in Section 6, two consecutive blocked Davidson steps (DAV2) are necessary to get a convergence speed that is comparable with the CG or RMM band-by-band methods. For large systems, where the orthogonalization is the leading factor, one blocked Davidson step (with only a single sweep over all bands) takes *more* time (and converges much slower) than one RMM-step (with two optimizations per band, one sub-space rotation and one Gram–Schmidt orthonormalization this is an order $3T^{\text{H}} + 3N_{\text{b}}^2 \times N_{\text{plw}}$ operation, to be compared with Eq. (76)). In addition we have found that two blocked Davidson steps for a fixed potential (DAV2) take generally more time than one CG sweep over all wavefunctions (approximately $3T^{\text{H}} + 5N_{\text{b}}^2 \times N_{\text{plw}}$ operations, to be compared with the sum of Eqs. (76) and (77)). For a comparison of the number of iterations required for each algorithm you might go to Section 6.1.

4. Charge density mixing

The second key step within our algorithm is an efficient mixing of the input and output charge densities. We have adopted the modified Broyden method proposed by Johnson [56]. This approach is flexible and for a special parameter setting the charge density mixing schemes of Pulay [24] and that proposed by Srivastava [57] and Blügel [58] are obtained. To improve the convergence further, we have adopted a special initial mixing matrix and a metric, which are both optimized for a plane wave basis set. In the next sections we will briefly discuss simple mixing, Pulay's and Johnson's approaches.

4.1. Simple mixing

The central quantity of all charge density mixing schemes is the charge density residual $R[\rho_{\text{in}}]$ (see Eq. (31))

$$R[\rho_{\text{in}}] = \rho_{\text{out}}[\rho_{\text{in}}] - \rho_{\text{in}}. \quad (78)$$

The norm of the residual vector

$$\langle R[\rho_{\text{in}}] | R[\rho_{\text{in}}] \rangle \quad (79)$$

must be zero for selfconsistency. Simple schemes take into account only information from the current iteration. Linear mixing for example adds a certain amount of R to the current input charge density

$$\rho_{\text{in}}^{m+1} = \rho_{\text{in}}^m + \gamma R[\rho_{\text{in}}^m]. \quad (80)$$

As in the case of the iterative matrix diagonalization (see Section 3.1), it is a good idea to improve the simple mixing by preconditioning the residual vector using knowledge about the Jacobian matrix. In this case the mixing equation is

$$\rho_{\text{in}}^{m+1} = \rho_{\text{in}}^m + G^1 R[\rho_{\text{in}}^m] \quad (81)$$

where $\mathbf{G}^1$ is a special preconditioning matrix. A simple but efficient scheme for a plane-wave basis set was proposed by Kerker [60], and we used this scheme with some success for the first calculations.

In the Kerker scheme the preconditioning matrix is diagonal in reciprocal space and given by

$$G^1_q = A\frac{q^2}{q^2+q_0^2}. \tag{82}$$

This scheme has the advantage of damping the oscillations in the low-q components of the charge density i.e. for small wave vectors the function behaves like Aq^2/q_0^2 and mixes only a small amount of the output charge density to the input charge density. For large wave vectors q, a simple linear mixing with the linear mixing parameter A is done. Generally A can be quite large and we found that $A = 0.8$ is always an acceptable choice, q_0 might be optimized for the actual system.

4.2. Pulay mixing

In the scheme of Pulay [24] the input charge density and the residual vectors are stored for a number of mixing steps. A new optimal input charge density is obtained in each step as a linear combination of the input charge densities of all previous steps

$$\rho_{\text{in}}^{\text{opt}} = \sum_i \alpha_i \rho_{\text{in}}^i. \tag{83}$$

Assuming linearity of the residual vector with respect to the input charge density ρ_{in}, the residual at $\rho_{\text{in}}^{\text{opt}}$ is given by

$$R[\rho_{\text{in}}^{\text{opt}}] = R\left[\sum_i \alpha_i \rho_{\text{in}}^i\right] = \sum_i \alpha_i R[\rho_{\text{in}}^i]. \tag{84}$$

The optimal new charge density must minimize the norm of the residual vector

$$\langle R[\rho_{\text{in}}^{\text{opt}}] | R[\rho_{\text{in}}^{\text{opt}}]\rangle \tag{85}$$

with respect to α_i under the constraint

$$\sum_i \alpha_i = 1, \tag{86}$$

which conserves the number of electrons. These equations are very similar to those given in Section 3.7, only the functional form of the constraint has changed. The optimal α_i is now given by

$$\alpha_i = \frac{\sum_j A_{ji}^{-1}}{\sum_{kj} A_{kj}^{-1}} \quad \text{with } A_{ij} = \langle R[\rho_{\text{in}}^j] | R[\rho_{\text{in}}^i]\rangle. \tag{87}$$

To improve the numerical stability and for of comparison with the formulas given in the next section it is convenient to transform for iteration m to a new set of independent variables defined by

$$\rho^m = \rho_{\text{in}}^m, \quad \Delta\rho^i = \rho_{\text{in}}^{i+1} - \rho_{\text{in}}^i,$$
$$R^m = R[\rho_{\text{in}}^m], \quad \Delta R^i = R[\rho_{\text{in}}^{i+1}] - R[\rho_{\text{in}}^i] \tag{88}$$

for $i < m$. The new optimal input charge density is then a linear combination

$$\rho_{\text{in}}^{\text{opt}} = \rho^m + \sum_{i=1}^{m-1} \bar{\alpha}_i \Delta\rho^i. \tag{89}$$

An one-to-one relationship between α_i and $\bar{\alpha}_i$ exists and it is evident that the transformation makes a constraint on $\bar{\alpha}_i$ unnecessary. It is easy to show that $\bar{\alpha}_i$ is given by

$$\bar{\alpha}_i = -\sum_{j=1}^{m-1} \bar{A}_{ji}^{-1}\langle \Delta R^j | R^m\rangle, \tag{90}$$

with

$$\bar{A}_{ij} = \langle \Delta R^j | \Delta R^i\rangle. \tag{91}$$

The charge density in the next step might be obtained via the equation

$$\begin{aligned}\rho_{\text{in}}^{m+1} &= \rho_{\text{in}}^{\text{opt}} + G^1 R[\rho_{\text{in}}^{\text{opt}}] \\ &= \rho^m + G^1 R^m + \sum_{i=1}^{m-1} \bar{\alpha}_i(\Delta\rho^i + G^1 \Delta R^i),\end{aligned} \tag{92}$$

where G^1 can be a constant corresponding to simple mixing or a matrix like that one given in Eq. (82).

4.3. Broyden mixing

Among the most sophisticated procedures to calculate the selfconsistent solution of the KS equations are the quasi-Newton algorithms proposed by Broyden [23]. These algorithms try to find an approximation for the Jacobian or the inverse Jacobian matrix by updating the Jacobian matrix at each iteration. Storing the full $N \times N$ Jacobian matrix is rarely possible for large selfconsistency problems, but in the last few years several authors were able to derive modified algorithms which require only the storage of a few N-dimensional vectors at each iteration; Srivastava [57] derived an algorithm for Broyden's second method (inverse Jacobian update) and similar results were obtained by Blügel for Broyden's first (Jacobian update) and second method [58]. Another important contribution goes back to Vanderbilt and Louie [59], who suggested a new more flexible version of Broyden's method, which avoids that information obtained in previous steps is lost during the update of the Jacobian matrix. Johnson [56] reformulated this method so that it requires only the storage of N-dimensional vectors. Here we will mainly concentrate on this approach because it is flexible and allows to obtain Blügel's and Pulay's methods for a special set of parameters.

The key point of quasi-Newton methods is the assumption that the residual vector can be linearized near the minimum,

$$R[\rho] = R[\rho_{\text{in}}^m] - \boldsymbol{J}^m(\rho - \rho_{\text{in}}^m), \tag{93}$$

where $\boldsymbol{J}^m$ is an approximation of the Jacobian matrix. If we require $R[\rho^*] = 0$ we obtain an optimal charge density ρ^* which makes the residual vector zero:

$$\rho^* = \rho_{\text{in}}^m + (\boldsymbol{J}^m)^{-1} R[\rho_{\text{in}}^m]. \tag{94}$$

In successive steps an improved approximation of the Jacobian matrix $\boldsymbol{J}^m$ or of the inverse Jacobian matrix $(\boldsymbol{J}^m)^{-1}$ is build up, and a new charge density is obtained from the current approximation of the inverse Jacobian matrix, the current charge density ρ_{in}^m and the current residual vector $R[\rho_{\text{in}}^m]$ using the equation

$$\rho_{\text{in}}^{m+1} = \rho_{\text{in}}^m + (\boldsymbol{J}^m)^{-1} R[\rho_{\text{in}}^m]. \tag{95}$$

The algorithms differ in the way how $\boldsymbol{J}^m$ is changed and updated in each iteration. To comply with the notation used by Johnson [56] we define $\boldsymbol{G}^m = (\boldsymbol{J}^m)^{-1}$. Johnson suggested a scheme in which information of all previous iterations is taken into account to calculate $\boldsymbol{G}^m$ for the current iteration. For iteration m this is done via a least square minimization of an error function

$$E = w_0 \|\boldsymbol{G}^{m+1} - \boldsymbol{G}^m\|^2 + \sum_{i=1}^{m} w_i \|\Delta\rho^i + \boldsymbol{G}^{m+1}\Delta R^i\|^2, \tag{96}$$

where $\|A\|^2 = \langle A|A\rangle$, and $\Delta\rho^i$ and ΔR^i were defined in the previous section in Eq. (88), and the w_i are weighting factors (see below). The definition of this error function can be understood easily in terms of the following arguments: (i) the first term corresponds to the requirement that the approximation for the inverse Jacobian matrix should not change too much between each iteration. Actually it turns out that this constraint is relatively unimportant and after obtaining the final formula we will concentrate on the case $w_0 \to 0$. (ii) The second term requires that the norm of

$$\Delta\rho^i + \boldsymbol{G}^{m+1}\Delta R^i \tag{97}$$

is as small as possible. If $R[\rho]$ is linear with respect to ρ and for the exact inverse Jacobian matrix $\boldsymbol{G}^{m+1} = \boldsymbol{G}^{\text{exact}}$ this quantity would be zero (compare with Eq. (94)).

Starting from Eq. (96) it is possible to derive an exact solution for $\boldsymbol{G}^{m+1}$. Because Ref. [56] contains a relatively large number of misprints we will give the final correct formulas once again:

$$\boldsymbol{G}^{m+1} = \boldsymbol{G}^1 - \sum_{k=1}^{m} |Z_k^m\rangle\langle\Delta R^k| \tag{98}$$

where

$$|Z_k^m\rangle = \sum_{n=1}^{m} \beta_{kn} w_k w_n |u^n\rangle + \sum_{n=1}^{m-1} \bar{\beta}_{kn} |Z_n^{m-1}\rangle \tag{99}$$

and

$$|u^n\rangle = \boldsymbol{G}^1|\Delta R^n\rangle + |\Delta\rho^n\rangle. \tag{100}$$

β_{kn} and $\bar{\beta}_{kn}$ are given by

$$\beta_{kn} = \left(w_0^2 + \bar{A}\right)_{kn}^{-1}, \qquad \bar{A}_{kn} = w_k w_n \langle\Delta R^n|\Delta R^k\rangle \tag{101}$$

and

$$\bar{\beta}_{kn} = \delta_{kn} - \sum_{j=1}^{m} w_k w_j \beta_{kj} \langle\Delta R^n|\Delta R^j\rangle. \tag{102}$$

If all iteration weights w_n are the same the equality $\bar{\beta}_{kn} = w_0^2 \beta_{kn}$ holds and the equations given in Ref. [56] are obtained (maybe this case was implicitly assumed in Ref. [56]).

It is now easy to show that Pulay's scheme can be obtained by evaluating the equations given above for $w_0 \to 0$ and $w_0 \ll w_n$. Interestingly, for the case $w_0 \to 0$ the choice of w_n does not influence G^{m+1} at all, which can be seen by showing that $w_k w_n \beta_{kn}$ is invariant under a change of an arbitrary weight w_n. Without loss of generality we therefore set w_n to 1 and obtain for the inverse Jacobian

$$G^m = G^1 - \sum_{k,n=1}^{m-1} \beta_{kn} |u^n\rangle\langle \Delta R^k|. \tag{103}$$

Some straightforward manipulation gives for the new input charge density $\rho_{\text{in}}^{m+1} = \rho_{\text{in}}^m + G^m R[\rho_{\text{in}}^m]$ (see Eq. (95)) exactly the same result as in Eq. (92). It is also possible to show that the inverse Jacobian obtained in this way makes Eq. (97) exactly zero for any $i < m$, therefore G^m might be considered as the best approximation of the exact inverse Jacobian matrix in the space searched up to now.

As a second case it is possible to derive Broyden's second method from the equations given above by setting $w_i = 0$ for $i < m$ and requiring $w_0 \ll w_m$. In this case the update equation is simply

$$|Z_k^m\rangle = |Z_k^{m-1}\rangle \quad \text{for } k < m \tag{104}$$

and

$$|Z_m^m\rangle = \frac{1}{\|\Delta R^m\|^2} \left(|u^m\rangle - \sum_{k=1}^{m-1} \langle \Delta R^k | \Delta R^m \rangle |Z_k^{m-1}\rangle \right) \tag{105}$$

in agreement with the formulas given by Blügel [58]. In Broyden's second method information of the current iteration is allowed to overwrite information of all previous iterations and Eq. (97) is zero only for the last iteration $i = m$.

We have found, that Broyden's second method is always slower for the charge density mixing than Pulay's method. The only problem for Pulay's method might be that the linear dependencies between consecutive search directions are too strong. In the context of charge density mixing this does not seem to happen, but we have also tried to use Pulay's and Broyden's second method in conjunction with the relaxation of the ionic degrees of freedom. For configurations with a small number of degrees of freedom linear dependencies between the forces for different positions develop and Pulay's method gets unstable. Broyden's second method seems to be more favorable in this case. For the ionic relaxation, another convenient choice is to take into account only information from a fixed small number n of previous steps (i.e. $w_k = 0$ for $k < m - n$, and $w_k \gg w_0$ for $m - n \leq k < m$).

At last we want to consider the case $w_0 \approx w_n$: This choice restricts changes in G between two iterations and we have found that this destroys most of the advantages of Broyden's scheme; the update of G does not work as expected. In this case G^1 must be close to the correct inverse Jacobian matrix for a reasonable convergence. In the spirit of the arguments given above it is also evident that a dynamic choice of w_n as proposed by Johnson is usually not applicable. Useful settings are only $w_n = 0$ or $w_n \gg w_0$, and we have already shown that for $w_n \gg w_0$ the actual choice of w_n does not influence G at all.

4.4. Preconditioning and metric

Two questions remain, first the choice of the initial matrix G^1, and second whether an optimized metric for evaluating the scalar products $\langle \cdot | \cdot \rangle$ can be found.

The initial mixing plays only a minor role, but for convenience we use the Kerker matrix G^1 (Eq. (82)) because it gives already good convergence in the first few steps. As we will show in Section 6.2.2, the technique is rather insensitive to the choice of the parameters for the initial mixing, and there is no need to optimize the parameters for different systems: $A = 0.8$ and $q_0 = 1.5$ Å^{-1} is always satisfactory. For magnetic systems and for some surfaces an initial linear mixing with $A = 0.1$ was convenient.

Second, a reasonable metric can help to reduce the number of iterations. We have found that the inclusion of a weighting factor

$$f_q = \frac{q^2 + q_1^2}{q^2} \tag{106}$$

in the evaluation of the scalar products

$$\langle A|B\rangle = \sum_q f_q A_q^* B_q \tag{107}$$

improves the results considerably for complex metallic systems. This function is inspired by the fact that the contributions for small wave vectors are more important than contributions for large wave vectors. The choice of q_1 is relatively unimportant and we set q_1 in a way that the shortest wave vector is weighted 20 times stronger than the longest wave vector. At this point, we also want to make clear that a considerable difference between charge density mixing and potential mixing exists. Taking into account only the Hartree term the potential is given by

$$V(q) \propto \frac{1}{q^2}\rho(q),$$

therefore the metric for the evaluation of scalar products differs by a factor of $1/q^4$ in both cases.

Third, we are frequently confronted with very large systems with FFT grids containing up to 64 × 64 × 64 points, which are necessary to describe the rather hard augmentation charges of transition metals. These large meshes exceed the storage possibilities even for the new efficient mixing schemes. A rather simple solution to this problem exists: We have found that no mixing is necessary for large wave vectors q, i.e. it is possible to set

$$\rho^{m+1}_{\text{in}\,q} = \rho^{m}_{\text{out}\,q} \tag{108}$$

without any loss of efficiency, and only a relatively small number of grid points must be treated with Broyden's method; usually we take all grid points which are also contained in the plane wave basis set $(h^2|q|^2/(2m_e) < E_{\text{cut}})$.

To summarize the results of this section: For the charge density mixing we usually use Pulay's method and we set G^1 to the matrix proposed by Kerker with the parameters $A = 0.8$ and $q_0 = 1.5$ Å^{-1}. For all cases treated up to now these parameters resulted in a very good convergence during the selfconsistent procedure, and optimizing the parameters never improved convergence speed by more than 10%. A comparison of different mixing methods can be found in Section 6.2.

5. Direct minimization of the KS-functional

As an alternative to the SC-iterative methods we also want to discuss briefly the direct minimization of the KS-functional. As in Eqs. (34) and (57) it is convenient to incorporate the orthonormality constraint using Lagrange multipliers. In the most general form, this results in a functional

$$\begin{aligned}\bar F[\{\phi_n\},\{\gamma_{nm}\},\{f_n\},\mu,\{R_N\}] &= F - \sum_{nm}\gamma_{nm}(\langle\phi_n|S|\phi_m\rangle - \delta_{nm}) \\ &\quad - \mu\Big(\sum_n f_n - N_{\text{el}}\Big),\end{aligned} \tag{109}$$

which has to be minimized with respect to all degrees of freedom. The gradient of this functional with respect to the wavefunctions is similar to Eq. (58)

$$\frac{\delta\bar F}{\delta\langle\phi_n|} = |g_n\rangle = f_n H|\phi_n\rangle - \sum_m S\gamma_{nm}|\phi_m\rangle. \tag{110}$$

but for a consistent definition of the gradient, we have to define the Lagrange multipliers in a different way: The gradient should describe energy differences up to first-order

$$\mathrm{d}F = \sum_m \langle\delta\phi_m|g_m\rangle + \text{c.c.}, \tag{111}$$

correctly, but now all bands are allowed to change simultaneously. If an unitary rotation of the wavefunctions $\{\phi_n\}$ is allowed,

$$\langle\phi_n|g_m\rangle + \langle\phi_m|g_n\rangle = 0 \quad \forall m, n, \tag{112}$$

has to be required and this results in

$$\gamma_{nm} = \tfrac{1}{2}\bar H_{nm}(f_n + f_m), \tag{113}$$

with

$$\bar H_{nm} = \langle\phi_m|H|\phi_n\rangle. \tag{114}$$

The explicit gradient is then given by

$$\begin{aligned}|g_n\rangle &= f_n\Big(1 - \sum_m S|\phi_m\rangle\langle\phi_m|\Big)H|\phi_n\rangle \\ &\quad + \sum_m \frac{1}{2}\bar H_{nm}(f_n - f_m)S|\phi_m\rangle.\end{aligned} \tag{115}$$

A similar result might be obtained – maybe in a more elegant way – by a generalization of the KS-functional to nonorthogonal orbitals [13]. The structure of Eq. (115) is very interesting. Clearly the first term describes changes which result from a change of the sub space spanned by the wavefunctions $\{\phi_n\}$ and was already obtained in Eq. (60), whereas the second term is new and corresponds to the energy change arising from an unitary transformation of the wavefunctions $\{\phi_n\}$. At the groundstate the energy change $\mathrm{d}F$ (Eq. (111)) must be zero for arbitrary variations $\delta\phi_m$, and the second term is only zero if the matrices $\overline{H}_{nm}$ and $F_{nm} = f_n \delta_{nm}$ commute. For materials with a gap this can be achieved by generating the eigenstates for the filled orbitals only (i.e. all $f_n = 1$), and the Lagrange multipliers at the groundstate are given by

$$\gamma_{nm} = \overline{H}_{nm} \tag{116}$$

(compare Eq. (11)), whereas for metals with $f_n \neq f_m$ both matrices only commute if $\overline{\mathrm{H}}_{nm}$ is diagonal, clearly demonstrating that the exact Kohn–Sham eigenstates have to be calculated for metals. In this case the Lagrange multipliers at the groundstate are given by

$$\gamma_{nm} = \delta_{nm} \epsilon_n f_n, \tag{117}$$

where ϵ_n are the exact Kohn–Sham eigenvalue. Finally, we want to point out, that the last term in Eq. (115) defines an unitary rotation matrix U,

$$U_{nm} = \delta_{nm} - \Delta \overline{H}_{nm}(f_n - f_m) \tag{118}$$

for small Δ, which might be used to rotate the wavefunction ϕ_n until the sub-space Hamiltonian (Eq. (114)) is diagonal.

5.1. Preconditioned search direction

To find a good search direction it is simplest to treat both terms in Eq. (115) independently. First, a correction vector to each state ϕ_n which changes the basis set $\{\phi_n\}$ has to be calculated. We use the correction vector already successfully applied in the sequential band by band methods (Section 3.4, Eq. (60)) but with a full inclusion of all Lagrange multipliers i.e.

$$|p_m\rangle = K\Big(1 - \sum_m S|\phi_m\rangle\langle\phi_m|\Big)H|\phi_n\rangle. \tag{119}$$

The explicit S orthogonalization of this vector can be avoided, because a Gram–Schmidt orthonormalization is done after updating all bands. Mind that it is very important to have a reasonable length for this correction vector $|p_m\rangle$, because the unitary transformation and the addition of the correction vectors are done at once in the all bands simultaneous scheme. Using the preconditioning function of Eq. (52) the rotation of the wavefunctions and the changes of the basis set are done in a well conditioned way.

Second, an unitary transformation of the wavefunctions ϕ_n has to be found, which makes the sub-space Hamiltonian (114) diagonal. Rotating the wavefunctions into the direction of the steepest descent (second term in Eq. (115) or Eq. (118)) turns out to be extremely slow. Much more efficient is a transformation based on second order Loewdin perturbation theory (this idea was first discussed by Gillan [12] in this context). In this case the rotation matrix is defined as

$$U_{nm} = \delta_{nm} - \Delta \frac{\overline{H}_{nm}}{\overline{H}_{mm} - \overline{H}_{nn}}. \tag{120}$$

For a start configuration far from the electronic groundstate, the matrix elements might become very large and perturbation theory fails, therefore we replace $x = \overline{H}_{nm}/(\overline{H}_{mm} - \overline{H}_{nn})$ by the quantity $\sin(\arctan(2x)/2)$ which is inspired by the exact treatment of a two by two matrix. This unitary matrix is used to rotate the wavefunctions according to the equation

$$\phi_n = \sum_m U_{nm} \phi_m^{\text{old}}. \tag{121}$$

As pointed out by Gillan [12] it might happen during the minimization procedure that the ordering of the partial occupancies is wrong i.e. $\overline{H}_{mm} > \overline{H}_{nn}$ but (incorrectly) $f_m > f_n$, in this case U_{nm} is set to zero to guarantee that the energy decreases along the search direction.

Finally we have to find a consistent update scheme for the partial occupancies. In principle a direct calculation of the gradient vector for f_n is possible if Fermi–Dirac statistics is used, because an explicit functional form for the entropy term $S(f)$ exists for this case (see for instance Ref. [61]). But no analytical form for the entropy term $S(f)$ is available for Gaussian smearing or the MP scheme, and within the tetrahedron method the partial occupancies f_n are no independent degrees of freedom. Therefore we optimize a new independent set of variables $\tilde{\epsilon}_n$ from which the partial occupancies are calculated directly using

$$f_n = f\left(\frac{\tilde{\epsilon}_n - \mu}{\sigma}\right).$$

The gradient vectors for these new variational degrees of freedom can be evaluated analytically, and are given for the smearing methods by

$$\frac{\partial F}{\partial \tilde{\epsilon}_n} = g_n \left((H_{nm} - \tilde{\epsilon}_n) - \frac{\Sigma_m g_m (H_{mm} - \tilde{\epsilon}_m)}{\Sigma_m g_m} \right) \tag{122}$$

with

$$g_n = \frac{\mathrm{d}f((\tilde{\epsilon}_n - \mu)/\sigma)}{\mathrm{d}\tilde{\epsilon}_n}$$

(we recently found that the same approach was used in Ref. [62]). The actual search direction used by us, however, is not this complicated expression for the exact gradient but simply the difference between $\tilde{\epsilon}_n$ and H_{nn}

$$H_{nm} - \tilde{\epsilon}_n. \tag{123}$$

At the groundstate the $\tilde{\epsilon}_n$ will converge to the exact KS eigenvalues, and the partial occupancies are correctly determined.

5.2. Steepest descent and conjugate gradient algorithm

The search direction discussed in the previous section can be used in an all bands simultaneous update scheme. Within the steepest descent algorithm it is no problem to re-orthogonalize the wavefunctions after each step, but we also do the re-orthonormalization within the conjugate gradient scheme. In principle this might slow down the CG algorithm, but we think that the orthonormalization causes only minor problems: It can be shown that the last search direction and the new gradient are orthogonal up to second-order in the trial step and this orthonormality is the most important condition for a stable and fast convergence of the CG routine. For the minimization we are using a standard (preconditioned) CG routine. The conjugated direction is simply given by

$$|f^i\rangle = |p^i\rangle + \frac{\langle p^i|g^i\rangle}{\langle p^{i-1}|g^{i-1}\rangle}|f^{i-1}\rangle, \tag{124}$$

where $|g^i\rangle$ is the gradient and $|p^i\rangle$ is the un-conjugated search direction, consisting of components of Eqs. (119), (120) and (123).

A second serious difficulty within the CG routine is the accuracy of the line minimization: We assume a quadratic behavior of the total energy along the search path, and evaluate the exact energy change for a finite step. Using this information and the expected first-order energy change (which is directly proportional to $\langle f^i|g^i\rangle$, i.e. the product of the conjugated search direction and the gradient), it is possible to determine the minimum. An improved treatment (especially an improved treatment of the orthonormality constraint, see Ref. [13]) might result in a better convergence, but even for our implementation the improvement over a steepest descent approach is considerable. Only for starting configurations far away from the exact groundstate problems might occur (a result of the way the orthonormality constraint and the line minimization are handled), but close to the groundstate the routine works very well.

To avoid any ambiguities we will refer to the algorithms which minimize the KS functional directly as CGa for the conjugate gradient scheme – and SDa for the steepest descent scheme. The CGa scheme should not be mixed up with the sequential CG scheme, discussed in the context of iterative matrix diagonalization (Section 3.6).

5.3. Other direct minimization methods

The standard CP method is also a direct method, and the most efficient version of the CP method was recently discussed by Tassone et al. [7]. In this case a damped and preconditioned second-order equation of motion is used to calculate the electronic groundstate. Tassone found that the second-order equation is far superior to a preconditioned steepest descent approach. In the limit of semiconducting systems (i.e. $N_b = N_{\mathrm{electrons}}/2$) their simple steepest descent approach is almost equivalent to our steepest descent approach. There are some small differences especially in the preconditioning and in the incorporation of the orthonormality constraint, which is done more consistently in Ref. [7]. It is interesting to point out that the damped second-order equation of motion proposed in Ref. [7] is closely related to an acceleration scheme for slowly converging series by Tchebycheff, which has been used for the mixing of charge densities by Akai and Dederichs [63].

Another efficient minimization algorithm, which includes partial occupancies was recently discussed by Grumbach, Hohl, Martin and Car [62], and first proposed by Gillan [12]. But both implementations suffer from the fact that the conjugate gradient algorithm is restricted to the wavefunctions (Eq. (119)). To us it seems that this is an unnecessary restriction: Grumbach et al. point out that they found a slower convergence at the end of the calculation, the energy-drop in the last part of the calculation of carbon was mainly due to the sub-space rotation (Eq. (120), or sub-space mixing using the terminology of Ref. [62]). To improve the convergence they had to perform additional exact sub-space diagonalizations after several standard steps. These problems are related to the simple steepest descent treatment of the subspace part in their work, showing that all degrees of freedom must be updated simultaneously and consistently.

6. Comparison of different techniques

6.1. Iterative matrix diagonalization

In this section we will compare different iterative diagonalization techniques considering the calculation of the eigenvalue spectrum for a fixed Hamiltonian only. The selfconsistent case will be covered in section 6.2. For all calculations the charge density was constructed from the atomic charge density of the constituents, therefore the calculated energy corresponds to the energy of the Harris–Foulkes (HF) functional [39–41] already introduced previously (see Section 2.3). In some cases like liquid Germanium (and most liquid metals) the calculation of the eigenvalues and eigenvectors of the Hamiltonian turns out to be the only limiting factor for a selfconsistent calculation, in other cases like hydrogen on a palladium surface the charge sloshing is so pronounced that the mixing procedure determines the overall performance.

6.1.1. Liquid metallic system

As a prototype for a liquid metallic system we consider Germanium at a temperature of $T = 1250$ K. This choice was influenced by two facts. First, we have done an extensive study of liquid and amorphous Germanium [17], and we have reasonable models for the liquid structure. Second, we want to compare the convergence of our techniques with the results for liquid Silicon published recently by Tassone et al. [7] and Grumbach et al. [62].

For the calculation we used a 64 atoms ensemble. The cutoff was 160 eV, Gaussian smearing with $\sigma = 0.2$ eV, the Γ point only and 148 bands were used for the calculation (20 bands more than necessary to hold all electrons). The initial electronic configuration was calculated doing a random initialization and 2 CG steps on the wavefunctions (i.e. 2 sub-space rotations, 2 CG sweeps over all bands, 2 optimizations of each wavefunction per sweep). This initial choice is mainly influenced by the fact, that the RMM scheme requires a reasonable electronic start configuration, because the algorithm traps the eigenvector closest to the initial trial vector. Results for the non-selfconsistent calculation are reported in Fig. 5. It is evident that the CG algorithm gives the best convergence, but the RMM makes up for this fact by requiring less time per step. Overall the performance is approximately the same on most processors, but the RMM scheme is much faster on machines with slow memory subsystems (like Silicon Graphics Power Challenge architecture or DEC Alpha machines). In the CG and RMM approximately two steps per band are made in each sweep. In addition one sub-space diagonalization is performed per step.

The blocked Davidson scheme results in the slowest convergence. One step in the plot corresponds to expanding the expansion set to $2N_b$ and collapsing it back to $1N_b$ twice (DAV2) i.e. one step in the plot corresponds to two simple blocked Davidson steps (see also Section 3.2). The convergence of the Davidson scheme would improve enormously by expanding the expansion set to $3N_b$ and collapsing

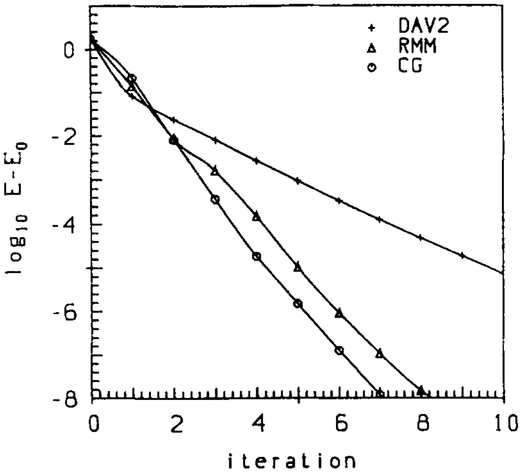

Fig. 5. Convergence of the total free energy per atom E (in eV) for different algorithms for 1 Ge (64 atoms), non-selfconsistent case. DAV2 corresponds to the blocked Davidson scheme, one step in the plot is equivalent to expanding the expansion set to $2N_b$ and collapsing it back to $1N_b$ twice, RMM it the sequential residual vector minimization, CG the sequential conjugate gradient scheme.

then to $1N_b$, but this requires additional computer memory and additional order N^3 operations, making the Davidson scheme even less favorable. In terms of computing time one DAV2 step takes approximately twice the time of one RMM step, making the Davidson scheme the slowest scheme for liquid Ge.

6.1.2. *Metallic surface*

As a second test system we consider the Pd (111) surface with a monolayer hydrogen located in the hollows between three Pd atoms. The supercell consists of 5 layers vacuum and 5 Pd atoms in a (1 × 1) unit cell, the geometry is not relaxed. The cutoff was 200 eV. Once again Gaussian smearing with $\sigma = 0.2$ eV, and a 5 × 5 × 1 Monkhorst Pack k point grid was used. 32 bands were included in the calculation (6 bands more than necessary to hold all electrons). The system is relatively small, but even for this small system the charge sloshing is extremely strong because the Fermi level lies at a rather steep point of the electronic density of states. We have also done calculations for larger systems containing up to 32 atoms, but the main points concerning the convergence are already captured by this small system.

Here we consider the non-selfconsistent case only, results are shown in Fig. 6. In this case the initial electronic configuration was calculated doing a random initialization and 3 CG steps on the wavefunctions. As previously, for the method DAV2 one step in the plot corresponds to 2 blocked Davidson steps. It is evident that the CG method is again most efficient, it is also better in terms of computer time, but the gain in comparison to the RMM and DAV method is modest (20%). For this calculation the performance and the time per step of CG, DAV2 and RMM are almost the same. This arises from the fact that order N^3 operations play only a minor role for this small system, and the number of evaluations of $(H - \epsilon_n S)|\phi_n\rangle$ is exactly 3 for DAV2 and using the dynamic break criterion approximately 3 for the RMM or CG scheme (sub-space rotation included). For a large system with more atoms per layer the RMM scheme is the fastest scheme and outperforms the DAV2 and CG schemes, we have verified this for calculations containing up to 32 atoms in the supercell.

6.1.3. *Diamond surface*

The test system is a clean C (100)(1 × 2) surface, modelled by a slab geometry containing 16 atoms (8 C-layers with 2 C-atoms per layer and 8 layers

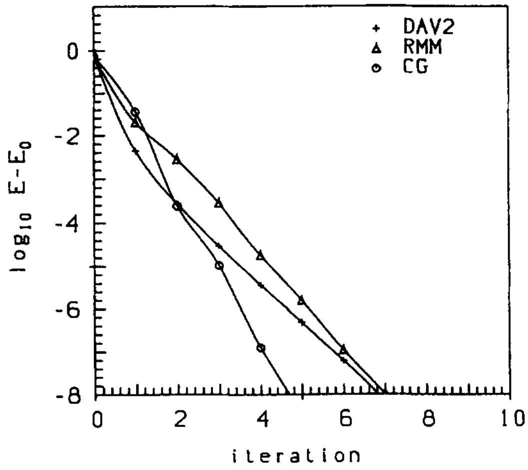

Fig. 6. Convergence of the total free energy per atom E (in eV) for the Pd(111) surface with a monolayer hydrogen located in the hollows between three Pd atoms, non-selfconsistent case. A slab consisting of 5 layers of Pd is used to model the surface.

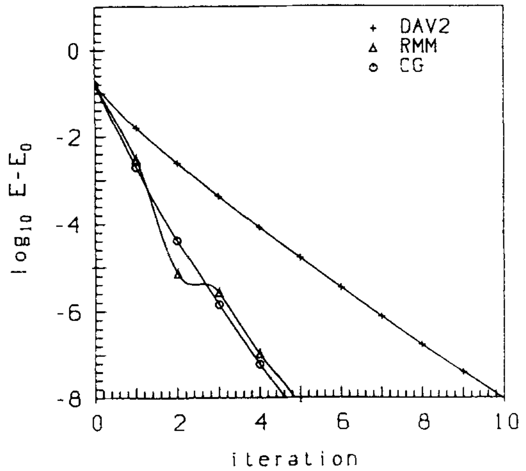

Fig. 7. Convergence of the total free energy per atom E (in eV) for a clean C(100)(1 × 2) surface, modeled by a slab containing 16 atoms, non-selfconsistent case.

vacuum). One side of the slab is reconstructed, resulting in a dimerization of the C atoms in the top layer, the other side is unreconstructed and possesses a single metallic surface band. Despite this metallic behavior, the convergence of the partial occupancies is fast and unproblematic, because well defined gaps separate the metallic surface band from the other bands (in contrast to the Pd surface).

Gaussian smearing with $\sigma = 0.2$ eV, and a 3 × 6 × 1 Monkhorst Pack k-point grid was used; 40 bands were included in the calculation (8 bands more than necessary to hold all electrons). The system is also only medium sized, but it captures all main points which are important for larger systems (calculations for systems containing up to 64 atoms have been done). The charge sloshing is not as important as for the Pd surface, but it is still considerable.

Results for the non-selfconsistent case are shown in Fig. 7. The initialization and the methods are the same as for l-Ge. The RMM and the CG scheme require the same number of steps, which is typical for semiconducting and insulating systems. Because orthonormalization (order N^3 operation) plays only a minor role the time per step is also the same for this system size. Going to large systems, where order N^3 operations are important, the RMM scheme becomes favorable. The DAV2 scheme is once again the slowest scheme, although the time per step for this small system is comparable to the RMM and CG scheme. Summarizing the results of the last three sections: We have found that

- the CG algorithm is fastest for very small systems, where order N^3 operations are negligible;
- the RMM algorithm is superior for large systems containing more than 20–30 atoms;
- and the DAV2 scheme is always outperformed by one of the other two techniques.

6.2. Comparison for selfconsistent calculations

6.2.1. Liquid metallic system

The convergence for the selfconsistent calculation of liquid Ge is mainly determined by the convergence of the iterative matrix diagonalization. Charge sloshing is negligible in this simple system. This is demonstrated in Fig. 8, where the convergence is compared for different matrix diagonalization schemes. In all cases we used the same initial wavefunctions as in the non-selfconsistent calculation, and the initial charge density is calculated from the charge density of the Ge atoms. It is evident that Fig. 8 shows the same behavior as Fig. 5. The mixing parameters are not important for this system, and therefore we used only the default parameters. Comparing Figs. 5 and 8 it might be seen that actually only one or two additional iterations are necessary

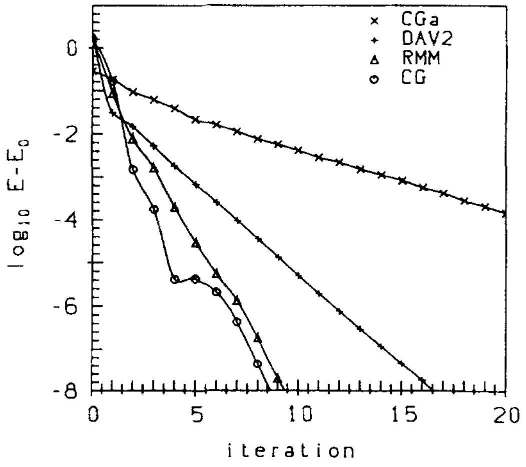

Fig. 8. Convergence of the total free energy per atom E (in eV) for different algorithms for l-Ge (61 atoms), selfconsistent case. RMM, CG and DAV2 are algorithms relying on the selfconsistency cycle. CGa is the conjugate gradient algorithm applied directly to the KS functional for 148 bands.

Table 3
Time necessary to do one iteration for a l-Ge ensemble containing 64 atoms for several algorithms on an IBM RS 6000/Model 590

	Complex	Real
DAV2	180 s	
CG sequent.	≈ 155 s	≈ 70 s
RMM sequent.	≈ 102 s	≈ 42 s
CGa-148	91 s	37 s
SDa-148	64 s	30 s
CGa-128	65 s	29 s
SDa-128	50 s	23 s

'Complex' is the timing for a full complex code, 'Real' for a code which takes into account that $C_q = C^*_{-q}$ if the Γ-point only is used for the k-point sampling. CG refers to the conjugate gradient band-by-band scheme with mixing, RMM refers to the residual minimization band-by-band scheme also with mixing, CGa and SDa are the conjugate gradient and steepest descent algorithms applied directly to the KS-functional for 128 or 148 bands. For the SDa method timing for a non-optimized trial step is given (this step is comparable to a Car-Parrinello step).

for the selfconsistent case. We found a similar behavior in most liquid, amorphous and bulk systems.

In Fig. 8 the convergence for the CGa routine, which minimizes the KS functional directly and updates all bands simultaneously, is also shown. To have approximately the same initial error in the start configuration the initial electronic configuration was calculated doing a random initialization and 3 sequential CG steps on the wavefunctions for a fixed charge density. One step in the CGa routine takes approximately the same time as one RMM step (see Table 3). The total convergence is reasonable but still slower by a factor 3–5 than for the RMM routine.

In Fig. 9 we compare the CGa approach with a simple steepest descent (SDa) approach. In the SDa-148 approach (148 corresponds to the number of bands included) the size of the final step was optimized, although an un-optimized step gives almost the same convergence (compare the two lines given for SDa-128). In addition we also show the results for the case of treating l-Ge as an insulator, i.e. including only $N_b = N_{\text{electron}}/2$ bands in the calculation (SDa-128, CGa-128). Of course the reduced number of bands also results in a reduced computational time. One steepest descent step without optimization of the trial step for 128 bands (this step is similar to a Car–Parrinello step) takes approximately half the time of a CGa or RMM step for 148 bands. Timings for all cases treated here are also compiled in Table 3. The additional computational time arising from the metallic treatment is more than made up by the improved convergence.

The results of this section can be compared with results recently published by Tassone et al. [7] and Grumbach et al. [62]. Tassone used two different algorithms a preconditioned steepest-descent (SD P) and a preconditioned damped second-order (D P) algorithm. The start configuration also differs from our calculation, but the total error in the energy at the beginning is comparable to our case. Our SDa-128 (without optimization of the step-size) shows almost the same convergence as the SD P scheme in Ref. [7], indicating that both algorithms are comparable. The CGa-128 scheme requires approximately half the number of steps as the D P scheme, but one CGa-128 step is probably more expensive, leading to a similar efficiency in terms of computing time. It is clear that the CGa-148 and especially the RMM scheme are much faster than any of the schemes discussed in Ref. [7]. Even if we take into account that one step in the methods relying on a selfconsistency cycle (RMM scheme) takes twice as long as

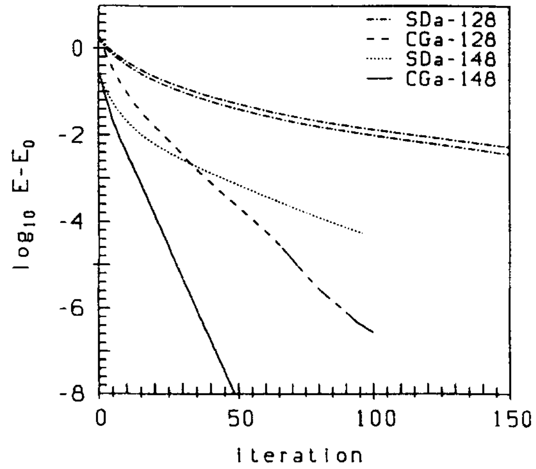

Fig. 9. Convergence of the total free energy per atom E (in eV) for different algorithms for l-Ge (64 atoms), selfconsistent case. 'CGa' and 'SDa' are the conjugate gradient and steepest descent algorithms applied directly to the KS-functional for 128 or 148 bands. For 'SDa-128' two lines are shown, for the lower line the step size was optimized at each iteration, for the upper line a fixed step size with $\Delta s = 0.3$ was used.

one D P step in Ref. [7] our algorithms should be faster by more than a factor 10. Similar conclusion might be drawn from a comparison with the technique proposed by Grumbach et al. [62], once again the algorithm based on the selfconsistency cycle outperforms other algorithms by more than a factor of 10.

We want to point out that the situation is different for insulating systems: The performance of the schemes using the selfconsistency cycle is almost the same for metallic and insulating systems and generally the standard algorithms (like the D P algorithm discussed by Tassone et al.) are fast and efficient for those systems. Therefore, the algorithm proposed in Ref. [7] should be comparable – maybe slower by a factor 2 – to our algorithms for insulating systems.

6.2.2. Metallic surface

The Pd(111) surface with hydrogen is the system with the strongest charge sloshing we have encountered up to now. The charge sloshing is the only limiting factor for the selfconsistent calculation and the convergence does not depend on the fact whether the CG, RMM or DAV2 algorithm is used for the calculation of the electronic eigenstates. In Fig. 10 results for the Pulay mixing are shown. In all cases we used the same initial wavefunctions as in the non-selfconsistent calculation, and the initial charge density is calculated from the charge density of the atomic constituents. The initial matrix G^1 was set to Kerker's mixing matrix with $q_0 = 1.5$ Å^{-1}, $A = 0.8$ (the default value), and to a linear mixing with $A = 0.1$. We have found that Pulay's scheme is robust with respect to changes in the initial mixing matrix. The value of q_0 may vary between 0.5 Å^{-1} and 2.0 Å^{-1} without a significant influence, and a simple linear mixing $G^1 = A$ with $A = 0.05$–0.5 works equally well. Nevertheless, it is very important to make A not too small in Pulay's method, because this might slow the convergence considerably ($A < 0.05$). The inclusion of a metric in the evaluation of the scalar products was found to be very important for this system, and in addition Broyden's second method was less stable than Pulay's algorithm (see Fig. 11).

Finally we have also tested the CGa scheme for this system and found a reasonable convergence behavior. To have a reasonable small error in the

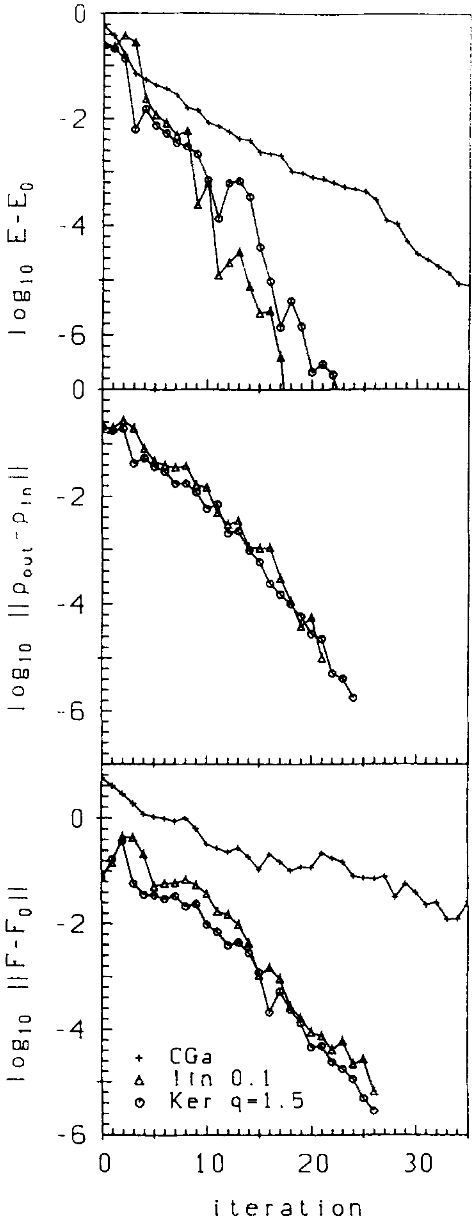

Fig. 10. Convergence for the Pd(111) surface, selfconsistent case. Top panel total free energy per atom E (in eV), second panel charge density residual vector, last panel forces (in eV/Å/atom). For the calculations of the charge density residual vector no metric was included (see text). Pulay's method was used for the mixing. For 'Ker 1.5' the initial matrix G^1 was set to Kerker's mixing matrix with $q_0 = 1.5$ Å^{-1}, and for 'lin 0.1' to a linear mixing $G^1 = 0.1$. Results for the conjugate gradient approach applied directly to the KS-functional (CGa) are also shown.

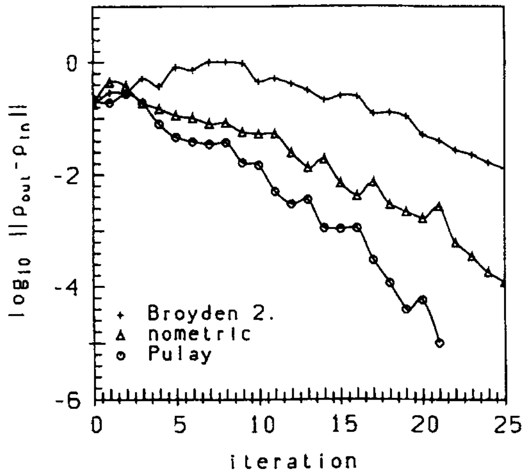

Fig. 11. Convergence of the charge density residual vector for different mixing algorithms for the Pd(111) surface. The initial matrix G^1 was set to $G^1 = 0.1$. For 'nometric' Pulay's method without metric was used, 'Pulay' corresponds to Pulay's method with metric, and 'Broyden 2.' to Broyden's second method with metric.

initial configuration we used the converged eigenfunctions of a non-selfconsistent calculation as the start configuration.

Especially interesting is the convergence of the forces. For the SC-methods the Harris Foulkes functional is used for the evaluation of the energy, whereas the exact KS energy is evaluated for the CGa scheme. Therefore it might be possible that the CGa scheme is superior with respect to the convergence of the forces. We found that this is not the case: For the SC-methods we have included the correction terms, discussed in section II D (method 'opt'). But even if no correction terms are included and if the mixed charge density is used for the calculation of the local contributions to the forces (method 'mix'), the SC methods are superior (see Fig. 4, there is only a small difference between methods 'mix' and 'opt', and only if the output charge density without corrections is used (method 'out') the convergence suffers seriously).

6.2.3. Diamond surface

The clean C(100)(1 × 2) surface shows also significant charge instabilities, although the problems are less severe than for the metallic surface. In Fig. 12 results for the Pulay mixing are shown. The initial

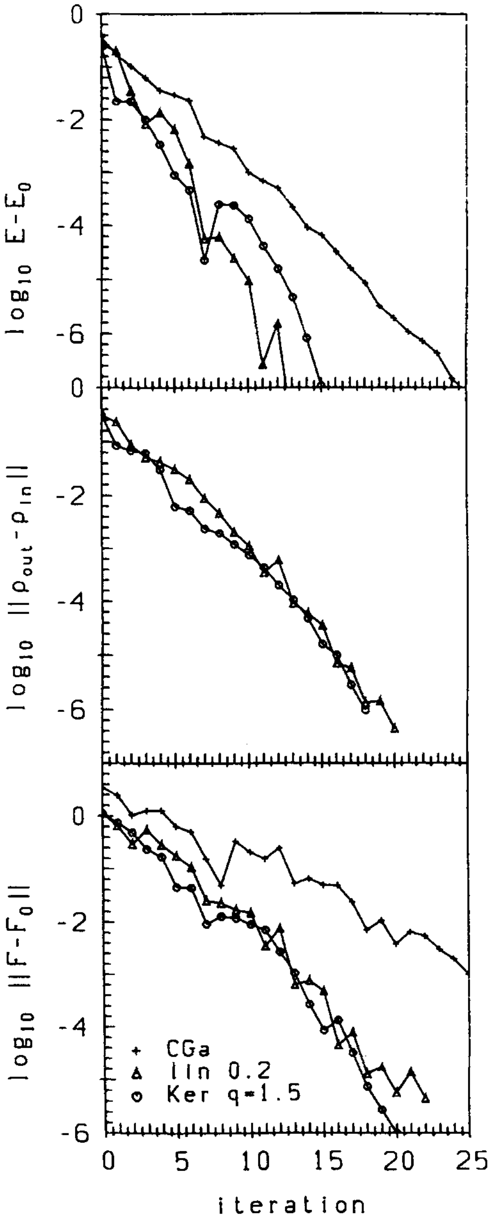

Fig. 12. Convergence for the C(100)(1×2) surface, selfconsistent case. Top panel total free energy per atom E (in eV), second panel charge density residual vector, last panel forces (in eV/Å/atom). In all cases Pulay's method was used for the mixing. For 'Ker 1.5' the initial matrix G^1 was set to Kerker's mixing matrix with $q_0 = 1.5$ Å^{-1}, and for 'lin 0.2' to a linear mixing $G^1 = 0.2$. Results for the conjugate gradient approach applied directly to the KS functional (CGa) are also shown.

matrix G^1 was set to Kerker's mixing matrix with $q_0 = 1.5$ Å^{-1}, $A = 0.8$ (the default value), and to a linear mixing with $A = 0.2$. As for Pd, Pulay's scheme is very robust with respect to changes in the initial mixing matrix. The inclusion of a metric was found to be less important for this case, and Broyden's second method was almost as efficient as Pulay's method, showing that the charge instabilities are less severe in insulating systems than in metallic systems. In Fig. 12 results for the conjugate gradient approach applied directly to the KS-functional (CGa) are also shown. The same number of bands as in the other approaches was included. It can be seen, that the convergence is quite good, but still not comparable to that of the CG or RMM scheme, especially forces converge significantly slower than within the SC-schemes.

7. Conclusion

We have presented a complete and detailed description of algorithms which allows to perform ab initio calculations of the electronic structure and the total energy using a plane-wave basis set. Although most ingredients of our method have been proposed by other authors in different contexts, we are not aware of any implementation which allows to achieve a similar efficiency especially if metallic systems are considered.

In obtaining this high stability and speed several points have been considered: First partial occupancies are extremely important for treating metallic systems. Only additional orbitals well above the Fermi level allow for an efficient calculation of the states close to and beneath the Fermi level. The main reason for this lies in the fact that the highest wavefunctions included in the calculation always converge slowest, if no gap exists right above these states. Inclusion of additional states above the Fermi level moves the 'difficult' region to the unoccupied states.

We have also shown that the selfconsistency cycle (SC) seems to be the most efficient way for calculating the KS-groundstate of metallic systems. Together with an efficient iterative matrix diagonalization scheme an algorithm which is close to an order $N_{\text{atoms}}^2 \ln N_{\text{atoms}}$ scaling is possible, at least for the problems currently tractable and of interest ($N_{\text{elect}} < 1000$). For instance we have performed calculations on liquid Te, and when we increased the number of atoms from 64 to 125, the computing time per molecular dynamics step increased by a factor 4.5 on an IBM workstation, to be compared with 3.8 if we had a pure order N^2 scaling. This rather favorable scaling is only possible in conjunction with the efficient RMM-DIIS method.

An alternative to the SC-method is the direct minimization of the KS functional if all degrees of freedom are treated consistently on the same footing. Despite the fact that the direct method is mathematically very appealing, it is currently still slower by a factor 1.5–10 than the SC method. The main problems in this scheme are: (i) The exact line minimization, which is necessary for an efficient conjugate gradient algorithm, is subtle. (ii) Different components (wavefunctions, sub-space rotation and partial occupancies) are optimized at the same time, which requires a careful choice of the search vectors of each component and of the length of each component with respect to the other components. We think that there is still some room left for further optimization of the direct methods, but it will be hard to beat the performance of the SC method. Finally one remaining problem of the direct methods – which we have not mentioned up to now – is that they require considerably more storage than SC methods. At the moment memory for computers is still relatively expensive, and ab-initio calculations are usually performed at the memory limit. Using the direct methods would sometimes force us to reduce the system size substantially.

In the introduction we have already pointed out, that the SC techniques discussed here have been used for a large number of problems. The essentials of the algorithm are unchanged since our first work [15], although several points have been improved (i.e. more efficient matrix diagonalization, Pulay mixing, ultrasoft pseudopotentials and tetrahedron method etc.). We are optimistic that our current implementation can be used for most problems tractable with the pseudopotential local density functional approach. We have also shown in previous papers, that ab-initio molecular dynamics can be performed with an efficiency that is comparable with the standard CP approach. But our method has several advantages in this respect: (i) The electrons are always in their instantaneous electronic groundstate, deviations from the Born–Oppenheimer surface can be controlled easily. (ii) Metals can be treated without the adiabaticity problem existing within the CP-method and there is no need to introduce artificial thermostats which control the temperature of the electrons.

It is also clear that the algorithms discussed here are easy to implement in any existing CP-like code. Our method is not limited to the pseudopotential approach, but can also be used within Blöchl's projector augmented-wave method (PAW) [64] or the linearized augmented plane wave method (LAPW) if projector functions are used [43]. For the efficiency of the iterative matrix diagonalization scheme based on the RMM-DIIS scheme the key point is the fast evaluation of $(\boldsymbol{H}-\epsilon S)|\phi\rangle$. If this operation is of the order $O(N^2)$ then the RMM scheme will outperform any other iterative technique for large systems, where order N^3 operations determine the overall performance.

For the mixing of charge densities, we also rely on Pulay's RMM method. The general advantages of the RMM-DIIS approach lie in the fact that this scheme does not require the evaluation of an exact gradient, but any search direction pointing to the 'right' direction is sufficient. Pulay's method is also not restricted to the minimization (or maximization) of a function but it can also be used to find saddle points of functions (in our case the stationary points of the Rayleigh quotient for the iterative matrix diagonalization and the stationary point of the Harris–Foulkes functional for the charge density mixing). Pulay's method allows to retain information from an unlimited number of previous steps. All together, we think that the RMM-DIIS technique – which lies at the heart of most of our computational algorithms is among the most powerful optimization techniques currently available.

Acknowledgements

We are indebted to J. Hafner; he has encouraged this work and he has supplied invaluable help in a lot of discussions. We also have to thank M.C. Payne, who has supplied a copy of his early ab-initio plane-wave program (local pseudopotentials only, version July, 1989). One of us (G.K.) has been supported by Siemens Nixdorf Austria within the contract of cooperation with the Technische Universität Wien.

Appendix A. Total derivatives of a constrained function

The electronic groundstate is defined as the minimum of a function $f[\{x\},\{R\}]$ under a constraint (for simplicity we assume only one constraint, but generalization to more than one constraint is straightforward)

$$g[\{x\},\{R\}]=0. \tag{A.1}$$

Using the Lagrange formalism a function $\bar{f}$

$$\bar{f}[\{x\},\{R\},\lambda]=f[\{x\},\{R\}]-\lambda g[\{x\},\{R\}] \tag{A.2}$$

has to be minimized with respect to λ and x to obtain the minimal x, under the constraint $g=0$. We now show that the total derivative of f with respect to R might be written as

$$\frac{\mathrm{d}\,\mathrm{Min}_{gx} f[\{f\},\{R\}]}{\mathrm{d}R_i} = \frac{\partial f[\{x\},\{R\}]-\lambda g[\{x\},\{R\}]}{\partial R_i}, \tag{A.3}$$

where $\mathrm{Min}_{gx} f$ denotes the value of the function f at its minimum with respect to x under the constraint $g=0$. To first-order $\mathrm{d}\bar{f}$ is given by

$$\mathrm{d}\bar{f}=\frac{\partial\bar{f}}{\partial R_i}\mathrm{d}R_i+\frac{\partial\bar{f}}{\partial x}\mathrm{d}x+\frac{\partial\bar{f}}{\partial\lambda}\mathrm{d}\lambda, \tag{A.4}$$

and at the minimum of $\bar{f}$ with respect to x and λ (denoted by $\mathrm{Min}_{x\lambda}\bar{f}$) this simply reduces to

$$\mathrm{d}\,\mathrm{Min}_{x\lambda}\bar{f}=\frac{\partial\bar{f}}{\partial R_i}\mathrm{d}R_i. \tag{A.5}$$

Second, from the definition of $\bar{f}$ we can write

$$\mathrm{d}\bar{f}=\mathrm{d}f-\lambda(\mathrm{d}g)-(\mathrm{d}\lambda)g. \tag{A.6}$$

At the minimum of $\bar{f}$ the constraint $g=0$ holds, and for a change along a direction with $\mathrm{d}g=0$ the equation

$$\mathrm{d}\,\mathrm{Min}_{x\lambda}\bar{f}=\mathrm{d}\,\mathrm{Min}_{gx}f \tag{A.7}$$

is obtained. Combining Eqs. (A.4) and (A.7) we obtain Eq. (A.3).

Appendix B. Forces within the linear tetrahedron method

We will show in this appendix, that the total energy is variational with respect to the partial occupancies f within the standard linear tetrahedron (LT) method, whereas the introduction of the correction terms by Blöchl (LT-C) results in a total energy which is not variational with respect to the partial occupancies. Within the LT-C method derivatives with respect to the partial occupancies have to be evaluated for the calculation of the forces acting on the ions. If these terms are omitted energy and forces are no longer consistent for a specific k-point mesh.

To address the problem we write the energy as a functional of the wavefunctions ϕ, the partial occupancies f, the chemical potential μ and the ionic positions $\boldsymbol{R}$. The partial occupancies depend on the chemical potential μ and via the eigenvalues of the Hamiltonian on the wavefunctions and on the ionic position $\boldsymbol{R}$. It is simplest to treat the chemical potential as an additional variational degree of freedom, and to incorporate the orthonormality constraint and the constraint on the occupancies using the Lagrange formalism (see also Section 2.4). In this way a new variational quantity

$$\bar{E}[\{C\},\{\gamma\},\{f\},\mu,\{\boldsymbol{R}\}] = E - \sum_{nqq'} \gamma_{nqq'} C^*_{nq'} S_{q'q} C_{nq} - \mu\left(\sum_n f_n - N\right). \tag{B.1}$$

can be defined, where C_{nq} is the expansion coefficient of the state $|\phi_n\rangle$ for the plane wave $|\boldsymbol{q}\rangle$ (compare with Eq. (32)). To first-order the change in total energy at the groundstate is given by the change of the Lagrangian $\bar{E}$ (see Appendix A) and can be written as

$$\mathrm{d}\bar{E} = \sum_{nq} \frac{\delta \bar{E}}{\delta C_{nq}} \delta C_{nq} + \sum_{N} \frac{\partial \bar{E}}{\partial \boldsymbol{R}_N} \mathrm{d}\boldsymbol{R}_N + \sum_{nn'} \frac{\partial \bar{E}}{\partial f_n} \frac{\partial f_n}{\partial \epsilon_{n'}} \mathrm{d}\epsilon_{n'} + \frac{\partial \bar{E}}{\partial \mu} \mathrm{d}\mu. \tag{B.2}$$

where n is a compound index for the band index n and the k-point index $\boldsymbol{k}$, the weighting factor $w_{\boldsymbol{k}}$ has been dropped for simplicity. The first term describes the change of the energy due to changes in the wavefunctions, and is zero at the KS-groundstate for arbitrary δC_{nq}. The third term corresponds to the energy change due to changes in the partial occupancies. We can rewrite

$$\frac{\partial \bar{E}}{\partial f_n} = \frac{\partial E}{\partial f_n} - \frac{\partial \mu(\Sigma_{n'} f_{n'} - N_{\mathrm{el}})}{\partial f_n}) = \epsilon_n - \mu, \tag{B.3}$$

where we have used the fact that $\partial E / \partial f_n = \epsilon_n$. The fourth term represents the energy change due to a change of the chemical potential, and can be written as

$$\frac{\partial \bar{E}}{\partial \mu} = \frac{\partial E}{\partial \mu} - \frac{\partial \mu(\Sigma_n f_n - N_{\mathrm{el}})}{\partial \mu} = \sum_n \frac{\partial f_n}{\partial \mu}(\epsilon_n - \mu) - \left(\sum_n f_n - N_{\mathrm{el}}\right) \tag{B.4}$$

where the last term is 0, considering that $\Sigma_n f_n - N_{\mathrm{el}} = 0$ for the correct occupancies.

Inserting the equations for the standard LT-method for f_n given in Blöchl's paper [36] it is easy to show that

$$\sum_n (\epsilon_n - \mu) \frac{\partial f_n}{\partial \epsilon_{n'}} = 0 \quad \forall n' \tag{B.5}$$

and

$$\sum_n (\epsilon_n - \mu) \frac{\partial f_n}{\partial \mu} = 0 \tag{B.6}$$

for each individual tetrahedron, and the only term which survives in Eq. (B.2) is $\partial E / \partial \boldsymbol{R}_N$. As expected, the evaluation of first-order energy changes is indeed very simple and for the calculation of forces the variational degrees of freedom ϕ and μ and the partial occupancies f_n can be kept fixed.

The left hand sides in Eqs. (B.5) and (B.6) are no longer equal to zero if the additional correction terms (LT-C), which take into account the curvature of the bands at the Fermi surface, are included. According to Blöchl [36] these correction terms are given by

$$\delta f_n = \sum_T \frac{1}{40} D_T(\mu) \sum_{j=1}^{4} (\epsilon_j - \epsilon_n), \qquad \text{(B.7)}$$

where $D_T(\mu)$ is the density of states for the tetrahedron T at the Fermi-level. The magnitude of this correction term depends on the spacing δ of the k-point mesh, and is of second-order in δ. The correction to the forces is therefore also of the order $O(\delta^2)$.

We want to point out that Blöchl's method is still variational with respect to the 'real' degrees of freedom, i.e. with respect to the wavefunctions: a first-order change in the wavefunctions will only result in a second-order change of the eigenvalues ϵ_n and second-order change of the chemical potential μ. From Eq. (B.2) it can be recognized that the change of the total energy with respect to arbitrary variations in the wavefunctions is therefore also of second order. The only inconvenience added by Blöchl's method is that the derivatives of the partial occupancies (and of the chemical potential μ) with respect to the ionic positions must be calculated in order to get exact forces. To first-order the change of the eigenvalue ϵ_n is given by (see Eq. 38)

$$\sum_{Nnqq'} C^*_{nq'} \frac{\partial(H[\rho, \{\boldsymbol{R}\}] - \epsilon_n S[\{\boldsymbol{R}\}])_{q'q}}{\partial \boldsymbol{R}_N} C_{nq} \delta \boldsymbol{R}_N . \qquad \text{(B.8)}$$

Eq. (B.8) makes the evaluation of forces possible at least in principle, but Eq. (B.8) is also extremely inconvenient — especially the evaluation of the local part $\partial V_{\mathrm{loc}}/\partial \boldsymbol{R}_N$ requires large additional work arrays to store the charge density of each band. For ultrasoft pseudopotentials the calculation of the augmentation part corresponding to each band is almost intractable, making Blöchl's method not applicable for US PP if accurate forces have to be calculated.

References

[1] W. Kohn and L. Sham, Phys. Rev. A 140 (1965) 1133.

[2] R.P. Feynman, Phys. Rev. 56 (1939) 340.

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

[4] G. Pastore, E. Smargiassi and F. Buda, Phys. Rev. A 44 (1991) 6334.

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

[6] M.C. Payne, J.D. Joannopoulos, D.C. Allan, M.P. Teter and D.H. Vanderbilt, Phys. Rev. Lett. B 56 (1986) 2656.

[7] F. Tassone, F. Mauri and R. Car. Phys. Rev. B 50 (1994) 10561.

[8] A. Williams and J. Soler, Bull. Am. Phys. Soc. B 32 (1987) 562.

[9] Guo-Xin Qian, M. Weinert, G.W. Fernando and J.W. Davenport, Phys. Rev. Lett. 64 (1990) 1146.

[10] M.P. Teter, M.C. Payne and D.C. Allan, Phys. Rev. B 40 (1989) 12255.

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

[12] M.J. Gillan, J. Phys.: Condens. Matter 1 (1989) 689.

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

[14] G. Kresse and J. Hafner, Phys. Rev. B 47 (1993) RC558.

[15] G. Kresse and J. Hafner, Phys. Rev. B 48 (1993) 13115.

[16] G. Kresse, Proc. 6th Int. Conf. on the Structure of Non-Cryst. Mater. (NCM6) (Prague, 1994), J. Non-Cryst. Solids 192/193 (1995) 222.

[17] G. Kresse and J. Hafner, Phys. Rev. B 49 (1994) 14251.

[18] G. Kresse, Proc. 9th Int. Conf. on Liquid and Amorphous Metals (LAM9) (Chicago, 1995).

[19] J. Furthmüller, J. Hafner and G. Kresse, Europhys. Lett. 28 (1994) 659, Phys. Rev. B 53 (1996) 7334.

[20] J. Furthmüller, G. Kresse, J. Hafner, R. Stumpf and M. Scheffler, Phys. Rev. Lett. 74 (1995) 5084.

[21] A. Eichler, J. Hafner, J. Furthmüller and G. Kresse, Surf. Sci. 346 (1996) 300.

[22] G. Kresse, J. Furthmüller and J. Hafner, Europhys. Lett. 32 (1995) 729.

[23] C.G. Broyden, Math. Comput. 19 (1965) 577.

[24] P. Pulay, Chem. Phys. Lett. 73 (1980) 393.

[25] D. Vanderbilt, Phys. Rev. B 41 (1990) 7892.

[26] K. Laasonen, A. Pasquarello, R. Car, C. Lee and D. Vanderbilt, Phys. Rev. B 47 (1992) 10142.

[27] G. Kresse, Thesis, Technische Universität Wien (1993).

[28] G. Kresse and J. Hafner, J. Phys.: Condens. Matter 6 (1994) 8245.

[29] N.D. Mermin, Phys. Rev. A 137 (1965) 1441.

[30] R.O. Jones and O. Gunnarsson. Rev. Mod. Phys. 61 (1989) 689.

[31] M. Weinert and J.W. Davenport. Phys. Rev. B. 45 (1992) 13709.

[32] R.M. Wentzcovitch, J.L. Martins and P.B. Allen, Phys. Rev B 45 (1992) 11372.

[33] C.-L. Fu and K.-M. Ho, Phys. Rev. B 28 (1983) 5480.

[34] A. Baldereschi, Phys. Rev. B 7 (1973) 5212;
D.J. Chadi and M.L. Cohen, Phys. Rev. B 8 (1973) 5747;
H.J. Monkhorst und J.D. Pack, Phys. Rev. B 13 (1976) 5188.

[35] O. Jepsen and O.K. Andersen, Solid State Commun. 9 (1971) 1763.

[36] P.E. Blöchl, O. Jepsen and O.K. Andersen, Phys. Rev. B 49 (1994) 16223.

[37] A. De Vita and M.J. Gillan, J. Phys.: Condens. Matter 3 (1991) 6225;
A. De Vita, PhD Thesis, Keele University (1992);
A. De Vita and M.J. Gillan, preprint (Aug. 1992).

[38] M. Methfessel and A.T. Paxton, Phys. Rev. B 40 (1989) 3616.

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

[40] W.M.C. Foulkes, Ph.D. Thesis, University of Cambridge (1987).

[41] W.M.C. Foulkes and R. Haydock, Phys. Rev. B 39 (1989) 12520.

[42] P. Pulay, in: Modern Theoretical Chemistry, ed. H.F. Schaefer (Plenum, New York, 1977); Mol. Phys. 17 (1969) 197.

[43] S. Goedecker and K. Maschke, Phys. Rev B 45 (1992) 1597.

[44] D.M. Wood and A. Zunger, J. Phys. A: Math. Gen. 18 (1985) 1343.

[45] D.M. Bylander, L. Kleinman and S. Lee, Phys. Rev. B 42 (1990) 1394.

[46] B.N. Parlett, The Symmetric Eigenvalue Problem (Prentice Hall, Englewood Cliffs, NJ, 1980).

[47] J. Furthmüller, Thesis, Universität Stuttgart (1991), unpublished.

[48] E.R. Davidson, in: Methods in Computational Molecular Physics, ed. G.H.F. Diercksen and S. Wilson, NATO Advanced Study Institute, Ser. C, Vol. 113 (Plenum, New York, 1983) p. 95; in: Rep. on the Workshop "Numerical Algorithms in Chemistry: Algebraic Methods", ed. C. Moler and I. Shavitt (Lawrence Berkley Lab. Univ. of California, 1978) p.49; J. Comput. Phys. 17, 87.

[49] B. Liu, in: Report on the Workshop "Numerical Algorithms in Chemistry: Algebraic Methods", ed. C. Moler and I. Shavitt (Lawrence Berkley Lab. Univ. of California, 1978) p.49.

[50] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes (Cambridge University Press, New York, 1986).

[51] E. Polak, Computational Methods in Optimization, (Academic Press, New York, 1971).

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

[53] P.E. Blöchl and R.M. Martin, private communication.

[54] T.A. Arias, M.C. Payne and J.D. Joannopoulos, Phys Rev. B 45 (1992) 1538.

[55] R.D. King-Smith, M.C. Payne and J.S. Lin, Phys. Rev. B 44 (1991) 13063.

[56] D.D. Johnson, Phys. Rev. B 38 (1988) 12087.

[57] G.P. Srivastava., J. Phys. A 17 (1984) L317.

[58] S. Blügel, PhD. Thesis, Aachen University (1988).

[59] D. Vanderbilt and S.G. Louie, Phys. Rev. B 30 (1984) 6118.

[60] G.P. Kerker, Phys. Rev. B 23 (1981) 3082.

[61] Y. Yamamoto and T. Fujiwara, Phys. Rev. B 46 (1992) 13596.

[62] M.P. Grumbach, D. Hohl, R.M. Martin and R. Car J. Phys.: Condens. Matter 1 (1994) 1999.

[63] H. Akai and P.H. Dederichs, J. Phys. C 18 (1985) 2455.

[64] P.E. Blöchl, Phys. Rev. B 50 (1994) 17953.