Publication IV

K. Baarman, V. Havu, and T. Eirola. Direct minimization for ensemble electronic structure calculations. *arXiv:1204.1205*, 20 pages, April 2012.

Direct Minimization for Ensemble Electronic Structure Calculations

K. Baarman*, V. Havu†, and T. Eirola*

July 11, 2012

Abstract

We consider a direct optimization approach for ensemble density functional theory electronic structure calculations. The update operator for the electronic orbitals takes the structure of the Stiefel manifold into account and we present an optimization scheme for the occupation numbers that ensures that the constraints remain satisfied. We also compare sequential and simultaneous quasi-Newton and nonlinear conjugate gradient optimization procedures, and demonstrate that simultaneous optimization of the electronic orbitals and occupation numbers improve performance compared to the sequential approach.

1 Introduction

Advances in computer power and numerical methods during the past few decades has dramatically increased the scope of electronic structure problems that can be computationally studied. Kohn-Sham density functional theory (DFT) methods can be used to reach precision comparable to experimental accuracy for insulators and semiconductors, while metallic systems remain more challenging. Metallic systems lack a gap between occupied and unoccupied electronic states in the energy spectrum, which leads to slower convergence compared to insulators and semiconductors. Smearing of the Fermi surface is often used to enable convergence of metallic systems as well as insulators at positive temperatures. Ensemble DFT permits direct computation of the occupation numbers of the orbitals based on the entropic term in the Helmholtz free energy. We consider an optimization problem where the target functional A corresponds to the Helmholtz free energy and the variables $\mathbf{X}$ and $\mathbf{f}$ to the electronic orbitals and occupation numbers respectively.

The optimization problem is therefore

$$\text{minimize } A(\mathbf{X}, \mathbf{f}), \tag{1}$$

*Department of Mathematics and Systems Analysis, Aalto University School of Science, Espoo, Finland, e-mail: kurt.baarman@aalto.fi

†Department of Applied Physics, Aalto University School of Science, Espoo, Finland

subject to

$$\mathbf{X} \in \mathcal{M} = \{\mathbf{X} \in \mathbb{R}^{m\times n} \mid \mathbf{X}^T\mathbf{X} = \mathbf{I}\}. \tag{2}$$

Furthermore $\mathbf{f} \in \mathbb{R}^n$ with $\sum_{i=1}^n f_i = n_e$ and $0 \leq f_i \leq 1$, where $n_e \in \mathbb{N}$ is the number of electrons in the system and $n_e \leq n$. We also assume that $\nabla_{\mathbf{X}} A(\mathbf{X}, \mathbf{f})$ and $\nabla_{\mathbf{f}} A(\mathbf{X}, \mathbf{f})$ are available, but expensive to compute. However, due to the form of $A(\mathbf{X}, \mathbf{f})$ the price to compute A, $\nabla_{\mathbf{X}} A$, and $\nabla_{\mathbf{f}} A$ simultaneously is comparable to computing one of them separately. Furthermore we assume that $m \gg n$, and that m is sufficiently large as to make storage of and operation with full $m \times m$ matrices prohibitively expensive.

The orthogonality constraint on $\mathbf{X}$ means that $\mathcal{M} \subset \mathbb{R}^{m\times n}$ defines the Stiefel manifold, which has the tangent space

$$\mathcal{T}_{\mathbf{X}}\mathcal{M} = \{\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{Z} \mid \mathbf{A}^T = -\mathbf{A} \text{ and } \mathbf{Z}^T\mathbf{X} = \mathbf{0}\}, \tag{3}$$

where $\mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{m\times n}$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$. We use the standard inner product

$$(\mathbf{X}, \mathbf{Y}) = \text{trace}(\mathbf{X}^T\mathbf{Y}). \tag{4}$$

Given an arbitrary matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ we can orthogonally project it onto $\mathcal{T}_{\mathbf{X}}\mathcal{M}$ with

$$\mathbf{Y} = \mathbf{P}_{\mathbf{X}}(\mathbf{W}) = (\mathbf{I} - \tfrac{1}{2}\mathbf{X}\mathbf{X}^T)\mathbf{W} - \tfrac{1}{2}\mathbf{X}\mathbf{W}^T\mathbf{X}. \tag{5}$$

Minimization approaches to non-temperature dependent DFT do not in general permit fractional occupation of electronic orbitals [4, 15, 24–26, 28]. In contrast, explicit minimization with regards to occupation numbers permits fractional occupation based on the entropy functional of the Helmholtz free energy and can improve convergence, especially for metallic systems [6,13,19]. It is also possible to transform Equation (1) into a nonlinear eigenvalue problem that can be solved through a self consistent field iteration [15, 18, 23, 24]. The absence of well separated occupied and unoccupied orbitals make metallic systems challenging to compute, and broadening of the Fermi surface is used to facilitate convergence [5, 27]. This broadening is often achieved by assigning the orbitals close to the Fermi level a fractional occupation number determined by the energy of the electronic orbital [14, 20, 21]. Direct minimization on the other hand does not require the orbital energies to be computed at every step, and these broadening schemes are therefore not well suited for minimization methods.

In [11] a framework for optimization methods on the Stiefel and Grassmann manifolds is presented, while [9] discusses a Newton-like iteration scheme on a more general manifold. Univariate optimization methods for the Stiefel manifold is presented in [7], where identity plus rank one Householder transforms are given as one possible choice for moving on the manifold. The choice of coordinates can also be based on a QR factorization and polar decompositions [8, 10] or Lie groups [16]. An overview of geometric numerical integration techniques can be found in [17].

In Section 2 we first recall the nonlinear conjugate gradient and the quasi-Newton methods adapted for use on the Stiefel manifold. We then present an optimization procedure for the occupation numbers and end the section by

presenting a simultaneous orbital-occupation optimization strategies. Then, in Section 3 we numerically demonstrate the method on a model problem that includes nonlinearities similar to a DFT problem. The conclusions are finally presented in Section 4.

2 Optimization with orthogonality constraints

2.1 Update and transport

We ensure that $\mathbf{X}_{k+1}$ satisfies the orthogonality constraint by using a unitary update operator $\mathbf{U}$ which maps $\mathcal{M} \to \mathcal{M}$. A search direction $\mathbf{Y} \in \mathcal{T}_{\mathbf{X}}\mathcal{M}$ given by an optimization procedure can be written

$$\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{Q}\mathbf{R}, \tag{6}$$

where $\mathbf{Q} \in \mathbb{R}^{m\times n}$, $\mathbf{A}, \mathbf{R} \in \mathbb{R}^{n\times n}$, $\mathbf{A}^T = -\mathbf{A}$, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and $\mathbf{Q}^T\mathbf{X} = \mathbf{0}$. If the terms in Equation (6) are not full rank the size of the matrices can be adjusted accordingly.

If we follow $\mathbf{Y}$ to update $\mathbf{X}$ along a Stiefel geodesic we obtain the update operator for $\mathbf{X}$ [11]

$$\mathbf{U} = \begin{bmatrix}\mathbf{X} & \mathbf{Q}\end{bmatrix} \exp\left(\tau \begin{bmatrix}\mathbf{A} & -\mathbf{R}^T \\ \mathbf{R} & \mathbf{0}\end{bmatrix}\right) \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}^T, \tag{7}$$

with step length parameter τ. The update operator generalized for an arbitrary matrix in $\mathrm{span}(\mathbf{X}, \mathbf{Q})$ is

$$\mathbf{U} = \begin{bmatrix}\mathbf{X} & \mathbf{Q}\end{bmatrix} \exp\left(\tau \begin{bmatrix}\mathbf{A} & -\mathbf{R}^T \\ \mathbf{R} & \mathbf{0}\end{bmatrix}\right) \begin{bmatrix}\mathbf{X} & \mathbf{Q}\end{bmatrix}^T, \tag{8}$$

where the orthogonality of $\mathbf{X}$ and $\mathbf{Q}$ has been exploited.

In order to use information gained from previous evaluations of A and ∇A we must take $\mathcal{M}$ into account. This requires us to transport vectors $\mathbf{Y} \in \mathcal{T}_{\mathbf{X}}\mathcal{M}$ to $\mathcal{T}_{\mathbf{U}\mathbf{X}}\mathcal{M}$ with the transport operator

$$\mathbf{T} = \mathbf{I}_m + \begin{bmatrix}\mathbf{X} & \mathbf{Q}\end{bmatrix} \left(\exp\left(\tau \begin{bmatrix}\mathbf{A} & -\mathbf{R}^T \\ \mathbf{R} & \mathbf{0}\end{bmatrix}\right) - \mathbf{I}_{2n}\right) \begin{bmatrix}\mathbf{X} & \mathbf{Q}\end{bmatrix}^T. \tag{9}$$

Here $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ and $\mathbf{I}_{2n} \in \mathbb{R}^{2n\times 2n}$, and $\mathbf{T}$ does not modify matrices $\mathbf{Z}$ that satisfy $\begin{bmatrix}\mathbf{X} & \mathbf{Q}\end{bmatrix}^T \mathbf{Z} = \mathbf{0}$.

Remark 1: The closely related Grassmann manifold is identical to the Stiefel manifold with the addition of the homogeneity condition $A(\mathbf{X}) = A(\mathbf{X}\mathbf{Q})$, where $\mathbf{Q}$ is orthogonal. The homogeneity condition is satisfied for orbitals with identical occupation numbers, but does not generally hold for ensemble DFT. A discussion of direct minimization with integer occupation numbers is presented in [1].

2.2 Nonlinear conjugate gradients

The conjugate gradient (CG) method can be viewed as an optimization method for a quadratic problem. Several generalizations of the CG method have been presented to solve optimization problems that are not quadratic [22]. Below, we review a nonlinear CG (NLCG) method adapted to account for the curvature of the manifold [11].

Given $\mathbf{X}_0$ which satisfies $\mathbf{X}_0^T\mathbf{X}_0 = \mathbf{I}$, the gradient projected onto $\mathcal{T}_{\mathbf{X}_0}\mathcal{M}$ is

$$\mathbf{F}_0 = \mathbf{P}_{\mathbf{X}_0}(\nabla_{\mathbf{X}} A(\mathbf{X}_0, \mathbf{f}_0)), \tag{10}$$

and the initial search direction is the direction of steepest descent

$$\mathbf{Y}_0 = -\mathbf{F}_0. \tag{11}$$

On the manifold the NLCG method then proceeds by minimizing A along the path defined by the search direction $\mathbf{Y}_k$. In practice we evaluate A once along the search direction and construct a quadratic approximation that we minimize. The step length, τ_k, that minimizes A along the search direction is then used to update $\mathbf{X}_k$ such that

$$\mathbf{X}_{k+1} = \mathbf{T}(\tau_k)\mathbf{X}_k, \tag{12}$$

and the gradient and search directions are transported to $\mathcal{T}_{\mathbf{X}_{k+1}}\mathcal{M}$ by $\mathbf{T}(\tau_k)$. The new projected gradient

$$\mathbf{F}_{k+1} = \mathbf{P}_{\mathbf{X}_{k+1}}(\nabla_{\mathbf{X}} A(\mathbf{X}_{k+1}, \mathbf{f}_{k+1})), \tag{13}$$

and search direction

$$\mathbf{Y}_{k+1} = -\mathbf{F}_{k+1} + \gamma_k \mathbf{T}(\tau_k)\mathbf{Y}_k, \tag{14}$$

are then computed where

$$\gamma_k = \frac{(\mathbf{F}_{k+1} - \mathbf{T}(\tau_k)\mathbf{F}_k, \mathbf{F}_{k+1})}{(\mathbf{F}_k, \mathbf{F}_k)}. \tag{15}$$

The step length is determined by the minimizer of a quadratic approximation of A along the search direction. The quadratic approximation is constructed by taking a trial step length $\tau_e = \frac{1}{10}\max(\tau_{\min}, \tau_{k-1})$, where $\tau_{\min}$ is a predefined minimum trial step length and computing

$$\begin{aligned} p(0) &= A(\mathbf{X}, \mathbf{f}), \\ p(\tau_e) &= A(\mathbf{T}(\tau_e)\mathbf{X}, \mathbf{f}), \\ p'(0) &= (\mathbf{Y}, \nabla_{\mathbf{X}} A(\mathbf{X}, \mathbf{f})). \end{aligned} \tag{16}$$

Then solve τ_k and limit it by $2\tau_{k-1}$, and construct the update $\mathbf{T}(\tau_k)$. This approximate line search requires one extra evaluation of A per step.

2.3 Quasi-Newton method

The quasi-Newton (QN) method is similar to Newton's method, but replaces the inverse Hessian with an approximation. This is frequently possible even when the Hessian is not available, and can still be used to improve performance for a badly conditioned minimization problem.

We base the QN method on Broyden's second or *bad* generalized update to construct the approximate inverse Hessian, $\mathbf{G}$, of A at $\mathbf{X}_k$. While Broyden's second update does not construct a symmetric approximation, or ensure that the approximation is positive definite it is a robust update choice for electronic structure calculations [2, 18]. Furthermore, $\mathbf{X}$ and $\nabla_{\mathbf{X}} A$ are $\mathbb{R}^{m\times n}$ matrices, which we take into account when constructing the generalized Broyden update. The secant condition is then

$$\mathbf{G}\Delta\mathbf{\Phi} = \Delta\mathbf{\Xi}, \tag{17}$$

where $\Delta\mathbf{\Phi}$ and $\Delta\mathbf{\Xi}$ are the collected orbital gradient and position differences projected onto the tangent space and transported to $\mathcal{T}_{\mathbf{X}_k}\mathcal{M}$. That is

$$\Delta\mathbf{\Phi} = \begin{bmatrix}\Delta\mathbf{F}_{k-1} & \mathbf{T}(\tau_{k-1})\Delta\mathbf{F}_{k-2} & \dots & \mathbf{T}(\tau_{k-1})\dots\mathbf{T}(\tau_{l+1})\Delta\mathbf{F}_{l}\end{bmatrix}, \tag{18}$$

and

$$\Delta\mathbf{\Xi} = \begin{bmatrix}\Delta\mathbf{X}_{k-1} & \mathbf{T}(\tau_{k-1})\Delta\mathbf{X}_{k-2} & \dots & \mathbf{T}(\tau_{k-1})\dots\mathbf{T}(\tau_{l+1})\Delta\mathbf{X}_{l}\end{bmatrix}, \tag{19}$$

for history length $k-l$. Here the gradient differences projected onto $\mathcal{T}_{\mathbf{X}_{i+1}}\mathcal{M}$ are

$$\Delta\mathbf{F}_i = \mathbf{F}_{i+1} - \mathbf{T}(\tau_i)\mathbf{F}_i, \tag{20}$$

and $\mathbf{F}_i$ is like in (13),

$$\mathbf{F}_i = \mathbf{P}_{\mathbf{X}_i}(\nabla_{\mathbf{X}} A(\mathbf{X}_i, \mathbf{f}_i)). \tag{21}$$

The projected occupation weighted orbital differences are

$$\Delta\mathbf{X}_i = \mathbf{P}_{\mathbf{X}_{i+1}}\big(\mathbf{X}_{i+1}\operatorname{diag}(\mathbf{f}_{i+1}) - \mathbf{X}_i\operatorname{diag}(\mathbf{f}_i)\big), \tag{22}$$

and the motivation for including the weight is that the unoccupied electronic orbitals do not contribute to the energy of the system. The no change condition is now

$$\mathbf{Z} = \mathbf{G}\mathbf{Z} \quad \forall\, \mathbf{Z} \text{ such that } \mathbf{Z}^T\Delta\mathbf{\Phi} = \mathbf{0}. \tag{23}$$

The secant and no change condition together correspond to the generalized Broyden's second update where all single orbital secant conditions are simultaneously enforced for the entire history length. We can therefore use the generalized update formula [12]

$$\mathbf{G} = \mu\mathbf{I} + (\Delta\mathbf{\Xi} - \mu\Delta\mathbf{\Phi})(\Delta\mathbf{\Phi}^T\Delta\mathbf{\Phi})^{-1}\Delta\mathbf{\Phi}^T, \tag{24}$$

where dropping the empty orbitals ensure that $\Delta\mathbf{\Phi}^T\Delta\mathbf{\Phi}$ is nonsingular in practice. The search direction given by the QN method is then

$$\mathbf{Y} = -\mathbf{G}\mathbf{F}, \tag{25}$$

and

$$\mathbf{X}_{k+1} = \mathbf{U}(\tau_k)\mathbf{X}_k, \tag{26}$$

where $\mathbf{Y}$ determines $\mathbf{U}$ as in Section (2.1). The line search is identical to the one described for the NLCG method in Section 2.2 with the addition of the constant underrelaxation $\beta_{\mathbf{X}} \in]0, 1]$ that we have included in the step length τ_k.

In practice only the last few history steps contribute significantly to the rate of convergence. Consequently, we discard the oldest trial solutions and gradient information when a predetermined history length is reached.

2.4 Optimization of occupation numbers

Given a set of electronic orbitals $\mathbf{X}$ it is possible to further reduce A by optimizing $\mathbf{f}$. Forcing occupation towards a uniform distribution increases contributions to A from higher energy states, while simultaneously increasing the entropy which contributes to a reduction of A at nonzero temperatures. The relative strength of both of these effects determine the ground state of the system, and can lead to nonzero occupation of higher energy states at positive temperatures or due to nonlinear effects.

Therefore, given $\mathbf{X}$, we want to find $\mathbf{f}$ that minimizes A. To keep the number of particles constant we determine the search direction $\mathbf{y}$ which is the vector closest $-\nabla_{\mathbf{f}} A(\mathbf{X}, \mathbf{f})$ that ensures that the conditions $\sum_{i=1}^{n} f_i = n_e$ and $0 \leq f_i \leq 1$ remain satisfied. To this end we solve

$$\text{minimize } \|\mathbf{y} + \nabla_{\mathbf{f}} A(\mathbf{X}, \mathbf{f})\|, \tag{27}$$

with the constraints $\sum_{i=1}^{n} y_i = 0$, $y_i \leq 0$ if $f_i = 1$, and $y_i \geq 0$ if $f_i = 0$. The first constraint on $\mathbf{y}$ ensures that the minimization step conserves electrons while the second and third condition prohibits unphysical occupation numbers. In practice we use the `quadprog` routine available in MATLAB to solve this problem. Given the search direction $\mathbf{y}$ we minimize A by constructing a quadratic approximation similar to (16).

After we have solved $\mathbf{y}$ the occupation step length σ_k is determined like in Section 2.2 with the addition of the constant underrelaxation $\beta_{\mathbf{f}} \in]0, 1]$ included in σ_k. In addition, we ensure that the occupation remains physical by limiting σ_k with σ_{M} such that $0 \leq f_i + \sigma_{\mathrm{M}} y_i \leq 1$ for all i.It is possible to take a longer step than σ_{M} by recomputing $\mathbf{y}$ from Equation (27) with the updated boundary information when an entry in $\mathbf{f}$ reaches the boundary of physical occupation, 0 or 1. However, convergence of occupation numbers is faster than orbital convergence, and the numbers of steps needed for convergence is therefore determined by the orbital convergence. Furthermore, if the occupation number of the least populated orbital has been less than 10^{-12} on two consecutive iterations we drop the associated orbitals.

2.5 Simultaneous step size selection

Typically an ensemble DFT problem is solved by sequentially optimizing the orbitals with fixed occupation numbers and then fixing the orbitals and optimizing the occupation numbers. This process is then repeated until a satisfactory solution is obtained.

The cost of evaluating A, $\nabla_{\mathbf{X}} A$, and $\nabla_{\mathbf{f}} A$ is comparable to evaluating one of them separately, and simultaneous optimization of A with respect to $\mathbf{X}$ and $\mathbf{f}$ can for this reason potentially reduce computational effort.

Given a pair of search directions $(\mathbf{Y},\ \mathbf{y})$ for the orbitals and occupation numbers respectively and starting guesses for step lengths, τ_{k-1} and σ_{k-1} we evaluate A and its gradients with the following trial step lengths

$$\tau_e = \tfrac{1}{10} \max(\tau_{\min}, \tau_{k-1}), \tag{28}$$

and

$$\sigma_e = \min(\sigma_{\mathrm{M}}, \tfrac{1}{10} \max(\sigma_{\min}, \sigma_{k-1})). \tag{29}$$

Here $\tau_{\min}$ and $\sigma_{\min}$ are minimum trial step lengths. With this we construct a quadratic surface approximation

$$p(\tau, \sigma) = c_1\tau^2 + c_2\sigma^2 + c_3\tau + c_4\sigma + c_5, \tag{30}$$

that we use to simultaneously update both $\mathbf{X}$ and $\mathbf{f}$ by evaluation in one trial point. This surface is determined by the system of equations

$$\begin{aligned}
p(0,0) &= A(\mathbf{X}, \mathbf{f}), \\
p_\tau(0,0) &= (\mathbf{Y}, \nabla_{\mathbf{X}} A(\mathbf{X}, \mathbf{f})), \\
p_\sigma(0,0) &= (\mathbf{y}_0, \nabla_{\mathbf{f}} A(\mathbf{X}, \mathbf{f})), \\
p_\tau(\tau_e, \sigma_e) &= (\mathbf{Y}, \nabla_{\mathbf{X}} A(\mathbf{T}(\tau_e)\mathbf{X}, \mathbf{f} + \sigma_e \mathbf{y})), \\
p_\sigma(\tau_e, \sigma_e) &= (\mathbf{y}_{\sigma_e}, \nabla_{\mathbf{f}} A(\mathbf{T}(\tau_e)\mathbf{X}, \mathbf{f} + \sigma_e \mathbf{y})).
\end{aligned} \tag{31}$$

Solving this system and finding the minimums gives the optimal step lengths $\hat{\tau}_k$ and $\hat{\sigma}_k$ for the quadratic approximation of the search directions. For the simultaneous NLCG method the step lengths are then $\tau_k = \hat{\tau}_k$ and $\sigma_k = \hat{\sigma}_k$ while the QN method uses $\tau_k = \beta_{\mathbf{X}} \hat{\tau}_k$ and $\sigma_k = \beta_{\mathbf{f}} \hat{\sigma}_k$, where $\beta_{\mathbf{X}}, \beta_{\mathbf{f}} \in \left]0, 1\right]$ are constant underrelaxation parameters. We then simultaneously update $\mathbf{X}$ and $\mathbf{f}$ with $\mathbf{X}_{k+1} = \mathbf{T}(\tau_k)\mathbf{X}_k$ and $\mathbf{f}_{k+1} = \mathbf{f}_k + \min(\sigma_{\mathrm{M}}, \sigma_k)\mathbf{y}$ respectively.

We then simultaneously update X and $\mathbf{f}$ with $\mathbf{X}_{k+1} = \mathbf{T}(\tau_k)\mathbf{X}_k$ and $\mathbf{f}_{k+1} = \mathbf{f}_k + \min(\sigma_{\mathrm{M}}, \sigma_k)\mathbf{y}$ respectively.

Remark 2: The surface (30) is determined by computing $\nabla_{\mathbf{X}} A$ and $\nabla_{\mathbf{f}} A$ at the trial step. The system of equations (31) could alternatively be determined by computing both A and $\nabla_{\mathbf{X}} A$ or A and $\nabla_{\mathbf{f}} A$ at $(\mathbf{T}(\tau_e)\mathbf{X}, \mathbf{f} + \sigma_e \mathbf{y})$.

Remark 3: Inclusion of the $\tau\sigma$ cross term would require an extra trial evaluation point for system (31) to be linearly independent.

3 Numerical experiments

We use a two dimensional model problem to compare the sequential and simultaneous NLCG and QN methods. This model problem is inspired by ensemble DFT, and corresponds to a three dimensional system constrained to two dimensions without spin effects and exchange-correlation terms while taking entropy into account. The model problem adapted from Reference [19] is

$$A(\mathbf{X}, \mathbf{f}) = -\tfrac{1}{2}\mathrm{tr}\big(\mathbf{X}^T\mathbf{L}\mathbf{X}\,\mathrm{diag}(\mathbf{f})\big) + \mathbf{v}_{\mathrm{ext}}^T\mathbf{n} + \tfrac{1}{2}\mathbf{v}_{\mathrm{int}}^T\mathbf{n} - TS(\mathbf{f}). \tag{32}$$

Here $\mathbf{L} \in \mathbb{R}^{m\times m}$ is the discretized Laplace operator, $\mathbf{v}_{\mathrm{ext}} \in \mathbb{R}^m$ the external potential, $\mathbf{n} \in \mathbb{R}^m$ the electron density, $\mathbf{v}_{\mathrm{int}} = \mathbf{V}\mathbf{n}$ the Hartree potential corresponding to the electron density $\mathbf{n}$, T to temperature, and S is the entropy. The electron density is

$$\mathbf{n} = (\mathbf{X} \circ \mathbf{X})\,\mathbf{f}, \tag{33}$$

where $\circ$ is the entrywise, or Hadamard, product. The entropy term is

$$S(\mathbf{f}) = -\sum_{i=1}^{n} f_i \ln(f_i + \delta(1 - f_i)) + (1 - f_i)\ln(1 - f_i + \delta f_i), \tag{34}$$

where $\delta > 0$ is a small regularization parameter that ensures that the derivative of S remains finite.

To calculate the potentials we use

$$(\mathbf{v}_{\mathrm{ext}})_i = -\sum_{j=1}^{N} \frac{Z_j}{\|\mathbf{r}_i - \mathbf{R}_j\| + \alpha}, \tag{35}$$

where the sum is over the nuclei with charge Z_j and position $\mathbf{R}_j$. The position corresponding to the discretization point i is $\mathbf{r}_i$, and the parameter α is used to regularize the potential. $\mathbf{V} \in \mathbb{R}^{m\times m}$ is similarly given by

$$\mathbf{V}_{ij} = \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\| + \alpha}. \tag{36}$$

We solve the problem in the unit square with zero boundary conditions corresponding to an infinite potential well. We use a uniform finite difference discretization with m inner points to obtain a system where $\mathbf{X} \in \mathbb{R}^{m\times n}$. Here n corresponds to the number of electronic orbitals in the calculation. As initial guess we use the solution of the quadratic problem using the first two terms of (32). The occupation numbers are initialized to

$$f_i = \tfrac{n_e}{n} + \tfrac{1}{2}\Delta\frac{n + 1 - 2i}{n + 1}, \tag{37}$$

where $\Delta = \min(n_e/n, 1 - n_e/n)$ and $n_e \leq n$ is the number of electrons. This choice ensures that the initial occupation of all orbitals is nonzero and emphasizes lower energy orbitals.

We demonstrate the methods for three external potentials. For all models we use potential regularization $\alpha = 5 \times 10^{-2}$ and entropy regularization $\delta = 10^{-3}$.

The first model is a single nucleus with charge $Z = 2$ centered at the center of the unit square with two electrons. For this system the second and third orbitals are degenerate. We calculate the model with 10 electronic orbitals and a first order finite difference discretization with 25 interior points in one dimension resulting in $m = 625$ spatial degrees of freedom. We will refer to this system as Z_2.

The second model, which we name Z_3-Z_2, consists of two nuclei, with a nuclei of charge $Z = 3$ placed at $(\frac{1}{3}, \frac{1}{3})$ and another with charge $Z = 2$ placed at $\frac{2}{3}, \frac{2}{3}$ and 5 electrons. This system has four well separated electronic orbitals, while the fifth and sixth are relatively close. The computation is initialized with 10 orbitals and 29 interior grid points in one dimension for $m = 841$.

The last model, Z_4-Z_3, consists of two nuclei, $Z = 4$ placed at $(\frac{1}{3}, \frac{1}{3})$, and $Z = 3$ at the grid point closest to $(\frac{2}{3}, \frac{13}{24})$ and 7 electrons. The off diagonal placement is chosen to break the symmetry of the system. This model initially has 14 orbitals and 29 interior grid points in one dimension ($m = 841$).

For the sequential QN orbital minimizer uses the parameters $\beta_{\mathbf{X}} = 0.4$, $\mu = 5 \times 10^{-5}$, and history length 6. The sequential QN and NLCG methods minimum trial step length $\tau_0 = 10^{-3}$ and we perform 6 orbital optimization steps before engaging the occupation number minimizer. Both sequential optimization routines use an identical SD routine with $\beta_{\mathbf{f}} = 0.5$, $\mu = 10^{-4}$ and $\sigma_0 = 10^{-4}$ for occupation number optimization with two optimization steps. We have tried several different combinations of orbital and occupation optimization steps and observed that this combination offers a good compromise. For the SD method μ only serves to scale the approximate line search.

We measure convergence by the energy difference to a reference energy computed by running the simultaneous methods for 3000 steps and the sequential methods for 3000 optimization rounds. We then use the lowest energy obtained as the reference energy.

The change in occupation numbers with rising temperature is graphically presented in Figures 1, 2, and 3. The same data is repeated in Tables 1, 2, and 3. At $T = 0$ the lowest electronic orbitals are fully occupied for Z_4-Z_3, while one electron is split between two degenerate orbitals for Z_2. Even though there is a small gap (1.69×10^{-2}) between the fifth and sixth electron orbitals for Z_3-Z_2 the fifth electron is split (0.55 vs 0.45) between these orbitals. We successfully replicated this split with a 3000 round sequential SD orbital occupation number optimization. Furthermore, restarting the SD iteration with a five orbital initial guess based on the split orbital reference solution results in convergence to a higher energy state.

Figures 4, 5, and 6 illustrate energy convergence for the different methods. The simultaneous methods generally perform better than the sequential methods, and the simultaneous NLCG method is more robust than the simultaneous QN approach. In the energy convergence for the sequential optimization routines the switch between orbital and occupation optimization is readily seen in the steplike energy convergence. Furthermore, the performance of the sequential

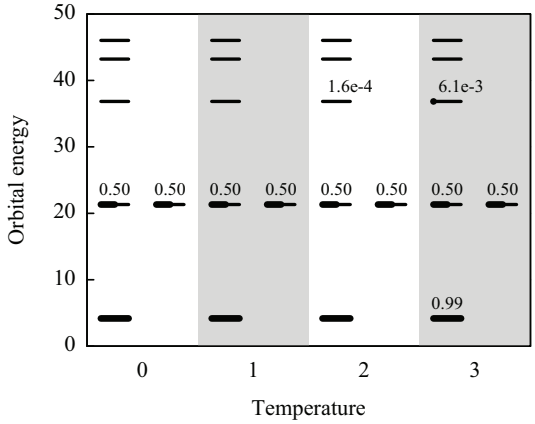

Figure 1: Orbital energy levels with occupation for Z_2 at varying temperatures. Fractional occupation numbers are indicated and the same data is also presented in Table 1.

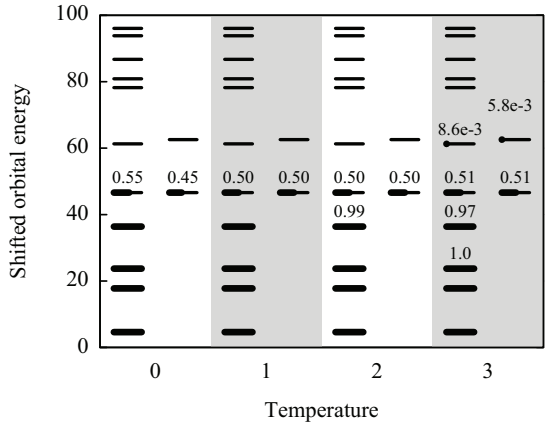

Figure 2: Orbital energy levels with occupation for Z_3-Z_2 at varying temperatures with orbital energy shifted by +5. Fractional occupation numbers are indicated and the same data is also presented in Table 2.

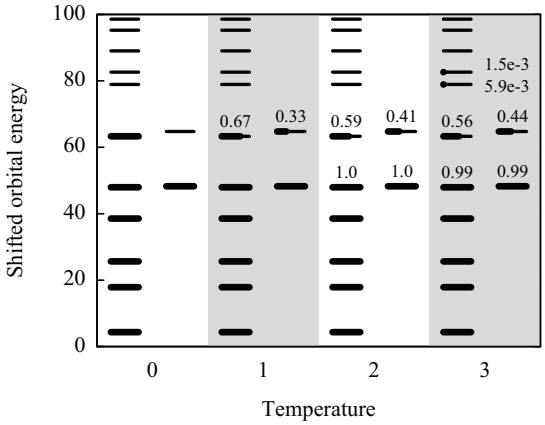

Figure 3: Orbital energy levels with occupation for Z_4-Z_3 at varying temperatures with orbital energy shifted by +10. Fractional occupation numbers are indicated and the same data is also presented in Table 3.

QN and NLCG methods is nearly identical for all models. This might be due to the limited number of step available for orbital optimization before occupation optimization is enabled.

The simultaneous NLCG method outperforms the QN method for the Z_4-Z_3 system shown in Figure 6. Increasing history generally improves the convergence rate of the QN method, but this did not significantly change the rate of convergence for this model. Frequent restarts limit history length and provide at least a partial explanation for this effect. Figure 7 presents Z_4-Z_3 restarts for the simultaneous QN method. Restarts are frequent for this model at all temperatures compared to Z_2 and Z_3-Z_2. However, for $T > 0$ there is generally sufficiently many steps between restarts for the history to grow to full length, and the rate of convergence does improve somewhat.

For the Z_4-Z_3 model the energy difference between the highest occupied and lowest unoccupied orbital is 1.69×10^{-2}, see Figure 3 and Table 3. This difference is comparatively small and could explain the poor performance of the QN method, particularly for $T = 0$. In Figure 8 the convergence rate of the optimization procedures for Z_4-Z_3 for $T = 0.3, 0.5, 0.7$, and the convergence rate for $T = 0$ is included for reference. At $T = 0.3$ the rate of convergence for the QN method is considerably improved and the convergence rate remains superior to $T = 0$ for $T = 0.5$ and $T = 0.7$. The elevated temperature broadens the Fermi surface and this could explain the improved convergence at $T = 0.3$, while the convergence of higher energy orbitals makes the problem more challenging at

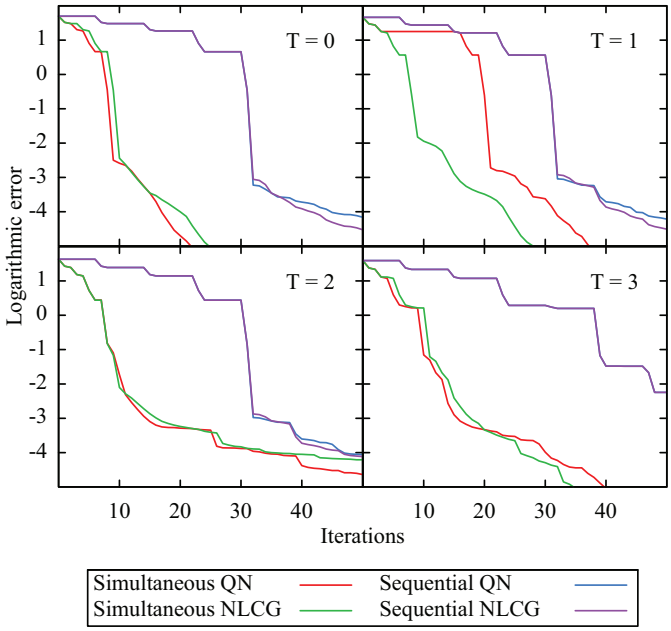

Figure 4: Energy convergence for Z_2 at varying temperatures.

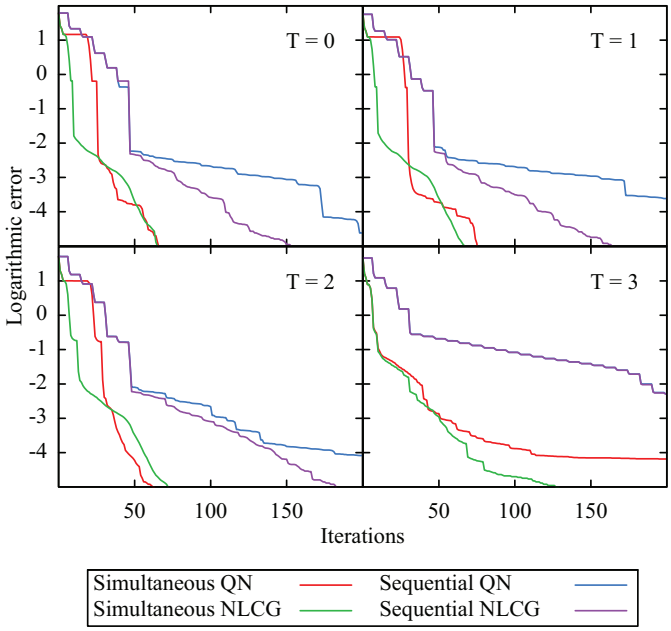

Figure 5: Energy convergence for Z_3-Z_2 at varying temperatures.

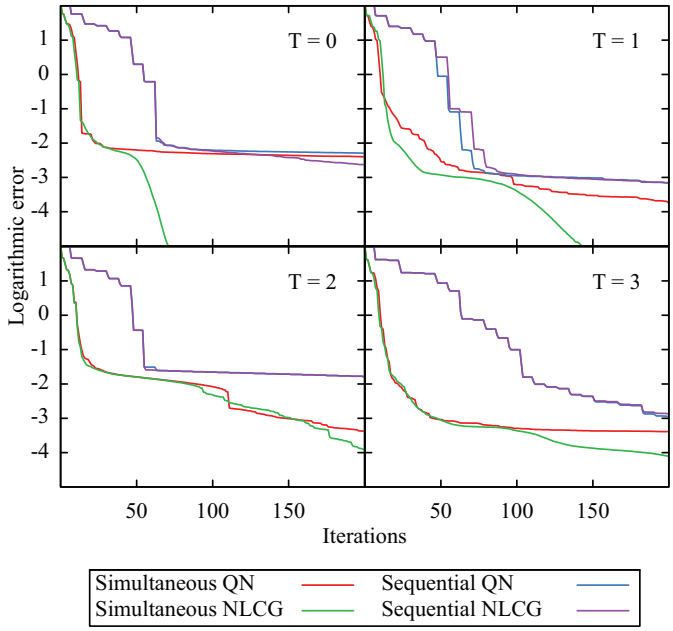

Figure 6: Energy convergence for Z_4-Z_3 at varying temperatures.

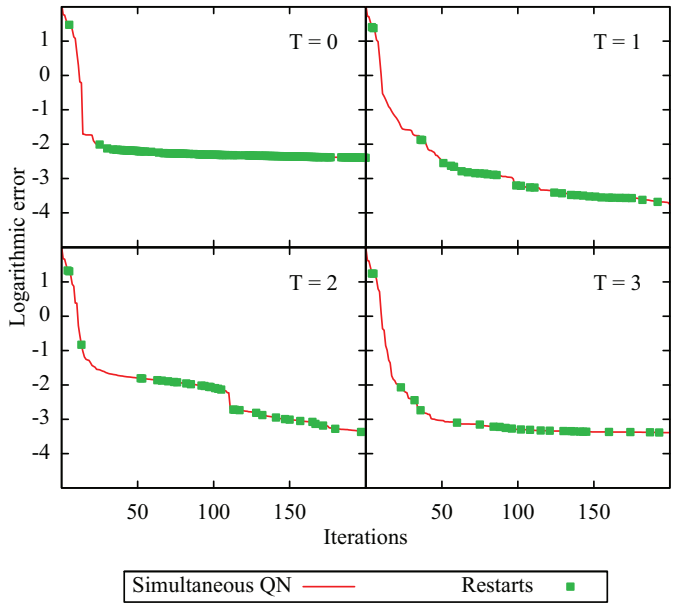

Figure 7: Energy convergence of the simultaneous QN method with restarts for Z_4-Z_3 at varying temperatures.

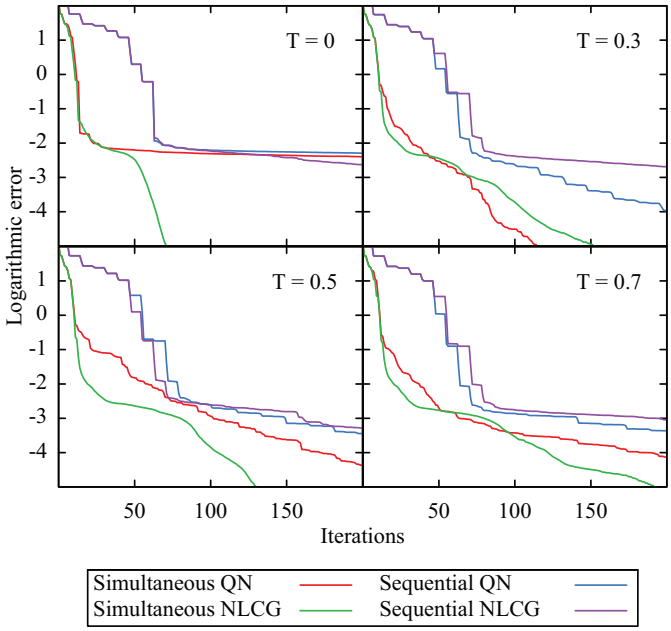

Figure 8: Energy convergence for Z_4-Z_3 at varying temperatures $T < 1$.

higher temperatures. This would also explain the decreasing performance of NLCG for higher temperatures.

Table 1: Orbital energy levels with occupation for Z_2 at varying temperatures. The same data is graphically presented in Figure 1.

E	Occ. (T=0)	Occ. (T=1)	Occ. (T=2)	Occ. (T=3)
4.172259	1.000000	1.000000	1.000000	0.996380
21.328241	0.500000	0.500000	0.499955	0.498751
21.328241	0.500000	0.500000	0.499880	0.498751
36.836577	0.000000	0.000000	0.000165	0.006117
43.225667	0.000000	0.000000	0.000000	0.000000
46.034373	0.000000	0.000000	0.000000	0.000000

Table 2: Orbital energy levels with occupation for Z_3-Z_2 at varying temperatures with orbital energies shifted by +5. The same data is graphically presented in Figure 2.

E	Occ. (T=0)	Occ. (T=1)	Occ. (T=2)	Occ. (T=3)
4.606322	1.000000	1.000000	1.000000	1.000000
17.773445	1.000000	1.000000	1.000000	1.000000
23.744218	1.000000	1.000000	1.000000	0.999833
36.378253	1.000000	1.000000	0.994738	0.970970
46.607469	0.554627	0.504114	0.504757	0.508795
46.624356	0.445373	0.495886	0.500505	0.506011
61.308726	0.000000	0.000000	0.000000	0.008575
62.566830	0.000000	0.000000	0.000000	0.005816
78.212923	0.000000	0.000000	0.000000	0.000000
80.870921	0.000000	0.000000	0.000000	0.000000
86.752316	0.000000	0.000000	0.000000	0.000000
93.853450	0.000000	0.000000	0.000000	0.000000
96.049712	0.000000	0.000000	0.000000	0.000000

4 Conclusion

We have presented two schemes for energy optimization of ensemble DFT computations. The updates take the problem constraints into account and permits us to use information obtained from previous evaluations of the target functional and gradients to improve rate of convergence. We have further demonstrated the methods numerically on a model problem inspired by the electronic structure theory and compared simultaneous and sequential schemes based on the QN and NLCG methods.

The ensemble model successfully concentrates occupation to low energy orbitals at low temperatures, and gradually increases occupation of higher energy orbitals at increasing temperature to increase the entropy of the system. Optimization of the occupation numbers also enables ensemble DFT calculations to automatically handle degenerate and near degenerate orbitals at $T = 0$, which are challenging for methods that construct the electron density by the Aufbau principle. Furthermore, is seems possible to broaden the Fermi surface by increasing temperature to accelerate convergence of small gap systems.

Simultaneous optimization schemes provide improved convergence compared to sequential approaches for both the NLCG and QN methods. While the NLCG and QN methods are often comparable in performance, the NLCG method is overall more robust. In contrast, Reference [3] found that QN method is more robust than the NLCG method. It is possible that the quadratic approximate line search gives a better result for the model problem. As the NLCG method depends heavily on a high quality line search this might provide a possible explanation. In the present case, the QN method performs poorly for problems with frequent restarts and while this effect does not fully explain the lack of convergence it can be used as a problem indicator.

Table 3: Orbital energy levels with occupation for Z_4-Z_3 at varying temperatures with orbital energies shifted by +10. The same data is graphically presented in Figure 3.

E	Occ. (T=0)	Occ. (T=1)	Occ. (T=2)	Occ. (T=3)
4.327524	1.000000	1.000000	1.000000	1.000000
17.873544	1.000000	1.000000	1.000000	1.000000
25.639021	1.000000	1.000000	1.000000	1.000000
38.541992	1.000000	1.000000	1.000000	1.000000
47.960214	1.000000	1.000000	0.999983	0.994464
48.278074	1.000000	1.000000	0.999841	0.993917
63.300484	1.000000	0.669980	0.591678	0.564432
64.759294	0.000000	0.330020	0.408498	0.439697
78.938961	0.000000	0.000000	0.000000	0.005937
82.626842	0.000000	0.000000	0.000000	0.001553
89.029063	0.000000	0.000000	0.000000	0.000000
95.252731	0.000000	0.000000	0.000000	0.000000
98.574017	0.000000	0.000000	0.000000	0.000000

References

[1] K. Baarman, T. Eirola, and V. Havu. Minimization by Householder transforms with orthogonality constraints. arXiv:1204.1204 [physics.comp-ph] 2012.

[2] K. Baarman, T. Eirola, and V. Havu. Robust acceleration of self consistent field calculations in density functional theory. *J. Chem. Phys.*, 134:134109, 2011.

[3] K. Baarman and J. VandeVondele. A compariston of accelerators for direct energy minimization in electronic structure calculations. *J. Chem. Phys.*, 134:244104, 2011.

[4] C. Bekas, E. Kokiopoulou, and Y. Saad. Computation of large invariant subspaces using polynomial filtered Lanczos iterations with applications in density functional theory. *SIAM J. Matrix Anal. Appl.*, 30:397, 2008.

[5] V. Blum, R. Gehrke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Comp. Phys. Commun.*, 180:2175, 2009.

[6] E. Cancès. Self-consistent field algorithms for Kohn-Sham models with fractional occupation numbers. *J. Chem. Phys.*, 114:10616, 2001.

[7] E. Celledoni and S. Fiori. Descent methods for optimization on homogeneous manifolds. *Mathematics and Computers in Simulation*, 79:1298, 2008.

[8] E. Celledoni and B. Owren. A class of intrinsic schemes for orthogonal integration. *SIAM J. Numer. Anal.*, 40:2069, 2002.

[9] M. T. Chu. On a numerical treatment for the curve-tracing of the homotopy method. *Numer. Math*, 42:323, 1983.

[10] L. Dieci and E. S. Van Vleck. Orthonormal integrators based on Householder and Givens transformations. *Future Gener. Comp. Sy.*, 19:363, 2003.

[11] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303, 1998.

[12] H.-r. Fang and Y. Saad. Two classes of multisecant methods for nonlinear acceleration. *Numer. Linear Algebra Appl.*, 16:197, 2009.

[13] C. Freysoldt, S. Boeck, and J. Neugebauer. Direct minimization technique for metals in density functional theory. *Phys. Rev. B*, 79:241103, 2009.

[14] C.-L. Fu and K.-M. Ho. First-principles calculation of the equilibrium ground-state properties of transition metals: Applications to Nb and Mo. *Phys. Rev. B*, 28:5480, 1983.

[15] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comp. Mat. Sci.*, 6:15, 1996.

[16] S. Krogstad. A low complexity lie group method on the Stiefel manifold. *BIT*, 43:107, 2003.

[17] C. Lubich, E. Hairer, and G. Wanner. *Geometric Numerical Integration*. Springer, 2006.

[18] L. D. Marks and D. R. Luke. Robust mixing for ab initio quantum mechanical calculations. *Phys. Rev. B*, 78:075114, 2008.

[19] N. Marzari, D. Vanderbilt, and M. C. Payne. Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators. *Phys. Rev. Lett.*, 79:1337, 1997.

[20] N. D. Mermin. Thermal properties of the inhomogeneous electron gas. *Phys. Rev.*, 137:A1441, 1965.

[21] M. Methfessel and A. T. Paxton. High precision sampling for Brillouin-zone integration in metals. *Phys. Rev. B*, 40:3616, 1989.

[22] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

[23] P. Pulay. Convergence acceleration in iterative sequences: The case of SCF iteration. *Chem. Phys. Lett.*, 73:393, 1980.

[24] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical methods for electronic structure calculations of materials. *SIAM Review*, 52:3, 2010.

[25] T. van Voorhis and M. Head-Gordon. A geometric approach to direct minimization. *Mol. Phys.*, 100:1713, 2002.

[26] J. VandeVondele and J. Hutter. An efficient orbital transformation method for electronic structure calculations. *J. Chem. Phys.*, 118:4365, 2003.

[27] F. Wagner, T. Laloyaux, and M. Scheffler. Errors in Hellmann-Feynman forces due to occupation-number broadening and how they can be corrected. *Phys. Rev. B*, 57:2102, 1998.

[28] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky. Parallel self-consistent-field calculations via Chebyshev-filtered subspace acceleration. *Phys. Rev. E*, 74:066704, 2006.