IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

# Density kernel optimization in the ONETEP code

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2008 J. Phys.: Condens. Matter 20 294207

(http://iopscience.iop.org/0953-8984/20/29/294207)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 18.111.89.111
The article was downloaded on 31/03/2011 at 14:28

Please note that terms and conditions apply.

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **20** (2008) 294207 (7pp) doi:10.1088/0953-8984/20/29/294207

# Density kernel optimization in the ONETEP code

**P D Haynes[1], C-K Skylaris[2], A A Mostofi[1] and M C Payne[3]**

[1] Departments of Physics and Materials, Imperial College London, Exhibition Road,
London SW7 2AZ, UK
[2] School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[3] Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, J J Thomson
Avenue, Cambridge CB3 0HE, UK

E-mail: p.haynes@imperial.ac.uk

Received 18 January 2008
Published 24 June 2008
Online at stacks.iop.org/JPhysCM/20/294207

**Abstract**
ONETEP is a linear scaling code for performing first-principles total energy calculations within density-functional theory (DFT). The method is based on the density-matrix formulation of DFT and involves the iterative minimization of the total energy with respect to a set of local orbitals and a density kernel. An overview is given of the kernel optimization methods proposed in the literature and implemented in ONETEP, focusing in particular on the constraints of compatibility, idempotency and normalization that must be applied. A method is proposed for locating the chemical potential which may be useful in applying the normalization constraint and analysing the electronic structure near the Fermi level.

## 1. Introduction

The popularity of density-functional theory (DFT) has grown enormously over the last two decades. This is largely due to the balance that it achieves between two competing requirements: on the one hand it gives a sufficiently accurate treatment of electron correlation for many purposes; on the other hand the computational effort required is relatively low. Both of these strengths derive from the mapping between the real many-electron system and a fictitious system of non-interacting particles that lies at the heart of DFT [1, 2]. This mapping provides the formal connection needed to treat exchange and correlation within the independent electron approximation while reducing the complexity of the problem to the solution of a single-particle Schrödinger equation.

The $O(N^3)$ asymptotic scaling of traditional DFT methods with system size $N$ arises from the cost of diagonalizing the single-particle Hamiltonian or, if that process is carried out iteratively, maintaining the orthogonality of the extended single-particle wave functions. While this cubic scaling is favourable when compared with methods based on correlated wave functions, it still does not permit calculations on the scale required to tackle nanostructures and biological macromolecules containing thousands of atoms. For this reason a considerable effort has been expended on the development of linear scaling or $O(N)$ methods which exploit the 'nearsightedness' of quantum many-body systems [3, 4] to ensure that the computational cost increases only linearly with the system size. The long-term investment by several groups is now resulting in a number of new codes, including ONETEP [5, 6], SIESTA [7] and CONQUEST [8, 9], which have been designed specifically for $O(N)$ calculations.

A detailed comparison of the various $O(N)$ methods proposed for insulators and semiconductors can be found in existing review articles [10, 11]. Briefly, the methods may be divided into four categories: projection methods such as the Fermi operator expansion [12, 13]; the divide and conquer approach [14, 15]; generalized energy functionals for non-orthogonal orbitals [16–21] and density-matrix (DM) minimization methods. The ONETEP code falls into the last of these categories, and this class of methods is the focus of this article.

Section 2 outlines the general approach taken by DM minimization methods. The constraints of idempotency and normalization are considered in sections 3 and 4 respectively, and the scheme currently implemented in ONETEP is described in section 5.

## 2. Density-matrix minimization

The single-particle DM is chosen as the central variable in many $O(N)$ methods because it provides a complete

0953-8984/08/294207+07$30.00 © 2008 IOP Publishing Ltd Printed in the UK

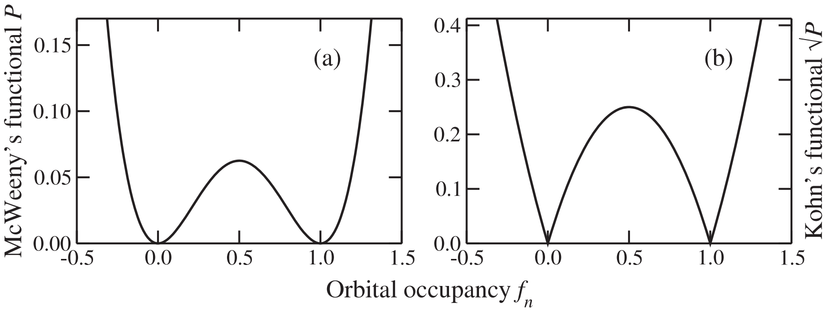

**Figure 1.** Penalty functionals for idempotency proposed by (a) McWeeny and (b) Kohn shown as a function of a single orbital occupancy $f_n$, all others being fixed at either zero or unity.

description of the fictitious Kohn–Sham system and demonstrates the property of nearsightedness explicitly. In the position representation, the DM $\rho(\boldsymbol{r}, \boldsymbol{r}')$ decays as the separation of its arguments $|\boldsymbol{r}-\boldsymbol{r}'|$ increases [22, 23]. Since a ground-state DM is always separable [24], this form is adopted generally:

$$\rho(\boldsymbol{r}, \boldsymbol{r}') = \sum_{\alpha\beta} \phi_\alpha(\boldsymbol{r}) K^{\alpha\beta} \phi^*_\beta(\boldsymbol{r}'). \quad (1)$$

The $\{\phi_\alpha(\boldsymbol{r})\}$ are a set of overlapping non-orthogonal orbitals [25] that are typically centred on atoms. These orbitals are often assumed to be real since the $\Gamma$-point is sufficient to sample the Brillouin zone of the large systems appropriate for O($N$) methods. Spin is not considered explicitly here. The overlap matrix $\mathbf{S}$ for these orbitals may be defined by its matrix elements $S_{\alpha\beta} = \langle\phi_\alpha|\phi_\beta\rangle$ and the set of dual orbitals $\{\phi^\alpha(\boldsymbol{r}) = \sum_\beta S^{\alpha\beta}\phi_\beta(\boldsymbol{r})\}$ defined by $\langle\phi^\alpha|\phi_\beta\rangle = \delta^\alpha_\beta$ may be generated where $S^{\alpha\beta}$ is a matrix element of the overlap matrix for the duals which is simply $\mathbf{S}^{-1}$. $K^{\alpha\beta}$ is a matrix element of the density kernel $\mathbf{K}$, which is the representation of the DM in terms of the duals.

Linear scaling is obtained by enforcing nearsightedness: the orbitals must be localized and the density kernel must be sparse. Following [26], the orbitals in ONETEP are truncated beyond a given radius (typically around 3.5 Å) and a longer independent cutoff is applied to the density kernel. Unique to ONETEP is the optimization of the local orbitals (known as non-orthogonal generalized Wannier functions [27]) in terms of a psinc basis set [28] equivalent to (and hence as accurate as) a set of plane waves, in which the 'FFT box' technique [29] is used to retain linear scaling even when fast Fourier transforms (FFTs) are employed.

Density-matrix minimization methods proceed by minimizing an energy functional of the DM with respect to the local orbitals and the density kernel, subject to the appropriate constraints. This article concerns the optimization of the density kernel so that from now on the local orbitals are assumed to be constant. What follows could therefore be equally applied to first-principles tight-binding [30, 31].

In order to find the ground state, three constraints must be satisfied during the minimization: compatibility, that the DM commute with the Hamiltonian $[\rho, H] = 0$; normalization, that the DM correspond to the correct number of electrons $\mathrm{tr}(\rho) = N_\mathrm{e}$; and idempotency, that powers of the DM are the same $\rho^2 = \rho$. It is the third, non-linear constraint of idempotency that is most difficult to enforce, and which is considered first.

## 3. Idempotency

In terms of the Kohn–Sham orbitals $\{\psi_n(\boldsymbol{r})\}$ (with eigenvalues $\{\varepsilon_n\}$) and their occupancies $\{f_n\}$ the ground-state DM takes the diagonal form (c.f. (1))

$$\rho(\boldsymbol{r}, \boldsymbol{r}') = \sum_n f_n \psi_n(\boldsymbol{r}) \psi^*_n(\boldsymbol{r}'). \quad (2)$$

From this form, idempotency is seen to correspond to the double constraint that the Kohn–Sham orbitals must be orthonormal $\langle\psi_n|\psi_m\rangle = \delta_{nm}$ and the occupancies must equal zero or unity. For the ground-state DM, states below the Fermi level $\mu$ ($\varepsilon_n < \mu$) must be occupied ($f_n = 1$) and states above the Fermi level ($\varepsilon_n > \mu$) unoccupied ($f_n = 0$). Hence idempotency enforces the Pauli exclusion principle and together with energy minimization applies the Aufbau principle. It is also noteworthy that idempotency derives from the orthonormality constraint that is the cause of the O($N^3$) scaling of traditional methods. Dealing with this awkward non-linear constraint may therefore be viewed as the price to be paid for avoiding orbital orthonormality.

### *3.1. Penalty functionals*

The vast majority of methods proposed in the literature for imposing idempotency find their origin in the penalty functional first proposed by McWeeny [24],

$$P[\rho] = \mathrm{tr}\left[\left(\rho^2 - \rho\right)^2\right] = \sum_n \left(f_n^2 - f_n\right)^2, \quad (3)$$

and illustrated in figure 1(a). It is clearly positive semi-definite, vanishing if and only if the DM is idempotent. Another attractive feature of this penalty functional is that the Hessian matrix of second derivatives (evaluated at idempotency) is the identity. Apart from being quartic rather than quadratic, this functional could not be easier to minimize. The method of steepest descents is already optimal: more sophisticated methods (e.g. conjugate gradients) and preconditioning schemes are unnecessary.

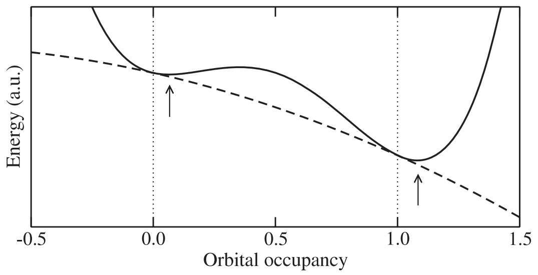

**Figure 2.** Variation with a single orbital occupancy (the rest being zero or unity) of the total energy $E$ with (solid line) and without (dashed line) the addition of the McWeeny penalty functional $P$ to enforce idempotency approximately. The positions of local minima of the total functional are indicated by arrows.

It might be thought that such a penalty functional could be used to enforce idempotency simply by minimizing a generalized energy functional $Q[\rho] = E[\rho] + \alpha P[\rho]$ where $E$ is the energy functional to be minimized (this might be the energy of the non-interacting Kohn–Sham system $\mathrm{tr}(\rho H)$ or the energy of the real interacting electronic system if the minimization is carried out self-consistently as in ONETEP) and $\alpha$ is an energy parameter used to control the strength of the penalty functional. Figure 2 illustrates how such a functional might vary as the occupancy of a single occupied orbital is varied. However since the slope of the energy at $f_n = 1$ is given by the Kohn–Sham eigenvalue $\varepsilon_n$ [32] (if the Kohn–Sham non-interacting energy is being minimized then the dashed curve in figure 2 would be a straight line), then the minimum of the total functional $Q$ cannot occur at $f_n = 1$ whatever the value of $\alpha$, so that this scheme can only impose idempotency approximately at best.

To circumvent this problem, Kohn [3] proposed to use the square root of $P$ as the penalty functional (see figure 1(b)) and thus to minimize $Q'[\rho] = E[\rho] + \alpha\sqrt{P[\rho]}$. As can be seen in figure 3, for sufficiently large $\alpha$ greater than some critical value $\alpha_{\mathrm{c}}$ (that depends on the Kohn–Sham eigenvalues $\{\varepsilon_n\}$) the total functional takes its minimum value for an idempotent DM. However, Kohn's functional is not differentiable at the desired ground-state minimum, making it wholly unsuitable for any practical minimization scheme [33].

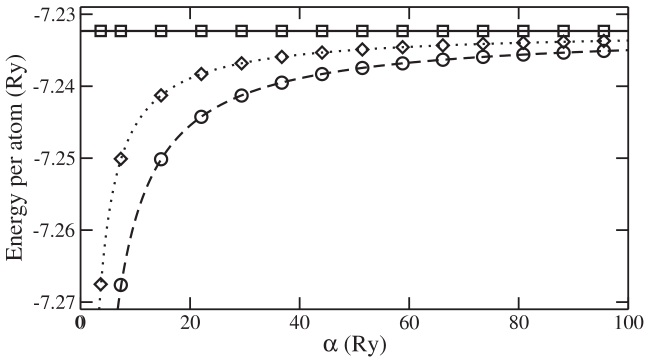

**Figure 4.** Variation of the minimized total energy $E$ (○), total functional $Q = E + \alpha P$ (◇) and corrected energy (□) as a function of the parameter $\alpha$ for crystalline silicon.

Another objection to the use of penalty functionals might be the existence of multiple local minima. However the normalization constraint eliminates most of these, and energy minimization drives the DM towards the desired ground-state minimum so that this is never a problem in practice if the DM is suitably prepared initially. In principle, optimization of the local orbitals provides a complete solution to the problem since it rotates the representation of the DM and can therefore convert a false local minimum into the global ground-state minimum.

Returning to the functional $Q[\rho] = E[\rho] + \alpha P[\rho]$, it can be shown that the error $\delta f_n$ in an orbital occupancy (i.e. its deviation from zero or unity) is given by

$$\delta f_n \approx -\frac{\varepsilon_n - \mu}{\alpha}. \tag{4}$$

The minimization of this functional is robust and straightforward and compatibility is guaranteed at its minimum. The error in the occupancies can be reduced by increasing the parameter $\alpha$, and the energy obtained at the minimum approaches the true ground-state energy from below with an error that also scales as $1/\alpha$, as shown in figure 4. However, since the functional $Q$ is differentiable at its minimum, a correction based upon a Taylor expansion can be made so that the total energy may be correctly calculated even when relatively small values of $\alpha$ are employed [34].

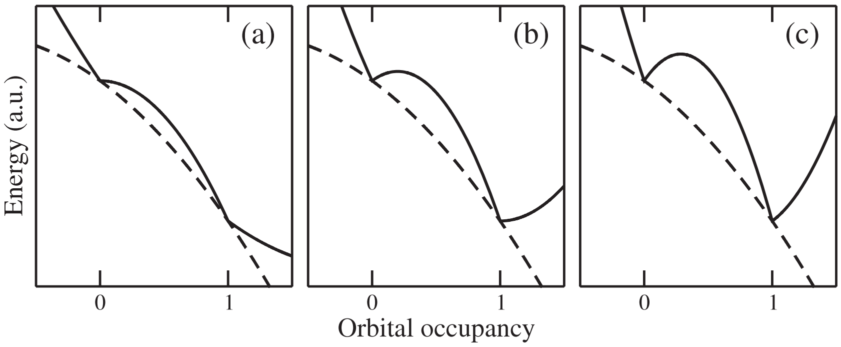

**Figure 3.** Variation with a single orbital occupancy (the rest being zero or unity) of the total energy $E$ with (solid line) and without (dashed line) the addition of the Kohn penalty functional $\sqrt{P}$ to enforce idempotency for three values of the energy parameter $\alpha$: (a) $\alpha < \alpha_{\mathrm{c}}$, (b) $\alpha = \alpha_{\mathrm{c}}$ and (c) $\alpha > \alpha_{\mathrm{c}}$.

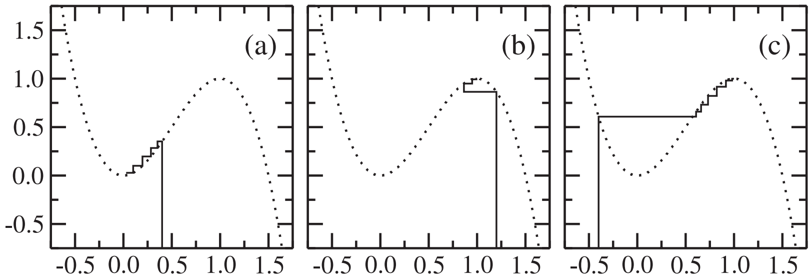

**Figure 5.** Illustration of the purification transformation in terms of orbital occupancies for three cases: (a) the initial occupancy $\frac{2}{5}$ converges to zero; (b) the initial occupancy $\frac{6}{5}$ converges to unity; (c) the initial occupancy $-\frac{2}{5}$ converges to unity.

### *3.2. Purification*

McWeeny [24] proved that steepest descents minimization of $P$ close to idempotency (where the step length may be fixed to $\frac{1}{2}$) results in the so-called purification transformation for iteratively improving a trial DM $\rho_0$ towards idempotency:

$$\rho_{k+1} = 3\rho_k^2 - 2\rho_k^3. \tag{5}$$

This iterative procedure, illustrated in figure 5, converges if the initial occupancies lie in the interval $(\frac{1-\sqrt{5}}{2}, \frac{1+\sqrt{5}}{2})$. However tighter bounds are desirable: if the occupancies lie in $[-\frac{1}{2}, \frac{3}{2}]$ then the purified DM will be 'weakly' idempotent, i.e. its occupancies will lie in [0, 1]. However this still allows the possibility, shown in figure 5(c), that the occupancy could 'flip' from unoccupied to occupied or *vice versa*, which might provoke instabilities or result in local minima during the minimization procedure. This can be avoided by ensuring that the occupancies all remain within the interval $(\frac{1-\sqrt{3}}{2}, \frac{1+\sqrt{3}}{2})$. In ONETEP the extremal occupancies are monitored to enforce this.

*3.2.1. Adaptive purification.* The occupancies $\{f_n\}$ are the eigenvalues of the density kernel $\mathbf{K}$ that satisfy the generalized eigenvalue equation $\mathbf{K}\mathbf{x}_n = f_n\mathbf{S}^{-1}\mathbf{x}_n$. The corresponding eigenvectors $\{\mathbf{x}_n\}$, which relate the Kohn–Sham orbitals $\{\psi_n\}$ to the local orbitals $\{\phi_\alpha\}$, are dense irrespective of the sparsity of $\mathbf{K}$. Nevertheless, a small fixed number of eigenvalue–eigenvector pairs may still be found in O($N$) operations using iterative methods. The extremal occupancies can be found by extremising the generalized Rayleigh quotient $\lambda(\mathbf{x}) = \mathbf{x}^\dagger\mathbf{K}\mathbf{x}/(\mathbf{x}^\dagger\mathbf{S}^{-1}\mathbf{x})$, which is far more accurate than Gershgorin estimates. In fact, the method is implemented using the equivalent function $\lambda(\mathbf{y}) = \mathbf{y}^\dagger\mathbf{SKSy}/(\mathbf{y}^\dagger\mathbf{Sy})$ that avoids the use of $\mathbf{S}^{-1}$ ($\mathbf{y}_n = \mathbf{S}^{-1}\mathbf{x}_n$ relates the $\{\psi_n\}$ to the duals $\{\phi^\alpha\}$).

Should the extremal occupancies lie outside the desired interval, then 'adaptive' purification is used to bring them back inside. This simply involves steepest descents minimization of $P$ where the optimal step length is calculated explicitly (rather than fixed to $\frac{1}{2}$) in order to avoid instabilities. This minimization procedure converges rapidly due to the properties of $P$ mentioned above and is more efficient than generalized purification transformations [35–37].

*3.2.2. Canonical purification.* The canonical purification method [38] is a non-self-consistent method for determined the ground-state DM of a fixed Hamiltonian. The eigenvalues of the Hamiltonian are inverted, shifted and scaled so that they lie in the interval [0, 1] i.e. $\varepsilon_n \rightarrow \frac{1}{2}(1 + (\mu - \varepsilon_n)/\varepsilon_{\max})$ where $\varepsilon_{\max} = \max(\{|\varepsilon_n - \mu|\})$. The extremal eigenvalues of the Hamiltonian matrix $\mathbf{H}$ in the representation of the local orbitals $\{\phi_\alpha\}$ can be found by extremising the quotient $\zeta(\mathbf{y}) = \mathbf{y}^\dagger\mathbf{H}\mathbf{y}/(\mathbf{y}^\dagger\mathbf{S}\mathbf{y})$. The purification transformation (5) is then repeatedly applied until the Kohn–Sham energy tr($\mathbf{KH}$) converges. In the absence of kernel truncation compatibility is guaranteed by construction. When truncation is applied, matrix products can only be approximately evaluated so that the Kohn–Sham energy eventually starts to increase, and the algorithm is terminated at this point. This method is used in ONETEP to generate the initial guess for the density kernel. A modified version that allows greater flexibility in the choice of purification transformation has also been developed [39].

### *3.3. Li–Nunes–Vanderbilt method*

By far the most widespread DM minimization method is that attributed to Li, Nunes and Vanderbilt (LNV) [40, 41] and reported simultaneously by Daw [42]. The purification transformation is used to define the DM $\rho$ in terms of an auxiliary matrix $\sigma$ as $\rho = 3\sigma^2 - 2\sigma^3$ where $\sigma$ is defined by:

$$\sigma(\boldsymbol{r}, \boldsymbol{r}') = \sum_{\alpha\beta} \phi_\alpha(\boldsymbol{r}) L^{\alpha\beta} \phi_\beta^*(\boldsymbol{r}') \tag{6}$$

and $L^{\alpha\beta}$ is a matrix element of the auxiliary kernel $\mathbf{L}$ which is related to the density kernel by $\mathbf{K} = 3\mathbf{LSL} - 2\mathbf{LSLSL}$.

Minimizing the total energy $E[\rho]$ with respect to $\sigma$ by optimizing the matrix elements of the auxiliary kernel $\mathbf{L}$ naturally drives the DM to idempotency. As long as the eigenvalues of $\mathbf{L}$ remain within the interval $[-\frac{1}{2}, \frac{3}{2}]$, then the purified DM $\rho$ will be weakly idempotent and a variational estimate of the ground-state energy is obtained. Moreover, because the energy functional is a cubic functional of $\mathbf{L}$ the method does not suffer from multiple minima. However this cubic dependence also means that the method is potentially unstable should any of the eigenvalues of $\mathbf{L}$ stray outside the range over which the purification transformation converges. A number of variants of the original LNV method have been proposed [43, 44].

## 4. Normalization constraint

All of the above methods for imposing the idempotency constraint need to be combined with a method for enforcing normalization. While this linear constraint is rather simpler to deal with, the manner in which this is done may affect the overall stability of the method. A Lagrange multiplier (the chemical potential $\mu$) may be employed to ensure that the minimum of the functional used corresponds to the correct number of electrons. This corresponds to minimizing the grand potential $\Omega = E - \mu N$ rather than the energy. The value of the chemical potential may vary during the calculation, and may not always be straightforward to determine.

There are three main approaches that have been taken to imposing normalization within the LNV method. The first is to constrain the purified electron number $\mathrm{tr}(\rho) = \mathrm{tr}(\mathbf{KS}) = \mathrm{tr}(3\mathbf{LSLS} - 2\mathbf{LSLSLS})$ which is a cubic function of the auxiliary kernel $\mathbf{L}$. Since the chemical potential is not easily determined, this approach involves projecting search directions to be orthogonal to the gradient of the purified electron number with respect to $\mathbf{L}$ to ensure that the purified electron number is fixed to first order. When a trial step is taken, $\mathbf{L}$ must be returned to the correct number of electrons by moving $\mathbf{L}$ along the electron number gradient [26, 45].

The second approach is to constrain the unpurified electron number $\mathrm{tr}(\sigma) = \mathrm{tr}(\mathbf{LS})$ [43]. Since this quantity is linear in $\mathbf{L}$ the chemical potential is straightforwardly determined. Choosing the correct chemical potential corresponds to projecting the search direction to be orthogonal to the unpurified electron number gradient, but since the constraint is linear no correction needs to be applied after changing $\mathbf{L}$.

The third approach is to generate a normalized and purified DM by construction, by rescaling i.e. modifying the purification transformation to

$$\rho = N_{\mathrm{e}} \frac{3\sigma^2 - 2\sigma^3}{\mathrm{tr}\big(3\sigma^2 - 2\sigma^3\big)}. \tag{7}$$

Although this is no longer a cubic dependence, and might therefore reintroduce multiple minima, in practice this has not been observed. The extra terms in the denominator generate terms in the search direction that automatically project out the electron number gradient, and effectively determine the chemical potential. Any correction to the electron number (only necessary when the local orbitals are optimized) is carried out by rescaling $\mathbf{L}$ appropriately.

### *4.1. Effect of truncation*

Since O($N$) is only achieved when the auxiliary and density kernels are truncated according to some spatial cutoff, it is vital to assess the effect of this truncation, which can be dramatic. In ONETEP, imposing the normalization constraint via the unpurified electron number has proved unstable for periodic systems, in which the purified and unpurified electron numbers can differ by a large amount. However it should be noted that this approach was proposed as part of a method that takes a rather different approach to truncation [43]. Instead of applying fixed sparsity patterns via spatial cutoffs on the local orbitals and kernels, thresholding is used to apply truncation via upper bounds on the values of matrix elements, which is necessary when the local orbitals are expanded in a Gaussian basis set. Within the regime of tight thresholds this approach to normalization may well be much more successful.

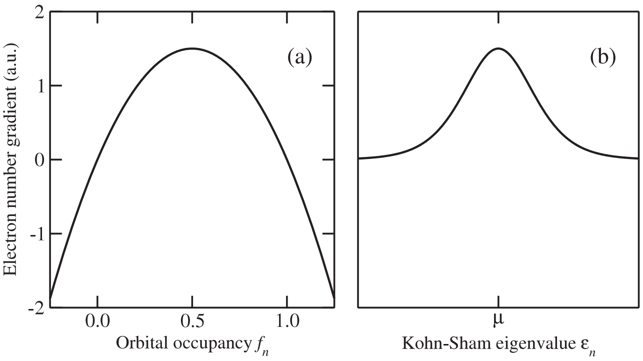

**Figure 6.** Schematic illustration of the purified electron number gradient as a function of (a) orbital occupancy and (b) Kohn–Sham eigenvalue.

Truncation can also have a marked effect on normalization imposed using the gradient of the purified electron number. Imposing a finite range on the density kernel is qualitatively similar to truncating a Fermi operator expansion [12, 13] or introducing a finite electronic temperature, in that while the occupancies remain clustered about zero and unity they are smeared out. This applies particularly to states close to the Fermi level. The purified electron number gradient $6\sigma(1-\sigma)$ is shown in figure 6 and it peaks markedly around the chemical potential. This means that the weight of the projection or correction can fall on a small number of states close to the Fermi level. Even a small change in the total electron number may require significant changes in the occupancies of these few states, and may cause them to be pushed outside the range of stability of the purification transformation. For this reason the rescaling approach was derived and implemented in ONETEP and does appear to confer stability upon the method.

### *4.2. Locating the chemical potential*

A method for determining the chemical potential $\mu$ is desirable for a number of the methods outlined above e.g. canonical purification and normalization within the LNV method. It is possible to take an empirical approach by running a number of simulations with different values of $\mu$ until the correct electron number is obtained at convergence, but this involves a significant amount of wasted effort. The difficulty stems from the fact that in general the chemical potential lies between interior rather than extremal eigenvalues of the Kohn–Sham Hamiltonian. The method of extremising a Rayleigh quotient cannot therefore be applied directly.

The folded spectrum method [46–48] has been used successfully to find the interior eigenvalues $\lambda$ of a given matrix $\mathbf{A}$ that lie closest to a given reference value $\lambda_{\mathrm{ref}}$. The eigenvalue spectrum of the matrix $\mathbf{A}$ is 'folded' up into the positive semi-definite spectrum of the matrix $(\mathbf{A} - \lambda_{\mathrm{ref}})^2$.

The smallest eigenvalues of the folded matrix correspond to the eigenvalues of $\mathbf{A}$ closest to $\lambda_{\mathrm{ref}}$. The problem of locating interior eigenvalues may therefore be converted into an extremal eigenvalue problem (albeit at the cost of some loss of efficiency due to the increase in condition number that the folding step produces). This method has been applied to the problem of finding the eigenvalues of a Hamiltonian closest to a given reference energy (usually the Fermi level).

When the density kernel is truncated, its eigenvalues (the orbital occupancies) closest to $\frac{1}{2}$ belong to those states nearest the chemical potential. These may be found by minimizing the Rayleigh quotient $\eta(\mathbf{y}) = \mathbf{y}^{\dagger}\mathbf{S}(\mathbf{KS} - \frac{1}{2})^2\mathbf{y}/(\mathbf{y}^{\dagger}\mathbf{S}\mathbf{y})$ with respect to the dense vector $\mathbf{y}$. The corresponding Kohn–Sham eigenvalues can then be accurately estimated by evaluating $\zeta(\mathbf{y}) = \mathbf{y}^{\dagger}\mathbf{H}\mathbf{y}/(\mathbf{y}^{\dagger}\mathbf{S}\mathbf{y})$ (this estimation becomes exact as the calculation converges since compatibility implies that $\mathbf{K}$ and $\mathbf{H}$ may be diagonalized simultaneously). This method may therefore be used to locate the energies of the states immediately above $\varepsilon_+$ and below $\varepsilon_-$ the chemical potential i.e. the LUMO and HOMO in molecular systems or the conduction band minimum and valence band maximum in extended systems. The chemical potential can therefore be estimated as $\mu = \frac{1}{2}(\varepsilon_+ + \varepsilon_-)$. When updating rather than initially locating the chemical potential, convergence may be accelerated by seeking the occupancies closest to those found during the previous iteration rather than $\frac{1}{2}$.

In the absence of truncation this method will not succeed since the orbital occupancies will all be zero or unity. However since the kernel is then dense linear scaling cannot be achieved and the $\mathrm{O}(N^3)$ cost of diagonalizing $\mathbf{H}$ directly will not add significantly to the cost of the calculation.

This scheme may be used to find and update the chemical potential needed by $\mathrm{O}(N)$ schemes such as the LNV method or canonical purification. In addition it may be used to analyse the electronic structure of a system once the ground state has been found e.g. to estimate a reference energy in the band gap prior to a folded spectrum calculation of the states closest to the Fermi level.

## 5. Implementation in ONETEP

In this section further computational details of the methods implemented in the ONETEP code are given, along with an outline of how these methods are combined in the kernel optimization part of the code. Further details of the parallel implementation [49] and examples of its applications to a variety of systems can be found elsewhere [50–52].

### *5.1. Non-orthogonality*

Throughout this article a distinction has been made between quantities that are covariant, such as the local orbitals $\{\phi_\alpha\}$, overlap $S_{\alpha\beta}$ and Hamiltonian $H_{\alpha\beta}$ matrix elements (all with Greek subscripts), and those that are contravariant, such as the dual orbitals $\{\phi^\alpha\}$ and the inverse overlap matrix elements $S^{\alpha\beta}$ (all with Greek superscripts). This is necessary because of the non-orthogonality of the local orbitals [53, 54].

In particular, when calculating search directions from gradients of functionals, it is necessary to use the metric tensors (the overlap matrix and its inverse) to convert between covariant gradients and contravariant search directions. For example, the gradient of the Kohn–Sham energy $E = \mathrm{tr}(\mathbf{KH})$ with respect to a matrix element of the contravariant density kernel $K^{\alpha\beta}$ is a covariant quantity $\partial E/\partial K^{\alpha\beta} = H_{\beta\alpha}$. The appropriate search direction is the contravariant quantity obtained by 'raising' both indices of $H_{\beta\alpha}$ using the metric tensor $S^{\alpha\beta}$ and may thus be written $\mathbf{S}^{-1}\mathbf{H}\mathbf{S}^{-1}$. This requires the inversion of the overlap matrix, which may be achieved using Hotelling's method [55], and results in a considerable improvement in the convergence of the method [56]. An alternative approach [44] is to transform the problem to an orthogonal representation by directly calculating the sparse inverse Cholesky factor of $\mathbf{S}$ [57, 58].

### *5.2. Overall scheme*

The combination of methods implemented in ONETEP is as follows:

(i) the local orbitals $\{\phi_\alpha\}$ are initially constructed by truncating pseudoatomic or Slater-type contracted Gaussian atomic orbitals;
(ii) the initial charge density is constructed by superposing atomic charge densities, from which the initial Hamiltonian in the representation of the local orbitals can be calculated;
(iii) the initial (non-self-consistent) density kernel is obtained using canonical purification [38];
(iv) this initial density kernel is then refined (self-consistently) using the approximate penalty functional method [34];
(v) the density kernel is further optimized using one of the variants of the LNV method described in section 3.3. During this process the extremal occupancies are monitored and adaptive purification is applied if necessary to ensure stability;
(vi) once the density kernel has been converged to the desired tolerance the local orbitals are updated, and the kernel optimization is repeated from step (v). The algorithm terminates when the energy is converged with respect to the density kernel and the local orbitals.

The density kernel optimization in steps (iv) and (v) involves direct minimization of the total energy of the real system of interacting electrons. No density or potential mixing is employed and the method is variational by construction [59].

A key feature is the combination of different methods to ensure that the algorithm is stable. A combination of the LNV method and purification has been proposed before [60] but the refinements of monitoring the occupancies so that additional purification steps are taken only when necessary, and the use of adaptive purification to ensure stability make the ONETEP approach particularly robust without any loss of efficiency.

## 6. Conclusions

An overview of the variety of $\mathrm{O}(N)$ methods for optimizing the density kernel within the representation of a fixed set of local orbitals has been given. The ONETEP code uses a combination of these methods to ensure robust and efficient minimization of the total energy with respect to the density kernel. The use of generalized Rayleigh quotients to monitor extremal occupancies and, in conjunction with the folded spectrum method, to locate the chemical potential, is a key tool in the successful application of these methods.

## Acknowledgment

PDH and C-KS acknowledge the support of University Research Fellowships from the Royal Society.

## References

[1] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** 864
[2] Kohn W and Sham L J 1965 *Phys. Rev.* **140** 1133
[3] Kohn W 1996 *Phys. Rev. Lett.* **76** 3168
[4] Prodan E and Kohn W 2005 *Proc. Natl Acad. Sci. USA* **102** 11635
[5] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2005 *J. Chem. Phys.* **122** 084119
[6] Haynes P D, Skylaris C-K, Mostofi A A and Payne M C 2006 *Phys. Status Solidi* b **243** 2489
[7] Soler J M, Artacho E, Gale J D, García A, Junquera J, Ordejón P and Sánchez-Portal D 2002 *J. Phys.: Condens. Matter* **14** 2745
[8] Bowler D R, Miyazaki T and Gillan M J 2002 *J. Phys.: Condens. Matter* **14** 2781
[9] Bowler D R, Choudhury R, Gillan M J and Miyazaki T 2006 *Phys. Status Solidi* b **243** 989
[10] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085
[11] Galli G 1996 *Curr. Opin. Solid State Mater. Sci.* **1** 864
[12] Goedecker S and Colombo L 1994 *Phys. Rev. Lett.* **73** 122
[13] Goedecker S and Teter M 1995 *Phys. Rev.* B **51** 9455
[14] Yang W 1991 *Phys. Rev. Lett.* **66** 1438
[15] Yang W and Lee T-S 1995 *J. Chem. Phys.* **103** 5674
[16] Mauri F, Galli G and Car R 1993 *Phys. Rev.* B **47** 9973
[17] Ordejón P, Drabold D A, Grumbach M P and Martin R M 1993 *Phys. Rev.* B **48** 14646
[18] Mauri F and Galli G 1994 *Phys. Rev.* B **50** 4316
[19] Ordejón P, Drabold D A, Martin R M and Grumbach M P 1995 *Phys. Rev.* B **51** 1456
[20] Kim J, Mauri F and Galli G 1995 *Phys. Rev.* B **52** 1640
[21] Yang W 1997 *Phys. Rev.* B **56** 9294
[22] des Cloizeaux J 1964 *Phys. Rev.* **135** A685
[23] He L and Vanderbilt D 2001 *Phys. Rev. Lett.* **86** 5341
[24] McWeeny R 1960 *Rev. Mod. Phys.* **32** 335
[25] Galli G and Parrinello M 1992 *Phys. Rev. Lett.* **69** 3547
[26] Hernández E and Gillan M J 1995 *Phys. Rev.* B **51** 10157
[27] Skylaris C-K, Mostofi A A, Haynes P D, Diéguez O and Payne M C 2002 *Phys. Rev.* B **66** 035119
[28] Mostofi A A, Haynes P D, Skylaris C-K and Payne M C 2003 *J. Chem. Phys.* **119** 8842
[29] Mostofi A A, Skylaris C-K, Haynes P D and Payne M C 2002 *Comput. Phys. Commun.* **147** 788
[30] Sankey O F and Niklewski D J 1989 *Phys. Rev.* B **40** 3979
[31] Horsfield A P 1997 *Phys. Rev.* B **56** 6594
[32] Janak J F 1978 *Phys. Rev.* B **18** 7165
[33] Haynes P D and Payne M C 1998 *Solid State Commun.* **108** 737
[34] Haynes P D and Payne M C 1999 *Phys. Rev.* B **59** 12173
[35] Kryachko E S 2000 *Chem. Phys. Lett.* **318** 210
[36] Holas A 2001 *Chem. Phys. Lett.* **340** 552
[37] Habershon S and Manby F R 2002 *Chem. Phys. Lett.* **354** 527
[38] Palser A H R and Manolopoulos D E 1998 *Phys. Rev.* B **58** 12704
[39] Niklasson A M N, Tymczak C J and Challacombe M 2003 *J. Chem. Phys.* **118** 8611
[40] Li X-P, Nunes R W and Vanderbilt D 1993 *Phys. Rev.* B **47** 10891
[41] Nunes R W and Vanderbilt D 1994 *Phys. Rev.* B **50** 17611
[42] Daw M S 1993 *Phys. Rev.* B **47** 10895
[43] Millam J M and Scuseria G E 1997 *J. Chem. Phys.* **106** 5569
[44] Challcombe M 1999 *J. Chem. Phys.* **110** 2332
[45] Hernández E, Gillan M J and Goringe C M 1996 *Phys. Rev.* B **53** 7147
[46] MacDonald J K L 1934 *Phys. Rev.* **46** 828
[47] Wang L-W and Zunger A 1994 *J. Phys. Chem.* **98** 2158
[48] Wang L-W and Zunger A 1994 *J. Chem. Phys.* **100** 2394
[49] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2006 *Phys. Status Solidi* b **243** 973
[50] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2005 *J. Phys.: Condens. Matter* **17** 5757
[51] Heady L, Fernandez-Serra Marivi, Mancera R L, Joyce S, Venkitaraman A R, Artacho E, Skylaris C-K, Colombi Ciacchi L and Payne M C 2006 *J. Med. Chem.* **49** 5141
[52] Skylaris C-K and Haynes P D 2007 *J. Chem. Phys.* **127** 164712
[53] Artacho E and Miláns del Bosch L 1991 *Phys. Rev.* A **43** 5770
[54] White C A, Maslen P, Lee M S and Head-Gordon M 1997 *Chem. Phys. Lett.* **276** 133
[55] Ozaki T 2001 *Phys. Rev.* B **64** 195110
[56] Gan C K, Haynes P D and Payne M C 2001 *Comput. Phys. Commun.* **134** 33
[57] Benzi M and Meyer C D 1995 *SIAM J. Sci. Comput.* **16** 1159
[58] Benzi M, Meyer C D and Tůma M 1996 *SIAM J. Sci. Comput.* **17** 1135
[59] Skylaris C-K, Diéguez O, Haynes P D and Payne M C 2002 *Phys. Rev.* B **66** 073103
[60] Bowler D R and Gillan M J 1999 *Comput. Phys. Commun.* **120** 95