Density functional theory for large molecular systems

Thesis submitted for the degree of *Philosophiae Doctor*

by

Simen Sommerfelt Reine

CTCC - Centre for Theoretical and Computational Chemistry
Department of Chemistry
Faculty of Mathematics and Natural Sciences
University of Oslo

Acknowledgments

I would first and foremost thank my supervisor Trygve Helgaker, for his invaluable guidance throughout this thesis. This work would not have been the same without his insights and ability to always push a project in the right direction. I would further like to thank my co-supervisor Filip Pawłowski, in particular for the help at the start of my thesis.

The list of people whom have contributed to the work of this thesis is long. I would in this respect like to thank to Poul Jørgensen for financial support for my half-year stay in Århus, Thomas Kjærgaard for the tight and valuable collaboration in the development of a new integral driver, Pawel Sałek for valuable computer assistance and for the contribution to the robust variational density-fitting paper, Erik Tellgren for the help in the early developments of the density-fitting driver, for the many nice and helpful literature reviews and discussions, Stinne Høst for her guidance and help with all problems related to density minimization, and for the development of the unrestricted code used for the density-fitting paper, Andreas Krapp for his contributions both to the density-fitting paper and to the efficient force evaluation paper, Maria Francesca Iozzi for all the help in the, at times frustrating, application of the geometry optimizer, Vebjørn Bakken for the computer assistance at the beginning of my thesis and for the invaluable geometry optimization experience and coding, Branislav Jansik for the help on parallel implementations and for extending the Löwdin decomposition to be used for the density-fitting paper, Andrew Teal and Ola Berg Lutnæs for the help and many nice discussions, Michal Johansson for running the turbomole comparison (showing that my new RI-code was indeed faster), Sonia Coriani for some final clarifications of response theory, Lea Thøgersen for the help with density-minimization issues and for providing nice molecular inputs, and finally Trond Saue, Lucas Visser, Andre Gomes and Radovan Bast for all the work on the two-component density-fitting paper (I hope we will be able to finalize this eventually).

I would also like to thank all the other co-workers for the many good times. Thanks

goes to Arne for the many cigars, Kjetil, Torgeir, Seema, Tarjeir, Jorun, Mangnus and Mette for the times at 'Cafe Erwin', and to John, Thomas, Harald, Peter, Anne, Einar, Valadia, Astrid, Peter, Claus and others for nice discussions and leisure at lunch-times.

I would also like to thank my friends for all their moral support, especially by Knut Johan at the many coffee-breaks. I would further like to thank Gjermund for the many early mornings at the squash court, and to Arne (and Knut) for the runs in the forest. Thanks also goes to Håkon, Bernt, Karl André, Stig and Runar for many nice distractions from the thesis.

A very special thanks goes to my lovely and loving wife Trine for the support and patience (at least most of the time) and to my adorable daughter Marita. I would also like to thank my mother Beate for all the (desperately needed) baby-sitting when I was at work, and to my father Erik for the valuable help and support, in particular when things went wrong with the renovation of the new house in the middle of my thesis.

Finally, I would like thank the Norwegian Research Council through the Strategic University Program in Quantum Chemistry (Grant No. 154011/420) and through the CeO Centre for Theoretical and Computational Chemistry (Grant No. 179568/V30) for financial support, and to acknowledge the NOTUR computing facilities which have been used to conduct most of the calculations presented in this thesis.

List of Papers

I *A unified scheme for the calculation of differentiated and undifferentiated molecular integrals over solid-harmonic Gaussians*

S. Reine, E. Tellgren and T. Helgaker

Physical Chemistry Chemical Physics, **9**, 4771-4779 (2007)

II *Linear-scaling implementation of molecular electronic self-consistent field theory*

P. Sałek, S. Høst, L. Thøgersen, P. Jørgensen, P. Manninen, J. Olsen, B. Jansik, **S. Reine**, F. Pawłowski, E. Tellgren, T. Helgaker and S. Coriani

The Journal of Chemical Physics, **126**, 114110 (2007)

III *Variational and robust density fitting of four-center two-electron integrals in local metrics*

S. Reine, E. Tellgren, A. Krapp, T. Kjærgaard, T. Helgaker, B. Jansik, S. Høst and P. Sałek

The Journal of Chemical Physics, **129**, 104101 (2008)

IV *An efficient density functional theory force evaluation for large molecular systems*

S. Reine, M. F. Iozzi, V. Bakken, A. Krapp, T. Helgaker, F. Pawłowski and P. Sałek

Manuscript

V *A ground-state-directed optimization scheme for the Kohn-Sham energy*

S. Høst, B. Jansik, J. Olsen, P. Jørgensen, **S. Reine** and T. Helgaker

Physical Chemistry Chemical Physics, **10**, 5344-5348 (2008)

VI *Towards black-box linear scaling optimization in Hartree-Fock and Kohn-Sham theories*

S. Høst, J. Olsen, B. Jansik, P. Jørgensen, **S. Reine**, T. Helgaker, P. Sałek and S. Coriani

Lecture Series on Computer and Computational Sciences, **1**, 1-10 (2006)

VII *Linear-scaling implementation of molecular response theory in self-consistent field electronic-structure theory*

S. Coriani, S. Høst, B. Jansik, L. Thøgersen, J. Olsen, P. Jørgensen, **S. Reine**, F. Pawłowski, T. Helgaker and P. Sałek

The Journal of Chemical Physics, **126**, 154108 (2007)

Contents

Chapter 1

Introduction

Since the development of quantum mechanics in the 1920s and with the introduction of the Schrödinger equation in 1926 [1], different approaches to solve the Schrödinger equation have received substantial attention by physicists and chemists around the world. With the discovery and development of computers we are now capable of solving the Schrödinger equation for systems one did not deem possible less than a century ago; illustrated by the famous statement of P. A. M. Dirac [2] in 1929,

> *The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole part of chemistry are thus completely known, and the difficulty is only that the applications of these laws leads to equations much too complicated to be soluble.*

Although exact solutions can be obtained only for a few systems, the introduction of computer technology and the development of computational methods have allowed the Schrödinger equation to be solved in an approximate fashion for a large variety of systems. The theoretical studies of molecular properties have developed into an important tool, used both for predictions and analysis of chemical and physical processes. As a result, the use of calculations and simulations, rigorously based on the laws of quantum mechanics, has increased dramatically in many branches of science in the recent years; not only in chemistry and physics, but also in related fields such as medicine and biology.

The traditional way to solve the Schrödinger equation is by the use of wave-function based methods. The hierarchical wave-function based methods can systematically be extended to any given level of accuracy, and state-of-the-art electronic structure calculations challenge the accuracy of experiments. The scaling behavior of these meth-

ods, however, currently limits their use to small systems. For larger systems density functional theory (DFT) have become very popular, since the DFT methods typically constitute a good compromise between cost and accuracy. The DFT approach is based upon the fact that the ground state electron density contains all the information necessary to obtain the potential, and vice versa. Thus all the information is in principle possible to obtain if one of the two is know. The exact universal functional is not known, and a multitude of different approximate functionals exists. Still, DFT is the most widely used quantum-mechanical approach today, and provide important qualitative information even for large systems. During the last decade linear-scaling developments have allowed DFT calculations for molecular systems containing more than 1000 atoms.

The aim of this thesis has been development towards routine calculation for large molecular systems. To reach this goal, it is necessary to reduce both the cost and scaling properties of the DFT approach, and to develop robust, black-box optimization schemes.

Chapter 2

A brief summary of included paper

In **Paper I** we develop a new integral evaluation scheme, in which the solid-harmonic Gaussian basis functions are expanded in Hermite rather than Cartesian Gaussian intermediate functions. This approach both reduces the cost and simplifies the evaluation of differentiated integrals, and of for example the two- and three-center two-electron integrals used in the so-called density-fitting approximation. This approach is used both in Paper III and Paper IV to improve computational performance and at the same time reduce programming efforts.

In **Paper II** we present a boxed density-fitting scheme for linear-scaling density-fitted Coulomb evaluation. This approach is based upon the partitioning of the electron density [3], and the approximation of each part individually. By adding a robust correction term to the density-fitted Coulomb contribution, the introduced errors are small - compared to for example the errors introduced by the numerical integration quadrature for the exchange-correlation term. Linear scaling of the density-fitted Coulomb contribution is demonstrated for polyalanine peptides containing up to 1200 atoms.

In **Paper III** we follow Dunlap [4] and use a robust and variational density-fitting formulation to approximate four-center two-electron integrals. The results of this paper clearly indicate that sparse metrics may in fact be used for linear-scaling density-fitting developments. As an example, obtaining the fitting coefficients using the overlap metric introduce errors within $50-100\%$ of the errors using the conventional Coulomb metric; instead of the previously reported order of magnitude larger errors in Refs. [5, 6].

In **Paper IV** we present an efficient DFT force evaluation - the forces are needed for traversing the potential energy surface of molecular systems, and are essential for the determination of equilibrium and transition state structures. For the density-fitted Coulomb force, the integral evaluation scheme of Paper I is combined with linear-scaling

multipole-moment far-field interactions. This is further combined with an efficient implementation of the exchange-correlation contribution and of the geometry-optimizer, and results are presented for systems containing up to 500 atoms.

In **Paper II** and **Paper V** we present a linear-scaling atomic-orbital (AO) based SCF optimization scheme. In the trust-region SCF (TRSCF) approach the AO density matrix is expanded utilizing an exponential parameterization [7], and, rather than a cubic-scaling diagonalization step, the Roothaan-Hall energy is minimized in each SCF iteration through a series of conjugate gradient iterations and combined with the density-subspace minimization (DSM) approach [8, 9] to obtain a new density matrix. By automating step size criteria, based on the trust-region approach [8, 9], the TRSCF approach can be used in a black-box manner (i.e. without the need for a common user to manually set a level-shift or damping parameter), and is further demonstrated to be more robust than the traditional Roothaan-Hall (RH) direct inversion in an iterative subspace (DIIS) approach [10]. This approach is applied to the optimization of polyalanine peptides containing up to 1200 atoms.

In **Paper VI** we develop an even more reliable and efficient linear-scaling optimization scheme, the augmented Roothaan-Hall (ARH) approach. In this approach, a local quadratic model of the KS energy, that is exact to second order in the subspace of the previous density matrices and constitute a good approximation in other directions, is minimized using the trust-region approach. The method differs from previous KS optimization methods in that it does not involve two separate steps, such as the RH diagonalization followed by the DIIS averaging. Instead, one single step is performed that exploits the curvature information spanned by the previous density matrices. Since the ARH contains information about the electronic Hessian, the method both enhances performance and converges by design to a minimum. This is demonstrated by sample calculations where the ARH approach finds a minimum and the traditional RH/DIIS approach either diverges or converges to a saddle-point.

Finally, in **Paper VII** we present a linear-scaling AO based linear-response implementation for HF and DFT. The response equations are solved iteratively in a subspace of paired trial vectors. The used of paired trial vectors preserve the algebraic structure of the response equations, both enhancing convergence and avoiding complex eigenvalues. A non-diagonal preconditioner combined with good initial guesses allows performance comparable with canonical molecular-orbital (MO) theory, with typically five to ten iterations needed for convergence. The computational time is dominated by the construction of the effective Fock/KS matrices, as in the canonical case, but

with linear complexity achieved using sparse-matrix algebra. Linear scaling, and robust convergence is demonstrated for the calculation of frequency-dependent polarizabilities and excitation energies of polyalanine peptides containing up to 1400 atoms.

Chapter 3

Theory

In this chapter we first give an introduction to the basic theory essential for this thesis. As the theory is considered fundamental, we will only occasionally provide references in this chapter. For a more thorough introduction, consult some of the many theory books on quantum chemistry, for example Refs. [7, 11, 12, 13] on which most of this introduction is based upon. We start with a brief introduction to theory essential for quantum chemistry in section 3.1. In section 3.2, we give an introduction to Hartree-Fock theory and briefly discuss the post Hartree-Fock methods. In section 3.3, we give an introduction to Kohn-Sham density functional theory, and finally in section 3.4 we give an introduction to response theory.

3.1 Introductory theory

This section give a brief introduction to some of the theory fundamental to molecular quantum chemistry; the Schrödinger equation, the Born-Oppenheimer approximation, the potential energy surface, the Slater determinant, the one-electron orbital basis-set expansion and the variation method.

3.1.1 The Schrödinger Equation

We are in this thesis interested in solving the non-relativistic, time-independent, N-electron Schrödinger equation for molecules,

$$\hat{H}\Psi_n = E_n\Psi_n, \tag{3.1}$$

where $\hat{H}$ is the molecular electronic Hamiltonian operator, Ψ_n the different eigenstates or wave functions, and E_n are the corresponding energies. The molecular electronic

Hamiltonian is given by

$$\hat{H} = \hat{h} + \hat{g} + \hat{h}_{\text{nuc}}, \tag{3.2}$$

where $\hat{h}_{\text{nuc}}$ is the nuclear repulsion, $\hat{h}$ is the one-electron part, consisting of the kinetic energy and the nuclear-electron attraction operators, and $\hat{g}$ the two-electron part of the Hamiltonian. In atomic units, we have

$$\begin{aligned} \hat{h}_{\text{nuc}} &= \frac{1}{2}\sum_{A\neq B}^{M}\frac{Z_A Z_B}{R_{AB}} \\ \hat{h} &= -\frac{1}{2}\sum_{i}^{N}\nabla_i^2 - \sum_{i}^{N}\sum_{A}^{M}\frac{Z_A}{r_{Ai}} \\ \hat{g} &= \frac{1}{2}\sum_{i\neq j}^{N}\frac{1}{r_{ij}}, \end{aligned} \tag{3.3}$$

with N the number of electrons, M the number of nuclei, Z_A the charge of nuclei A, R_{AB} the distance between nuclei A and B, r_{Ai} the distance between nuclei A and electron i, and r_{ij} the distance between the two electrons i and j. The N-electron wave function $\Psi_n = \Psi_n(\mathbf{x}_1, ..., \mathbf{x}_N)$ depends on the $3N$ spatial $\{\mathbf{r}_i = (x_i, y_i, z_i)\}$ and N spin coordinates $\{s_i\}$; jointly written in the compact notation $\{\mathbf{x}_i = (\mathbf{r}_i, s_i)\}$. The complexity of the electronic Schrödinger equation stems from the fact that the N-electron wave functions depend on the coupled $4N$ spatial and spin coordinates $\{\mathbf{x}_i\}$.

3.1.2 The Born-Oppenheimer approximation

To arrive at the molecular electronic Hamiltonian, the Born-Oppenheimer approximation has been adopted. Although the state of a many-particle system depends on all particles involved (in the molecular case, both electrons and nuclei), the motion of the nuclei is slow compared to the motion of the electrons, due to the three or more orders of magnitude difference in their masses. In the Born-Oppenheimer approximation the electronic state is therefore taken to be independent of the motion of the nuclei, depending only on their positions. For high accuracy, the motion of the nuclei should be accounted for by adding vibrational corrections to the electronic energy. Also note that when two different states cross, the Born-Oppenheimer approximation breaks down. But, as stated in Ref. [11] the Born-Oppenheimer approximation introduce only very small errors for the majority of systems.

3.1.3 The potential energy surface

The energy as a function of the nuclear coordinates is denoted the potential energy surface (PES). Information about the PES of a molecular system is crucial for the theoretical study of molecules and their interactions. The different minima, or equilibrium geometries, are important in for example the determination of reaction enthalpies, and also forms the basis for the calculation of several other chemical properties - like vibrational spectra and various electric and magnetic properties. Saddle points are also of great importance, as they represent transitional structures, which are important in the determination of possible reaction pathways and for the determination of reaction barriers. Efficient geometry-optimization procedures, involving energy, gradient and possibly Hessian evaluations, are therefore essential for the efficient application to quantum chemical methods to problems in chemistry.

3.1.4 Slater determinants

For a system of identical *fermions*, such as electronic systems, the wave function is, according to *the Pauli principle*, anti-symmetric with respect to an interchange of two fermions,

> **The Pauli Principle** *The total wave function must be antisymmetric under the interchange of any pair of identical fermions and symmetric under the interchange of any pair of identical bosons.*

One way to fulfill the Pauli principle is by expanding the total N-electron wave function in a linear combination of N-electron *Slater determinants* $|\mathrm{SD}\rangle$. The Slater determinants are anti-symmetrized linear combinations of the products of one-electron functions, for example

$$|\mathrm{SD}\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \Phi_1(\mathbf{x}_1) & \Phi_1(\mathbf{x}_2) & \dots & \Phi_1(\mathbf{x}_N) \\ \Phi_2(\mathbf{x}_1) & \Phi_2(\mathbf{x}_2) & \dots & \Phi_2(\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_N(\mathbf{x}_1) & \Phi_N(\mathbf{x}_2) & \dots & \Phi_N(\mathbf{x}_N) \end{vmatrix}, \tag{3.4}$$

where the one-electron *spin orbitals* $\Phi_p(\mathbf{x})$ depend both on the three spatial coordinates and the spin coordinate.

There are many possible approaches for solving the molecular electronic Shrödinger equation of Eq. (3.1), among others the HF and post HF approaches, DFT and quantum

Monte Carlo methods. In principle these methods are quite different, but in practice a linear combination of Slater determinants (possibly only one) typically forms the basis of the approximate wave function.

3.1.5 Basis sets

The spin orbitals $\Phi_p(\mathbf{x})$, from which the Slater determinants are constructed, are products of *orbitals* $\phi_p(\mathbf{r})$ and *spin functions* $\sigma(s)$,

$$\Phi_p(\mathbf{x}) = \phi_p(\mathbf{r})\sigma(s). \tag{3.5}$$

The spin functions are either the spin up or the spin down functions $\alpha(s)$ or $\beta(s)$, respectively, whereas the orbitals could in principle be three-dimensional functions of any given form.

The orbitals are expanded in a *basis* $\{\chi_a\}$ of three-dimensional functions of known form, according to

$$\phi_p(\mathbf{r}) = \sum_a C_{ap}\chi_a(\mathbf{r}). \tag{3.6}$$

When the basis functions $\{\chi_a\}$ are atomic orbitals, the above expansion is known as the linear-combination of atomic-orbitals (LCAO) approach. The orbital coefficients C_{ap} are for molecules known as the MO coefficients, and the basis functions χ_a are typically taken to be atom-centered functions. These functions are somewhat loosely denoted atomic orbitals (AOs) - although not actually atomic orbitals their form typically resembles that of the atomic orbitals. Throughout this thesis we will denote these functions as AOs. The set of all AOs in a basis is called a *basis set*. To be able to reproduce the form of the MOs, the basis set must in principle be complete. In practice, however, a truncation must be made.

Many different forms of these AO basis functions are possible but, in quantum chemistry, the spherical harmonic Slater type orbitals (STOs) and Gaussian type orbitals (GTOs) have proven successful. Of these two, most quantum chemistry software programs today use the GTOs rather that the STOs, although some use is made of STOs in DFT. In this thesis, we only use GTOs. In chapter 4, we will see how this choice reduces the six-dimensional two-electron integrals to one-dimensional integrals (and recurrence relations on these).

3.1.6 The variation method

Before addressing the variation method, we take a look at the properties of the exact solutions to the Schrödinger equation, Eq. (3.1). The energy expectation value is a functional of the trial wave function Ψ,

$$E[\Psi] = \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle}. \tag{3.7}$$

All exact solutions $\Psi_n(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of Eq. (3.1) are *variational* in the sense that, for all possible variations $\delta\Psi_n$ that are orthogonal to Ψ_n, the energy is stable, which is easily verified because

$$\begin{aligned} \langle \delta\Psi_n | \hat{H} | \Psi_n \rangle &\equiv \int \delta\Psi_n^*(\mathbf{x}_1, \ldots, \mathbf{x}_N) \hat{H} \Psi_n(\mathbf{x}_1, \ldots, \mathbf{x}_N) d\mathbf{x}_1 \ldots d\mathbf{x}_N \\ &= E_n \langle \delta\Psi_n | \Psi_n \rangle = 0. \end{aligned} \tag{3.8}$$

Indeed, solving the Schrödinger equation of Eq. (3.1) and the variational optimization of Eq. (3.7) are identical. This is known as *the variation principle*,

> **The Variation Principle** *The ground state solution of the time-independent Schrödinger equation (Eq. 3.1) is equivalent to the variational optimization of the energy functional (Eq. 3.7).*

This principle leads us to the powerful procedure of *the variation method*. According to the variation method, the best set of parameters $\mathbf{C}$ of a given trial wave function $\Psi(\mathbf{C})$ are the values that gives stationary points of the energy function

$$E(\mathbf{C}) = \frac{\langle \mathbf{C} | \hat{H} | \mathbf{C} \rangle}{\langle \mathbf{C} | \mathbf{C} \rangle}. \tag{3.9}$$

A wave function that has been optimized according to the variation method is termed *variational*. Variational optimization has several advantages. First, the energy $E(\mathbf{C})$ is always greater than or equal to the true ground-state energy E_0. This ensures that improvements of the wave function always result in a decrease in the energy, with the ground-state energy as a lower bound. Second, the error in the energy is second order in the error in the wave function,

$$E(\mathbf{C}) = E[\Psi_0 + \delta\Psi] = E[\Psi_0] + \mathcal{O}(\delta\Psi^2), \tag{3.10}$$

where Ψ_0 is the true ground state wave function and where $\delta\Psi$ is the difference between the true ground state and the trial wave function $\Psi(\mathbf{C})$. Third, energy derivatives needed in for example response theory are greatly simplified because the derivatives

with respect to the variational parameters $\mathbf{C}$ are zero. This leads to the Wigner $2n+1$-rule in response theory; provided the energy has been optimized variationally one only needs to solve the order n response equations in order to obtain $2n+1$-order response functions.

Note that there are several restrictions on the wave function that need to be met, for instance that the MOs are orthogonal, so the optimization is not entirely free. There are different ways to impose conditions on the wave function. One way is to use *Lagranges method of undetermined multipliers*, in which a *Lagrangian* is constructed as the sum of the energy and of the different constraints multiplied with undetermined Lagrange multipliers. The Lagrangian is optimized variationally with respect to both the variational parameters $\mathbf{C}$ and the Lagrange multipliers $\boldsymbol{\lambda}$.

3.2 Hartree-Fock theory

In restricted HF (RHF) theory, the wave function $|\mathrm{HF}\rangle$ is taken to be a single *configuration state function* $|\mathrm{CSF}\rangle$; which is a fixed (and minimal) linear combination of Slater determinants $|\mathrm{SD}\rangle$, constructed in such a way as to provide the correct spin symmetry. Note that in unrestricted Hartree-Fock (UHF) theory, in which no symmetry constraints are imposed on the total spin, and for closed-shell systems, the Hartree-Fock wave function is always a single Slater determinant.

3.2.1 The Hartree-Fock equations

If a RHF wave function $|\mathrm{HF}\rangle$, constructed from n molecular spin orbitals Φ_i, is optimized according to Lagranges method of undetermined multipliers (see section 3.1.6) under the constraint that the MOs are orthonormal, we arrive at the Hartree-Fock equations [11],

$$\hat{f}\Phi_i \equiv \left[\hat{h} + \sum_{j}^{n_{\mathrm{occ}}}(\hat{J}_j - \hat{K}_j)\right]\Phi_i = \sum_{ij}^{n}\epsilon_{ij}\Phi_j, \quad \forall\Phi_i. \tag{3.11}$$

Here $\hat{f}$ is the *Fock operator*, n_{occ} refers to the number of occupied spin orbitals and ϵ_{ij} are the Lagrange multipliers. The Coulomb operator $\hat{J}_i$ and the exchange operator $\hat{K}_i$ operating on an arbitrary one-electron spin orbital $g(\mathbf{x})$ are defined as

$$\begin{aligned}\hat{J}_i g(\mathbf{x}_1) &= \left(\int \Phi_i^*(\mathbf{x}_2)\Phi_i(\mathbf{x}_2)\frac{1}{r_{12}}d\mathbf{x}_2\right) g(\mathbf{x}_1) \\ \hat{K}_i g(\mathbf{x}_1) &= \left(\int \Phi_i^*(\mathbf{x}_2) g(\mathbf{x}_2)\frac{1}{r_{12}}d\mathbf{x}_2\right) \Phi_i(\mathbf{x}_1).\end{aligned} \tag{3.12}$$

The exchange operator K_i only gives non-vanishing contributions when the spin function of g and Φ_i are identical. For a closed-shell system, the RHF equations therefore reduce to

$$\left[\hat{h} + \sum_{j}^{n_{\text{occ}}}(2\hat{J}_j - \hat{K}_j)\right]\phi_i = \sum_{ij}^{n}\epsilon_{ij}\phi_j, \quad \forall\phi_i, \tag{3.13}$$

where n and n_{occ} now refers to the number of orbitals and of doubly occupied orbitals ϕ_i, respectively, and where the Coulomb and the exchange operators operate instead on a one-electron function $g(\mathbf{r})$. Note that in the *canonical Hartree-Fock* representation, the orbitals undergo a unitary transformation to the canonical MOs in which the Lagrange multipliers constitute a diagonal matrix $\epsilon_{ij} = \delta_{ij}\epsilon_i$. In the canonical representation the MOs are the eigenfunctions of the Fock operator $\hat{f}$, and the corresponding eigenvalues ϵ_i are denoted *orbital energies*.

3.2.2 The Roothaan-Hall equations

For simplicity, we will in the following restrict ourselves to closed-shell HF theory and assume real basis functions. When the MOs $\phi_i(\mathbf{r})$ are expanded in a linear combination of AO basis functions $\chi_a(\mathbf{r})$, in accordance with Eq. (3.6), we arrive at the Roothaan-Hall equations,

$$\mathbf{FC} = \mathbf{SCE}. \tag{3.14}$$

To arrive at Eq. (3.14) we have in addition multiplied from the left with the different basis functions and integrated. The *Fock matrix* $\mathbf{F}$ is the sum of the one-electron matrix $\mathbf{h}$ and the two-electron Coulomb $\mathbf{J}$ and exchange $\mathbf{K}$ matrices, given by

$$\begin{aligned} F_{ab} &= h_{ab} + 2J_{ab} - K_{ab} \\ &= \langle a|\hat{h}|b\rangle + 2\sum_{i}^{n_{\text{occ}}}(ab|ii) - \sum_{i}^{n_{\text{occ}}}(ai|bi) \\ &= \langle a|\hat{h}|b\rangle + 2\sum_{cd}(ab|cd)D_{cd} - \sum_{cd}(ac|bd)D_{cd}, \end{aligned} \tag{3.15}$$

and the overlap matrix $\mathbf{S}$ is given by

$$S_{ab} = \langle ab\rangle. \tag{3.16}$$

In Eq.(3.15) we have used the Mulliken like notation

$$\begin{aligned} \langle f\rangle &= \int \chi_f(\mathbf{r})d\mathbf{r} \\ \langle f|\hat{w}|g\rangle &= \int \chi_f(\mathbf{r})\hat{w}\chi_g(\mathbf{r})d\mathbf{r} \\ (f|g) &= \int \chi_f(\mathbf{r}_1)\frac{1}{r_{12}}\chi_g(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2, \end{aligned} \tag{3.17}$$

and where the AO density matrix $\mathbf{D}$ is the sum of the product of occupied MO-coefficient pairs according to

$$D_{ab} = \sum_{i}^{n_{\text{occ}}} C_{ai} C_{bi}. \tag{3.18}$$

3.2.3 The self-consistent field approach

In the canonical representation, in which $\mathbf{E}$ is diagonal, the Roothaan-Hall equations has the form of a generalized eigenvalue problem - from which the diagonal elements ϵ_i and the MO coefficients $\mathbf{C}$ can be found by diagonalization. However, since the Fock matrix depends on the MO coefficients, or the density matrix, the Roothaan-Hall equations defines a (non-linear) pseudo-eigenvalue problem, where the solutions are found in a *self consistent field* (SCF).

In general, a self consistent solution can be obtained through an iterative SCF optimization. For instance, when solving the Roothaan-Hall equations, the initial MO-coefficients are used to construct a density matrix. From the density matrix, a Fock matrix is built and diagonalized. The new MO coefficients are again used to build a density matrix and so forth. The procedure is repeated until the MO coefficients (or the density matrix) are reproduced to a given accuracy - the solutions are then said to be self-consistent. SCF convergence can be difficult - convergence is not guaranteed and in some cases the converged solution does not represent a minimum.

A highly successful approach to improve SCF convergence has been the DIIS approach of Pulay [10]. In this scheme, subsequent sequences of Fock matrices $\left\{\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \dots\right\}$, density matrices $\left\{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \dots\right\}$ and error-estimates $\left\{\mathbf{E}^{(1)}, \mathbf{E}^{(2)}, \dots\right\}$ are stored, and at a given iteration k, the error in the subspace of error-estimates $\mathbf{E}_{\text{int}}^{(k)}$,

$$\mathbf{E}_{\text{int}}^{(k)} = \sum_{i=1}^{k} e_i \mathbf{E}^{(i)}, \tag{3.19}$$

is interpolated by minimization of the scalar error function $\text{SE}(\mathbf{e})$, given by

$$\text{SE}(\mathbf{e}) = \text{Tr}\left(\mathbf{E}_{\text{int}}^{(k)} \mathbf{E}_{\text{int}}^{(k)}\right), \tag{3.20}$$

under the normalization constraint

$$\sum_{i=1}^{k} e_i = 1. \tag{3.21}$$

Once the coefficients have been obtained an extrapolated Fock matrix $\mathbf{F}_{\text{ext}}^{(k)}$ is constructed according to

$$\mathbf{F}_{\text{ext}}^{(k)} = \sum_{i=1}^{k} e_i \mathbf{F}^{(i)}, \tag{3.22}$$

from which the new density matrix $\mathbf{D}^{(k+1)}$ is obtained. The new density matrix is again used to construct the next Fock matrix $\mathbf{F}^{(k+1)}$, and so on.

Is some cases, oscillations make convergence problematic. Then *damping* or *level shifting* may be introduced to improve the convergence. With damping, the SCF step is limited by taking as you new density $\mathbf{D}^{(k+1)}$ a linear combination of the current density $\mathbf{D}^{(k)}$ and the predicted, undamped density matrix $\mathbf{D}_{\text{pred}}^{(k+1)}$,

$$\mathbf{D}^{(k+1)} = \alpha\mathbf{D}^{(k)} + (1-\alpha)\mathbf{D}_{\text{pred}}^{(k+1)}. \tag{3.23}$$

The damping parameter α can in principle be any number between zero (no damping) and one (full damping). Level shifting limits the SCF step by increasing the energy of the unoccupied orbitals. This effectively reduces how much the orbitals rotate (see section 6.1), or mix occupied and unoccupied orbitals, and may therefore reduce or remove oscillations from the SCF cycles. Increasing the level shift reduces the oscillations, but at the same time larger shifts reduce the convergence rate.

We have now had a look at different techniques to achieve SCF convergence. But, as mentioned, SCF convergence does not necessarily mean that the energy has reached a minimum - only that the converged wave function is stationary with respect to orbital rotations. In order to identify the nature of the solution, one must analyze the electronic Hessian (which is the second derivative of the energy with respect to the variational parameters). If the eigenvalues of the Hessian are all positive the solution has reached a minimum, otherwise the solution represents a *saddlepoint*. Evaluation of the Hessian is typically quite demanding, and such an analysis is typically not carried out.

In *second order SCF* theory both the gradient and the Hessian are calculated at each SCF cycle. This leads to quadratic convergence near the minimum, and convergence is therefore obtained in only a few iterations. This method is computationally demanding due to the expensive evaluation of the Hessian and seldom used in practice. If the computationally demanding Hessian is replaced by an approximate Hessian, it is possible to enhance convergence (for an example see Paper VI). Note that such approaches are no longer quadratically convergent.

3.2.4 Electron correlation

The Hartree-Fock method typically accounts for about 99.5% of the electronic energy, and several chemical properties, like dipole moments, polarizabilities, excitation energies, magnetizabilities and force constants, are typically off by less than 10%. The problem with HF is that it does not include all *electron correlation* effects. The broad-

est definition of electron correlation is to say that the position of one electron depends on the position of all other electrons. In the mean-field HF approach, however, each electron interacts with the other electrons only through an averaged potential, and fails to account for the *instantaneous* electron-electron repulsion effects. It does however, include the electron correlation enforced by the anti-symmetry of the HF wave function. This correlation is termed the *Fermi correlation* and accounts for most of the correlation. Note also that whenever the HF wave function is composed of more than one Slater-determinant, this leads to the inclusion of additional electron correlation effects.

In wave-function theory, the term electron correlation is normally reserved to describe the correlation that occurs upon superpositions of configuration-state functions - that is, the difference between the HF and the exact result. This leads to the Löwdin definition of the *electron correlation energy*: "the correlation energy for a certain state with respect to a specified Hamiltonian is the difference between the exact eigenvalue of the Hamiltonian and its expectation value in the Hartree-Fock approximation for the state under consideration." Note that it is implicit in the above definition that the basis-set limit is taken.

There exist different hierarchical wave-function methods, like the different orders of Møller-Plesset perturbation theory (MPPT), configuration-iteration (CI), coupled-cluster (CC) and multi-configurational self-consistent field (MCSCF) theory, that incorporates electron correlation effects by making linear combinations of CSFs. When the full CI wave function, in which all possible configuration state-functions are included, is dominated by a single reference CSF function, the CC approaches in particular provide highly accurate results. When the full CI is dominated by more than a single reference state, for instance when looking at bond breaking, the MCSCF methods typically works well - in particular the complete active space (CAS) approach. The problem with these methods, however, is their poor scaling with system size, and therefore these methods are currently fairly limited with respect system size. In this thesis the focus is on developing theory for treating large systems. Therefore, we will not discuss the different wave-function approaches further. For large systems, DFT has proven highly successful, as it constitute a good compromise between cost and accuracy. In the next subsection we will give a brief introduction to Kohn-Sham DFT.

3.3 Density functional theory

Today, the most widely used method in quantum chemistry is the DFT approach. In this section, we will first consider the basis of density functional theory by the introduction of the Hohenberg-Kohn theorems and the Kohn-Sham equations, followed by an overview of some of the most common exchange-correlation functionals.

3.3.1 The Hohenberg-Kohn theory

Fundamental to density functional theory is the electron density. The N-electron density $\rho(\mathbf{r})$ depends only on three spatial coordinates $\mathbf{r}$,

$$\rho(\mathbf{r}) = N \int |\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)|^2 \, ds_1 d\mathbf{x}_2 \ldots d\mathbf{x}_N. \tag{3.24}$$

The electron density is non-negative, vanishes at infinity and gives the total number of electrons N when integrated over the real-space coordinates $\mathbb{R}^3$,

$$\int \rho(\mathbf{r}) d\mathbf{r} = N. \tag{3.25}$$

In 1964 Hohenberg and Kohn [14] established that the exact ground-state electron density $\rho(\mathbf{r})$ may be uniquely associated with one external potential $v_{\text{ext}}(\mathbf{r})$ (up to an additive constant),

> **The Hohenberg-Kohn theorem** *Each v-representable N-electron density $\rho(\mathbf{r})$ is the ground state density of at most one external potential $v_{ext}(\mathbf{r}) + C$, which is determined up to an additive constant C.*

An electron density $\rho(\mathbf{r})$ is termed *v-representable* if it is associated with the ground state wave function of an electronic Hamiltonian of the form

$$\begin{aligned} \hat{H} &= -\frac{1}{2}\sum_i^N \nabla_i^2 + \frac{1}{2}\sum_{i\neq j}^N \frac{1}{r_{ij}} + \sum_i^N v_{\text{ext}}(\mathbf{r}_i) \\ &= \hat{T} + \hat{V}_{ee} + \sum_i^N v_{\text{ext}}(\mathbf{r}_i), \end{aligned} \tag{3.26}$$

where $v_{\text{ext}}(\mathbf{r})$ defines the external potential. It follows from the Hohenberg-Kohn theorem that the potential $v(\mathbf{r})$ is a functional of the electron density, $v[\rho]$, and that the ground state energy is a functional of the electron density $\rho(\mathbf{r})$, in the sense that the density uniquely determines the external potential (up to an additive constant), which

in turn determines the energy $E[v]$. Thus the Hohenberg-Kohn theorem provide, at least in principle, a means of obtaining the ground state energy.

The *Hohenberg-Kohn functional* $F[\rho]$ defined by

$$F[\rho] = E[v[\rho]] - \int \rho(\mathbf{r}) v[\rho] d\mathbf{r}, \tag{3.27}$$

does not depend on the potential, and is therefore universal. To see this, we rewrite the Hohenberg-Kohn functional in terms of the wave function $\Psi_0[\rho]$ associated with the density ρ. For non-degenerate systems, there is only one such wave function, and we may then write Eq. (3.27) uniquely as

$$\begin{aligned} F[\rho] &= \langle \Psi_0[\rho] | T + V_{\text{ee}} + \sum_i^N v(\mathbf{r}_i) | \Psi_0[\rho] \rangle - \int \rho(\mathbf{r}) v[\rho] d\mathbf{r} \\ &= \langle \Psi_0[\rho] | T + V_{\text{ee}} | \Psi_0[\rho] \rangle = T[\rho] + V_{\text{ee}}[\rho]. \end{aligned} \tag{3.28}$$

Hohenberg and Kohn further established that the variational principle may be recast in terms of the electron density $\rho(\mathbf{r})$ rather than the wave function $\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$,

> **Hohenberg-Kohn variation principle** *The ground state energy can be obtained by the density minimization*
>
> $$E_0[v] = \min_\rho \left(F[\rho] + \int \rho(\mathbf{r}) v[\rho] d\mathbf{r} \right), \tag{3.29}$$
>
> *where the minimization is constrained to densities that are v-representable.*

In the original formulation by Hohenberg and Kohn, the ground state was assumed to be non-degenerate and the variational optimization was constrained to densities that where v-representable. It has been shown that certain reasonable densities are not v-representable [15, 16]. Levy [17, 15] solved both these problems by showing that it was sufficient for the density to be *N-representable*, which means that the density can be obtained from an N-electron ground state. This gives the energy functional

$$E_0[v] = \inf_{\Psi \to \rho} \langle \Psi | \hat{T} + V_{ee} | \Psi \rangle + \int v_{\text{ext}}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r}, \tag{3.30}$$

where the infimum inf is the greatest lower bound (rather than a minimum).

Hohenberg and Kohn thus established that the ground state electron density contains all the information that is needed to reconstruct the external potential, and therefore the different wave functions and energies. The functional form of the energy functional $F[\rho]$, however, is not known.

3.3.2 Kohn-Sham density functional theory

In 1965 Kohn and Sham [18] derived a set of equations for finding the density in a self consistent fashion. They began by partitioning the energy according to

$$E[\rho,\phi] = F[\rho,\phi] + \int v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r}) = T_s[\phi] + J[\rho] + E_{\text{xc}}[\rho] + \int v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r}) \tag{3.31}$$

where T_s is the kinetic energy of a non-interacting system that gives raise to the density ρ, v_{ext} is the external potential and E_{xc} is the *exchange-correlation energy*. Here the only unknown is the exchange-correlation energy, which contains the exchange and correlation energies, and the correction to the kinetic energy T_s (which is obtained in the non-interacting system). A single Slater determinant is the exact eigenfunction of the Hamiltonian of a non-interacting system. Furthermore, any density $\rho(\mathbf{r})$ can be obtained from a single Slater determinant. As a consequence, a scheme similar to Hartree-Fock can be adapted for DFT; know as the Kohn-Sham DFT approach. If the Kohn-Sham (KS) energy functional of Eq. (3.31) is optimized under the constraint that the the KS orbitals $\phi_i(\mathbf{r})$ are orthonormal, we arrive at the Kohn-Sham orbital equations,

$$\left[-\frac{1}{2}\nabla^2 - v_{\text{ext}} + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}'-\mathbf{r}|} + \frac{\delta E_{\text{xc}}[\rho]}{\delta\rho(\mathbf{r})}\right]\phi_i(\mathbf{r}) = \sum_j \epsilon_{ij}\phi_j(\mathbf{r}), \tag{3.32}$$

with the density given by

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2 . \tag{3.33}$$

We proceed in the exact same fashion as when deriving the Roothaan-Hall equations for Hartree-Fock, i.e. make an expansion of the orbitals according to Eq. (3.6), multiplying from the left with an AO basis function $\chi_a(\mathbf{r})$ and integrating, to arrive at the Kohn-Sham equations

$$\mathbf{F}^{\text{KS}}\mathbf{C} = \mathbf{SCE}. \tag{3.34}$$

For close-shell systems the KS matrix $\mathbf{F}^{\text{KS}}$ is given by

$$\begin{aligned} F_{ab}^{\text{KS}} &= h_{ab} + 2J_{ab} + X_{ab} \\ &= \langle a|\hat{h}|b\rangle + 2\sum_{cd}(ab|cd)D_{cd} + \int \chi_a(\mathbf{r})\chi_b(\mathbf{r})v_{\text{xc}}(\mathbf{r})d\mathbf{r}, \end{aligned} \tag{3.35}$$

where $\mathbf{X}$ is the exchange-correlation matrix and where the *exchange-correlation potential* $v_{\text{xc}}(\mathbf{r})$ is given by

$$v_{\text{xc}}(\mathbf{r}) = \frac{\delta E_{\text{xc}}[\rho]}{\delta\rho(\mathbf{r})}. \tag{3.36}$$

The solutions to the KS equations are, in direct parallel to solving the Roothaan-Hall equations of Eq. (3.14), found in a self-consistent field as outlined in section 3.2.3.

There are obvious similarities between HF and KS theory. In both methods, a set of orbitals is determined self consistently, the cost of the two methods is similar, and both methods use an effective potential. Still, the two methods are also different by design. The HF orbitals gives an approximate wave function, whereas the KS orbitals in principle give the exact density. In both approaches, the energy is determined variationally, but the DFT energy, although in principle exact, does not provide an upper-bound to the true ground state energy for approximate exchange-correlation potentials.

3.3.3 Exchange-correlation functionals

The quality of the KS approximation depends solely on the functional form of the exchange-correlation energy. In this subsection, we provide a basic overview of the general form and provide a few examples of the most common functionals.

The *local density approximation* (LDA) is based on the uniform electron gas, as proposed by Kohn and Sham in their initial paper [18], where the electrons are evenly distributed on a positive background charge. The LDA energy is given as

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int \rho(\mathbf{r})\epsilon_{\mathrm{xc}}[\rho(\mathbf{r})]d\mathbf{r}, \tag{3.37}$$

where ϵ_{xc} is the exchange-correlation energy per electron in a uniform electron gas. For a uniform electron gas, the exchange-correlation energy is split into separate exchange and correlation energies. The exchange energy for a uniform electron gas was derived by Dirac [19],

$$E_{\mathrm{x}}^{\mathrm{LDA}}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int \rho^{4/3}(\mathbf{r})d\mathbf{r}. \tag{3.38}$$

There is no such simple expression for the correlation energy of a uniform electron gas. However, in 1980 Ceperley and Alder [20] performed highly accurate Monte-Carlo simulations on the uniform electron gas, resulting in the development of several analytical forms by Vosko, Wilk and Nussair (VWN) [21]. For the correlation energy, one of these VWN functional forms (number V), here denoted E_{c}^{VWN}, remains the most commonly used for the LDA approach to date. The LDA functional thus has the following form,

$$E_{\mathrm{xc}}^{\mathrm{LDA}} = E_{\mathrm{x}}^{\mathrm{LDA}} + E_{\mathrm{c}}^{VWN}. \tag{3.39}$$

Despite the rather simple form, LDA provides fairly accurate results, in most cases comparable or better than the HF approach. With the LDA functional, it is possible

to achieve good vibrational frequencies, equilibrium structures and dipole moments. However, bond energies are systematically too high.

In the *generalized gradient approximation* (GGA), the energy functional $E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho]$ depends on the gradient of the density $\nabla\rho(\mathbf{r})$ in addition to the density,

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho] = \int F_{\mathrm{xc}}\left(\rho, \nabla\rho\right) d\mathbf{r}, \tag{3.40}$$

where the form of F_{xc} varies for the different functionals. As was the case for LDA, the GGA exchange-correlation energy $E_{\mathrm{xc}}^{\mathrm{GGA}}$ is also split into separate exchange and correlation parts. The use of GGA can lead to significant improvements compared to the properties calculated with LDA. There exists a variety of different GGA functionals, including functionals specifically designed for the calculation of certain chemical properties - commonly used functionals are the BLYP [22, 23], BPW91 [22, 24] and PBE [25]. To give an example, the BLYP functional has the following form

$$E_{\mathrm{xc}}^{\mathrm{BLYP}} = E_{\mathrm{x}}^{\mathrm{LDA}} + E_{\Delta\mathrm{x}}^{\mathrm{B88}} + E_{\mathrm{c}}^{\mathrm{LYP}}, \tag{3.41}$$

where the first term is the LDA exchange energy, the second term is the GGA correction term to the exchange energy as suggested by Becke [22], and the third term is the GGA correlation term of Lee, Yang and Perdew (LYP) [23].

In HF theory, the Fermi correlation is accounted for by the exchange contribution, which accounts for the bulk (about 90%) of the correlation energy, and neither the LDA or the GGA approaches include Fermi correlation in a fully satisfactory manner The Coulomb repulsion energy

$$J = \frac{1}{2}(\rho|\rho) = \frac{1}{2}\sum_{ij}^{n_{\mathrm{occ}}}(ii|jj), \tag{3.42}$$

appearing in both HF and DFT includes the *self-interaction* term, $(ii|ii)$. An electron in an occupied orbital moves in the averaged potential generated by all electrons - including itself. In HF this self interaction is canceled by an equivalent interaction of the opposite sign appearing in the exchange term,

$$K = \frac{1}{2}\sum_{ij}^{n_{\mathrm{occ}}}(ij|ij). \tag{3.43}$$

Hybrid functionals combines a part x_k of the HF exchange K together with GGA functional contributions,

$$E_{\mathrm{xc}}^{\mathrm{Hybrid}} = E_{\mathrm{x}}^{\mathrm{GGA}} - x_k K + E_{\mathrm{c}}^{\mathrm{GGA}}. \tag{3.44}$$

Actually, K differs from the HF exchange because it is based on the KS orbitals rather than the HF orbitals. Still it is often denoted HF exchange, or sometimes as the non-local exchange or the exact exchange.

The inclusion of the full HF exchange ($x_k = 1$) was proposed by Kohn and Sham in their original paper [18], but deriving an appropriate and accurate correlation functional for use with the non-local exchange has been difficult. Including only a proportion of the HF exchange has proven much more successful. The B3LYP functional of Stevens *et al.* [26] is the most popular and widely used hybrid functional to date. It combines the correlation and the exchange terms of the BLYP and LDA functionals in a semi-empirical manner,

$$E_{\mathrm{xc}}^{\mathrm{B3LYP}} = (1 - x_k)E_{\mathrm{x}}^{\mathrm{LDA}} + aE_{\Delta\mathrm{x}}^{\mathrm{B88}} - x_k K + bE_{\mathrm{c}}^{\mathrm{LYP}} + (1 - b)E_{\mathrm{c}}^{\mathrm{VWN}}. \tag{3.45}$$

In the B3LYP functional 20% of the HF is included ($x_k = 0.2$), and the two parameters a and b are 0.72 and 0.81, respectively. Other examples of hybrid functional are the PBE0 [27] and the B97 series [28, 29, 30].

Finally, we would like to mention the class of *range separated functionals*, in which the exchange interaction is described by different mechanisms for short and long range interactions. An example of such a functional is the CAM-B3LYP functional [31].

3.4 Response theory

In response-function theory, we determine the time-development of an observable when the molecular system is subjected to, for example, an external electric or magnetic field. The response of the observable may be expanded in powers of the field strength: the linear response is determined by the linear response function, the quadratic response by the quadratic response function, and so on [32]. Molecular response properties, for example the frequency-dependent polarizability, may be calculated from the response functions by specifying operators for the observable in question as well as the applied field. From the poles and residues of the response functions, additional molecular properties can be obtained, including for example excitation energies and the corresponding transition moments. At most frequencies of the external field, the interaction imposes a small change in the wave function. If the frequency matches an excitation energy, the external field may introduce an excitation that gives rise to large changes in the wave function. We will in the following give an introduction to the basics of response theory, for a more thorough discussion consult for example Ref. [32], upon which this introduction is based.

3.4.1 From the time domain to the frequency domain

The time development of the exact wave function $|\bar{0}\rangle$ is governed by the time-dependent Schrödinger equation

$$H|\bar{0}(t)\rangle = i\frac{\partial}{\partial t}|\bar{0}(t)\rangle. \tag{3.46}$$

The Hamilton operator is decomposed into a time-independent part H_0 and a time-dependent perturbation V^t, according to

$$H = H_0 + V^t. \tag{3.47}$$

We further assume that the perturbation V^t is switched on adiabatically at $t = -\infty$, In the unperturbed limit $t = -\infty$ the time evolution approaches $|\bar{0}(t)\rangle = \exp(-iE_0t)|0\rangle$, where $|0\rangle$ is an eigenfunction of the unperturbed Hamiltonian H_0,

$$H_0|0\rangle = E_0|0\rangle. \tag{3.48}$$

In the frequency domain, the perturbation operator can be written in terms of the Fourier transformation

$$V^t = \int_{-\infty}^{\infty} V^\omega \exp[(-i\omega + \epsilon)t]d\omega, \tag{3.49}$$

where the positive infinitesimal ϵ ensures that the field is switched on adiabatically. The perturbation V^t is required to be Hermitian, which imposes the condition

$$(V^\omega)^\dagger = V^{-\omega}. \tag{3.50}$$

on the frequency components of V^t. At finite time t, we can write the perturbed phase-isolated wave function $|\tilde{0}(t)\rangle = \exp(iF(t))|\bar{0}(t)\rangle$ as a perturbation expansion

$$\begin{aligned} |\tilde{0}(t)\rangle = |0\rangle &+ \int_{-\infty}^{\infty} |0_1^\omega\rangle \exp[(-i\omega + \epsilon)t]d\omega \\ &+ \int_{-\infty}^{\infty} |0_2^{\omega_1,\omega_2}\rangle \exp[(-i(\omega_1 + \omega_2) + 2\epsilon)t]d\omega_1 d\omega_2 + \ldots, \end{aligned} \tag{3.51}$$

where $|0_1^\omega\rangle$ and $|0_2^{\omega_1,\omega_2}\rangle$ contain terms that are linear and quadratic in the perturbations, respectively.

3.4.2 Response functions

Similarly, the averaged expectation value $A_{\text{av.}}(t)$ of an operator A can be expanded to different orders of the perturbation according to

$$\begin{aligned} A_{\text{av.}}(t) &= \langle \bar{0}(t)|A|\bar{0}(t)\rangle \\ &= \langle 0|A|0\rangle + \int_{-\infty}^{\infty} \langle\langle A; V^{\omega}\rangle\rangle_{\omega} \exp[(-i\omega+\epsilon)t] d\omega \\ &\quad + \frac{1}{2}\int_{-\infty}^{\infty} \langle\langle A; V^{\omega_1}, V^{\omega_2}\rangle\rangle_{\omega_1\omega_2} \exp[(-i(\omega_1+\omega_2)+\epsilon)t] d\omega_1 d\omega_2 + \ldots, \end{aligned} \tag{3.52}$$

where $\langle\langle A; V^{\omega}\rangle\rangle_{\omega}$ is the linear response function and $\langle\langle A; V^{\omega_1}, V^{\omega_2}\rangle\rangle_{\omega_1\omega_2}$ is the quadratic response function. In the definition of the response functions we assume that the limit $\epsilon \to 0$ has been taken.

It can be shown [32] that in the basis of exact eigenfunctions $\{|n\rangle\}$ of H_0, for which the normalized time-dependent wave function is given by

$$|\tilde{0}(t)\rangle = \frac{|0\rangle + \sum_n d_n|n\rangle}{\sqrt{1+\mathbf{d}^{\mathrm{T}}\mathbf{d}}}, \tag{3.53}$$

the linear response function can be written as a sum over eigenstates

$$\langle\langle A; V^{\omega}\rangle\rangle_{\omega} = \sum_{n\neq 0} \frac{\langle 0|A|n\rangle\langle n|V^{\omega}|0\rangle}{\omega - (E_n - E_0)} - \sum_{n\neq 0} \frac{\langle 0|V^{\omega}|n\rangle\langle n|A|0\rangle}{\omega + (E_n - E_0)}. \tag{3.54}$$

This equation is called the spectral resolution of the linear response function. In Eq. (3.54) E_n is the energy corresponding to state $|n\rangle$.

3.4.3 Poles and residues

The linear response function has poles at frequencies equal to plus or minus the excitation energies $\omega_n = E_n - E_0$ of the unperturbed system. The corresponding residues, given by

$$\begin{aligned} \lim_{\omega\to\omega_n} (\omega-\omega_n)\langle\langle A; V^{\omega}\rangle\rangle_{\omega} &= \langle 0|A|n\rangle\langle n|V^{\omega}|0\rangle \\ \lim_{\omega\to\omega_n} (\omega+\omega_n)\langle\langle A; V^{\omega}\rangle\rangle_{\omega} &= -\langle 0|V^{\omega}|n\rangle\langle n|A|0\rangle, \end{aligned} \tag{3.55}$$

involve the transition matrix elements. The linear response function thus contains information about the excitation energies from the reference state $|0\rangle$ to an exited state $|n\rangle$, and the corresponding transition matrix elements, which is sufficient information to describe all one-photon processes. Multi-photon processes can similarly be described by means of the residues of higher-order response-functions.

3.4.4 Response equations

By expanding the time-dependent coefficients $\mathbf{d}$ of Eq. (3.53) to different orders of the perturbation, $\mathbf{d} = \mathbf{d}^{(1)} + \mathbf{d}^{(2)} + \ldots$, the responses to different orders can be obtained by solving the corresponding response-equations, obtained from the Schrödinger equation by collecting the terms of the various orders. In the frequency domain the first- and second-order response equations are given by

$$\begin{aligned} (\mathbf{E}^{[2]} - \omega\mathbf{I})\mathbf{d}^{(1)}(\omega) &= -\mathbf{V}^{\omega[1]} \\ (\mathbf{E}^{[2]} - (\omega_1+\omega_2)\mathbf{I})\mathbf{d}^{(2)}(\omega_1,\omega_2) &= -\frac{1}{2}\left(\mathbf{V}^{\omega_1[2]}\mathbf{d}^{(1)}(\omega_2) + \mathbf{V}^{\omega_2[2]}\mathbf{d}^{(1)}(\omega_1)\right), \end{aligned} \tag{3.56}$$

respectively, with the elements of $\mathbf{V}^{\omega[1]}$, $\mathbf{V}^{\omega[2]}$ and $\mathbf{E}^{[2]}$ according to

$$\begin{aligned} V_i^{\omega[1]} &= \langle i|V^\omega|0\rangle \\ V_{ij}^{\omega[2]} &= \langle i|V^\omega|j\rangle - \delta_{ij}\langle 0|V^\omega|0\rangle \\ E_{ij}^{[2]} &= \langle i|H_0|j\rangle - \delta_{ij}\langle 0|H_0|0\rangle = \delta_{ij}(E_i - E_0). \end{aligned} \tag{3.57}$$

The corresponding response-functions are given by

$$\begin{aligned} \langle\langle A; V^\omega\rangle\rangle_\omega &= \mathbf{d}^{(1)\dagger}(-\omega)\mathbf{A}^{[1]} + \mathbf{A}^{[1]\dagger}\mathbf{d}^{(1)}(\omega) \\ \langle\langle A; V^{\omega_1}, V^{\omega_2}\rangle\rangle_{\omega_1,\omega_2} &= \mathbf{d}^{(2)\dagger}(-\omega_1,-\omega_2)\mathbf{A}^{[1]} + \mathbf{A}^{[1]\dagger}\mathbf{d}^{(2)}(\omega_1,\omega_2) \\ &\quad + \mathbf{d}^{(1)\dagger}(-\omega_1)\mathbf{A}^{[2]}\mathbf{d}^{(1)}(\omega_2) + \mathbf{d}^{(1)\dagger}(-\omega_2)\mathbf{A}^{[2]}\mathbf{d}^{(1)}(\omega_1), \end{aligned} \tag{3.58}$$

with

$$\begin{aligned} A_i^{[1]} &= \langle i|A|0\rangle \\ A_{ij}^{[2]} &= \langle i|A|j\rangle - \delta_{ij}\langle 0|A|0\rangle. \end{aligned} \tag{3.59}$$

Hence, to obtain the response functions to different orders one needs to solve response equation of the general form

$$(\mathbf{E}^{[2]} - \omega\mathbf{I})\mathbf{d} = -\mathbf{A}_{\text{rhs}}, \tag{3.60}$$

where the right-hand side $\mathbf{A}_{\text{rhs}}$ varies for different perturbations and response orders. The theory discussed here is for exact theory. For approximate methods, derivations based on the wavefunction parameterization are needed.

Chapter 4

Integral evaluation

Efficient integral evaluation is central to quantum chemistry. There are several different schemes and formalisms to enhance the performance of integral evaluation. In the first section of this chapter, section 4.1, we give a brief overview of the McMurchie-Davidson formalism for integral evaluation, and then, in section 4.2, we outline the integral evaluation scheme presented in Paper I. This scheme enhances performance and simplifies the implementation of differentiated integrals, and further reduces the cost of two- and three-center integral evaluation. In section 4.3 we discuss the efficient and linear-scaling evaluation of the Coulomb contribution, followed by the linear-scaling evaluation of the exchange contribution in section 4.4. Finally, we will give an introduction to the numerical quadrature typically used for the evaluation of the exchange-correlation contribution in section 4.5.

4.1 The McMurchie-Davidson scheme

Most quantum chemistry softwares today use GTOs rather than STOs for integral evaluation. A greater number of GTOs than STOs are needed to obtain the same level of accuracy, but integration over GTOs is both simpler and faster than integration over STOs. The main advantage of using Gaussian basis functions is that they are separable in the Cartesian directions according to the Gaussian product rule [7], which greatly simplifies the integration.

There are several formalisms for molecular integral evaluation over GTOs, for instance the Rys scheme [33], the McMurchie-Davidson scheme [34], the Obara-Saika scheme [35], as well as modifications to these schemes [36, 37]. The different schemes typically has certain advantages and disadvantages, see for instance [7]. We will not

go into the different schemes or their differences in this thesis, but rather focus on the McMurchie-Davidson scheme; which forms the basis of the integral evaluation in DALTON [38]. In this section we summarize some of the key points regarding integral evaluation using the McMurchie-Davidson scheme. A more thorough introduction is given in Ref. [7].

4.1.1 Solid-harmonic Gaussian basis functions

In the McMurchie-Davidson scheme, the integral evaluation over contracted real solid-harmonic GTO basis functions $\chi^{lm}_{a,\mathbf{A}}(\mathbf{r})$ is first carried out over primitive Cartesian GTOs $G^{\mathbf{i}}_{a_m,\mathbf{A}}(\mathbf{r})$ of the form

$$G^{\mathbf{i}}_{a_m,\mathbf{A}}(\mathbf{r}) = x_A^{i_x} y_A^{i_y} x_A^{i_y} \exp(-a_m r_A^2), \tag{4.1}$$

with Cartesian powers $\mathbf{i} = (i_x, i_y, i_z)$, primitive exponent a_m, the center of the Gaussian $\mathbf{A} = (A_x, A_y, A_z)$, and with the distances $r_A = |\mathbf{r} - \mathbf{A}|$, $x_A = x - A_x$ and similarly for y_A and z_A. Then, the primitive Cartesian integrals are both contracted with contraction coefficients C_m to a contracted basis and spherical transformed to the real solid-harmonicl basis. Note that in Eq. (4.1) we have omitted the normalization constant which we include into the contraction coefficients. The contraction and the spherical transformation steps are independent of each other, which means we are free to choose the order of these steps

$$\chi^{l_a m_a}_{a,\mathbf{A}}(\mathbf{r}) = \sum_{\mathbf{i}} \mathcal{S}^{l_a m_a}_{\mathbf{i}} \sum_m C_m G^{\mathbf{i}}_{a_m,\mathbf{A}}(\mathbf{r}) = \sum_m C_m \sum_{\mathbf{i}} \mathcal{S}^{l_a m_a}_{\mathbf{i}} G^{\mathbf{i}}_{a_m,\mathbf{A}}(\mathbf{r}). \tag{4.2}$$

where $\mathcal{S}^{l_a m_a}_{\mathbf{i}}$ denotes the Cartesian to solid-harmonic transformation coefficients [7], with magnetic quantum number m_a.

4.1.2 The expansion of Cartesian overlap distributions in Hermite Gaussians

The *overlap distribution* $\Omega_{ab}(\mathbf{r})$ is the product between two contracted real solid harmonic Gaussians, and can thus be expanded in primitive Cartesian overlap distributions $\Omega^{\mathbf{ij}}_{a_m b_n}(\mathbf{r})$ according to

$$\Omega_{ab}(\mathbf{r}) = \chi^{l_a m_a}_{a,\mathbf{A}}(\mathbf{r}) \chi^{l_b m_b}_{b,\mathbf{B}}(\mathbf{r}) = \sum_{\mathbf{ij}} \mathcal{S}^{ab}_{\mathbf{ij}} \sum_{mn} C_{mn} \Omega^{\mathbf{ij}}_{a_m b_n}(\mathbf{r}), \tag{4.3}$$

with the joint spherical transformation

$$\mathcal{S}^{ab}_{\mathbf{ij}} = \mathcal{S}^{l_a m_a}_{\mathbf{i}} \mathcal{S}^{l_b m_b}_{\mathbf{j}}, \tag{4.4}$$

and with the joint contraction coefficient

$$C_{mn} = C_m C_n. \tag{4.5}$$

The primitive Cartesian overlap distribution is by the Gaussian product theorem [7] given as

$$\Omega_{ab}^{\mathbf{ij}}(\mathbf{r}) = G_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r}) G_{b,\mathbf{B}}^{\mathbf{j}}(\mathbf{r}) = x_A^{i_x} y_A^{i_y} z_A^{i_z} x_B^{j_x} y_B^{j_y} z_B^{j_z} \exp(-\mu R_{AB}^2) \exp(-p r_P^2), \tag{4.6}$$

with $p = a + b$, $\mu = ab/p$, $R_{AB} = |\mathbf{A} - \mathbf{B}|$ and $\mathbf{P} = (a\mathbf{A} + b\mathbf{B})/p$. In the McMurchie-Davidson scheme this overlap distribution is again written as a linear combination of Hermite Gaussians [34]

$$\Omega_{ab}^{\mathbf{ij}}(\mathbf{r}) = \sum_{t+u+v=0}^{l_a+l_b} E_{tuv}^{\mathbf{ij}} \Lambda_{tuv}(\mathbf{r}), \tag{4.7}$$

with $l_a = i_x + i_y + i_z$ and similarly for l_b, with the Hermite to Cartesian transformation coefficients $E_{tuv}^{\mathbf{ij}}$, and where the Hermite Gaussians are given by

$$\Lambda_{tuv}(\mathbf{r}) = \frac{\partial^{t+u+v}}{\partial P_x^t \partial P_y^u \partial P_z^v} \exp(-p r_P^2). \tag{4.8}$$

Note that when calculating the overlap distribution between real solid harmonic Gaussians, the intermediate transformation to Cartesian primitives is not necessary, as we can transform directly to the real solid harmonic basis according to,

$$\Omega_{ab}(\mathbf{r}) = \sum_{mn} C_{mn} \sum_{tuv}^{l_a+l_b} E_{tuv}^{a_m b_n} \Lambda_{tuv}^{mn}(\mathbf{r}), \tag{4.9}$$

with $\Lambda_{tuv}^{mn}(\mathbf{r})$ the Hermite Gaussian for primitive pair mn, and with

$$E_{tuv}^{ab} = \sum_{\mathbf{ij}} \mathcal{S}_{\mathbf{ij}}^{ab} E_{tuv}^{\mathbf{ij}}. \tag{4.10}$$

The differentiations in the three different Cartesian directions are independent of each other, which leads to simple recurrence relations for integrals over Hermite Gaussians. Note that since the Gaussians are separable in the three Cartesian directions, we can separate the E-coefficients $E_{tuv}^{\mathbf{ij}} = E_t^{i_x j_x} E_u^{i_y j_y} E_v^{i_z j_z}$. The recurrence for the E_t^{ij} is given as [7]

$$\begin{aligned} E_0^{i+1,j} &= X_{PA} E_0^{ij} + E_1^{ij} \\ E_0^{i,j+1} &= X_{PB} E_0^{ij} + E_1^{ij} \\ E_t^{i,j} &= \frac{1}{2pt} \left(i E_{t-1}^{i-1,j} + j E_{t-1}^{i,j-1} \right), \quad t > 0, \end{aligned} \tag{4.11}$$

starting from $E_0^{00} = \exp(-\mu X_{AB}^2)$.

4.1.3 One-electron integrals

We are now ready to look at the integral evaluation of the different integrals over real solid-harmonic Gaussians $\chi_{a,\mathbf{A}}^{l_a m_a}(\mathbf{r})$. We start out with the multipole-moment integrals from which the overlap integrals are a special case. The overlap integrals are fundamental for AO-based DFT, and the multipole-moment integrals are used in for example the fast multipole-moment method, see section 4.3, to achieve linear-scaling Coulomb evaluation. The Cartesian multipole-moment integrals $S_{ab}^{\mathbf{e}}$ between two contracted real solid harmonic GTOs, $\chi_{a,\mathbf{A}}^{l_a m_a}(\mathbf{r})$ and $\chi_{b,\mathbf{B}}^{l_b m_b}(\mathbf{r})$, expanded at center $\mathbf{C}$ with the Cartesian powers $x_C^{e_x} y_C^{e_y} z_C^{e_z}$, can be written as a linear combination of primitive Cartesian multipole-moment integrals $S_{a_m b_n}^{\mathbf{ije}}$, using Eq. (4.2),

$$\begin{aligned} S_{ab}^{\mathbf{e}} &\equiv \int \Omega_{ab}(\mathbf{r}) x_C^{e_x} y_C^{e_y} z_C^{e_z} d\mathbf{r} \\ &= \sum_{\mathbf{ij}} \mathcal{S}_{\mathbf{ij}}^{ab} \sum_{mn} C_{mn} \int \Omega_{a_m b_n}^{\mathbf{ij}}(\mathbf{r}) x_C^{e_x} y_C^{e_y} z_C^{e_z} d\mathbf{r} \\ &= \sum_{\mathbf{ij}} \mathcal{S}_{\mathbf{ij}}^{ab} \sum_{mn} C_{ml} S_{a_m b_n}^{\mathbf{ije}}. \end{aligned} \tag{4.12}$$

These primitive Cartesian multipole-moment integrals can again be separated into the Cartesian directions according to

$$S_{ab}^{\mathbf{ije}} = S_{ab}^{i_x j_x e_x} S_{ab}^{i_y j_y e_y} S_{ab}^{i_z j_z e_z}, \tag{4.13}$$

which following the McMurchie-Davidson scheme are given by

$$S_{ab}^{ije} = \int \Omega_{ab}^{ij}(\mathbf{x}) x_C^e d\mathbf{x} = \sum_{t=0}^{i+j} E_t^{ij} \int \Lambda_t(\mathbf{x}) x_C^e d\mathbf{x} = \sum_{t=0}^{i+j} E_t^{ij} M_t^e. \tag{4.14}$$

The multipole-moment integrals over the Hermite Gaussians M_t^e can be found by recurrence according to [7],

$$M_t^{e+1} = t M_{t-1}^e + X_{PC} M_t^e + \frac{1}{2p} M_{t+1}^e, \tag{4.15}$$

starting from

$$M_t^0 = \delta_{t0} \sqrt{\frac{\pi}{p}}, \tag{4.16}$$

which follows directly by taking the differentiation outside the integration, and by noting that $M_t^e = 0$ for all $t > e$. For the primitive Cartesian overlap integrals this gives the simple expression

$$S_{ab}^{\mathbf{ij}} \equiv S_{ab}^{\mathbf{ij0}} = \sum_{t+u+v=0}^{l_a+l_b} E_{tuv}^{\mathbf{ij}} M_t^0 M_u^0 M_v^0 = E_{000}^{\mathbf{ij}} \left(\frac{\pi}{p}\right)^{3/2}. \tag{4.17}$$

4.1.4 Nuclear attraction integrals

We now turn our attention to the evaluation of Coulomb integrals. Similarly to the overlap integrals, the overlap distribution between two contracted real solid-harmonic Gaussians, Eq. (4.3), are expanded in primitive Cartesian overlap distributions, which are again expanded in Hermite Gaussians according to Eq. (4.7). The nuclear attraction part h_{ab}^{na}, of the one-electron part of the Fock/KS matrix Eqs. (3.15) and (3.35), is given by

$$h_{ab}^{\mathrm{na}} = -\sum_{K} Z_K V_{ab}(\mathbf{C_K}) = -\sum_{\mathbf{ij}} \mathcal{S}_{\mathbf{ij}}^{ab} \sum_{mn} C_{mn} \sum_{K} Z_K V_{a_m b_n}^{\mathbf{ij}}(\mathbf{C}_K). \tag{4.18}$$

where $\mathbf{C_K}$ is the center and Z_K the nuclear charge of nuclei K, and where the primitive Cartesian Coulomb potential $V_{ab}^{\mathbf{ij}}(\mathbf{C})$ is given by

$$V_{ab}^{\mathbf{ij}}(\mathbf{C}) = \int \Omega_{ab}^{\mathbf{ij}}(\mathbf{r}) \frac{1}{r_C} d\mathbf{r} = \frac{2\pi}{p} \sum_{tuv} E_{tuv}^{\mathbf{ij}} R_{tuv}(p, \mathbf{R}_{PC}). \tag{4.19}$$

4.1.5 The Hermite Coulomb integrals

The Hermite Coulomb integrals $R_{tuv}(p, \mathbf{R}_{PC})$ are given as the derivatives of the zeroth order Boys Function $F_0(x)$, according to [34]

$$R_{t,u,v}\left(p, \mathbf{R}_{PC}\right) = \frac{\partial^{t+u+v} F_0(pR_{PC}^2)}{\partial P_x^t \partial P_x^u \partial P_x^v}, \tag{4.20}$$

with the nth-order Boy's function given as

$$F_n(x) = \int_0^1 e^{-xt^2} t^{2n} \mathrm{d}t. \tag{4.21}$$

The integrals $R_{t,u,v} \equiv R_{t,u,v}^0$ can be found by recursion from the spherical integrals

$$R_{000}^n(p, \mathbf{R}_{PC}) = (-2p)^n F_n(pR_{PC}^2) \tag{4.22}$$

of orders $n \leq t+u+v$, according to the recurrence relations

$$\begin{aligned}
R_{t+1,u,v}^n(p, \mathbf{R}_{PC}) =& t R_{t-1,u,v}^{n+1}(p, \mathbf{R}_{PC}) + X_{PC} R_{t,u,v}^{n+1}(p, \mathbf{R}_{PC}) \\
R_{t,u+1,v}^n(p, \mathbf{R}_{PC}) =& u R_{t,u-1,v}^{n+1}(p, \mathbf{R}_{PC}) + Y_{PC} R_{t,u,v}^{n+1}(p, \mathbf{R}_{PC}) \\
R_{t,u,v+1}^n(p, \mathbf{R}_{PC}) =& v R_{t,u,v-1}^{n+1}(p, \mathbf{R}_{PC}) + Z_{PC} R_{t,u,v}^{n+1}(p, \mathbf{R}_{PC}).
\end{aligned} \tag{4.23}$$

4.1.6 Two-electron Coulomb repulsion integrals

We now turn our attention to the four-center two-electron Coulomb repulsion integrals $(ab|cd)$, which play a central role in quantum chemistry. The contracted four-center

two-electron integrals $(ab|cd)$ can be written as a linear combination of primitive four-center two-electron $[a_m b_n|c_r d_s]$ according to

$$(ab|cd) = \sum_{mn} C_{mn} \sum_{rs} C_{rs}[a_m b_n|c_r d_s]. \tag{4.24}$$

Here, the primitive solid harmonic Coulomb integrals $[ab|cd]$ are, similarly to the primitive Cartesian Coulomb potential of Eq. (4.19), given as

$$\begin{aligned}
[ab|cd] &= \sum_{\mathbf{ij}} \mathcal{S}^{ab}_{\mathbf{ij}} \sum_{\mathbf{kl}} \mathcal{S}^{cd}_{\mathbf{kl}} \int \Omega^{\mathbf{ij}}_{ab}(\mathbf{r}_1) \frac{1}{r_{12}} \Omega^{\mathbf{kl}}_{cd}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\
&= \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \sum_{\mathbf{ij}} \mathcal{S}^{ab}_{\mathbf{ij}} \sum_{tuv} E^{\mathbf{ij}}_{tuv} \sum_{\mathbf{kl}} \mathcal{S}^{cd}_{\mathbf{kl}} \sum_{\tau\nu\phi} (-1)^{\tau+\nu+\phi} E^{\mathbf{kl}}_{\tau\nu\phi} R_{t+\tau,u+\nu,v+\phi}(\alpha, \mathbf{R}_{PQ}) \\
&= \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \sum_{tuv} E^{ab}_{tuv} \sum_{\tau\nu\phi} (-1)^{\tau+\nu+\phi} E^{cd}_{\tau\nu\phi} R_{t+\tau,u+\nu,v+\phi}(\alpha, \mathbf{R}_{PQ}),
\end{aligned} \tag{4.25}$$

with $q = c + d$, $\mathbf{Q} = (c\mathbf{C} + d\mathbf{D})/q$, $\mathbf{R}_{PQ} = \mathbf{P} - \mathbf{Q}$ and $\alpha = pq/(p+q)$. Note that to improve performance, contractions that reduce the number of intermediates should be carried out as early as possible. This means that we do not calculate the primitive real solid-harmonic Coulomb integrals $[ab|cd]$ explicitly, but rather carry out contractions and transformations on the two electrons separately. The McMurchie-Davidson algorithm for four-center two-electron integrals is outlined in figure 4.1. The loops are carried out over *shell pairs* ab and cd, as integrals between such pairs share intermediates. A shell pair ab consists of basis functions from the *shells* of two atoms, each shell sharing both primitives and angular momentum.

The permutational symmetries of the four-center two-electron integrals

$$\begin{aligned}
(ab|cd) &= (ba|cd) = (ba|dc) = (ab|dc) \\
&= (cd|ab) = (cd|ba) = (dc|ba) = (dc|ab)
\end{aligned} \tag{4.26}$$

is exploited by limiting the two loops in figure 4.1 so that for example $a \geq b \geq c \geq d$, which reduce the number of computations by (up to) a factor eight. Note that as a special case two- and three-center two-electron integrals can be calculated using the algorithm outlined in figure 4.1. This can be done by expanding the primitive solid-harmonic Gaussian basis functions in Hermite Gaussians, instead of the overlap distribution of Eq. (4.7), according to

$$\chi^{l_a m_a}_{a,\mathbf{A}}(\mathbf{r}) = \sum_{t+u+v=0}^{l_a} E^{a}_{tuv} \Lambda_{tuv}(\mathbf{r}). \tag{4.27}$$

Loop ab shell pairs

  Set p, $\mathbf{P}$, E_{tuv}^{ab} (for all primitive pairs mn)

    Loop cd shell pairs

      Set q, $\mathbf{Q}$, $E_{\tau\nu\phi}^{cd}$ (for all primitive pairs rs)

      Set α, $\mathbf{R}_{PQ}$ (for all primitive quadruples $mnrs$)

      Build $F_n(\alpha, \mathbf{R}_{PQ})$ and $R_{t+\tau,u+\nu,v+\phi}(\alpha, \mathbf{R}_{PQ})$

      Contract $(tuv|c_r d_s] = \sum_{\tau\nu\phi} (-1)^{\tau+\nu+\phi} E_{\tau\nu\phi}^{c_r d_s} R_{t+\tau,u+\nu,v+\phi}(\alpha, \mathbf{R}_{PQ})$

      Contract $(tuv|cd) = \sum_{rs} (tuv|c_r d_s] C_{rs}$

      Contract $[a_m b_n|cd) = \sum_{tuv} E_{tuv}^{a_m b_n} (tuv|cd)$

      Contract $(ab|cd) = \sum_{mn} C_{mn} [a_m b_n|cd)$

    End loop cd

End loop ab

Figure 4.1: Outline of the McMurchie-Davidson algorithm for four-center two-electron integrals.

4.2 McMurchie-Davidson using Hermite primitives

In Paper I we develop a new scheme for integral evaluation in which the intermediate integration is carried out over Hermite Gaussians $H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ rather than the Cartesian Gaussians $G^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ of Eq. (4.1). This reduces computational efforts for both the evaluations of differentiated integrals and for the evaluation of two- and three-center integrals, in addition to making the expressions for differentiated integrals simpler.

4.2.1 Solid-harmonic Gaussians expanded in Hermite rather than Cartesian Gaussian primitives

The new integration scheme follows from the fact that the solid-harmonic combinations of Cartesian, $G^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$, and Hermite Gaussians, $H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$, are identical,

$$\chi^{l_a m_a}_{a,\mathbf{A}} = \sum_{i+j+k=l_a} \mathcal{S}^{l_a m_a}_{\mathbf{i}} H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r}) = \sum_{i+j+k=l_a} \mathcal{S}^{l_a m_a}_{\mathbf{i}} G^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r}). \tag{4.28}$$

The primitive Hermite Gaussians $H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ are taken to be scaled versions of the Hermite Gaussians $\Lambda_{tuv}(\mathbf{r})$ of Eq. (4.8), according to

$$H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r}) = \frac{1}{(2a)^{l_a}} \Lambda_{\mathbf{i}}(\mathbf{r}) = \frac{1}{(2a)^{l_a}} \mathcal{H}_{\mathbf{i}}(2a, \mathbf{r}) \exp(-a r_A^2), \tag{4.29}$$

with the Hermite polynomial $\mathcal{H}_{ijk}(2a, \mathbf{r})$ given by

$$\mathcal{H}_{ijk}(2a, \mathbf{r}) = \mathcal{H}_i(2a x_A) \mathcal{H}_j(2a y_A) \mathcal{H}_k(2a z_A), \tag{4.30}$$

and with $\mathcal{H}_t(x)$ defined by the Rodrigues expression

$$\mathcal{H}_t(x) = (-1)^t \exp(x^2) \frac{d^t}{dx^t} \exp(-x^2). \tag{4.31}$$

With the above definition, it follows that the leading terms of the polynomials in front of the exponent $\exp(-a r_A^2)$ of both the Cartesian Gaussian $G^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ and the scaled Hermite Gaussian $G^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ are equal to $x_A^{i_x} y_A^{i_y} z_A^{i_z}$. Following Paper I, the identity of the two solid-harmonic transformations of Eq. (4.28) follows because only the homogeneous polynomial of degree $l_a = i_x + i_y + i_z$, with unity prefactor, remains after the transformation.

Although the new integration scheme in Paper I is developed in both the Obara-Saika and the McMurchie-Davidson formalisms, we will limit the scope in this section to the McMurchie-Davidson formalism, and demonstrate the benefits of this new scheme using the McMurchie-Davidson formalism.

4.2.2 The expansion of Hermite primitive overlap distributions

Similarly to the overlap distribution between two primitive Cartesian Gaussians $\Omega_{ab}^{\mathbf{ij}}(\mathbf{r})$, Eq. (4.7), the overlap distribution between two primitive Hermite Gaussians $\Omega_{ab}^{\mathbf{ij}}(\mathbf{r})$ can be expressed as

$$\Omega_{ab}^{\mathbf{ij}}(\mathbf{r}) = H_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r}) H_{b,\mathbf{B}}^{\mathbf{j}}(\mathbf{r}) = \sum_{t+u+v=0}^{l_a+l_b} \mathcal{E}_{tuv}^{\mathbf{ij}} \Lambda_{tuv}(\mathbf{r}). \tag{4.32}$$

Note that the overlap distribution has here been expanded in the Hermite Gaussians $\Lambda_{tuv}(\mathbf{r})$ to coincide with the expansion of Eq. (4.7), whereas in Paper I the expansion is instead in the scaled Hermite Gaussians $H_{p,\mathbf{P}}^{tuv}(\mathbf{r})$. This choice has been made in order for the intermediate integrals over Hermite Gaussians to be identical to the integrals developed in the previous section. The Hermite E-coefficients $\mathcal{E}_{tuv}^{\mathbf{ij}}$ are also separated into E-coefficients for the three Cartesian directions $\mathcal{E}_t^{ij}$,

$$\mathcal{E}_{tuv}^{\mathbf{ij}} = \mathcal{E}_t^{i_x j_x} \mathcal{E}_u^{i_y j_y} \mathcal{E}_v^{i_z j_z}, \tag{4.33}$$

which can again be obtained by the recurrence relations

$$\begin{aligned} \mathcal{E}_0^{i+1,j} &= X_{PA} \mathcal{E}_0^{i,j} + \mathcal{E}_1^{ij} - \frac{i}{2a} \mathcal{E}_0^{i-1,j} \\ \mathcal{E}_0^{i,j+1} &= X_{PB} \mathcal{E}_0^{i,j} + \mathcal{E}_1^{ij} - \frac{j}{2b} \mathcal{E}_0^{i,j-1} \\ \mathcal{E}_t^{ij} &= \frac{1}{2pt} \left(i \mathcal{E}_{t-1}^{i-1,j} + j \mathcal{E}_{t-1}^{i,j-1} \right), \quad t > 0. \end{aligned} \tag{4.34}$$

The recurrence relations for the E-coefficients in Hermite compared to the Cartesian representation, Eq. (4.34) and Eq. (4.11) respectively, are equal except for the additional third term in the two first expressions of Eq. (4.34). The integrals using primitive Hermite Gaussians basis functions follows exactly the same formulas as for the integrals using primitive Cartesian integrals, by replacing all $E_{tuv}^{\mathbf{ij}}$ and E_{tuv}^{ab} by $\mathcal{E}_{tuv}^{\mathbf{ij}}$ and $\mathcal{E}_{tuv}^{ab}$ in Eqs. (4.17), (4.19) and (4.25). We note, that computationally these recurrence relations are fast compared to other steps in the integral evaluation, so the added terms of Eq. (4.34) are negligible with respect to computational time.

4.2.3 Differentiated integrals

The advantage of using Hermite Gaussians becomes clear for differentiated integrals and for two- and three-center integrals. When differentiating the primitive Hermite

Gaussian of Eq. (4.29) with respect to the Gaussian center we get another single primitive Hermite Gaussian, according to

$$H^{\mathbf{i},\mathbf{I}}_{a,\mathbf{A}}(\mathbf{r}) \equiv \frac{\partial^{Ia}}{\partial \mathbf{A}^{\mathbf{I}}} H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r}) = (2a)^I H^{\mathbf{i}+\mathbf{I}}_{a,\mathbf{A}}(\mathbf{r}), \tag{4.35}$$

with $\mathbf{I} = (I_x, I_y, I_z)$, $I = I_x + I_y + I_z$ and with

$$\frac{\partial^I}{\partial \mathbf{A}^{\mathbf{I}}} = \frac{\partial^I}{\partial A_x^{I_x} \partial A_y^{I_y} \partial A_z^{I_z}}. \tag{4.36}$$

The corresponding derivative of a primitive Cartesian Gaussian, Eq. (4.1), gives a linear combination of several primitive Cartesian Gaussians. Clearly, expanding the solid-harmonic Gaussians in Hermite rather than Cartesian Gaussians is preferable for geometrical derivatives. To give an example, let us consider a differentiated four-center two-electron integral. First, consider the differentiation of overlap distribution of Eq. (4.32), which gives

$$\Omega^{\mathbf{ij},\mathbf{IJ}}_{ab}(\mathbf{r}) = \frac{\partial^{I+J} \Omega^{\mathbf{ij}}_{ab}(\mathbf{r})}{\partial \mathbf{A}^{\mathbf{I}} \partial \mathbf{B}^{\mathbf{J}}} = (2a)^I (2b)^J \sum_{tuv}^{l_a+I+l_b+J} \mathcal{E}^{\mathbf{i}+\mathbf{I},\mathbf{j}+\mathbf{J}}_{tuv} \Lambda_{tuv}. \tag{4.37}$$

The differentiated overlap distribution use the same expansion coefficients $\mathcal{E}^{\mathbf{ij}}_{tuv}$ as in Eq. (4.32), only expanded to higher order. This gives the differentiated primitive solid-harmonic four-center two-electron integral

$$\begin{aligned}
[ab|cd]^{\mathbf{IJKL}} =& \frac{\partial^{I+J+K+L}}{\partial \mathbf{A}^{\mathbf{I}} \partial \mathbf{B}^{\mathbf{J}} \partial \mathbf{C}^{\mathbf{K}} \partial \mathbf{D}^{\mathbf{L}}} [ab|cd] \\
=& \sum_{\mathbf{ij}} \mathcal{S}^{ab}_{\mathbf{ij}} \sum_{\mathbf{kl}} \mathcal{S}^{cd}_{\mathbf{kl}} \int \Omega^{\mathbf{ij},\mathbf{IJ}}_{ab}(\mathbf{r}_1) \frac{1}{r_{12}} \Omega^{\mathbf{kl},\mathbf{KL}}_{cd}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\
=& \frac{2\pi^{5/2}}{pq\sqrt{p+q}} (2a)^I (2b)^J (2c)^J (2d)^K \sum_{\mathbf{ij}} \mathcal{S}^{ab}_{\mathbf{ij}} \sum_{tuv} \mathcal{E}^{\mathbf{i}+\mathbf{I},\mathbf{j}+\mathbf{J}}_{tuv} \\
& \sum_{\mathbf{kl}} \mathcal{S}^{cd}_{\mathbf{kl}} \sum_{\tau\nu\phi} (-1)^{\tau+\nu+\phi} \mathcal{E}^{\mathbf{k}+\mathbf{K},\mathbf{l}+\mathbf{L}}_{\tau\nu\phi} R_{t+\tau,u+\nu,v+\phi}(\alpha, \mathbf{R}_{PQ}).
\end{aligned} \tag{4.38}$$

The differentiated integrals of Eq. (4.38) can be obtained (to any order) making only minor modification of an existing code for the undifferentiated integrals of Eq. (4.25). Note that the summation over tuv goes from $0 \leq t+u+v \leq l_a + I + l_b + J$ and similarly for the summation over $\tau\nu\phi$, so the order of expansion of both the expansion coefficients $\mathcal{E}^{\mathbf{ij}}_{tuv}$ and the Hermite Coulomb integrals $R_{tuv}(\alpha, \mathbf{R}_{PQ})$ are increased accordingly for the differentiated integrals. Thus, the number of intermediate Hermite integrals $R_{t+\tau,u+\nu,v+\phi}(\alpha, \mathbf{R}_{PQ})$ is the same as when using the Cartesian Gaussian intermediates, but the number of contractions with the E-coefficients is greatly reduced

with increasing angular momentum. For instance the first derivatives with respect to both center **C** and **D** of a *cd* shell pair consisting of two p-orbitals, gives a total of 54 solid-harmonic derivative components (two centers, three Cartesian directions, three components per p-orbital). Using Cartesian Gaussians all these components are generated by contracting (both differentiated and undifferentiated) E-coefficients with the Hermite integrals, $R_{tuv}(\alpha, \mathbf{R}_{PQ})$, whereas when using Hermite Gaussians only the modified E-coefficients of the 36 Hermite Gaussian pairs (the two combinations of dp- and pd-orbital pairs) are contracted with the Hermite integrals. The spherical transformation to the solid-harmonic basis can be postponed to a later stage.

4.2.4 Two- and three-center integrals

For the two- and three-center integrals the expressions become even simpler. The differentiated two-center two-electron integrals become

$$\begin{aligned}
[p|q]^{\mathbf{MN}} =& \frac{\partial^{M+N}}{\partial \mathbf{P}^{\mathbf{M}} \partial \mathbf{Q}^{\mathbf{N}}} [p|q] \\
=& \sum_{\mathbf{m}} \mathcal{S}^{p}_{\mathbf{m}} \sum_{\mathbf{n}} \mathcal{S}^{q}_{\mathbf{n}} \int H^{\mathbf{m},\mathbf{M}}_{p,\mathbf{P}}(\mathbf{r}_1) \frac{1}{r_{12}} H^{\mathbf{n},\mathbf{N}}_{q,\mathbf{Q}}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\
=& \frac{2\pi^{5/2}}{pq\sqrt{p+q}} (2p)^{M-l_p} (-2q)^{N-l_q} \sum_{\mathbf{m}} \mathcal{S}^{p}_{\mathbf{m}} \sum_{\mathbf{n}} \mathcal{S}^{q}_{\mathbf{n}} R_{\mathbf{m}+\mathbf{M}+\mathbf{n}+\mathbf{N}}(\alpha, \mathbf{R}_{PQ}),
\end{aligned} \tag{4.39}$$

with $\mathbf{M} = (M_x, M_y, M_z)$, $M = M_x + M_y + M_z$ and similarly for $\mathbf{N}$ and N, and the differentiated three-center two-electron integrals become

$$\begin{aligned}
[p|cd]^{\mathbf{MKL}} =& \frac{\partial^{M+K+L}}{\partial \mathbf{P}^{\mathbf{M}} \partial \mathbf{C}^{\mathbf{K}} \partial \mathbf{D}^{\mathbf{L}}} [p|cd] \\
=& \sum_{\mathbf{m}} \mathcal{S}^{p}_{\mathbf{m}} \sum_{\mathbf{kl}} \mathcal{S}^{cd}_{\mathbf{kl}} \int H^{\mathbf{m},\mathbf{M}}_{p,\mathbf{P}}(\mathbf{r}_1) \frac{1}{r_{12}} \Omega^{\mathbf{kl},\mathbf{KL}}_{cd}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\
=& \frac{2\pi^{5/2}}{pq\sqrt{p+q}} (2p)^{M-l_p} (2c)^{J} (2d)^{K} \sum_{\mathbf{m}} \mathcal{S}^{p}_{\mathbf{m}} \sum_{\mathbf{kl}} \mathcal{S}^{cd}_{\mathbf{kl}} \sum_{\tau\nu\phi} \\
& (-1)^{\tau+\nu+\phi} \mathcal{E}^{\mathbf{k}+\mathbf{K},\mathbf{l}+\mathbf{L}}_{\tau\nu\phi} R_{m_x+M_x+\tau, m_y+M_y+\nu, m_z+M_z+\phi}(\alpha, \mathbf{R}_{PQ}),
\end{aligned} \tag{4.40}$$

Note the significant reduction in the number of terms both for higher angular momentum and for higher order of differentiation. For instance for the undifferentiated three-center integrals of Eq. (4.40), with $M = K = L = 0$, the number of intermediates $(tuv|cd)$, see figure 4.1, for a shell p is reduced from $(l_p+M+1)(l_p+M+2)(l_p+M+3)/6$ terms to only $(l_p+M+1)(l_p+M+2)/2$ terms, per solid-harmonic component of the shell pair *cd*, by using Hermite rather than Cartesian primitives. This reduce the number

of innermost contractions (cd) over E-coefficients accordingly. Finally, the outermost contraction (p) with E-coefficients is replaced by a simple scaling.

4.3 The Coulomb contribution

The Coulomb contribution to the Fock or KS matrices of Eqs. (3.15) and (3.35), the Coulomb matrix

$$J_{ab} = \sum_{cd} (ab|cd) D_{cd} = (ab|\rho), \tag{4.41}$$

can be obtained in a linear scaling fashion by combining integral screening [39, 40] and the continuous fast multipole-method (CFMM) [41]. Several approaches to further accelerate the construction of the Coulomb matrix exists in the literature. These include J-engine based integral evaluation [42, 43, 44, 45], the Fourier transform Coulomb [46], Cholesky decomposition [47] and density-fitting approximations [48, 49, 5, 50]. We will in this section limit the scope to integral screening, CFMM and J-engine, and later, in chapter 5, we will discuss the density-fitting approach.

4.3.1 Integral screening

The overlap distribution $\Omega_{ab}(\mathbf{r})$ decays exponentially with the square of the distance R_{AB} between the centers of the two solid-harmonic Gaussians, according to Eqs. (4.3) and (4.6), and thus becomes negligible with distance. Therefore, the number of non-negligible overlap distributions $\Omega_{ab}(\mathbf{r})$ scales as $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$. A rigorous upper bound to the absolute value of two-electron Coulomb repulsion integrals can be attained by Cauchy-Schwarz (CS) screening of Häser and Ahlrichs [39],

$$|(f|g)| \leq \sqrt{(f|f)}\sqrt{(g|g)}. \tag{4.42}$$

Note that there is also a nice proof of this inequality by Whitten [48]. When applied to the four-center two-electron integrals the scaling is reduced from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^2)$. The CS screening does not, however, account for the $1/R$ distance decay (or faster) between the centers of the two non-overlapping charge-distributions $g(\mathbf{r}_1)$ and $f(\mathbf{r}_2)$. This distance decay can be incorporated using the multipole-based integral estimate of Lambrecht and Ochsenfeld [40].

4.3.2 The fast multipole method

The fast multipole-method (FMM) of Greengard and Rokhlin [51] was introduced for gravitational forces in astronomy to achieve linear scaling for long-range particle interactions, and can straightforwardly be applied to the interaction between charges. In FMM a distribution of particles is expanded in multipole moments around some shared point in space, and the approximate interaction between two well separated particle distributions can accurately be represented by their multipole moment interactions. Larger and larger distributions can be treated to the same level of accuracy when the distance between the two particle distributions increase, and by systematically enlarging the size of the distributions with distance, linear scaling can be achieved.

More specifically, the FMM procedure starts by forming a parent box that contains all particles. This parent box is then bisected along each Cartesian axis. Each child box is further subdivided and so forth forming a computational *family tree*. The number of subdivisions is chosen such that the number of particles at the lowest level is (approximately) independent of the total number of particles. Each particle is placed within a box on the lowest level, and the multipole moment of the charges contained in a given box is then expanded in multipole moments about the center of the box; all empty boxes are removed. The multipole moments of the lowest level boxes are then translated to the parent boxes at the next level and so forth up through the family tree. The next step constitutes building up the potential from all boxes at a given level that are well separated and which are not already included at a higher level. Then, the potentials at the different levels are translated from the parent boxes and added to their children all the way down the tree. The next step constitutes the calculation of the far-field potential at the position of each particle from the Taylor expansion in the center of the lowest level boxes. Finally, the interactions between the particles and the far-field potential are calculated.

4.3.3 The continuous fast multipole method

FMM was first used for molecular systems by Ding, Karasawa and Goddard [52] in their cell multipole method, and was extended to continuous charge distributions in the CFMM approach by White and Head-Gordon [41]. Contrarily to point charges, the continuous distributions overlap in space. To illustrate how FMM is extended to continuous distributions, we consider the Coulomb interaction U_{pq} between two

spherical charge distribution $\exp(-pr_P^2)$ and $\exp(-qr_Q^2)$

$$U_{pq} = \int \frac{\exp(-pr_{1P}^2)\exp(-qr_{2P}^2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2, \tag{4.43}$$

which, following Watson *et* al. [45], can be separated into a classical U_{pq}^{cls} and a non-classical contribution U_{pq}^{non}

$$U_{pq} = U_{pq}^{\text{cls}} + U_{pq}^{\text{non}}. \tag{4.44}$$

Although, in principle a Gaussian extend over the entire space, it decays exponentially with the square of the distance from its center. Therefore, it is sufficient to define some extent r_p for each charge distribution $\exp(-pr_P^2)$. Using the definition of the extent by Watson *et* al.

$$r_p = \frac{1}{\sqrt{p}} \text{erfc}^{-1}(10^{-k}), \tag{4.45}$$

it follows [7] that for two spherical Gaussians, separated by more than the sum of their extents, we have

$$\left| \frac{U_{pq}^{\text{non}}}{U_{pq}^{\text{cls}}} \right| < 10^{-k}, \tag{4.46}$$

for some choice k. This rigorous upper bound to the ratio between the non-classical and the classical contributions allows for a division of the contributions that must be treated explicitly and those that can be treated with multipole-moment interactions. Note from Eq. (4.45) that, as expected, it follows that the smaller the exponent the longer the extent. In the CFMM approach [41], distributions with differing extents are classified into *branches* according to how many boxes that must separate two distributions (at a given level of the FMM tree structure) in order for the interactions to be treated classically. This gives rise to multiple branches of varying extents that makes the CFMM both more cumbersome and slower than regular FMM. Note that by Eq. (4.44) it is possible to separate each individual interactions into purely classical and non-classical contributions, so that all well separated classical contributions can be treated directly with FMM rather than CFMM [45]. Note, that in this approach the classical contributions are not only charges, but rather multipoles.

4.3.4 The *J*-engine approaches

In the *J*-engine schemes, the density-matrix elements are contracted in an early stage to bypass the explicit calculation of four-center two-electron Coulomb integrals, as proposed independently by Ahmadi and Almlöf [42] and by White and Head-Gordon [43]. The two methods differ only by the formalism used for integral evaluation - the

McMurchie-Davidson scheme [34] by Ahmadi and Almlöf and a variation of the Obara-Saika scheme [35] by White and Head-Gordon. White and Head-Gordon later presented an improved J-engine scheme based on the McMurchie-Davidson integral formalism. The speed-up of J-engine compared to integral evaluation schemes based on explicit four-center integrals typically range from factor 2-10, and is greater with increasing angular momentum functions. The implementation adopted in `DALTON` [38] is presented in Watson et al. [45], and is based on the scheme by Ahmadi and Almlöf. In this scheme the electron density $\rho(\mathbf{r})$ is expanded according to

$$\rho(\mathbf{r}) = \sum_{ab} \Omega_{ab}(\mathbf{r}) D_{ab} = \sum_{p,\mathbf{P}} \sum_{t+u+v=0}^{l_a+l_b} (-1)^{t+u+v} F_{tuv}^{p,\mathbf{P}} \Lambda_{\mathbf{t}}(p, \mathbf{r}_P), \tag{4.47}$$

with

$$F_{tuv}^{p,\mathbf{P}} = (-1)^{t+u+v} \sum_{mn \in p,\mathbf{P}} C_m C_n E_{tuv}^{a_m b_n} D_{a_m b_n} \tag{4.48}$$

with $mn \in p, \mathbf{P}$ meaning all primitive overlap distributions sharing both exponent p and center $\mathbf{P}$. For the Coulomb matrix of Eq. (4.41) this then gives

$$J_{ab} = \sum_{mn} C_m C_n \sum_{t+u+v=0}^{l_a+l_b} E_{tuv}^{a_m b_n} \sum_{q,\mathbf{Q}} \frac{2\pi^{5/2}}{p_{mn} q \sqrt{p_{mn}+q}} \sum_{\tau\nu\phi} F_{\tau\nu\phi}^{q,\mathbf{Q}} R_{t+\tau,u+\nu,v+\phi}(\alpha_{mn}, \mathbf{R}_{P_{mn}Q}), \tag{4.49}$$

where we have used Eq. (4.25), and where the subscript mn on p_{mn}, $\mathbf{P}_{mn}$ and α_{mn} indicate that the overlap distribution arise from the two primitive solid-harmonic Gaussians with exponents a_m and b_n, respectively.

4.4 The exchange contribution

The exchange matrix $\mathbf{K}$ appearing in both HF and hybrid KS theory,

$$K_{ab} = \sum_{cd} (ac|bd) D_{cd}, \tag{4.50}$$

is intrinsically linear scaling provided the density-matrix elements D_{cd} decay with distance. This can be seen by noting that the density-matrix elements couple basis functions belonging to different electrons, according to Eq. (3.17), thus effectively damping the long ranged $1/r_{12}$ interaction. For insulators, the first order-reduced density matrix $\rho(\mathbf{r}, \mathbf{r}')$, given by

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{cd} D_{cd} \chi_c(\mathbf{r}) \chi_d(\mathbf{r}'), \tag{4.51}$$

decays exponentially with distance at long range, with the exponent proportional to the band gap of the system. For conductors, which do not have a band gap, the decay rate is only proportional to some power of $1/r_{12}$ (which at the length scale of molecules or solids is *very* slow). For finite systems the situation is less clear cut. However, systems with small HOMO-LUMO gaps behaves similarly to conductors, whereas systems with large HOMO-LUMO gaps behaves similarly to insulators.

Although screening typically shows a small prefactor, it becomes significant for large systems. Linear-scaling calculation of exchange contribution (for insulator like systems) is achieved by combining integral screening and proper reorganization of the integral loop structure, as first implemented in the order N exchange (ONX) [53] and then later in the linear-scaling exchange (LinK) [54]. At the heart of both ONX and LinK is the sorting of shell pairs by decreasing values, so that the different loops can be exited whenever an integral estimate becomes smaller than a certain threshold. This allows the screening to be performed in time proportional to system size. The main difference between the ONX and the LinK schemes is that in the original ONX approach the full permutational symmetries of the four-center two-electron integrals $(ab|cd)$, Eq. (4.26), is not exploited, as is done in the LinK scheme. Both approaches reduce naturally to quadratic dependence in systems where the HOMO-LUMO gap is small.

Figure 4.4 gives an outline of the LinK scheme. The first step constitutes both a removal of small elements and a sorting of the shell pairs ab - according to decreasing values of $G_{ab} = \sqrt{(ab|ab)}$. The second step constitutes making lists L_a of all shells c that give significant estimates of the couplings with shell a through the density -matrix D_{ac}, sorted according to decreasing values of $|D_{ac}|G_c^{\max}$; with $G_c^{\max}$ the maximum element of G_{cd} with shell c fixed. Third, within the loop of significant "bra" shell pairs ab, one sets up the mini-list ML of all "ket" shell pairs to interact with. This is done by merging the mini-list ML_a of all shell pairs cd the shell a interact with and the mini-list ML_b of all shell pairs the shell b interacts with. Finally, all integrals $(ab|cd)$ between the current "bra" shell pair ab, and all "ket" shell pairs cd in the mini-list ML are calculated and contracted with density-matrix elements (exploiting fully the permutational symmetry of the four-center two-electron integrals $(ab|cd)$).

Remove shell pairs ab with maximum element $G_{ab} < \text{thresh}/G_{cd}^{\max} D_{cd}^{\max}$ and sort shell pairs according to decreasing values $G_{ab} = \sqrt{(ab|ab)}$
Loop over a's in significant "bra" shell pair list ab
  Loop over c's in significant "ket" shell pair list cd
    If ($|D_{ac}|G_a^{\max} G_c^{\max} \geq$ thresh) then
      Store significant c's for each a in list L_a
    Else
      Exit c loop
    End if
  End c loop
End a loop
Sort L_a list according to decreasing values $G_c^{\max}|D_{ac}|$
Loop over significant "bra" shell pairs ab
  Loop over c's in L_a
    Loop over significant d's in "ket" shell pair list cd
      If ($|D_{ac}|G_{ab}G_{cd} \geq$ thresh) then
        Add cd to ML_a
      Else
        Exit d loop
      End if
    End d loop
    If resulting number of d's are zero exit c loop
  End c loop
  Create similarly list ML_b
  Merge the two lists ML_a and ML_b into ML
  Loop over significant "ket" shell pairs cd in list ML
    Form $(ab|cd)$ and contract with D_{ac}, D_{ad}, D_{bc} and D_{bd}
  End cd loop
End ab loop

Figure 4.2: Outline of the LinK algorithm.

4.5 The exchange-correlation contribution

Due to the complex expressions of available exchange-correlation (XC) functionals F_{xc} the integration of F_{xc} is performed numerically in the course of the energy evaluation. The exchange-correlation contribution the numerical quadratures are intrinsically linear scaling due to the fast decaying nature of the basis functions used [55, 56] and by using linear-scaling grid-generation [57]. The DFT exchange-correlation evaluation consists of three main step, namely the grid generation, the evaluation of the electron density and the evaluation of the exchange-correlation contribution to the KS matrix and energy.

The molecular grid is broken down into atomic grids, separated into radial and angular components, followed by a weight correction by means of a space partitioning. For the space partitioning we follow Becke [58], in which the grid weights w_i are evaluated according to

$$w_i = w_i^A w_A(\mathbf{r}_i) \tag{4.52}$$

where w_i^A are the atomic grid weights, and $w_A(\mathbf{r})$ the partitioning function [58, 57] depending on the nuclear positions and of the spatial position $\mathbf{r}_i$ of the grid point i. The partitioning function is close to unity near the atom center A and zero near other atoms. Although the partitioning function in principle depends on all nuclear coordinates, in practice the corrected weights are unaffected by excluding contributions from atoms B far from A. The density at each grid point i is evaluated according to

$$\rho(\mathbf{r}_i) = \sum_{ab} D_{ab}\chi_a(\mathbf{r}_i)\chi_b(\mathbf{r}_i). \tag{4.53}$$

For the efficient evaluation, the space is partitioned into spatial boxes. For each box only non-vanishing AOs need to be evaluated, and for large molecular systems linear scaling is achieved since the number of non-vanishing AOs becomes saturated in each box. Once the density at each grid point has been evaluated, the exchange-correlation contribution to the KS matrix and energy can be evaluated. The KS energy is determined according to

$$E_{\mathrm{xc}} = \int F_{\mathrm{xc}}[\rho(\mathbf{r})]\mathrm{d}\mathbf{r} = \sum_i w_i F_{\mathrm{xc}}[\rho(\mathbf{r}_i)], \tag{4.54}$$

with $F_{\mathrm{xc}}[\rho(\mathbf{r})]$ the functional form, and with the XC contribution to the KS matrix $\mathbf{X}$ according to

$$X_{ab} = \int \chi_a(\mathbf{r})\chi_b(\mathbf{r})v_{\mathrm{xc}}[\rho(\mathbf{r})]\mathrm{d}\mathbf{r} = \sum_i w_i\chi_a(\mathbf{r}_i)\chi_b(\mathbf{r}_i)v_{\mathrm{xc}}[\rho(\mathbf{r}_i)]. \tag{4.55}$$

The extension to the GGA approach follows the same structure as outlined for LDA above, but requires the additional evaluation of the AO gradient elements and the gradient of the density.

Chapter 5

Density fitting

The evaluation of molecular integrals is central to quantum chemistry, and is often one of the time-limiting steps. Therefore approximations of these integrals have, in addition to improved integral evaluation schemes, been a concern from the early developments of quantum chemistry. The density-fitting methods, or alternatively the resolution-of-the-identity (RI) methods, have today been established as highly successful for approximating the Coulomb contribution. In these approaches the expensive evaluation of four-center integrals is replaced by the evaluation of two- and three-center integrals, and a set of linear equations for the fitting coefficients. Speed-ups in the range of 3-30 are commonly observed, with errors well within the basis-set errors. Typical errors due to density fitting are about two orders of magnitude smaller than the basis-set errors.

Linear-scaling density-fitting developments is one of the main topics of this thesis. In this chapter we therefore start in section 5.1 with a somewhat detailed overview of the density-fitting approximation and at the same time introduce some important concepts to be used in the next section, section 5.2, in which we discuss linear-scaling density-fitting approaches. In section 5.3 we present the boxed density-fitting scheme for accurate linear-scaling density-fitted Coulomb matrix formation. In section 5.4 we present a robust variational formulation and implementation. We establish that the robust variational formulation can be adopted to solve for the fitting coefficient in sparse, rather than the Coulomb metric, at little loss of chemical accuracy, allowing the presented formalism to be applied in linear-scaling density-fitting developments. Finally, in section 5.5, we present a linear-scaling density-fitted Coulomb force evaluation, accelerated using the novel integral evaluation scheme presented in Paper I. The presented force and energy evaluation is efficient, and used for geometry optimization

for molecules containing up to 400 atoms.

5.1 Historical overview

The density-fitting approximations as we know them today have gone through several different phases, starting with the axial expansion of Boys and Shavitt's [59] in 1959, in which a product between two STOs was expanded by least-square-fitting in twenty single STOs distributed on the line connecting the two centers, followed by the projection of diatomic differential overlap (PDDO) method [60, 61] in 1968 and 1969, and by the limited expansion of diatomic overlap (LEDO) [62, 63] in 1969 and 1971. The density-fitting approximation is today most often tributed to the 1973 contributions of Whitten [48] and of Baerends, Ellis and Ros [49], and to the following developments by Dunlap, Connolly and Sabin [5, 50] in 1979.

5.1.1 Whitten paper

In his 1973 paper, Ref. [48], Whitten developed a mathematical framework for approximation of two-electron integrals $(f|g)$ according to

$$(f|g) \approx (\tilde{f}|\tilde{g}), \tag{5.1}$$

in terms of approximate densities $\tilde{f}(\mathbf{r}_1)$ and $\tilde{g}(\mathbf{r}_2)$, by providing rigorous error-bounds between the true and the approximated integrals

$$\Big|(f|g) - (\tilde{f}|\tilde{g})\Big| \leq \delta, \tag{5.2}$$

for given tolerance δ. Following the LEDO approach of Billingsley and Bloor [62, 63], the four-center two-electron integrals $(ab|cd)$ are approximated according to

$$(ab|cd) \approx \widetilde{(ab|cd)} = (\widetilde{ab}|\widetilde{cd}). \tag{5.3}$$

In this approach the product between two GTOs is expanded in single GTOs auxiliary basis functions according to

$$\Omega_{ab}(\mathbf{r}) \approx \tilde{\Omega}_{ab}(\mathbf{r}) = \sum_{\alpha} c_{\alpha}^{ab} \chi_{\alpha}(\mathbf{r}), \tag{5.4}$$

and similarly for the product $\Omega_{cd}(\mathbf{r})$. Whitten realized that minimization of the residual density Coulomb repulsion integral Δ_{ab},

$$\Delta_{ab} = (ab - \widetilde{ab}|ab - \widetilde{ab}), \tag{5.5}$$

and similarly for Δ_{cd}, minimize the error in the approximation $(ab|cd) \approx (\widetilde{ab}|\widetilde{cd})$. Such a minimization actually leads to the same set of linear equations for the fitting coefficients c_α^{ab}

$$\sum_\beta (\alpha|\beta) c_\beta^{ab} = (\alpha|ab). \tag{5.6}$$

as those used in the LEDO method. Note that one can also view the *fitting equations* of Eq. (5.6) as imposing that the Coulomb interaction between the auxiliary functions $\chi_\alpha(\mathbf{r})$ and both the true $\Omega_{ab}(\mathbf{r})$ and the *fitted* distributions $\widetilde{\Omega}_{ab}(\mathbf{r})$ should be equal. Harris and Rein [64] originally used this idea, of enforcing that the approximated charge distributions yield correct values for certain integrals, to calibrate the fitting coefficients - an idea that was adapted into the LEDO procedure.

The above formulation of Eqs. (5.3-5.6) is more or less identical to the formulas most people use today. However, Whittens formulation deviates from the conventional formulation used today on a few of points. First, Whitten proposed to approximate the four-center two-electron integrals only when their estimated error was below a certain threshold δ (as later used in an SCF implementation by Jafri and Whitten [65]). Second, he did not make any assumptions about the auxiliary functions - whether they where atom-centered, or if the same set of functions where used to approximate all the different products of GTOs $\Omega_{ab}(\mathbf{r})$. Third, the formulae for the integral approximation, Eq. (5.3), often takes a slightly different form, namely

$$(ab|cd) \approx \widetilde{(ab|cd)} = (ab|\widetilde{cd}), \tag{5.7}$$

in particular when used to approximate the Coulomb matrix $\mathbf{J}$ used for the construction of both the Fock- and KS-matrices, Eqs. (3.15) and (3.35), respectively.

5.1.2 Baerends, Ellis and Roos paper

In the paper by Baerends *et* al [49], the full electronic density

$$\rho(\mathbf{r}) = \sum_{ab} D_{ab} \Omega_{ab}(\mathbf{r}) \tag{5.8}$$

is approximated by least-square-fitting of the expansion

$$\widetilde{\rho}(\mathbf{r}) = \sum_\alpha c_\alpha \chi_\alpha(\mathbf{r}). \tag{5.9}$$

The least-square-fit is obtained by minimization of the residual density norm Δ_ρ^δ

$$\Delta_\rho^\delta = \int \left(\rho(\mathbf{r}) - \widetilde{\rho}(\mathbf{r})\right)^2 d\mathbf{r}, \tag{5.10}$$

subject to the charge constraint

$$\int \widetilde{\rho}(\mathbf{r})d\mathbf{r} = \int \rho(\mathbf{r})d\mathbf{r} = N_e, \tag{5.11}$$

with N_e the number of electrons. This gives the linear set of equations for the fitting-coefficients c_α

$$\sum_\beta \langle\alpha\beta\rangle c_\beta = \langle\alpha\rho\rangle + \lambda\langle\alpha\rangle, \tag{5.12}$$

with the Lagrange multiplier λ given by

$$\lambda = \frac{N_e - \sum_{\alpha\beta}\langle\alpha\rangle\langle\alpha\beta\rangle^{-1}\langle\beta\rho\rangle}{\sum_{\alpha\beta}\langle\alpha\rangle\langle\alpha\beta\rangle^{-1}(\beta\rangle}. \tag{5.13}$$

The approximated density is again used to build an approximate Coulomb potential $\tilde{V}_C(\mathbf{r})$ according to

$$V_C(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' \approx \tilde{V}_C(\mathbf{r}) = \int \frac{\widetilde{\rho}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'. \tag{5.14}$$

This gives the *density fitted* Coulomb matrix $\tilde{J}_{ab}$

$$\tilde{J}_{ab} = \int \Omega_{ab}(\mathbf{r})\tilde{V}_C(\mathbf{r})d\mathbf{r} = (ab|\tilde{\rho}), \tag{5.15}$$

to be compared with the regular Coulomb matrix of Eq. (4.41).

As an alternative Baerends *et.* al [49] also suggested to fit the individual pair-atomic densities $\rho_{AB}(\mathbf{r})$ instead of the full electron density,

$$\rho_{AB}(\mathbf{r}) = \sum_{a\in A, b\in B} \Omega_{ab}(\mathbf{r})D_{ab}, \tag{5.16}$$

where $a \in A$ means all basis functions $\chi_a(\mathbf{r})$ sharing atomic center $\mathbf{A}$, by using auxiliary basis functions centered on the two centers $\mathbf{A}$ and $\mathbf{B}$, only. In this manner the *fitted* electron density

$$\widetilde{\rho}(\mathbf{r}) = \sum_{AB} \widetilde{\rho}_{AB}(\mathbf{r}), \tag{5.17}$$

is obtained in a more economical manner.

The approach by Baerends *et.* al is in many ways similar to the PDDO method, but differs on certain points. First, each integral is not approximated individually (by approximating individual diatomic distributions $\Omega_{ab}(\mathbf{r})$), rather the full Coulomb interaction is approximated (by approximating the full electron density $\rho(\mathbf{r})$). Second, and perhaps more importantly, was the choice by Baerends *et.* al to use an independent (atom-centered) auxiliary basis set $\{\chi_\alpha\}$. This was contrary to the PDDO and

LEDO approximations. In both these approaches the auxiliary basis set was taken as a union between basis functions centered on the two atomic centers, as in the alternative approach of Baerends *et.* al. However, the atom-centered set of auxiliary basis functions was in the PDDO and LEDO approximations taken to be the combination of 1) all possible products between two regular basis functions sharing center and 2) a few additional functions. This choice of basis functions both limited the size of the regular basis sets to be used, and led to problems with linear dependencies when solving for the expansion coefficients.

5.1.3 Dunlap, Connolly and Sabin papers

The density-fitting approach presented in the two 1979 papers by Dunlap, Connolly and Sabin [5, 50] greatly resembles the approach of Baerends *et.* al in form, but, following the LEDO approach and the Whitten paper, uses the Coulomb rather than the overlap metric for the determination of the fitting coefficients. So, instead of performing a least-square-fit (in the overlap metric), the residual density Coulomb repulsion integral

$$\Delta_\rho = (\rho - \widetilde{\rho}|\rho - \widetilde{\rho}), \tag{5.18}$$

is minimized subject to the charge conserving constraint of Eq. (5.11). This gives the slightly modified linear equation set for the fitting coefficients

$$\sum_\beta (\alpha|\beta)c_\beta = (\alpha|\rho) + (\alpha)\lambda, \tag{5.19}$$

with the Lagrange multiplier λ now given by

$$\lambda = \frac{N_e - \sum_{\alpha\beta}(\alpha)(\alpha|\beta)^{-1}(\beta|\rho)}{\sum_{\alpha\beta}(\alpha)(\alpha|\beta)^{-1}(\beta)}. \tag{5.20}$$

Dunlap *et.* al further note that by Eq. (5.18) we have

$$\frac{1}{2}(\rho|\rho) = (\rho|\widetilde{\rho}) - \frac{1}{2}(\widetilde{\rho}|\widetilde{\rho}) - \frac{1}{2}\Delta_\rho, \tag{5.21}$$

Therefore, the first order correction ΔJ to the fitted Coulomb repulsion energy

$$\tilde{J} = \frac{1}{2}(\rho|\widetilde{\rho}), \tag{5.22}$$

is given by

$$\Delta J = \frac{1}{2}\left[(\rho|\widetilde{\rho}) - (\widetilde{\rho}|\widetilde{\rho})\right]. \tag{5.23}$$

This correction term is both used to analyze the errors introducing by fitting the density in either the Coulomb or the overlap metric [5], and as a correction to the fitted

Coulomb repulsion energy [50]. For the example calculations presented in Ref. [5], typical errors in the fitted energies are about one order of magnitude smaller when using the Coulomb rather than the overlap metric.

5.1.4 Robust and variational fitting

In another development from 1979, Mintmire [66] argued that the charge conserving constraint should be lifted. By taking

$$\tilde{J} = (\rho|\widetilde{\rho}) - \frac{1}{2}(\widetilde{\rho}|\widetilde{\rho}), \tag{5.24}$$

as the approximated Coulomb repulsion energy, this then gives a *variational* fitting procedure that at the same time minimizes the error Δ_ρ. Note that differentiation of Eq. (5.24) with respect to the density-matrix elements D_{ab} and the fitting coefficients c_α gives, respectively, the fitted Coulomb matrix of Eq. (5.15) and the unconstrained version of Eq. (5.19), namely

$$\sum_\beta (\alpha|\beta)c_\beta = (\alpha|\rho). \tag{5.25}$$

Dunlap later [4, 67] denoted a fitting method that corrects the target function to first order in the error made by the fit as *robust*. Note that although there are many ways to obtain the fitting coefficient c_α, and that Eq. (5.24) should in general be used to obtain a robust fit, the fitting procedure of Eqs. (5.15) and (5.25) produces an approximated energy that is robust even when using Eq. (5.22), as can be seen by noting that by Eq. (5.25) the first order correction to the Coulomb repulsion energy of Eq. (5.23) is zero.

Although the original ideas leading to the density-fitting approximation of Refs. [49, 5] were developed for the approximation of individual two-electron integrals, the density-fitting approximation was for a long time only used to approximate the Coulomb repulsion term of Eq. (5.15). For four-center two-electron integrals, Dunlap [4] emphasizes that in general the robust approximation

$$\widetilde{(ab|cd)} = (ab|\widetilde{cd}) + (\widetilde{ab}|cd) - (\widetilde{ab}|\widetilde{cd}) \tag{5.26}$$

should be used. The minimization of either Δ_{ab} of Eq. (5.5) or $\widetilde{(ab|cd)}$ of Eq. (5.26) with respect to the fitting coefficients c_α^{ab} gives the linear equation set of Eq. (5.6). Therefore, provided the same set of auxiliary basis functions is used to approximate both $\Omega_{ab}(\mathbf{r})$ and $\Omega_{cd}(\mathbf{r})$, the three approximations $(ab|\widetilde{cd})$, $(\widetilde{ab}|cd)$ and $(\widetilde{ab}|\widetilde{cd})$ become identical, and

thus the three approximations of Eqs. (5.3), (5.7) and (5.26). For derivatives however, Eq. (5.26) should be used.

Dunlap further notes [4] that by employing Eq. (5.6) any function of the robust approximation of Eq. (5.26), $E[\widetilde{(ab|cd)}]$, is automatically variational with respect to the fitting coefficients c_α^{ab} since by Eq. (5.26) we have

$$\frac{dE[\widetilde{(ab|cd)}]}{dc_\alpha^{ab}} = \frac{\partial E[\widetilde{(ab|cd)}]}{\partial \widetilde{(ab|cd)}} \frac{\partial \widetilde{(ab|cd)}}{\partial c_\alpha^{ab}} = 0. \tag{5.27}$$

5.1.5 Density fitting of the exact exchange

Density fitting of the exact exchange was first introduced by Weigend [68] as late as in 2002. In this paper, the exchange matrix of Eq. (4.50) was approximated by combining Eqs. (5.6) and (5.7), according to

$$\tilde{K}_{ab} = \sum_{cd} \sum_{\alpha\beta} (ac|\alpha)(\alpha|\beta)^{-1}(\beta|bd) D_{cd} = \sum_{cd} \sum_{\alpha'} (ac|\alpha')(\alpha'|bd) D_{cd}. \tag{5.28}$$

Here the last step constitutes a transformation to an orthogonal auxiliary basis

$$\chi_{\alpha'}(\mathbf{r}) = \sum_{\alpha} (\alpha'|\alpha)^{-\frac{1}{2}} \chi_\alpha(\mathbf{r}). \tag{5.29}$$

Note that this approach scales as $\mathcal{O}(N^4)$ in both the transformation and contraction steps.

5.1.6 Considerations

As we have now hoped to have demonstrated, the density-fitting approximation which uses the full set of atom-centered auxiliary basis functions for each individual overlap distribution $\Omega_{ab}(\mathbf{r})$ and follows Eqs. (5.6) and (5.26) (leading to Eqs. (5.6) and (5.7)) is the best choice of fitting with respect to the variational property and the automatic robustness. The problem with this way of fitting becomes apparent for large systems due to 1) the cubic scaling nature of solving the fitting equations and 2) the number of significant auxiliary functions included in the expansion of the overlap distribution $\Omega_{ab}(\mathbf{r})$. When approximating the Coulomb matrix, the first point 1) only becomes a problem for large systems (typically more than 10000 auxiliary basis functions), due to the very fast matrix-libraries routines - like the lapack DPOSV, which uses standard LU-decomposition of the metric-matrix (Cholesky-decomposition), followed first by a forward-substitution step and then by a backward-substitution step, see for instance

Ref. [69]. Note that the LU factors only need to be calculated once at the beginning of the SCF cycle. For the exchange matrix, although 1) takes longer than for the Coulomb matrix, since a set of linear equations must be solved for each orbital pair ab, the second point 2) becomes the computational bottleneck due to the $\mathcal{O}(N^4)$ scaling of the different contractions and transformation steps.

Before we proceed with the linear-scaling density-fitting developments we note that the important subject of auxiliary basis sets have been studied by Eichkorn et al. [70, 71] for polarized split-valence (SVP) and polarized triple-zeta valence (TZVP) basis sets, and by Weigend [68, 72] for correlation consistent basis sets and for the fitting of exact exchange. Further note that the density-fitting approximation can be accelerated using the Poisson equation [73], and that is can be applied to MP2 theory [74].

5.2 Linear-scaling density fitting

The recent developments toward large systems have highlighted the need for a linear-scaling density-fitting scheme. In this section, we give a brief overview of different linear-scaling density-fitting schemes presented in the literature. We first discuss methods based on the use of a local metric; next we consider methods based on the spatial partitioning of the electron density.

5.2.1 Density fitting using local metrics

For the Coulomb contribution, density-fitting methods based on the use of a local metric has been explored by Refs. [49, 6, 75] and is further developed in paper Paper III, using a robust, variational formulation. In the approach of Baerends *et al.* [49], the electron density is fitted in the overlap metric, giving errors one order of magnitude greater than in the Coulomb metric [5]. This result was confirmed by Vahtras *et al.* [6], who compare three different ways of fitting the four-center integrals in the overlap metric to the corresponsing ones fitted in the Coulomb metric. In the paper by Jung *et al.* [75], the expansion coefficients obtained in the Coulomb metric, overlap metric, and attenuated metric $w(\mathbf{r}_1, \mathbf{r}_2) = \mathrm{erfc}(\omega r_{12})/r_{12}$ are compared. The attenuated metric bridges the Coulomb and the overlap metrics by varying the value of the damping parameter ω. The coefficients obtained in the overlap metric decay more or less exponentially with distance, whereas the coefficients obtained in the Coulomb metric decay more slowly at long distances. For a one-dimensional test system studied in that paper, the fitting coefficients decay as $\sim r^{-1.25}$ in the Coulomb metric, with a faster decay observed

for two- and three-dimensional systems. The authors further provide statistics on atomization energies for the G2 benchmark set using RI second-order Møller–Plesset (MP2) perturbation theory in the cc-pVDZ basis, reporting errors six to seven times larger in the overlap metric than in the Coulomb metric. Note that none of the above schemes are either robust or variational; which means Lagrange multipliers are needed for a proper description already for first derivatives and that a far larger number of auxiliary basis functions must be employed to obtain the same level of accuracy as for the robust approaches.

5.2.2 The partitioning approach

We now turn our attention to the partitioning approach; in which the density is written as a linear combination of subsystem densities and where each subsystem density is approximated by auxiliary functions in some local region. For the Coulomb contribution, this approach has been explored by several authors [49, 3, 76] and is also used in Paper II. In the paper by Baerends *et.* al [49] the pair densities $\rho_{AB}(\mathbf{r})$ can, as already described in the previous section, be fit individually - including auxiliary basis functions centered only on the two parent atoms A and B. By standard screening techniques, only a linear number of non-negligible pair densities need to be approximated and the density can thus be approximated in a linear scaling fashion. The PES of this approximation is continuous, but the resulting energy is neither robust nor variational. It is worth noting that STOs are used rather than GTOs, and that the fitted density is used to build an approximate Coulomb potential that is included in the numerical evaluation together with the exchange–correlation contribution.

In the paper by Gallant and St-Amant [3], the density is partitioned using Yangs partitioning [77], according to

$$\rho(\mathbf{r}) = \sum_s \rho^s(\mathbf{r}) = \sum_{ab} x^s_{ab} D_{ab} \Omega_{ab}(\mathbf{r}), \tag{5.30}$$

with

$$x^s_{ab} = \begin{cases} 1, & \text{if both } a \in s \text{ and } b \in s \\ 1/2, & \text{if either } a \in s \text{ or } b \in s \\ 0, & \text{otherwise.} \end{cases} \tag{5.31}$$

Each subsystem density $\rho^s(\mathbf{r})$ is fitted separately by including fitting functions centered on all atoms within some predefined vicinity of the density, with the charge constraint

$$\int \widetilde{\rho}^s(\mathbf{r}) d\mathbf{r} = \int \rho^s(\mathbf{r}) d\mathbf{r} = Q^s \tag{5.32}$$

enforced for each subsystem. The resulting errors can be made arbitrarily small by enlarging the buffer size to include (atom-centered) auxiliary functions. The approach of Gallant and St-Amant is neither robust nor variational, nor does it provide a continuous PES.

In the local atomic density fitting (LADF) or the atomic resolution of the identity (ARI), of Sodt *et al.* [76] the density is partitioned into atomic regions by localizing the individual overlap distributions $\Omega_{ab}(\mathbf{r})$ to one of the atoms that the basis functions originate from. Following Gallant and St.-Amant, these atomic densities are fitted individually by including fitting functions in some buffer zone around the atom. In addition a 'bump' function is introduced on the boundary of the buffer zone to smoothly turn off which fitting functions to include. A robust correction term is added for the individual Coulomb matrix element, which makes the algorithm both robust and variational.

Of the above partitioning schemes, the (alternative) pair-atomic fitting scheme of Baerends *et.* al is perhaps the most appealing, but the LADF scheme offers the best compromise between cost and accuracy. The bump function does, however, represent an artifact, which for instance can create artificial minima on the PES. Also, in all the above partitioning schemes, except the pair-atomic fitting, some cut-off scheme must be adopted. A criticism of such cut-off schemes is that the impact of the fitting error on the calculated properties is difficult to predict.

5.2.3 Linear-scaling density fitting of the exchange contribution

Linear-scaling aspects of density fitting of exchange contribution were first considered by Polly *et al.* [78]. The fitted exchange matrix

$$\tilde{K}_{ab} = \sum_{i}^{n_{\mathrm{occ}}} (ai|\alpha) c_{\alpha}^{bi}, \tag{5.33}$$

is computed in linear time. This is achieved using localized orbitals $\chi_a(\mathbf{r})$ and $\phi_i(\mathbf{r})$ that interact with auxiliary functions $\xi_\alpha(\mathbf{r})$ only in some local domain. Localization of the molecular orbitals is achieved through Pipek-Mezey localization Ref. [79], but can in principle be localized by any localization procedure - see Ref. [80] and references therein. The final density-fitted exchange energy

$$\tilde{K} = \sum_{ij}^{n_{\mathrm{occ}}} \sum_{\alpha} (ij|\alpha) c_{\alpha}^{ij}, \tag{5.34}$$

is computed without use of local fitting domains. It is argued that the fitted exchange energy depends sensitively on the size of the fitting domains, whereas the optimized

MO coefficients do not. Reported errors are in the micro-Hartree range. In effect the MOs are not optimized variationally, although the energy is corrected through first order. It should be noted that the final step of Eq. (5.34) does not scale linearly with system size; i.e. without the use of local fitting domains. Furthermore, this scheme does not provide continuous potential energy surfaces.

The ARI exchange method (ARI-K) of Sodt *et al.* [81] is an extension of the LADF or ARI approach of Ref. [76], applied to the exchange rather than the Coulomb contribution. In this approach, the product overlaps $\Omega_{ai}(\mathbf{r})$ are approximated by auxiliary basis functions $\xi_\alpha(\mathbf{r})$ in the local domain $[A]$ near the parent atom of AO $\chi_a(\mathbf{r})$,

$$\tilde{\Omega}_{ai}(\mathbf{r}) = \sum_{\alpha\in[A]} c_\alpha^{ai}\xi_\alpha(\mathbf{r}), \tag{5.35}$$

with

$$c_\alpha^{ai} = \sum_{\beta\in[A]} (\alpha|\beta)_A^{-1}(\beta|ai). \tag{5.36}$$

As in the LADF scheme, continuity of the potential energy surface is ensured by the use of a bump function which is incorporated into the individual inverses $(\alpha|\beta)_A^{-1}$ associated with the centers A, see Ref. [81] for details. The exchange matrix of Eq. (4.50) is further approximated according to

$$\tilde{K}_{ab} = \frac{1}{2}\sum_i \Big(\sum_{\alpha\in[A]} c_\alpha^{ai}(\alpha|bi) + \sum_{\beta\in[B]} (ai|\beta)c_\beta^{bi} \Big). \tag{5.37}$$

We note that this approach is non-variational, which is justified by reporting errors in energies using Eq. (5.37) that are typically only twice those of regular density fitting of exchange.

5.3 Boxed density fitting

The boxed density-fitting approach presented in Paper II follows the approach of Gallant and St-Amant, with a few exceptions. First the charge constraint has been lifted, second the partitioning is not performed by selection of functional groups but by automatically generated boxes, finally a first order correction to the energy, according to Eq. (5.23), is added in order to make the energy robust. For the generation of the boxes, the total system is put into a rectangular box, which is recursively bisected until no sub-box contains more than some fixed number of auxiliary basis functions. Each subsystem is fitted using auxiliary basis functions located within an extended subsystem s, comprising the original subsystem s padded with a buffer zone δs. The error in

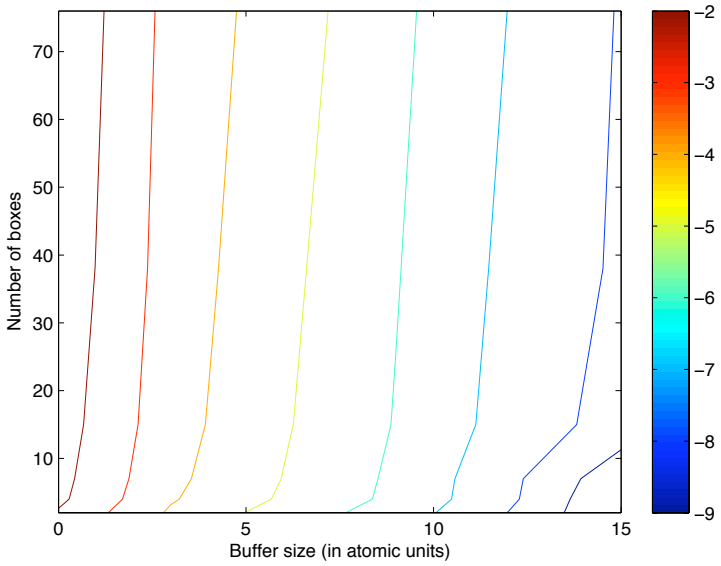

Figure 5.1: Contour plot of the boxed density-fitting error compared to regular Coulomb density fitting, as a function of the size of the buffer zone and the number of boxes for the valinomycine peptide ($C_{54}N_6O_{18}H_{90}$), using 6-31G basis and with auxiliary basis given in Refs. [70, 71]. Absolute errors in the energy in logarithmic scale, from 10^{-2} to 10^{-9} Hartree.

the energy using the boxed density-fitting approach compared to regular fitting using the Coulomb metric, is plotted in figure 5.1 for the valinomycine peptide. As can be seen from the figure, the errors can be made arbitrarily small by extending the buffer size. Although the error increases with the number of boxes, the approach is most sensitive to the buffer size. For the calculations presented in Paper II a buffer zone of 5 Bohr was used, and the total system was recursively bisected until no sub-box contained more than 5000 auxiliary basis functions.

5.4 Robust and variational fitting using local metrics

In Paper III we formulate the robust and variational fitting of four-center two-electron integrals in a general metric, and present results for the fitting of both the Coulomb and exchange contributions using the Coulomb, attenuated Gaussian damped Coulomb and overlap metrics to obtain the fitting coefficients.

5.4.1 Robust and variational fitting of two-electron four-center integrals

These coefficients are, in a general metric $w(\mathbf{r}_1, \mathbf{r}_2)$, obtained by minimizing the residual density interaction Δ^w_{ab}, according to

$$\Delta^w_{ab} = \langle \Delta ab | w | \Delta ab \rangle, \tag{5.38}$$

with $|\Delta ab\rangle = |ab\rangle - |\widetilde{ab}\rangle$ and with the Mulliken like notation

$$\langle f|w|g\rangle = \int f(\mathbf{r}_1) w(\mathbf{r}_1, \mathbf{r}_2) g(\mathbf{r}_1 d\mathbf{r}_1 d\mathbf{r}_2. \tag{5.39}$$

This lead to the set of linear equations

$$\langle \Delta ab|w|\beta\rangle = 0, \quad \langle \alpha|w|\Delta cd\rangle = 0 \tag{5.40}$$

for the fitting coefficients. These equations are sparse when local metric and basis functions are used, allowing for a solution in time proportional to system size. Following Dunlap [4], and introducing Lagrange multipliers $\bar{c}^{ab}_\alpha$ and $\bar{c}^{cd}_\beta$ to make the integrals variational with respect to the fitting coefficients, we obtain

$$\widetilde{(ab|cd)} = (ab|\widetilde{cd}) + (\widetilde{ab}|cd) - (\widetilde{ab}|\widetilde{cd}) - \langle \overline{ab}|w|\Delta cd\rangle - \langle \Delta ab|w|\overline{cd}\rangle\,, \tag{5.41}$$

in the notation

$$\langle \overline{ab}| = \sum_\alpha \bar{c}^{ab}_\alpha \langle \alpha|, \quad |\overline{cd}\rangle = \sum_\beta \bar{c}^{cd}_\beta |\beta\rangle. \tag{5.42}$$

The last two terms of Eq. (5.41) are not needed for the unperturbed integrals $\widetilde{(ab|cd)}$, but become important for the calculation of molecular properties.

5.4.2 The Coulomb contribution using local metric

For the approximate Coulomb matrix the above approximation gives

$$\tilde{J}_{ab} = (ab|\widetilde{\rho}) + (\widetilde{ab}|\Delta\rho), \tag{5.43}$$

with $|\Delta\rho\rangle = |\rho\rangle - |\widetilde{\rho}\rangle$, which compared to Eq. (5.15) has a first order correction term. Note that in the Coulomb metric this term is zero by Eq. (5.25). Also note that the Coulomb integrals appearing in Eq. (5.43) are the same as those appearing in

the standard density-fitting approximation of Eqs. (5.25) and (5.15), except that the two-center two-electrons integrals $(\alpha|\beta)$ now appears in a form,

$$(\alpha|\widetilde{\rho}) = \sum_{\beta}(\alpha|\beta)c_{\beta}, \tag{5.44}$$

that can also be obtained in a linear scaling fashion using CFMM. Finally note that when evaluating the approximate Coulomb matrix one does not need to calculate the three-index fitting coefficients c_{α}^{ab}, rather on can first solve

$$\sum_{\beta}\langle\alpha|w|\beta\rangle c_{\alpha} = \langle\alpha|w|\rho\rangle \tag{5.45}$$

then build the first contribution $(ab|\widetilde{\rho})$, and the intermediate $(\alpha|\Delta\rho)$ used to obtain Δc_{α} according to

$$\sum_{\beta}\langle\alpha|w|\beta\rangle\Delta c_{\beta} = (\alpha|\Delta\rho), \tag{5.46}$$

and finally build

$$(\widetilde{ab}|\Delta\rho) = \sum_{\alpha}\langle ab|w|\alpha\rangle\Delta c_{\alpha}. \tag{5.47}$$

Similar methodology is also possible for the perturbed contributions needed for property evaluations. To summarize, the cubic-scaling linear solver step for obtaining the fitting coefficients using the Coulomb metric can be replaced by the additional evaluation of the sparse two- and three-center integrals, and two sets of sparse linear equations, when using a sparse metric $w(\mathbf{r}_1, \mathbf{r}_2)$. The sparse linear equations of Eq. (5.40) can be solved using for example the linear-scaling Lowdin decomposition as outlined in Ref. [82]. For the benchmark set of Peach *et.* al [83] the robust variational fitting of the Coulomb contribution using overlap metric to obtain the fitting coefficients gave errors in the energy due to density fitting within approximately a factor 2 larger than when using the regular Coulomb metric; even when using auxiliary basis sets optimized for fitting in the Coulomb metric.

5.4.3 The exchange contribution using local metric

Linear scaling of the exchange matrix is more intricate as it involves the three-index fitting coefficients c_{α}^{ab} or the half-transformed coefficients c_{α}^{ai}, where i denotes a molecular orbital. The robust variational fitting of the exchange matrix can be written as

$$\tilde{K}_{ab} = \sum_{cd}\widetilde{(ac|bd)}D_{cd} = \sum_{i}\widetilde{(ai|bi)} = \sum_{i}\left[(ai|\widetilde{bi}) + (\widetilde{ai}|\Delta bi)\right], \tag{5.48}$$

where i denotes an occupied molecular orbital, and with

$$|\widetilde{ai}\rangle = \sum_c C_{ci}|\widetilde{ac}\rangle = \sum_\alpha c_\alpha^{ai}|\alpha\rangle. \tag{5.49}$$

The three-index fitting coefficients are either found directly by solving for the half-transformed coefficients c_α^{ai} or by first solving for the coefficients in AO-basis c_α^{ab} and then contract with either the density-matrix elements or half-transform to c_α^{ai}.

For insulators, linear-scaling density-fitted exchange-matrix construction can be achieved in a local metric by following the same arguments as for the regular exchange matrix, and by pretabulating which three-center Coulomb repulsion integrals $(ab|\alpha)$ (or $(ai|\alpha)$) to calculate. First, we note that, in a local metric, the number of fitting coefficients c_α^{ab} scales linearly with system size, as auxiliary basis functions $\xi_\alpha(\mathbf{r})$ sufficiently far away from the product overlaps $\Omega_{ab}(\mathbf{r})$ do not contribute to the fitted product overlap $\tilde{\Omega}_{ab}(\mathbf{r})$ [75]. Second, since the density-matrix elements D_{cd} couple basis functions on two different electrons, $\chi_c(\mathbf{r}_1)$ and $\chi_d(\mathbf{r}_2)$, we can neglect all integrals $(ac|bd)$ where the density-matrix elements become sufficiently small; for example, using Cauchy-Schwarz screening

$$|(ac|bd)D_{cd}| \leq \sqrt{(ac|ac)}\sqrt{(bd|bd)}\,|D_{cd}|\,. \tag{5.50}$$

Therefore, the fitted integrals $\widetilde{(ac|bd)}$ of $(ac|bd)$ need only be calculated whenever

$$\sqrt{(ac|ac)}\sqrt{(bd|bd)}\,|D_{cd}| \geq \epsilon, \tag{5.51}$$

for a given threshold ϵ. For insulators, the density-matrix decrease exponentially with increasing distance, which means, for instance, that $\Omega_{ac}(\mathbf{r}_1)$ only interact with $\widetilde{\Omega}_{bd}(\mathbf{r}_2)$ provided $\chi_c(\mathbf{r}_1)$ and $\chi_d(\mathbf{r}_2)$ are within some finite distance of each other. As a result, $\chi_a(\mathbf{r}_1)$ and $\chi_b(\mathbf{r}_2)$ must also be close to each other. The same argument applies to the fitting functions since $\xi_\alpha(\mathbf{r}_2)$, included in $\widetilde{\Omega}_{bd}(\mathbf{r}_2)$, have a limited extent from the center of $\Omega_{bd}(\mathbf{r}_2)$, from which $\widetilde{\Omega}_{bd}(\mathbf{r}_2)$ originates. The combined effects of locality in the density matrix and locality in the fit imply that the number of contributing three-center integrals $(ac|\alpha)$ scales linearly with system size. The same argument holds for the term including the two-center integrals $(\alpha|\beta)$. Also note that for the half-transformed fitting coefficients c_α^{ai} linear scaling can be achieved provided local molecular-orbitals (LMOs) are used—see Ref. [80] and references therein. Linear-scaling follows by combining LMOs and Cauchy–Schwarz screening, since, provided the AOs χ_a and χ_b are sufficiently far away from each other, a given LMO will not overlap with both AOs. To see this, we apply the Cauchy–Schwarz inequality twice

$$|(ai|bi)| \leq \sqrt{(ai|ai)}\sqrt{(bi|bi)} \leq \Big[\sum_c |C_{ci}|\,\sqrt{(ac|ac)}\Big]\Big[\sum_c |C_{ci}|\,\sqrt{(bc|bc)}\Big], \tag{5.52}$$

where we have used

$$(ai|ai) = \sum_{cd} C_{ci}C_{di}(ac|ad) \leq \sum_{cd} |C_{ci}|\,|C_{di}|\,\sqrt{(ac|ac)}\sqrt{(ad|ad)}$$
$$= \Big[\sum_{c} |C_{ci}|\,\sqrt{(ac|ac)}\Big]^2 \qquad (5.53)$$

and similarly of $(bi|bi)$.

5.5 Density-fitted Coulomb force evaluation

The evaluation of the density-fitted Coulomb force has been considered in Refs. [84, 85, 86], and for the multipole-moment treatment in Refs. [87, 88]. In Paper IV we combine the density-fitted Coulomb force evaluation with multipole-moment treatment using the CFMM approach [41], and integral screening techniques [39, 48], for the first implementation of linear-scaling density-fitted Coulomb force evaluation. The construction of the density fitted Coulomb force contributions are accelerated using the McMurchie–Davidson J–engine like integral scheme presented in Ref. [45], in combination the novel integral evaluation scheme presented in Paper I. Expanding the solid harmonic Gaussians in Hermite rather than Cartesian Gaussians, has the benefits of reducing the cost of the differentiated integral evaluation and simplifying the implementation. In Paper IV we demonstrate the efficient implementation of molecular forces for systems containing up to 500 atoms.

5.5.1 The density-fitted Coulomb force contributions

Differentiation of the fitted electronic Coulomb repulsion energy of Eq. (5.24) with respect to the nuclear coordinate R_e gives [85]

$$\tilde{J}^{\mathbf{e}} = \frac{d\tilde{J}}{dR_e} = \sum_{ab} D^{\mathbf{e}}_{ab}\tilde{J}_{ab} + \sum_{ab} D_{ab}\tilde{J}^{\mathbf{e}}_{ab} + \sum_{\alpha} c_\alpha(g^{\mathbf{e}}_\alpha - \tilde{g}^{\mathbf{e}}_\alpha) \qquad (5.54)$$

with $e = x, y, z$, and where we have introduced $\mathbf{e}$ as the first, second or third row of the three by three identity matrix for differentiation with respect to the x, y or z Cartesian directions, respectively. The first term is the density-fitted Coulomb contribution to the so-called Pulay force [89], with the differentiated density matrix is given [90] as $\mathbf{D}^{\mathbf{e}} = -\mathbf{D}\mathbf{S}^{\mathbf{e}}\mathbf{D}$. The Pulay force is evaluated by contracting the differentiated density matrix $\mathbf{D}^{\mathbf{e}}$ with the (converged) KS matrix. This leave the three terms for the density-

fitted Coulomb force

$$\begin{aligned} \tilde{J}^{\mathbf{e}}_{ab} &= (\{ab\}^{\mathbf{e}}\,|\tilde{\rho}) \\ g^{\mathbf{e}}_{\alpha} &= (\alpha^{\mathbf{e}}|\rho) \\ \tilde{g}^{\mathbf{e}}_{\alpha} &= (\alpha^{\mathbf{e}}|\tilde{\rho}). \end{aligned} \tag{5.55}$$

5.5.2 Linear-scaling density-fitted force evaluation

Similarly to the undifferentiated Coulomb contribution $\tilde{\mathbf{J}}$ the three differentiated contributions of Eq. (5.55) are obtained in a linear scaling fashion combining Cauchy-Schwartz screening [39, 48] of Eq. (4.42), which for the derivative case includes second derivative integrals, and the CFMM approach [41].

There are two possible ways to obtain the far-field (FF) Coulomb gradient contribution - either by differentiating a given multipole moment expansion of the classical part of the interaction energy [88], $J^{\text{cls.}}$, given by [7],

$$J^{\text{cls.}} = \frac{1}{2}\sum_{p}\sum_{q\in \text{FF}_{\mathbf{P}}} D_p \mathbf{q}_p(\mathbf{P})^{\text{T}}\mathbf{W}(\mathbf{R}_{\bar{\mathbf{P}}\mathbf{P}})^{\text{T}}\mathbf{T}(\mathbf{R}_{\bar{\mathbf{Q}}\bar{\mathbf{P}}})\mathbf{W}(\mathbf{R}_{\bar{\mathbf{Q}}\mathbf{Q}})\mathbf{q}_q(\mathbf{Q})D_q, \tag{5.56}$$

or by first taking the analytical derivative and then introducing the CFMM approximation [87]. Differentiation of $J^{\text{cls.}}$, gives and exact gradients for a given order of multipole moment expansion, provided the partitioning of the global system into a hierarchical family of boxes remains the same throughout the optimization, and provided the centers $\mathbf{P}$ of the charge-distributions $\Omega_p(\mathbf{r})$ remain within the same boxes (with centers $\bar{\mathbf{P}}$). During the course of the optimization, however, the charge-distributions can move between different boxes. Furthermore, keeping the boxes fixed throughout the optimization is not a good alternative, since for instance different starting geometries then would converge to different minima. Therefore, both ways of obtaining the CFMM contribution to the gradient are limited by the accuracy of the CFMM expansion. We choose the second approach for obtaining the gradient, namely to first do the analytical derivative of the density-fitted energy according to Eq. (5.54), and then introduce the CFMM approximation for each of the three terms of Eq. (5.55) afterwards. This approach is simpler to implement as it only includes the multipole moments $\mathbf{q}^{\mathbf{e}}_p$ of the differentiated charge-distributions $\Omega^{\mathbf{e}}_p(\mathbf{r})$, keeping the FF potential fixed; and thus also leaving the translation matrices $\mathbf{W}(\mathbf{R})$ and the interaction matrices $\mathbf{T}(\mathbf{R})$ unchanged.

5.5.3 Acceleration of the near-field force contributions

Construction of the three contributions of Eq. (5.55) is accelerated using J-engine based integral evaluation of Ref [45], and we further utilize the fact that solid-harmonic combinations of Cartesian $G^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ and Hermite Gaussian $H^{\mathbf{i}}_{a,\mathbf{A}}(\mathbf{r})$ atomic orbitals are identical according to Paper I. This gives the three differentiated contributions of Eq. (5.55)

$$\begin{aligned}
\tilde{J}^{\mathbf{e}}_{ab} =& \sum_{\mathbf{ij}} S^a_{\mathbf{i}} S^b_{\mathbf{j}} \sum_{mn} C_m C_n \sum_{|\mathbf{t}|=0}^{l_a+l_b+1} \left(\delta_{\mathbf{EA}} \mathcal{E}^{\mathbf{i}+\mathbf{e},\mathbf{j}}_{\mathbf{t}} + \delta_{\mathbf{EB}} \mathcal{E}^{\mathbf{i},\mathbf{j}+\mathbf{e}}_{\mathbf{t}} \right) \\
& \qquad \sum_{q,\mathbf{Q}} \sum_{|\mathbf{u}|=l_q} \tilde{F}^{q,\mathbf{Q}}_{\mathbf{u}} R_{\mathbf{t}+\mathbf{e}+\mathbf{u}}(\gamma, \mathbf{R}_{PQ}) \\
g^{\mathbf{e}}_{\alpha} =& \delta_{\mathbf{EP}} \sum_{|\mathbf{t}|=l_\alpha} S^{\alpha}_{\mathbf{t}} \sum_{m} C_m (2\alpha_m)^{1-l_\alpha} \sum_{q,\mathbf{Q}} \sum_{|\mathbf{u}|=0}^{l_q} F^{q,\mathbf{Q}}_{\mathbf{u}} R_{\mathbf{t}+\mathbf{e}+\mathbf{u}}(\gamma, \mathbf{R}_{PQ}) \\
\tilde{g}^{\mathbf{e}}_{\alpha} =& \delta_{\mathbf{EP}} \sum_{|\mathbf{t}|=l_\alpha} S^{\alpha}_{\mathbf{t}} \sum_{m} C_m (2\alpha_m)^{1-l_\alpha} \sum_{q,\mathbf{Q}} \sum_{|\mathbf{u}|=l_q} \tilde{F}^{q,\mathbf{Q}}_{\mathbf{u}} R_{\mathbf{t}+\mathbf{e}+\mathbf{u}}(\gamma, \mathbf{R}_{PQ}),
\end{aligned} \tag{5.57}$$

where the Dirac delta function $\delta_{\mathbf{AB}}$ is zero if the centers $\mathbf{A}$ and $\mathbf{B}$ are different, and one if they are identical, and where the primitive Hermite repulsion integrals $R_{tuv}(\gamma, R_{PQ})$ are found using the recurrence relations of Eq. (4.20). As can be seen from Eq. (5.57) the use of Hermite rather than Cartesian Gaussians has two advantages. First, for the one-center auxiliary functions, the number of contractions is reduced, which can be seen for instance from the first term of $\tilde{J}^{\mathbf{e}}_{ab}$ where the innermost summation only contains terms for which $|\mathbf{u}| = l_q$, rather than stating from $|\mathbf{u}| = 0$. Second, there are no differentiated E-coefficients involved, instead the $E^{\mathbf{ij}}_{\mathbf{t}}$'s are incremented by one order in the quantum numbers.

5.5.4 Results and considerations

In Paper IV we present computational timings for the force evaluation for linear alkene chains, containing up to 502 atoms, and demonstrate efficient linear scaling formation of the density-fitted Coulomb matrix and gradient evaluation, at the BP86/6-31G** level of theory. Both the evaluation of the FF and NF contributions to the density-fitted Coulomb force, takes about a factor two longer than the contribution to a single Coulomb matrix construction. We further report averaged timings for the energy, force and geometry optimization steps for the geometry optimization of the taxol and valinomycine molecules at BP86/6-31G and BP86/6-31G** level of theory, and for titin molecule at the BP86 level of theory, in a combination of the 6-31G and 6-31G* bases. The density-fitted Coulomb NF forces are only a factor 2.1 to 2.5 slower than

the corresponding construction of the density-fitted Coulomb matrix, whereas the FF evaluation is only 20 to 30 percent slower.

For the full geometry optimization, using the incremental scheme [53] for the KS-matrix construction, the forces takes 24 to 28 percent of the full calculation, for the given systems. The forces are balanced between the density-fitted Coulomb and the XC forces, with the quadratic scaling one-electron part taking about 20 percent of the force evaluation for the largest system titin (with 392 atoms and 2221 contracted basis function). For the same system, a single point calculation takes on average about 1 hour for the energy evaluation, a little less than 20 minutes for the force evaluation, and less than two minutes for the geometry-optimization step.

The computational time of the energy evaluation is dominated by the FF evaluation of the density-fitted Coulomb contributions. For the titin molecule this step takes on average 38 percent of the energy evaluation time. The XC contribution takes 22, the grid evaluation 15, NF contribution 9.5, RH/DIIS optimization 7.8 and the construction of the screening matrices 4.4 percent.

The results presented in Paper IV clearly demonstrates the efficiency of the presented implementation. By removal of the quadratic scaling density-fitted force evaluation by using the CFFM approach, and through the implementation of the new integral evaluation scheme of Paper I for the NF contribution and an efficient XC correlation implementation.

Chapter 6

Wave-function optimization

The traditional SCF optimization combines the Roothaan-Hall (RH) diagonalization step of Eqs. (3.14) and (3.34) with the DIIS approach as outlined in section 3.2.3. Although highly successful, there are two problems with the RH/DIIS approach. The first is the cubic-scaling diagonalization step. The second problem concerns the convergence properties and the quality of the converged solution. The RH energy, which is minimized upon diagonalization, represents only a crude model to the true SCF energy. At the expansion point the RH energy has the correct gradient but only an approximate Hessian. As a consequence, convergence can at times be difficult to obtain and there is no guarantee that the converged solution actually represents a minimum.

In this chapter we look at two linear-scaling alternatives to the RH/DIIS approach, which incorporates additional information of Hessian. Both the linear-scaling trust-region SCF (LS-TRSCF) and the augmented Roothaan-Hall (ARH) approaches presented here are based on an exponential parameterization of the AO density matrix. In the next section we start by introducing the exponential parameterization, and continue with the LS-TRSCF and ARH approaches in the following two sections.

6.1 Parameterization of the density matrix

Let $\mathbf{D}$ be a valid AO density matrix, satisfying the trace

$$\mathrm{Tr}(\mathbf{DS}) = \frac{N}{2}, \tag{6.1a}$$

symmetry

$$\mathbf{D}^T = \mathbf{D}, \tag{6.1b}$$

and idempotency conditions

$$\mathbf{DSD} = \mathbf{D}. \tag{6.1c}$$

Any valid N-electron AO density matrix can then be obtained through *orbital rotations* acting on $\mathbf{D}$, according to the exponential parameterization of Helgaker *et.* al [7, 91, 92] in 2000,

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX}), \tag{6.2}$$

where $\mathbf{X}$ is an arbitrary anti-hermitian matrix (or anti-symmetric in the case of real orbital rotations). By inspection, it is easy to verify that the orbital-rotated density matrix of Eq. (6.2) satisfies the trace, symmetry and idempotency conditions of Eqs. (6.1a-c). Only the occupied-virtual and virtual-occupied rotations are non-redundant, therefore in order to avoid redundancies, the projection $\mathcal{P}(\mathbf{X})$, given by

$$\mathcal{P}(\mathbf{X}) = \mathbf{P}_\mathrm{o}\mathbf{X}\mathbf{P}_\mathrm{v}^\mathrm{T} + \mathbf{P}_\mathrm{v}\mathbf{X}\mathbf{P}_\mathrm{o}^\mathrm{T}, \tag{6.3}$$

with the projections onto the occupied $\mathbf{P}_\mathrm{o}$ and virtual $\mathbf{P}_\mathrm{v}$ spaces

$$\begin{aligned} \mathbf{P}_\mathrm{o} &= \mathbf{DS} \\ \mathbf{P}_\mathrm{v} &= \mathbf{I} - \mathbf{DS}, \end{aligned} \tag{6.4}$$

replaces the anti-hermitian matrix $\mathbf{X}$ in Eq. (6.2). The density matrix can further be expanded in orders of $\mathbf{X}$ according to the Baker-Campbell-Hausdorff expansion

$$\mathbf{D}(\mathbf{X}) = \mathbf{D} + [\mathbf{D}, \mathcal{P}(\mathbf{X})]_\mathbf{S} + \frac{1}{2}\left[[\mathbf{D}, \mathcal{P}(\mathbf{X})]_\mathbf{S}, \mathcal{P}(\mathbf{X})\right]_\mathbf{S} + \ldots, \tag{6.5}$$

with the S commutator $[\mathbf{A}, \mathbf{B}]_\mathbf{S}$ given as

$$[\mathbf{A}, \mathbf{B}]_\mathbf{S} = \mathbf{ASB} - \mathbf{BSA}. \tag{6.6}$$

Note that the exponential $\exp(\mathbf{A})$ is evaluated as a Taylor expansion according to

$$\exp(\mathbf{A}) = \sum_{n=0}^{\infty} \frac{\mathbf{A}^n}{n!}. \tag{6.7}$$

The convergence of the Taylor expansion of Eq. (6.7) is rapid only for small values of $\mathbf{A}$, and in order to accelerate convergence for large arguments one can apply the scaled relation

$$\exp(\mathbf{A}) = \left[\exp(2^{-k}\mathbf{A})\right]^{2^k}. \tag{6.8}$$

In this way the density matrix can be evaluated in about ten matrix multiplies. In this chapter the orbital rotations will be used for the formulation of AO-based HF/KS theory, which allows for linear-scaling SCF optimization. Orbital rotations based formalism is also used in connection with response theory in the next chapter.

6.1.1 AO based HF/KS theory

Solving the RH/KS equations is intrinsically cubic scaling in time due to the RH diagonalization step. In an AO-based formulation, however, we may replace the diagonalization step by a series of matrix multiplies, thereby allowing linear-scaling SCF optimization in the limit of large systems when the matrices are sparse. Since all valid AO densities can be obtained through the orbital rotations of Eq. (6.2), it is not necessary to go via the MO coefficients - we can instead parameterize the HF/KS wave function in terms of the orbital rotation parameters $\mathbf{X}$. It is worth noting that the canonical MOs found when solving the RH/KS equations is just one set out of infinitely many other choices of MOs, related by unitary transformations. The density matrix however is unique.

In terms of the AO density matrix $\mathbf{D}(\mathbf{X})$, the exact closed-shell SCF energy E^{SCF} can be expressed directly as

$$E^{\mathrm{SCF}}(\mathbf{X}) = \mathrm{Tr}\left[\mathbf{D}(\mathbf{X})\mathbf{h}\right] + \mathrm{Tr}\left(\mathbf{D}(\mathbf{X})\mathbf{G}[\mathbf{D}(\mathbf{X})]\right) + E_{\mathrm{xc}}\left[\mathbf{D}(\mathbf{X})\right] + \hat{h}_{\mathrm{nuc}}, \tag{6.9}$$

with the two-electron matrix $\mathbf{G}(\mathbf{D})$ defined by

$$G_{ab}(\mathbf{D}) = \sum_{cd} \left[2(ab|cd) - x_k(ac|bd)\right] D_{cd}. \tag{6.10}$$

Here $E_{\mathrm{xc}}\left[\mathbf{D}(\mathbf{X})\right]$ does not include the exact exchange; which is instead included in the two-electron matrix $\mathbf{G}(\mathbf{D})$ with fraction x_k. The RH diagonalization of Eqs. (3.14) and (3.34) and the minimization of the Roothaan-Hall energy [91, 92]

$$E^{\mathrm{RH}}(\mathbf{X}) = \mathrm{Tr}\left[\mathbf{D}(\mathbf{X})\mathbf{F}\right], \tag{6.11}$$

give rise to the same density matrix; so in effect the two approaches are identical. At the expansion point ($\mathbf{X} = \mathbf{0}$) both the true SCF energy E^{SCF} and the RH energy E^{RH} have the same gradient $\mathbf{g}$, but the RH energy has noly an approximate Hessian $\mathbf{H}$. Note that the Fock- or KS-matrix $\mathbf{F}$ of Eq. (6.11) is determined at the point of expansion (with density matrix $\mathbf{D}$). The minimization of Eq. (6.11) is carried out by first making an expansion to different orders of the non-redundant parameter $\mathbf{X}$, and then minimizing with respect to $\mathbf{X}$. Inserting the parameterization of Eq. (6.5) into Eq. (6.11), and ignoring third- and higher-order terms, minimization gives the set of linear equations

$$\mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} - \mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}} - \mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{F}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{F}^{\mathrm{vv}} = -\left(\mathbf{F}^{\mathrm{ov}} - \mathbf{F}^{\mathrm{vo}}\right) \tag{6.12}$$

where we have used

$$\frac{\partial \mathrm{Tr}(\mathbf{AB})}{\partial \mathbf{B}} = \mathbf{A}^{\mathrm{T}}, \tag{6.13}$$

and the anti-symmetry relation $\mathbf{X}^{\mathrm{T}} = -\mathbf{X}$, and where we have introduced the matrix-notation $\mathbf{M}^{\mathrm{ab}} = \mathbf{P}_{\mathrm{a}}^{\mathrm{T}}\mathbf{M}\mathbf{P}_{\mathrm{b}}$, with a and b either the occupied or virtual spaces o or v. The quasi-Newton equations of Eq. (6.12), for which the product of the approximate Hessian $\mathbf{H}$ with $\mathbf{X}$ is given by the left-hand side of Eq. (6.12), can be solved using for example the conjugate gradient approach. In the AO basis such an approach encounters difficulties [92], due to the high condition number of the Hessian. This problem was addressed in the curvy step method of Shao *et.* al [93] in 2003 using the preconditioned conjugate gradient (PCG) approach, which was achieved by transforming the Newton equations to the orthogonal Cholesky basis in which the condition number is much smaller. In the next two sections we discuss improvements to the curvy step method.

6.2 Trust-region SCF

The RH energy E^{RH} constitute only a crude model to the true SCF energy E^{SCF}, which may lead to steps that are to large to be trusted. In the trust-region approach [8, 9] the Newton step is only taken if the Hessian is positive definite and the Newton step is inside the trust-region; otherwise the minimum is determined on the boundary of the trust region. This approach was adapted into the TRRH approach by Thøgersen *et.* al [8, 9] in 2003, in conjunction with the diagonalization of the Fock/KS-matrix. In Paper II and Paper V this approach is extended further to a formalism suitable for linear scaling (LS-TRRH), in which the diagonalization step is replaced by a minimization procedure involving only matrix multiplications. In the linear-scaling trust-region SCF (LS-TRSCF) method, the LS-TRRH approach is combined with the trust-region density-subspace minimization (TRDSM), also presented in Refs. [8, 9].

6.2.1 The Roothaan-Haal Newton equations

In the LS-TRRH approach, the RH energy is minimized subject to the constraint that the new occupied space does not differ appreciably from the old occupied space, according to

$$\begin{aligned} ||\mathbf{D}(\mathbf{X}) - \mathbf{D}||_S^2 &= \mathrm{Tr}\left\{[\mathbf{D}(\mathbf{X}) - \mathbf{D}]\,\mathbf{S}\,[\mathbf{D}(\mathbf{X}) - \mathbf{D}]\,\mathbf{S}\right\} \\ &= 2N - 2\mathrm{Tr}\left[\mathbf{DSD}(\mathbf{X})\mathbf{S}\right] \leq \delta, \end{aligned} \tag{6.14}$$

for some maximal step-size δ. Introducing the constraint into the RH-energy of Eq. (6.11) yields the Lagrangian

$$L^{\mathrm{RH}}(\mathbf{X}) = E^{\mathrm{RH}}(\mathbf{X}) - 2\mu \left\{ N - \mathrm{Tr}\left[\mathbf{DSD}(\mathbf{X})\mathbf{S}\right] - \delta \right\} \tag{6.15}$$

with Lagrange multiplier μ. Expanding the Lagrangian in powers of $\mathbf{X}$ gives

$$\begin{aligned} L^{\mathrm{RH}}(\mathbf{X}) =& \mathrm{Tr}(\mathbf{FD}(\mathbf{X}) + \mathrm{Tr}(\mathbf{F}^{\mathrm{vo}}\mathbf{X} - \mathbf{F}^{\mathrm{ov}}\mathbf{X}) \\ &+ \mathrm{Tr}(\mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\mathbf{X} - \mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}}\mathbf{X}) \\ &+ 2\mu \left[\mathrm{Tr}(\mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\mathbf{X} - \delta\right] + \mathcal{O}(\mathbf{X}^3) \end{aligned} \tag{6.16}$$

Minimization with respect to $\mathbf{X}$, ignoring the higher order terms, gives the quasi-Newton equations

$$\begin{aligned} \mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} - \mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}} - \mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{F}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{F}^{\mathrm{vv}} \\ -2\mu \left(\mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\right) = -\left(\mathbf{F}^{\mathrm{ov}} - \mathbf{F}^{\mathrm{vo}}\right), \end{aligned} \tag{6.17}$$

where the level-shift μ can, according to Eq. (6.14), be used to limit the step-size of the orbital rotations based on tolerance δ. For each non-redundant solution $\mathbf{X} = \mathcal{P}(\mathbf{X})$, Eq. (6.17) has redundant solutions $\mathbf{X} + \mathbf{X}_R$, where $\mathbf{X}_R$ contains only redundant elements. Restricting ourselves to only the non-redundant solutions, and introducing the notation

$$\begin{aligned} \mathbf{g} &= \mathbf{F}^{\mathrm{ov}} - \mathbf{F}^{\mathrm{vo}} \\ \mathbf{H}(\mu) &= \mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}} - \mu\mathbf{S}, \end{aligned} \tag{6.18}$$

for the RH gradient and level-shifted Hessian, we can write the RH Newton equations of Eq. (6.17) more compactly as

$$\mathbf{H}(\mu)\tilde{\mathbf{X}}\mathbf{S} + \mathbf{S}\tilde{\mathbf{X}}\mathbf{H}(\mu) = -\mathbf{g}. \tag{6.19}$$

It is here assumed that $\tilde{\mathbf{X}}$ is pure in the sense that $\tilde{\mathbf{X}} = \mathcal{P}(\tilde{\mathbf{X}})$. Before looking into how Eq. (6.19) is solved, it can be instructive to recast the RH Newton equations in terms of a vectorized linear equation form. Applying the vec operator

$$\mathrm{vec}\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \end{pmatrix} \tag{6.20}$$

to both sides of Eq. (6.19), noting the relationship

$$\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}} \otimes \mathbf{A})\mathrm{vec}\,\mathbf{B} \tag{6.21}$$

with $\otimes$ the Kronecker (or direct) product, we arrive at the RH Newton linear equations

$$\mathcal{H}(\mu)\mathrm{vec}\,\tilde{\mathbf{X}} = -\mathrm{vec}\,\mathbf{g}, \tag{6.22}$$

with the level-shifted Hessian matrix given by

$$\mathcal{H}(\mu) = \mathbf{H}(\mu) \otimes \mathbf{S} + \mathbf{S} \otimes \mathbf{H}(\mu). \tag{6.23}$$

6.2.2 Preconditioner

To accelerate the convergence of the conjugate gradient method, a proper preconditioner $\mathcal{W}$, that approximates the Hessian $\mathcal{H}(\mu)$ and is easy to invert, is essential. The preconditioned RH linear equations is given by

$$\mathcal{W}^{-1}\mathcal{H}(\mu)\mathrm{vec}\,\tilde{\mathbf{X}} = -\mathcal{W}^{-1}\mathrm{vec}\,\mathbf{g}. \tag{6.24}$$

Factorizing the preconditioner according to $\mathcal{W} = \mathcal{V}^{\mathrm{T}}\mathcal{V}$, with $\mathcal{V} = \mathbf{V} \otimes \mathbf{V}$, we arrive at the preconditioned RH Newton equations

$$\mathbf{H}_{\mathrm{V}}(\mu)\tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{S}_{\mathrm{V}} + \mathbf{S}_{\mathrm{V}}\tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{H}_{\mathrm{V}}(\mu) = -\mathbf{g}_{\mathrm{V}}, \tag{6.25}$$

with

$$\begin{aligned} \mathbf{g}_{\mathrm{V}} &= \mathbf{F}_{\mathrm{V}}^{\mathrm{ov}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{vo}} \\ \mathbf{H}_{\mathrm{V}}(\mu) &= \mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}} - \mu\mathbf{S}_{\mathrm{V}}, \end{aligned} \tag{6.26}$$

in notation

$$\begin{aligned} \mathbf{A}_{\mathrm{V}} &= \mathbf{V}^{-\mathrm{T}}\mathbf{A}\mathbf{V}^{-1} \\ \mathbf{A}^{\mathrm{V}} &= \mathbf{V}\mathbf{A}\mathbf{V}^{\mathrm{T}}. \end{aligned} \tag{6.27}$$

In the limit of large μ the Newton equations of Eq. (6.25) takes the form $\mu\mathbf{S}_{\mathrm{V}}\tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{S}_{\mathrm{V}} = -\mathbf{g}_{\mathrm{V}}$. which suggests that a suitable preconditioner $\mathbf{V}$ is obtained by factorizing the overlap matrix

$$\mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{V} \tag{6.28}$$

which then gives $\mathbf{S}_{\mathrm{V}} = \mathbf{I}$. Such a factorization may be achieved in infinitely many ways. In Paper II we tested both the Cholesky factorization [69], $\mathbf{V}_{\mathrm{C}} = \mathbf{U}$, and the Löwdin decomposition [94], $\mathbf{V}_{\mathrm{S}} = \mathbf{S}^{1/2}$. Both factorizations yield the same condition number for the preconditioned Hessian $\kappa\left[\mathcal{W}^{-1}\mathcal{H}(\mu)\right]$, as they are related by a (condition number conserving) orthonormal transformation. Since the structures of $\mathbf{F}$ and $\mathbf{S}$ are broadly similar, these preconditioners typically reduce the condition number by several

orders of magnitude, greatly enhancing the conjugate gradient convergence and thus reducing overall computational efforts. Of all possible orthogonal bases, the Löwdin basis is the one that most closely resembles the AO basis, ensuring that locality is preserved to the greatest possible extent [95]. Although the Cholesky and Löwdin decomposed preconditioners show similar behavior, we use as default the Löwdin basis. A further improvement is possible by a diagonal preconditioning; in which the set of linear equations is scaled so that the diagonal part of the Hessian becomes the identity matrix. This is achieved by scaling the factorization matrix $\mathbf{V}$ by the square-root diagonal of the Hessian, according to

$$\mathbf{V}_{\mathrm{H}} = \mathrm{diag}\left([\mathbf{H}_{\mathrm{V}}(\mu)]_{11}, [\mathbf{H}_{\mathrm{V}}(\mu)]_{22}, \dots\right) \mathbf{V}. \tag{6.29}$$

6.2.3 The level-shifted Newton equations in the canonical MO basis

To better understand the convergence of the PCG algorithm, and how the level-shift parameter is to be chosen, we express Eq. (6.25) in the unoptimized canonical MO basis; in which the diagonal of the Fock/KS matrix consists of the orbital energies ϵ_P and the occupied-virtual and virtual-occupied blocks are non-zero. The level-shifted Hessian elements are given by

$$H_{AIBI}(\mu) = \delta_{AB}\delta_{IJ}(\epsilon_A - \epsilon_I - \mu), \tag{6.30}$$

which for the virtual-occupied elements of Eq. (6.25) gives

$$(\epsilon_A - \epsilon_I - \mu)X_{AI} = F_{AI}, \tag{6.31}$$

where X_{AI} is the solution vector in canonical MO basis, and with occupied indices I, J and unoccupied indices A, B. The step-length function

$$||\mathbf{X}||_{\mathrm{S}}^2 = \sum_{AI} \frac{F_{AI}^2}{(\epsilon_A - \epsilon_I - \mu)^2} \tag{6.32}$$

has $k+1$ branches, with k the number of eigenvalues $\epsilon_A - \epsilon_I$ of the unshifted Hessian, as illustrated in figure 6.1. The function is positive for all values μ and has asymptotes equal to the eigenvalues. For $\mu < \min(\epsilon_A - \epsilon_I)$ the level-shifted Hessian is positive definite and the RH energy is lowered to first and second orders [7]. Note that with too large level-shifts $|\mu|$ each step is small and the convergence slow, with too small level-shifts we may take steps that are too long to be trusted. The level shift parameter

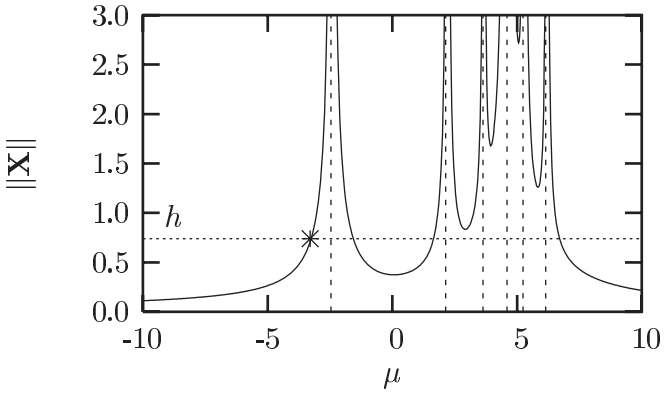

Figure 6.1: The step length $||\mathbf{X}||$ as a function of the level shift parameter μ. The trust-radius h is marked by the horizontal dotted line, and the step length function marked with an asterisk represents the chosen level shift.

μ that limits the step length $||\mathbf{X}^{\mathrm{V}}||$ to some maximal value h, is the lower bound to the intersection between the step-length function and the value h marked by an asterisk in figure 6.1 No level-shifting is performed if the intersection, μ, is positive, since the Hessian in such cases is positive definite.

6.2.4 The level-shifted Newton equaitons as an eigenvalue problem

In the AO basis the Hessian is not diagonal, and the iterations are determined iteratively. The level-shift parameter is determined by first solving the augmented eigenvalue problem

$$\mathbf{A}(\alpha)\begin{pmatrix}1\\ \tilde{\mathbf{x}}\end{pmatrix} = \mu\begin{pmatrix}1\\ \tilde{\mathbf{x}}\end{pmatrix} \tag{6.33}$$

in a reduced space R of $n+1$ trial vectors

$$\begin{pmatrix}1\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}0\\ \mathbf{b}_1\end{pmatrix}, \begin{pmatrix}0\\ \mathbf{b}_2\end{pmatrix}, \ldots, \begin{pmatrix}0\\ \mathbf{b}_n\end{pmatrix}, \tag{6.34}$$

by dynamically adjusting the parameter α to satisfy

$$||\alpha^{-1}\tilde{\mathbf{x}}^{\mathrm{R}}||^2 = h^2. \tag{6.35}$$

Here, the trial vectors $\mathbf{b}_i = \mathrm{vec}\,\mathbf{B}_i$ are orthonormal, and the first trial vector $\mathbf{b}_1$ is the normalized gradient vector

$$b_1 = ||\mathbf{g}_{\mathrm{V}}||^{-1}\mathbf{g}_{\mathrm{V}}. \tag{6.36}$$

To determine the lowest eigenvalue of the augmented Hessian efficiently, a good initial guess is required, but since the Hessian is not strongly diagonally dominant, such a guess is usually not available. In practice, therefore, the augmented Hessian is only used to update α, to ensure that the level shift is in the proper interval, and of the correct size. The improved trial vectors are obtained by solving the reduced space level-shifted Newton equations

$$\mathcal{H}_{\mathrm{V}}^{\mathrm{R}}(\mu)\tilde{\mathbf{x}}^{\mathrm{R}} = -\mathbf{g}_{\mathrm{V}}^{\mathrm{R}} = -||\mathbf{g}_{\mathrm{V}}|| \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{6.37}$$

where the reduced space Hessian elements are given by

$$\left[\mathcal{H}_{\mathrm{V}}^{\mathrm{R}}(\mu)\right]_{ij} = \mathbf{b}_i^{\mathrm{T}} \mathcal{H}_{\mathrm{V}}(\mu) \mathbf{b}_j. \tag{6.38}$$

6.2.5 Summary and concluding remarks

To summarize, each time we want to solve the preconditioned RH Newton equations of Eq. (6.25) we run a series of PCG iterations. In each PCG iteration, we first solve the augmented eigenvalue problem of Eq. (6.33), by adjusting the value α, until the condition $||\alpha^{-1}\tilde{\mathbf{x}}^{\mathrm{R}}||^2 \approx h^2$ is met. This gives a level-shift λ, equal to the lowest eigenvalue of the reduced space augmented eigenvalue problem, which is used to find the solution vector of the reduced space level-shifted Newton equation, using the basis $\mathbf{b}_i$. To give a more robust convergence in the global region, the Newton equations are first converged by constraining the Frobenius norm of the step $||\mathbf{X}^{\mathrm{V}}||$ followed by constraining the size-intensive maximum absolute element size $X_{\mathrm{max}}^{\mathrm{V}}$. We have found $||\mathbf{X}^{\mathrm{V}}|| = 0.6$ and $X_{\mathrm{max}}^{\mathrm{V}} = 0.35$ to be suitable parameters, and converge the preconditioned RH Newton, Eq. (6.25), until the residual $||\mathbf{R}||$ is reduced by a factor 100 and 50, respectively, from the initial two-dimensional reduced space solution. Note that in the local region both two step size conditions are fulfilled without imposing a level-shift. Here, convergence is achieved when the residual $||\mathbf{R}||$ is reduced by a factor 100. Note, that the overall SCF convergence is not sensitive to the choice of these convergence thresholds. Note that as a convergence criteria for the SCF optimization we used the size-intensive gradient norm $||\mathbf{g}^{\mathbf{V}}||/\sqrt{N}$, with N the number of electrons.

The sample calculations presented in Paper II, demonstrate that the LS-TRSCF scheme is both fast and robust. When compared to the curvy step method, Ref. [93], the main differences are the diagonal preconditioning of the PCG approach, which gives

significant reductions in the number of PCG iterations needed, and the level-shifting of the SCF iterations. An important feature of the LS-TRSCF method is that the level shift is determined dynamically; without any additional effort to a common user. This can be seen from comparison of RH/DIIS and LS-TRSCF calculations on some selected difficult cases. The RH/DIIS shows at times erratic behavior without the adding of a level-shift, whereas the LS-TRSCF scheme converges in all cases. Finally, linear scaling is demonstrated using sparse-matrix algebra for polyalanine peptides including up to 119 alanine residues.

6.3 Augmented Roothaan-Hall

The standard method of optimization consists of a two-step procedure. First, in the RH step, a new density is constructed by diagonalization of the KS matrix, or alternatively, as we have outlined in the previous section, by an energy minimization. Second, an improved density is determined, by combining this new density with the density matrices of the previous iterations; like in the DIIS or the TR-DSM step. Although this two-step procedure has been very successful, it sometimes fails, either by converging to a saddle point or by diverging. Whereas divergence is an obvious failure, a convergence to a saddle point leaves the user unaware that the solution does not represent the electronic ground state, unless a stability analysis of the stationary point is performed. Such a stability test is rarely performed, as the computational cost is comparable to that of the whole optimization. Previous attempts at improving the RH-DIIS convergence, Refs. [96, 8, 9] and in Paper II and Paper V, have retained the two-step framework, modifying the two steps separately, and have not constituted a dramatic improvement. In Paper VI we present the augmented Roothaan-Hall (ARH) method, in which the two step procedure is replaced a single step that fully exploits the Hessian information from the previous iterations. At each iteration we construct a local quadratic model of the KS energy that is exact to second order in the directions of the previous iterations and a good approximation in the remaining directions. The new density matrix is obtained by applying the trust-region minimization method, as described in the previous section, thereby ensuring that the energy is lowered at each iteration. Since the algorithm exploits information about the Hessian it converges by design to a minimum. Like the LS-TRSCF method is does not rely upon diagonalization and is based on matrix multiplications which enables linear scaling for large systems when the sparsity of the matrices is exploited.

6.3.1 The augmented Roothaan-Hall energy function

Assuming that we have carried out n iterations, in which we have generated sequences of density matrices $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_n$ and KS matrices $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_n$, with $\mathbf{F}_i = \mathbf{F}(\mathbf{D}_i)$. Expanding the KS energy to second order about $\mathbf{D}_n$ gives

$$E^{\mathrm{KS}}(\mathbf{D}) = E(\mathbf{D}_n) + \langle \mathbf{D} - \mathbf{D}_n | \mathbf{E}^{[1]}(\mathbf{D}_n) \rangle + \frac{1}{2} \langle \mathbf{D} - \mathbf{D}_n | \mathbf{E}^{[2]}(\mathbf{D}_n) | \mathbf{D} - \mathbf{D}_n \rangle, \tag{6.39}$$

where $\langle \mathbf{A} | \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^{\mathrm{T}} \mathbf{B})$. Ignoring the second order term we retain the RH energy of Eq. (6.11) with the constant offset $E(\mathbf{D}_n) - \langle \mathbf{D}_n | \mathbf{E}^{[1]}(\mathbf{D}_n) \rangle$. Therefore, the minimization of Eq. (6.39), with the second order term removed, is identical to the minimization of the RH energy. To make better use of the information in the density-matrix subspace spanned by the n former density matrices, we retain the second order term of Eq. (6.39) by invoking the quasi-Newton condition. First, the new density matrix is expanded using the exponential parameterization $\mathbf{D}(\mathbf{X})$ of Eq. (6.5) around the current density matrix $\mathbf{D}_n$. In an orthonormal basis this gives, by ignoring third and higher order terms,

$$\begin{aligned} E^{\mathrm{KS}}(\mathbf{X}) = & E(\mathbf{D}_n) + \langle [\mathbf{D}_n, \mathbf{X}] | \mathbf{E}^{[1]}(\mathbf{D}_n) \rangle \\ & + \frac{1}{2} \langle [[\mathbf{D}_n, \mathbf{X}], \mathbf{X}] | \mathbf{E}^{[1]}(\mathbf{D}_n) \rangle \\ & + \frac{1}{2} \langle [\mathbf{D}_n, \mathbf{X}] | \mathbf{E}^{[2]}(\mathbf{D}_n) | [\mathbf{D}_n, \mathbf{X}] \rangle. \end{aligned} \tag{6.40}$$

Note that both $\mathbf{E}^{[1]}(\mathbf{D})$ and $\mathbf{E}^{[2]}(\mathbf{D})$ still are the variations with respect to the density matrix elements, not the elements of $\mathbf{X}$.

6.3.2 The augmented Roothaan-Hall Netwon equaitons

The Newton equations are obtained by differentiating Eq. (6.40) with respect to $\mathbf{X}$, yielding

$$\begin{aligned} \frac{1}{2} \langle [[\mathbf{D}_n, \mathbf{X}], \mathbf{X}]^{\mu} | \mathbf{E}^{[1]}(\mathbf{D}_n) \rangle + \langle [\mathbf{D}_n, \mathbf{X}]^{\mu} | \mathbf{E}^{[2]}(\mathbf{D}_n) | [\mathbf{D}_n, \mathbf{X}] \rangle \\ = \langle [\mathbf{D}_n, \mathbf{X}]^{\mu} | \mathbf{E}^{[1]}(\mathbf{D}_n) \rangle, \end{aligned} \tag{6.41}$$

where the superscript μ denotes differentiation with respect to X_μ of $\mathbf{X}$. The left-hand side represents a multiplication of the Hessian with $\mathbf{X}$ and the right-hand side the gradient.

When the Newton equations are solved iteratively, each new trial vector is transformed by the Hessian. The first Hessian contribution of Eq. (6.41) is easy to evaluate,

whereas the second contribution requires a revaluation of both the Coulomb and the exchange-correlation contribution. We retain an approximation to the second order term of Eq. (6.39) by invoking the quasi-Newton condition

$$\begin{aligned}\mathbf{E}^{[2]}(\mathbf{D}_n)(\mathbf{D}_i - \mathbf{D}_n) &= \mathbf{E}^{[1]}(\mathbf{D}_i) - \mathbf{E}^{[1]}(\mathbf{D}_n) \\ &= 2\mathbf{F}(\mathbf{D}_i) - 2\mathbf{F}(\mathbf{D}_n) = 2\mathbf{F}_{in},\end{aligned} \tag{6.42}$$

and restrict $\mathbf{E}^{[2]}(\mathbf{D}_n)$ in the second Hessian term in Eq. (6.41) to operate only on the density-matrix subspace. This is achieved by the introduction of the density-subspace projector

$$\mathcal{P}_n = \sum_{i,j=1}^{n-1} |\mathbf{D}_{in}\rangle [\mathbf{T}^{-1}]_{ij} \langle \mathbf{D}_{jn}|, \quad T_{ij} = \langle \mathbf{D}_{in}|\mathbf{D}_{jn}\rangle \tag{6.43}$$

with $\mathbf{D}_{in} = \mathbf{D}_i - \mathbf{D}_n$, which gives the ARH approximation to the Hessian trial-vector transformation

$$\begin{aligned}&\langle [\mathbf{D}_n, \mathbf{X}]^\mu | \mathbf{E}^{[2]}(\mathbf{D}_n) \mathcal{P}_n | [\mathbf{D}_n, \mathbf{X}] \rangle \\ &\quad = 2 \sum_{ij} \langle [\mathbf{D}_n, \mathbf{X}]^\mu | \mathbf{F}_{in} \rangle [\mathbf{T}^{-1}]_{ij} \langle \mathbf{D}_{jn} | [\mathbf{D}_n, \mathbf{X}] \rangle\end{aligned} \tag{6.44}$$

An explicit expression for the non-redundant ARH quasi-Newton equations is given by

$$\begin{aligned}&(\mathbf{F}_n^{\mathrm{vv}} - \mathbf{F}_n^{\mathrm{oo}})\mathbf{X} + \mathbf{X}(\mathbf{F}_n^{\mathrm{vv}} - \mathbf{F}_n^{\mathrm{oo}}) \\ &\quad + \sum_{ij} (\mathbf{F}_{in}^{\mathrm{ov}} - \mathbf{F}_{in}^{\mathrm{vo}}) [\mathbf{T}^{-1}]_{ij} \mathrm{Tr}(\mathbf{D}_{jn}[\mathbf{D}_n, \mathbf{X}]) \\ &\quad = \mathbf{F}_n^{\mathrm{vo}} - \mathbf{F}_n^{\mathrm{ov}}.\end{aligned} \tag{6.45}$$

The above expression is equal to the RH Newton equations of Eq. (6.25), in an orthogonal basis, except for the addition of the third term. This additional contribution to the product of the Hessian with the trial vector goes beyond the RH Hessian. Within the density-matrix subspace the Hessian becomes exact to within the finite-difference error of the quasi-Newton condition, whereas in the orthogonal complement of the density-matrix subspace, the ARH Hessian reverts to the RH Hessian. The RH Hessian is in itself quite accurate, except in the directions that represents orbitals of similar energies. Since the density-matrix subspace spans primarily such directions, the ARH Hessian constitute a good approximation to the true KS Hessian. Similarly to the LS-TRSCF method we can straightforwardly apply the trust-region method also for the ARH method.

6.3.3 Concluding remarks

The presented calculations of Paper VI clearly demonstrates the benefits of the ARH method. Foremost, for difficult cases where the RH/DIIS converge to a saddlepoint, the ARH method locates the minimum. Second, the ARH exhibit fast convergence due to the benefits of the second order approximation, noting that with an exact Hessian quadratic convergence is obtained.

Chapter 7

Linear response theory

In section 3.4 we gave a brief introduction to response theory in the exact case. Response theory has been implemented for various approximate methods including HF, CC and DFT, traditionally in the MO basis. Recently, AO-based response theory has been explored by several authors.

Linear-scaling AO-based evaluation of static molecular properties has previously been considered by Ochsenfeld and Head Gordon [97], where the idempotency is taken care of by replacing the density-matrix by its McWeeny-purified counterpart [98]. Using this approach, Ochsenfeld *et.* al have reported a linear-scaling implementation of NMR shifts for linear alkanes and presented results for three-dimensional systems with more than 1000 atoms [99]. A linearly scaling time-independent response theory had also been presented by Niklasson and co-workers [100, 101] within a purification framework.

Larsen *et* al. [102] presented in 2000 an AO-based parameterization of HF and KS response-function theory. In paper Paper VII we used the derivation of Larsen *et* al. to obtain and implement a linear-scaling algorithm for solving the response eigenvalue and linear equations and for evaluating frequency-dependent second-order molecular properties, which we will discuss in the following. In 2007 Kussmann and Ochsenfeld [103] have also reported time-dependent Hartree-Fock and Kohn-Sham calculations of the frequency-dependent polarizability and hyper-polarizability using a linear-scaling framework.

7.1 Linear-scaling response theory

Here we start by giving a brief introduction to the time dependent AO based response functional formulation of Ref. [102], and then describe the linear-scaling linear-

response-function implementation of Paper VII.

7.1.1 AO-based SCF linear response theory

In the AO basis the linear response function of Eq. (3.58) can be written as [102]

$$\langle\langle A; B\rangle\rangle_\omega = \mathrm{Tr}\left[\mathbf{A}^{[1]}\mathbf{X}(\omega)\right] \tag{7.1}$$

with the corresponding response equation

$$(\mathbf{E}^{[2]} - \omega\mathbf{S}^{[2]})\mathrm{vec}\,\mathbf{X}(\omega) = -\mathrm{vec}\,\mathbf{B}^{[1]}, \tag{7.2}$$

which can be compared to the linear response equation of Eq. (3.56). Here $\mathbf{M}^{[1]}$ is the property gradient

$$\mathbf{M}^{[1]} = \mathbf{SDM} - \mathbf{MDS} = \mathbf{P}_\mathrm{o}^\mathrm{T}\mathbf{M} - \mathbf{MP}_\mathrm{o}, \tag{7.3}$$

of the operator M represented by the AO matrix $\mathbf{M}$, with elements

$$M_{ab} = \langle a|M|b\rangle. \tag{7.4}$$

As in Paper II and Paper VI, the generalized Hessian matrix $\mathbf{E}^{[2]}$ and the metric matrix $\mathbf{S}^{[2]}$ are not needed explicitly, but may instead be defined in terms of their linear transformations on a trial vector $\mathrm{vec}\,\mathbf{b}$ [102],

$$\begin{aligned}\boldsymbol{\sigma} &= \mathbf{E}^{[2]}(\mathbf{b}) = \mathcal{P}_\mathrm{T}\left[\mathbf{FD}_b\mathbf{S} - \mathbf{SD}_b\mathbf{F} + \mathbf{G}(\mathbf{D}_b)\mathbf{DS} - \mathbf{SDG}(\mathbf{D}_b)\right] \\ \boldsymbol{\rho} &= \mathbf{S}^{[2]}(\mathbf{b}) = -\mathcal{P}_\mathrm{T}\left[\mathbf{SD}_b\mathbf{S}\right],\end{aligned} \tag{7.5}$$

with the Fock/KS matrix given by

$$\mathbf{F} = \mathbf{h} + \mathbf{G}(\mathbf{D}). \tag{7.6}$$

Here $\mathbf{G}(\mathbf{D})$ includes the Coulomb, the exact exchange and the exchange-correlation contributions, and where we have introduced, in analogy with Eq. (6.3),

$$\mathcal{P}_\mathrm{T}(\mathbf{M}) = \mathbf{P}_\mathrm{o}^\mathrm{T}\mathbf{MP}_\mathrm{v} + \mathbf{P}_\mathrm{v}^\mathrm{T}\mathbf{MP}_\mathrm{o}, \tag{7.7}$$

and with the transformed density matrix

$$\mathbf{D}_b = [\mathcal{P}(\mathbf{b}), \mathbf{D}]_S = \mathcal{P}([\mathbf{b}, \mathbf{D}]_S) = \mathbf{P}_\mathrm{v}\mathbf{bP}_\mathrm{o}^\mathrm{T} - \mathbf{P}_\mathrm{o}\mathbf{bP}_\mathrm{v}^\mathrm{T}. \tag{7.8}$$

Under the assumption that $\mathcal{P}(\mathbf{b}) = \mathbf{b}$, we may also write the linear transformations of Eq. (7.5) in the form

$$\begin{aligned}\boldsymbol{\sigma} &= \mathbf{E}^{[2]}(\mathbf{b}) = (\mathbf{F}^\mathrm{vv} - \mathbf{F}^\mathrm{oo})\mathbf{bS} + \mathbf{Sb}(\mathbf{F}^\mathrm{vv} - \mathbf{F}^\mathrm{oo}) + \mathbf{G}^\mathrm{vo}(\mathbf{b}) - \mathbf{G}^\mathrm{ov}(\mathbf{b}) \\ \boldsymbol{\rho} &= \mathbf{S}^{[2]}(\mathbf{b}) = -\mathbf{S}^\mathrm{vv}\mathbf{bS}^\mathrm{oo} + \mathbf{S}^\mathrm{oo}\mathbf{bS}^\mathrm{vv}.\end{aligned} \tag{7.9}$$

The excitation energies ω_{n0} from the ground state $|0\rangle$ to the excited state $|n\rangle$ are the eigenvalues of the generalized eigenvalue problem

$$(\mathbf{E}^{[2]} - \omega_{n0}\mathbf{S}^{[2]})\mathrm{vec}\,\mathbf{X}_n = \mathbf{0}. \tag{7.10}$$

The corresponding transition moment of A is obtained from the residue of the linear response function

$$\langle 0|A|n\rangle = \mathrm{Tr}[\mathbf{A}^{[1]}\mathbf{X}_n]. \tag{7.11}$$

7.1.2 Iterative solution of response equations

Noting that

$$\begin{aligned} [\mathbf{E}^{[2]}(\mathbf{b})]^{\mathrm{T}} &= \mathbf{E}^{[2]}(\mathbf{b}^{\mathrm{T}}) \\ [\mathbf{S}^{[2]}(\mathbf{b})]^{\mathrm{T}} &= -\mathbf{S}^{[2]}(\mathbf{b}^{\mathrm{T}}), \end{aligned} \tag{7.12}$$

it follows that if the transformations of Eq. (7.9) are known for a given trial matrix $\mathbf{b}_i$,

$$\begin{aligned} \boldsymbol{\sigma}_i &= \mathbf{E}^{[2]}(\mathbf{b}_i) \\ \boldsymbol{\rho}_i &= \mathbf{S}^{[2]}(\mathbf{b}_i), \end{aligned} \tag{7.13}$$

they are also known for the transposed trial matrix $\mathbf{b}_i^{\mathrm{T}}$,

$$\begin{aligned} \boldsymbol{\sigma}_i^{\mathrm{T}} &= \mathbf{E}^{[2]}(\mathbf{b}_i^{\mathrm{T}}) \\ -\boldsymbol{\rho}_i^{\mathrm{T}} &= \mathbf{S}^{[2]}(\mathbf{b}_i^{\mathrm{T}}). \end{aligned} \tag{7.14}$$

Since the transformations of $\mathbf{b}_i$ and $\mathbf{b}_i^{\mathrm{T}}$ are related in such a simple manner, new trial matrices are always added in pairs.

Similarly to the solution of the level-shifted RH Newton equation of Eq. (6.25), the solution to the response and generalized equation of Eq. (7.2), and the generalized eigenvalue equation for the excitation energies of Eq. (7.10), are obtained in the reduced space formed by the basis of the paired orthonormal trial matrices $\left\{\mathbf{b}_1, \mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2, \mathbf{b}_2^{\mathrm{T}}, \ldots, \mathbf{b}_n, \mathbf{b}_n^{\mathrm{T}}, \right\}$, satisfying the projection relation $\mathbf{b}_i = \mathcal{P}(\mathbf{b}_i)$. The basis of trial matrices and their transformed counterparts, given by Eqs. (7.13) and (7.14), are used to set up both the response equations

$$(\mathbf{E}_{\mathrm{R}}^{[2]} - \omega\mathbf{S}_{\mathrm{R}}^{[2]})\mathbf{X}_{\mathrm{R}}(\omega) = -\mathbf{B}_{\mathrm{R}}^{[1]}, \tag{7.15}$$

and the generalized eigenvalue equation for the excitation energies

$$(\mathbf{E}_{\mathrm{R}}^{[2]} - \omega_{n0}\mathbf{S}_{\mathrm{R}}^{[2]})\mathbf{X}_{\mathrm{R},n} = \mathbf{0}. \tag{7.16}$$

in a reduced space R of dimension $2n$. In the PCG approach, the residual is used to obtain the next conjugate vector. To accelerate convergence the residuals $\mathbf{R}$ and $\mathbf{R}_n$, of Eqs. (7.2) and (7.10), given by

$$\mathbf{R} = \mathbf{E}^{[2]}(\mathbf{X}) - \omega\mathbf{S}^{[2]}(\mathbf{X}) + \mathbf{B}^{[1]} \tag{7.17}$$

and

$$\mathbf{R}_n = \mathbf{E}^{[2]}(\mathbf{X}_n) - \omega_{n0}\mathbf{S}^{[2]}(\mathbf{X}_n), \tag{7.18}$$

are preconditioned as will be discussed in the next subsection.

7.1.3 Preconditioning

The preconditioner $\mathbf{M}$ should be a good approximation to the response matrix, in the sense that the condition number of $\mathbf{M}^{-1}(\mathbf{E}^{[2]} - \omega\mathbf{S}^{[2]})$ should be significantly smaller than that of $\mathbf{E}^{[2]} - \omega\mathbf{S}^{[2]}$. Moreover, the cost of solving the preconditioning equation

$$\mathbf{M}\mathrm{vec}\mathbf{R}_p = \mathrm{vec}\mathbf{R}, \tag{7.19}$$

should be significantly smaller than the cost of solving the response equations Eq. (7.2). The most expensive step in the solution of the response equations is the evaluation of $\mathbf{G}(\mathbf{D})$, corresponding to the two last terms of Eq. (7.9). Since these two terms are small compared to the other terms in Eq. (7.9), a good preconditioner is given by

$$\mathbf{M} = \mathbf{E}_{\mathrm{F}}^{[2]} - \omega\mathbf{S}^{[2]}, \tag{7.20}$$

where $\mathbf{E}_{\mathrm{F}}^{[2]}$ is an approximation to $\mathbf{E}^{[2]}$ with the two last terms in Eq. (7.9) neglected. In the AO basis, the solution to the preconditioning equation Eq. (7.19) is difficult since the condition number of $\mathbf{E}_{\mathrm{F}}^{[2]} - \omega\mathbf{S}^{[2]}$ is large. The conditioning number may be greatly reduced by a transformation to an orthogonal AO (OAO) basis, like the Cholesky and Löwdin bases, in the same fashion as for the TRSCF approach of Paper II. In the OAO basis the linear transformations of Eq (7.5), take the form

$$\begin{aligned}(\boldsymbol{\sigma}_{\mathrm{F}})_{\mathrm{V}} &= (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}})\mathbf{X}^{\mathrm{V}} + \mathbf{X}^{\mathrm{V}}(\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}}) \\ \boldsymbol{\rho}_{\mathrm{V}} &= \mathbf{D}^{\mathrm{V}}\mathbf{X}^{\mathrm{V}} - \mathbf{X}^{\mathrm{V}}\mathbf{V}^{\mathrm{V}},\end{aligned} \tag{7.21}$$

in notation according to Eq. (6.27). The preconditioning of the residual for the response equations Eq. (7.2) is performed in the OAO basis, using a diagonal preconditioner, with elements

$$M_{\alpha\beta,\alpha\beta} = (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}})_{\alpha\alpha} + (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}})_{\beta\beta} - \omega\left[(\mathbf{D}^{\mathrm{V}})_{\alpha\alpha} - (\mathbf{D}^{\mathrm{V}})_{\beta\beta}\right. \tag{7.22}$$

The preconditioning of the residual of the eigenvalue equations may be carried out in the same manner but with the frequency ω replaced by the excitation energy ω_{n0}.

7.1.4 Initial vectors for the response equations

When solving the preconditioned response equations, the property gradient $-\mathbf{B}^{[1]}$ is straightforwardly taken as the trial conjugate vector for the PCG approach. For the eigenvalue equations however, another starting guess must be adopted. In the MO basis, the initial guess of an excitation vector has previously been successfully obtained as the solution to the simplified response eigenvalue equations

$$\left[\begin{pmatrix} \mathbf{\Delta}\epsilon & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}\epsilon \end{pmatrix} - \omega \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}\right] \mathrm{vec}\mathbf{Y}_{\mathrm{MO}} = \mathbf{0}, \tag{7.23}$$

with $\mathbf{\Delta}\epsilon_{AI} = \epsilon_A - \epsilon_I$. The left-hand side matrix of Eq. (7.23) is the simplified response matrix $\mathbf{M}_{\mathrm{MO}} = (\mathbf{E}_{\mathrm{F}}^{[2]})_{\mathrm{MO}} - \omega(\mathbf{S}^{[2]})_{\mathrm{MO}}$, and the solution matrix

$$\mathbf{Y}_{\mathrm{MO}} = \begin{pmatrix} \mathbf{0} & \mathbf{Y}_{\mathrm{vo}} \\ \mathbf{Y}_{\mathrm{ov}} & \mathbf{0} \end{pmatrix}, \tag{7.24}$$

has zero elements in the $\mathbf{Y}_{\mathrm{vo}}$ and $\mathbf{Y}_{\mathrm{ov}}$ blocks except for a unit element in $\mathbf{Y}_{\mathrm{vo}}$ corresponding to the considered orbital-energy difference $\epsilon_A - \epsilon_I$.

In the OAO basis, the initial vector becomes $\mathbf{Y}_{\mathrm{OAO}} = \mathbf{C}\mathbf{Y}_{\mathrm{MO}}\mathbf{C}^{\mathrm{T}}$ with elements $(\mathbf{Y}_{\mathrm{OAO}})_{ab} = C_{aA}C_{bI}$. The eigenvectors $\mathbf{C}$ of the Fock/KS matrix in OAO basis, can be obtained using iterative techniques for the highest occupied and lowest virtual orbitals, to form the initial guesses for the excitation vectors.

7.1.5 Results and considerations

Calculations of frequency-dependent polarizabilities and excitation energies were carried out for polyalanine peptides of increasing size. The largest peptide considered contained 1392 atoms, demonstrating the efficiency and robustness of the presented algorithm. As for the optimization of HF and KS density matrices, the solution of the response equations is dominated by the construction of the Fock/KS matrix, once in each iteration of the subspace algorithm, with performance similar to that in the MO basis. Important features include the paired-structure of the trial vectors, a non-diagonal preconditioner and good start vectors. The solution of the preconditioning equations is dominated by matrix multiplications, for which sparse-matrix algebra is applied to approach linear scaling. The preconditioning is carried out in the Löwdin basis.

Chapter 8

Concluding remarks

The further development of quantum-chemical methods is important, as they help understanding complicated chemical processes and provide important qualitative and quantitative information. Of the different quantum-chemical methods available, the DFT approach constitutes a good compromise between cost and accuracy. In this thesis we have investigated several strategies to improve the existing DFT methodologies:

- We have presented a novel integral evaluation scheme (Paper I), in which the solid-harmonic Gaussian are expanded in Hermite rather than Cartesian Gaussians. The presented scheme simplifies the evaluation of derivative integrals since differentiation merely increments the quantum numbers of the Hermite integrals. Consequently, the differentiation can be carried out to arbitrary order using the same code as for the undifferentiated integrals. Moreover, the presented scheme simplifies the evaluation of two- and three-center integrals, bypassing the time-consuming transformation to Cartesian basis.

- We have developed a boxed density-fitting scheme (Paper II) that corrects the fitted Coulomb energy to first order, for linear-scaling density-fitted Coulomb matrix evaluation. This approach is both efficient and accurate, and has been applied to molecular energy optimizations (Paper II) and to the calculation of frequency-dependent molecular response properties for polyalanine peptides containing up to 1400 atoms (Paper VII).

- We have presented a robust variational density-fitting formulation for the fitting of four-center two-electron integrals, applied to the density-fitted Coulomb and exchange matrix constructions, by solving the fitting equation in local metrics instead of the traditional Coulomb metric (Paper III). The reported results

demonstrate that local metrics can be used for linear-scaling density-fitting developments, without jeopardizing the accuracy of the calculations. The formalism is suitable for the extension to molecular properties and to other quantum-chemistry methods involving four-center two-electron integrals. The errors of performing the fitting in the overlap metric are, for benchmark calculations, shown to be only factor $1.5 - 2.0$ larger than in the traditional Coulomb metric, rather than the order of magnitude larger errors reported previously.

- An efficient evaluation of the molecular forces has been developed and applied to systems containing up to 500 atoms (Paper IV). The different contributions to the density-fitted Coulomb force is demonstrated to scale linearly with system size by combining screening with multipole moment far-field interactions. The evaluation of the near-field is particularly effective, using the novel integration scheme of Paper I. The forces have further been applied to the geometry optimization of systems containing up to 400 atoms.

- We have further implemented an efficient linear-scaling AO-based SCF optimization scheme, the TRSCF approach (Paper II), based on the exponential parameterization of the AO density matrix. The RH energy is minimized in each SCF through a series of preconditioned conjugate-gradient iterations, using the Löwdin orthogonal AO basis, bypassing the traditional cubic-scaling diagonalization step, and combined with the trust-region DSM approach for density averaging. By automating step size criteria, based on the trust-region approach, the TRSCF approach can be used in a black-box manner (i.e. without the need for a common user to manually set a level-shift or damping parameter), and is further demonstrated to be more robust than the traditional RH/DIIS approach. Linear-scaling SCF optimization is reported for polyalanine peptides containing up to 1200 atoms.

- In the linear-scaling ARH approach (Paper VI), a local quadratic model of the KS energy, that is exact to second order in the subspace of the previous density matrices and constitute a good approximation in other directions, is minimized using the trust-region approach. The method differs from previous KS optimization methods in that it does not involve two separate steps, such as the RH diagonalization followed by the DIIS averaging. Instead, one single step is performed that exploits the curvature information spanned by the previous density matrices. Since the ARH contains information about the electronic Hessian, the method

both enhances performance and converges by design to a minimum, resulting in a robust and efficient optimization scheme. This is demonstrated by sample calculations where the ARH approach finds a minimum and the traditional RH/DIIS approach either diverges or converges to a saddle-point.

- We have finally presented a linear-scaling AO-based linear response implementation for HF and DFT (Paper VII). The response equations are solved iteratively in a subspace of paired trial vectors. The used of paired trial vectors preserve the algebraic structure of the response equations, both enhancing convergence and avoiding complex eigenvalues. A non-diagonal preconditioner combined with good initial guesses allows performance comparable with canonical MO theory, with typically five to ten iterations needed for convergence. The computational time is dominated by the construction of the effective Fock/KS matrices, as in the canonical case, but with linear complexity achieved using sparse-matrix algebra. Linear scaling, and robust convergence is demonstrated for the calculation of frequency-dependent polarizabilities and excitation energies of polyalanine peptides containing up to 1400 atoms.

To briefly summarize, we have explored different approaches to improve the DFT methodology. The improvements include 1) enhancing computational performance, 2) reducing scaling behavior, 3) development of black-box methods and 4) extending the applicability of existing methodology.

The developments have been carried out in a development branch of the quantum chemistry package DALTON, in a collaboration including many developers. My main contribution has been associated with reducing the computational prefactor of the DFT method, by the development and implementation of efficient integral evaluation schemes combined with linear-scaling and density-fitting developments. Significant effort has also been directed at code optimizations. The combined efforts by myself and my co-workers have resulted in orders of magnitude speed-ups for the evaluation of the Coulomb and XC contributions, compared to the available code when I started my thesis in 2004.

During the course of this thesis, it has become apparent that we would benefit from a new integral driver tailored for large systems, and we have implemented a new integral driver the last year of my thesis. The flexibility of the new driver will allow us to explore different approaches for efficient integral evaluation more easily. There are several approaches I would like to explore in the future. As demonstrated by Paper IV, the evaluation of the density-fitted Coulomb contribution is dominated by the

FF evaluation. It is therefore natural to explore different approaches for the efficient evaluation of the FF density-fitting Coulomb contributions. Possible improvements include the splitting of the Coulomb interaction into classical and non-classical contributions [45], which allows for efficient FF evaluation through FMM rather than CFFM and with possible extensions for tailoring the FMM approach for the density-fitting approach. Additionally, the implementation of the Poisson density-fitting approach [73] will reduce the number of Coulomb interactions to be treated with FFM. The Poisson density-fitting approach can also be used for the efficient NF evaluation and for density-fitting exchange developments.

The main computational bottleneck in the current implementation is the exact exchange, evaluated using the LinK scheme [54]. The exact exchange is particularly important in the DFT treatment of large systems, as the pure functionals tend to underestimate the HOMO-LUMO gaps, resulting in density optimization difficulties, in addition to a poor description of for example long range excitations and polarizabilities. Therefore the efficient evaluation of exact exchange is important. Efficient density-fitted exchange evaluation has been developed as part of this thesis for small to medium sized systems, and I would like to extend these development for an efficient linear-scaling approach by combining linear-scaling density-fitting approaches with the LinK scheme. The main computational challenge for efficient density-fitting treatment of the exact exchange for larger systems, is the number of auxiliary basis functions included in the fitting expansion of the charge distributions - due to the long decaying nature of the fitting coefficients. Highly local fitting schemes will reduce the prefactor of the different transformation and contraction steps involved, and one way forward may be the diatomic fitting [104] in combination with a robust variational density-fitting formulation as the one presented in Paper III. Such developments can also prove important for correlated treatments, like for the RI-MP2 approach [74].

The developments for efficient integral evaluation schemes is not fruitful without efficient and reliable density optimization and response solver methodology. Therefore the continued collaboration with colleagues in these fields is important, and I hope to broaden my scope through such collaboration in the future. Naturally, developed methodology for integral evaluation also needs to be extended to allow efficient property evaluations.

Bibliography

[1] E. Schrödinger, Phys. Rev. **28**, 1049 (1926).

[2] P. A. M. Dirac, Proceedings of the Royal Society A. **A123**, 714 (1929).

[3] R. T. Gallant and A. St-Amant, Chem. Phys. Lett. **256**, 569 (1996).

[4] B. I. Dunlap, J. Mol. Struct. (Theochem). **501**, 221 (2000).

[5] B. I. Dunlap, J. W. D. Connolly and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).

[6] O. Vahtras, J. Almlöf and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).

[7] T. Helgaker, P. Jørgensen and J. Olsen. *Molecular Electronic–Structure Theory.* Wiley, Chichester, 2000.

[8] L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek and T. Helgaker, J. Chem. Phys. **121**, 16 (2003).

[9] L. Thøgersen, J. Olsen, A. K. hn, P. Jørgensen, P. Sałek and T. Helgaker, J. Chem. Phys. **123**, 074103 (2004).

[10] P. Pulay, J. Comp. Chem. **3**, 556 (1982).

[11] F. Jensen. *Introduction to Computational Chemistry.* Wiley, Chichester, 1999.

[12] P. W. Atkins and R. S. Friedman. *Molecular Quantum Mechanics, Third edition.* Oxford University Press, Oxford, 1997.

[13] W. Koch and M. C. Holthausen. *A Chemist's Guide to Density Functional Theory, 2nd ed.* Wiley-VCG, Weinheim, 2000.

[14] P. Hohenberg and W. Kohn, Phys. Rev. B. **136**, 864 (1964).

[15] M. Levy, Phys. Rev. A. **26**, 1200 (1982).

[16] E. H. Lieb, Int. J. Quantum Chem. **24**, 243 (1983).

[17] M. Levy, Proc. Natl. Acad. Sci. **76**, 6062 (1979).

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

[19] P. A. M. Dirac, Proc. Camb. Phil. Soc. **26**, 376 (1930).

[20] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[21] S. J. Vosko, L. Wilk and M. Nusair, Can. J. Phys. **58**, 1200 (1980).

[22] A. D. Becke, Phys. Rev. A. **38**, 3098 (1988).

[23] C. Lee, W. Yang and R. G. Parr, Phys. Rev. B. **37**, 785 (1988).

[24] J. Perdew, J. A. Chevary, S. Vosko, K. Jackson, M. Pederson, and C. Fiolhais, Phys. Rev. B. **46**, 6671 (1992).

[25] J. Perdew, K. Burke and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[26] P. J. Stevens, J. F. Devlin, C. F. Chabalowski and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).

[27] C. Adamo and V. Barone, J. Phys. Chem. **110**, 6158 (1999).

[28] A. D. Becke, J. Chem. Phys. **107**, 8554 (1997).

[29] P. J. Wilson, T. J. Bradley and D. J. Tozer, J. Chem. Phys. **115**, 9233 (2001).

[30] T. W. Keal and D. J. Tozer, J. Chem. Phys. **123**, 121103 (2005).

[31] T. Yanai, D. P. Tew and N. C. Handy, Chem. Phys. Lett. **393**, 51 (2004).

[32] J. Olsen and P. Jørgensen. *Modern Electronic Structure Theory,* Ch. 13*, edited by D. R. Yarkony.* World Scientific, Singapore, 1995.

[33] M. Dupuis, J. Rys and H. F. King, J. Chem. Phys. **65**, 111 (1976).

[34] L. E. McMurchie and E. R. Davidson, J. Comp. Phys. **26**, 218 (1978).

[35] S. Obara and A. Saika, J. Chem. Phys. **84**, 3963 (1986).

[36] M. Head-Gordon and J. A. Pople, J. Chem. Phys. **89**, 5777 (1988).

[37] A. M. Köster, J. Chem. Phys. **118**, 9943 (2003).

[38] DALTON, *an ab initio electronic structure program, Release 2.0*, 2005. See http://www.kjemi.uio.no/software/dalton/dalton.html.

[39] M. Häser and R. Ahlrichs, J. Comp. Chem. **10**, 104 (1989).

[40] D. S. Lambrecht and C. Ochsenfeld, J. Chem. Phys. **123**, 184101 (2005).

[41] C. A. White, B. G. Johnson, P. M. W. Gill and M. Head-Gordon, Chem. Phys. Lett. **230**, 8 (1994).

[42] G. R. Ahmadi and J. Almlöf, Chem. Phys. Lett. **246**, 364 (1995).

[43] C. A. White and M. Head-Gordon, *J. Chem. Phys.* **104**, 2620 (1996).

[44] Y. Shao and M. Head-Gordon, *Chem. Phys. Lett.* **323**, 425 (2000).

[45] M. A. Watson, P. Sałek, P. Macak and T. Helgaker, J. Chem. Phys. **121**, 2915 (2004).

[46] L. Füsti-Molnár and P. Pulay, J. Chem. Phys. **117**, 7827 (2002).

[47] N. H. F. Beebe and J. Linderberg, Int. J. Quantum Chem. **7**, 683 (1977).

[48] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).

[49] E. J. Baerends, D. E. Ellis and P. Ros, Chem. Phys. **2**, 41 (1973).

[50] B. I. Dunlap, J. W. D. Connolly and J. R. Sabin, J. Chem. Phys. **71**, 4993 (1979).

[51] L. Greengard and V. I. Rokhlin, J. Comp. Phys. **73**, 325 (1987).

[52] H. Ding, N. Karasawa and W. A. Goddard, J. Chem. Phys. **97**, 4309 (1992).

[53] E. Schwegler, M. Challacombe and M. Head-Gordon, J. Chem. Phys. **106**, 9708 (1997).

[54] C. Ochsenfeld, C. A. White and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).

[55] van Wüllen, Chem. Phys. Lett. **219**, 8 (1994).

[56] J. M. Perez-Jorda and W. Jang, Chem. Phys. Lett. **241**, 469 (1995).

[57] R. E. Stratmann, Chem. Phys. Lett. **257**, 213 (1996).

[58] A. D. Becke, J. Chem. Phys. **88**, 2547 (1988).

[59] S. F. Boys and I. Shavitt. *A fundamental calculation of the energy surface for the system of three hydrogen atoms.* University of Wisconsin Rept. WIS-AF-13, 1959.

[60] M. D. Newton, N. S. Ostlund and J. A. Pople, J. Chem. Phys. **49**, 5192 (1968).

[61] M. D. Newton, J. Chem. Phys. **51**, 3917 (1969).

[62] F. P. Billingsley II and J. E. Bloor, Chem. Phys. Lett. **4**, 48 (1969).

[63] F. P. Billingsley II and J. E. Bloor, J. Chem. Phys. **55**, 5178 (1971).

[64] F. E. Harris and R. Rein, Theor. Chim. Acta. **6**, 73 (1966).

[65] J. A. Jafri and J. L. Whitten, J. Chem. Phys. **61**, 2116 (1974).

[66] J. W. Mintmire, Int. J. Quantum Chem. Quantum Chem. Symp. **13**, 163 (1979).

[67] B. I. Dunlap, J. Mol. Struct. (Theochem). **529**, 37 (2000).

[68] F. Weigend, Phys. Chem. Chem. Phys. **4**, 4285 (2002).

[69] W. H. Press, S. A. Teukolsky, W. T. Wetterling and B. P. Flannery. *Numerical Recipes in Fortran, 2nd ed.* Cambridge University Press, Cambridge, 1992.

[70] K. Eichkorn, O. Treutler, H. Öhm, M. Häser and R. Ahlrichs, Chem. Phys. Lett. **240**, 283 (1995).

[71] K. Eichkorn, F. Weigend, O. Treutler and R. Ahlrichs, Theor. Chim. Acta. **97**, 119 (1997).

[72] F. Weigend, J. Comp. Chem. **29**, 167 (2008).

[73] F. R. Manby and P. J. Knowles, Phys. Rev. Lett. **87**, 163001 (2000).

[74] M. Feyereisen and G. Fitzgerald, Chem. Phys. Lett. **208**, 359 (1993).

[75] Y. Jung, A. Sodt, P. M. W. Gill and M. Head-Gordon, Proc. Natl. Acad. Sci. USA. **102**, 6692 (2005).

[76] A. Sodt, J. E. Subotnik and M. Head-Gordon, J. Chem. Phys. **125**, 194109 (2006).

[77] W. T. Yang and T.-S. Lee, J. Chem. Phys. **103**, 5674 (1995).

[78] R. Polly, H.-J. Werner, F. R. Manby and P. J. Knowles, Mol. Phys. **102**, 2311 (2004).

[79] J. Pipek and P. G. Mezey, J. Chem. Phys. **90**, 4916 (1989).

[80] F. Aquilante, T. B. Pedersen and R. Lindh, J. Chem. Phys. **126**, 194106 (2007).

[81] A. Sodt and M. Head-Gordon, J. Chem. Phys. **128**, 104106 (2008).

[82] B. Jansík, S. Høst, P. Jørgensen, J. Olsen and T. Helgaker, J. Chem. Phys. **126**, 124104 (2007).

[83] M. J. G. Peach, P. Benfield, T. Helgaker and D. J. Tozer, J. Chem. Phys. **128**, 044118 (2008).

[84] R. Fournier, J. Andzelm and D. R. Salahub, J. Chem. Phys. **90**, 6371 (1989).

[85] B. I. Dunlap, J. Andzelm and J. W. Mintmire, Phys. Rev. A. **42**, 6354 (1990).

[86] D. Rappoport and F. Furche, J. Chem. Phys. **122**, 064105 (2005).

[87] J. C. Burant, M. C. Strain, G. E. Scuseria and M. J. Frisch, Chem. Phys. Lett. **248**, 43 (1996).

[88] Y. Shao, C. A. White and M. Head-Gordon, J. Chem. Phys. **114**, 6572 (2001).

[89] P. Pulay, Mol. Phys. **17**, 197 (1969).

[90] H. Larsen, , T. Helgaker, J. Olsen and P. Jørgensen, J. Chem. Phys. **115**, 10344 (2001).

[91] T. Helgaker, H. Larsen, J. Olsen and P. Jørgensen, Chem. Phys. Lett. **327**, 397 (2000).

[92] H. Larsen, J. Olsen, P. Jørgensen and T. Helgaker, J. Chem. Phys. **115**, 9685 (2001).

[93] Y. Shao, C. Saravanan, M. Head-Gordon and C. A. White, J. Chem. Phys. **118**, 6144 (2003).

[94] P.-O. Löwdin, J. Chem. Phys. **18**, 365 (1950).

[95] B. C. Carlson and J. M. Keller, Phys. Rev. **105**, 102 (1957).

[96] K. N. Kudin, G. E. Scuseria and E. Cancés, J. Chem. Phys. **116**, 8255 (2002).

[97] C. Ochsenfeld and M. Head-Gordon, Chem. Phys. Lett. **270**, 399 (1997).

[98] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).

[99] C. Ochsenfeld, J. Kussmann and F. Koziol, Angew. Chem. Int. Ed. **43**, 4485 (2004).

[100] V. Weber, A. M. N. Niklasson and M. Challacombe, Phys. Rev. Lett. **92**, 193002 (2004).

[101] A. M. N. Niklasson and V. Weber, J. Chem. Phys. **127**, 064105 (2007).

[102] H. Larsen, P. Jørgensen, J. Olsen and T. Helgaker, J. Chem. Phys. **113**, 8908 (2000).

[103] J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **127**, 204103 (2007).

[104] C. Fonseca Guerra, J. G. Snijders, G. te Velde and E. J. Baerends, Theor. Chem. Acc. **99**, 391 (1998).

Papers

Paper I

A unified scheme for the calculation of differentiated and undifferentiated molecular integrals over solid-harmonic Gaussians

S. Reine, E. Tellgren and T. Helgaker

Physical Chemistry Chemical Physics, **9**, 4771-4779 (2007)

A unified scheme for the calculation of differentiated and undifferentiated molecular integrals over solid-harmonic Gaussians

Simen Reine,† Erik Tellgren‡ and Trygve Helgaker*§

Received 12th April 2007, Accepted 10th May 2007
First published as an Advance Article on the web 4th July 2007
DOI: 10.1039/b705594c

Utilizing the fact that solid-harmonic combinations of Cartesian and Hermite Gaussian atomic orbitals are identical, a new scheme for the evaluation of molecular integrals over solid-harmonic atomic orbitals is presented, where the integration is carried out over Hermite rather than Cartesian atomic orbitals. Since Hermite Gaussians are defined as derivatives of spherical Gaussians, the corresponding molecular integrals become the derivatives of integrals over spherical Gaussians, whose transformation to the solid-harmonic basis is performed in the same manner as for integrals over Cartesian Gaussians, using the same expansion coefficients. The presented solid-harmonic Hermite scheme simplifies the evaluation of derivative molecular integrals, since differentiation by nuclear coordinates merely increments the Hermite quantum numbers, thereby providing a unified scheme for undifferentiated and differentiated four-center molecular integrals. For two- and three-center two-electron integrals, the solid-harmonic Hermite scheme is particularly efficient, significantly reducing the cost relative to the Cartesian scheme.

1. Introduction

In molecular electronic-structure theory, an essential step is the evaluation of molecular one- and two-electron integrals over one-electron basis functions, which are typically taken to be linear combinations of solid-harmonic Gaussians. Over the years, several efficient schemes have been developed for the evaluation of such integrals: the Rys scheme,[1] the McMurchie–Davidson scheme,[2] the Obara–Saika scheme,[3] as well as modifications to these schemes.[4,5] In all these schemes, the integration is carried out over Cartesian Gaussians, followed by a transformation to the solid-harmonic basis (or by a series of partial transformations to this basis, in the course of the integration). A disadvantage of this approach is that derivatives of Cartesian Gaussians with respect to the orbital centers are linear combinations of undifferentiated Gaussians, making the evaluation of derivative integrals cumbersome.

In the present paper, we observe that solid-harmonic combinations of Cartesian Gaussian atomic orbitals are in fact identical to the corresponding combinations of Hermite Gaussians, generated by differentiation of spherical Gaussians with respect to the orbital center. Based on this observation, we propose to evaluate molecular integrals over Hermite rather than Cartesian Gaussians or, equivalently, to generate molecular integrals by differentiation of integrals over spherical Gaussians. In this manner, we obtain derivative integrals (as needed for the evaluation of molecular gradients and Hessians) and integrals involving the momentum operator (as needed for the kinetic energy and for kinetically-balanced small components in relativistic theory) by a simple modification of the scheme for undifferentiated integrals, consisting only in the raising of the Hermite quantum numbers. As a bonus, the use of Hermite rather than Cartesian Gaussians simplifies the evaluation of two- and three-center integrals significantly, relative to the scheme based on Cartesian Gaussians.[6] Živković and Maksić have considered the use of Hermite Gaussian basis functions but not in solid-harmonic form.[7]

The remainder of this paper consists of four sections. First, in Section 2, we demonstrate that solid-harmonic Gaussians may be expanded in Hermite Gaussians, using the same expansion coefficients as for the Cartesian Gaussians. Next, in Section 3, we expand molecular integrals over solid-harmonic Gaussians in terms of two-, three- and four-center Hermite integrals, whose evaluation by the Obara–Saika and McMurchie–Davidson schemes is described in Section 4. Section 5 contains some concluding remarks.

2. Solid-harmonic Gaussians

In the present section, we discuss solid-harmonic Gaussian functions, noting that these may be constructed equally well from Cartesian and Hermite Gaussians, using the same solid-harmonic expansion coefficients. The properties of the Cartesian and Hermite Gaussians are compared and it is pointed out that Hermite Gaussians, defined as (scaled) derivatives of spherical Gaussians, are better suited than Cartesian Gaussians in applications where orbital derivatives are needed—for example, in calculations of molecular gradients and Hessians and in relativistic calculations where the small-component

Centre of Theoretical and Computational Chemistry, Department of Chemistry, University of Oslo, P.O. Box 1033 Blindern, N-0315 Oslo, Norway
† Present address: Department of Chemistry, University of Aarhus, DK-8000 Århus C, Denmark.
‡ Present address: Department of Chemistry, University of Durham, South Road, Durham, DH1 3LE, UK.
§ Present address: Department of Chemistry, University of Durham, South Road, Durham, DH1 3LE, UK.

functions are obtained from the large-component ones by differentiation.

2.1 Solid-harmonic Gaussians expanded in Cartesian Gaussians

In this paper, we consider the evaluation of molecular integrals over solid-harmonic Gaussians of the form

$$G_{lm}(\boldsymbol{r}, a, \boldsymbol{A}) = S_{lm}(\boldsymbol{r}_A)\exp(-ar_A^2) \tag{1}$$

where $a > 0$ is the Gaussian exponent, $\boldsymbol{r}_A = \boldsymbol{r} - \boldsymbol{A}$ is the position of the electron $\boldsymbol{r}$ relative to the center of the Gaussian $\boldsymbol{A}$, and $S_{lm}(\boldsymbol{r}_A)$ with $0 \leq |m| \leq l$ is a real-valued solid-harmonic function of $\boldsymbol{r}_A$. The real-valued solid-harmonic functions satisfy Laplace's equation and are eigenfunctions of the total angular momentum and (when linearly combined) of the projected angular momentum:[8]

$$\nabla^2 S_{lm}(\boldsymbol{r}_A) = 0 \tag{2}$$

$$L^2\, S_{lm}(\boldsymbol{r}_A) = l(l+1)\, S_{lm}(\boldsymbol{r}_A) \tag{3}$$

$$L_z[S_{lm}(\boldsymbol{r}_A) + \mathrm{i}S_{l,-m}(\boldsymbol{r}_A)] = m[S_{lm}(\boldsymbol{r}_A) + \mathrm{i}S_{l,-m}(\boldsymbol{r}_A)] \tag{4}$$

In atomic units, the angular-momentum operators about $\boldsymbol{A}$ are as usual given by

$$L^2 = L_x^2 + L_y^2 + L_z^2 \tag{5}$$

$$L_z = -\mathrm{i}\left(x_A \frac{\mathrm{d}}{\mathrm{d}y_A} - y_A \frac{\mathrm{d}}{\mathrm{d}x_A}\right) \tag{6}$$

and similarly for L_x and L_y. Rewriting the operator for the total angular momentum about $\boldsymbol{A}$ in the form

$$L^2 = -r_A^2\nabla^2 + (\boldsymbol{r}_A \cdot \boldsymbol{\nabla})^2 + \boldsymbol{r}_A \cdot \boldsymbol{\nabla} \tag{7}$$

and invoking eqn (2), we find that the solid harmonics satisfy the eigenvalue equation

$$(\boldsymbol{r}_A \cdot \boldsymbol{\nabla})\, S_{lm}(\boldsymbol{r}_A) = l\, S_{lm}(\boldsymbol{r}_A) \tag{8}$$

Since the eigenfunctions of $\boldsymbol{r}_A \cdot \boldsymbol{\nabla}$ belonging to the eigenvalue l are the set of homogeneous polynomials of degree l in $\boldsymbol{r}_A$, we conclude that the solid harmonics $S_{lm}(\boldsymbol{r}_A)$ are homogeneous polynomials of degree l in $\boldsymbol{r}_A$:

$$S_{lm}(\zeta \boldsymbol{r}_A) = \zeta^l S_{lm}(\boldsymbol{r}_A) \tag{9}$$

Consequently, we may expand the solid harmonics in Cartesian monomials in the form

$$S_{lm}(\boldsymbol{r}_A) = \sum_{i+j+k=l} S_{ijk}^{lm} x_A^i y_A^j z_A^k \tag{10}$$

where all terms vanish except those for which $i + j + k = l$. An explicit expression for this expansion is given in ref. 8 and 9 but is not needed for the present development.

In the calculation of one- and two-electron molecular integrals over solid-harmonic Gaussians eqn (1), the integration is commonly performed over Cartesian Gaussians

$$G_{ijk}(\boldsymbol{r}, a, \boldsymbol{A}) = x_A^i y_A^j z_A^k \exp(-ar_A^2), \tag{11}$$

with "quantum numbers" $i \geq 0, j \geq 0, k \geq 0$, using for example the Rys scheme,[1] the McMurchie–Davidson scheme,[2] or the Obara–Saika scheme[3] followed by a transformation to solid-harmonic form

$$G_{lm}(\boldsymbol{r}, a, \boldsymbol{A}) = \sum_{i+j+k=l} S_{ijk}^{lm} G_{ijk}(\boldsymbol{r}, a, \boldsymbol{A}) \tag{12}$$

Here, we shall consider an alternative approach, where the integration is first performed over the Hermite Gaussians

$$H_{ijk}(\boldsymbol{r}, a, \boldsymbol{A}) = \frac{\partial^{i+j+k}\exp(-ar_A^2)}{(2a)^{i+j+k}\partial A_x^i \partial A_y^j \partial A_z^k} \tag{13}$$

followed by the transformation

$$G_{lm}(\boldsymbol{r}, a, \boldsymbol{A}) = \sum_{i+j+k=l} S_{ijk}^{lm} H_{ijk}(\boldsymbol{r}, a, \boldsymbol{A}) \tag{14}$$

using the same coefficients as in eqn (12). In Subsection 2.2, we shall demonstrate that the resulting functions eqn (14) are in fact identical to the standard solid-harmonic Gaussians eqn (12). Consequently, we may choose for our integration those functions that are best suited to the task: the Cartesian Gaussians eqn (11) or the Hermite Gaussians eqn (13).

Let us briefly compare the properties of the Cartesian and Hermite Gaussians eqns (11) and (13). Both functions may be factorized in the Cartesian directions

$$G_{ijk}(\boldsymbol{r}, a, \boldsymbol{A}) = G_i(x, a, A_x)\, G_j(y, a, A_y)\, G_k(z, a, A_z) \tag{15}$$

$$H_{ijk}(\boldsymbol{r}, a, \boldsymbol{A}) = H_i(x, a, A_x)\, H_j(y, a, A_y)\, H_k(z, a, A_z) \tag{16}$$

where the x components are given by

$$G_i(x, a, A_x) = x_A^i \exp(-ax_A^2) \tag{17}$$

$$H_i(x, a, A_x) = \frac{\mathrm{d}^i \exp(-ax_A^2)}{(2a)^i \mathrm{d}A_x^i} \tag{18}$$

and likewise for the other components. We also note that the Cartesian Gaussians satisfy the recurrence relations

$$x_A G_i(x, a, A_x) = G_{i+1}(x, a, A_x) \tag{19}$$

$$\frac{\mathrm{d}G_i(x, a, A_x)}{\mathrm{d}A_x} = 2aG_{i+1}(x, a, A_x) - iG_{i-1}(x, a, A_x) \tag{20}$$

whereas the corresponding relations for the Hermite Gaussians are given by

$$x_A H_i(x, a, A_x) = H_{i+1}(x, a, A_x) + \frac{i}{2a} H_{i-1}(x, a, A_x) \tag{21}$$

$$\frac{\mathrm{d}H_i(x, a, A_x)}{\mathrm{d}A_x} = 2aH_{i+1}(x, a, A_x) \tag{22}$$

where eqn (21) follows from $[(\mathrm{d}/\mathrm{d}A_x)^i, x_A] = -i(\mathrm{d}/\mathrm{d}x_A)^{i-1}$. Note that the first (incremented) term is the same in the Cartesian and Hermite recurrence relations eqns (19)–(22). The only difference occurs in the second (decremented) term, which vanishes upon multiplication by x_A in the Cartesian Gaussian and upon differentiation by A_x in the Hermite Gaussian.

Because of eqn (22), the Hermite Gaussians are particularly well suited to integration over differentiated solid-harmonic functions since differentiation of eqn (14) merely raises the

quantum numbers of the Hermite Gaussians:

$$\frac{\partial^{I+J+K} G_{lm}(\boldsymbol{r}, a, \boldsymbol{A})}{\partial A_x^I \partial A_y^J \partial A_z^K} = (2a)^{I+J+K} \sum_{i+j+k=l} S_{ijk}^{lm} H_{i+I,j+J,k+K}(\boldsymbol{r}, a, \boldsymbol{A}) \tag{23}$$

whereas differentiation of the Cartesian Gaussians produces linear combinations of Gaussians eqn (20). Thus, provided the solid-harmonic Gaussians are expanded in Hermite Gaussians, there is no need for a special derivative code, to any order in differentiation. We shall also see that Hermite Gaussians are better suited than Cartesian Gaussians to integration over charge distributions consisting of linear combinations of single Gaussians rather than product Gaussians, as occur, for example, in the evaluation of Coulomb energies by density fitting.[10, 11] Special methods for two- and three-center Coulomb integrals have previously been considered by Köster[6] (for Hermite functions) and by Ahlrichs.[12]

2.2 Solid-harmonic Gaussians expanded in Hermite Gaussians

In Subsection 2.1, we expressed the solid-harmonic function $S_{lm}(\boldsymbol{r}_A)$ as a homogeneous polynomial of degree l in x_A, y_A, and z_A, see eqn (10). We shall now establish the equivalence of the real solid-harmonic functions expressed as linear combinations of either Hermite or Cartesian Gaussians. For this purpose, consider the transformed solid harmonics

$$\mathscr{S}_{lm}(\mathbf{r}_A) = \sum_{i+j+k=l} S_{ijk}^{lm} \mathscr{H}_i(x_A) \mathscr{H}_j(y_A) \mathscr{H}_k(z_A) \tag{24}$$

where the coefficients S_{ijk}^{lm} are the same as in eqn (10) but where the monomial x_A^i of degree i has been replaced by the Hermite polynomial of the same degree:

$$\mathscr{H}_i(x_A) = (-2)^{-i} \exp(x_A^2) \frac{\mathrm{d}^i}{\mathrm{d}x_A^i} \exp(-x_A^2) \tag{25}$$

and likewise for y_A^j and z_A^k. As seen by induction on the Rodrigues expression eqn (25), the Hermite polynomials may be generated recursively as

$$\mathscr{H}_{i+1}(x_A) = x_A \mathscr{H}_i(x_A) - \frac{i}{2} \mathscr{H}_{i-1}(x_A) \tag{26}$$

beginning with $\mathscr{H}_0(x_A) = 1$. Note that we have normalized the Hermite polynomials such that the leading term is equal to x_A^i: 1, x_A, $x_A^2 - \frac{1}{2}$, $x_A^3 - \frac{3}{2}x_A$, and so on.

Clearly, the transformed solid harmonics eqn (24) have the same leading terms as the standard solid harmonics eqn (10). Therefore, since the standard solid harmonics are homogeneous eqn (9) (which means that all terms are leading terms), the equivalence of the standard and transformed solid harmonics eqns (10) and (24) is established if we can show that the transformed solid harmonics eqn (24) are also homogeneous. For this purpose, we introduce the differential operator

$$D_x = \exp\left(-\frac{1}{4}\frac{\mathrm{d}^2}{\mathrm{d}x_A^2}\right) = \sum_{k=0}^{\infty} \frac{1}{(-4)^k k!} \frac{\mathrm{d}^{2k}}{\mathrm{d}x_A^{2k}} \tag{27}$$

which commutes with $\mathrm{d}/\mathrm{d}x_A$ and satisfies the commutator relation

$$[x_A, D_x] = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x_A} D_x \tag{28}$$

Using this relation, we find that

$$\begin{aligned} D_x x_A^{i+1} &= x_A D_x x_A^i - [x_A, D_x] x_A^i \\ &= x_A D_x x_A^i - \frac{i}{2} D_x x_A^{i-1} \end{aligned} \tag{29}$$

Comparing with eqn (26) and noting that $D_x 1 = 1$, we conclude that the operator D_x, when applied to x_A^i, generates the Hermite polynomial $\mathscr{H}_i(x_A)$ of eqn (25):

$$\mathscr{H}_i(x_A) = D_x x_A^i \tag{30}$$

Next, introducing this relation and the corresponding relations for y_A^j and z_A^k in eqn (24), we find that the transformed and standard solid harmonics are related in the following manner:

$$\begin{aligned} \mathscr{S}_{lm}(\boldsymbol{r}_A) &= \sum_{i+j+k=l} S_{ijk}^{lm} D_x x_A^i D_y y_A^j D_z z_A^k \\ &= D_x D_y D_z S_{lm}(\boldsymbol{r}_A) \end{aligned} \tag{31}$$

Finally, applying $\boldsymbol{r}_A \cdot \boldsymbol{\nabla}$ to the transformed solid harmonics eqn (31), we obtain

$$\begin{aligned} &(\boldsymbol{r}_A \cdot \nabla) \mathscr{S}_{lm}(\boldsymbol{r}_A) = (\boldsymbol{r}_A \cdot \nabla) D_x D_y D_z S_{lm}(\boldsymbol{r}_A) \\ &= D_x D_y D_z (\boldsymbol{r}_A \cdot \nabla) S_{lm}(\boldsymbol{r}_A) + [\boldsymbol{r}_A \cdot \nabla, D_x D_y D_z] S_{lm}(\boldsymbol{r}_A) \\ &= D_x D_y D_z l S_{lm}(\boldsymbol{r}_A) + \tfrac{1}{2} D_x D_y D_z \nabla^2 S_{lm}(\boldsymbol{r}_A) \\ &= l D_x D_y D_z S_{lm}(\boldsymbol{r}_A) = l \mathscr{S}_{lm}(\boldsymbol{r}_A) \end{aligned} \tag{32}$$

where we have used the homogeneity of the solid harmonics eqn (8), the commutator relation eqn (28), and Laplace's equation eqn (2), thereby demonstrating that the $S_{lm}(\boldsymbol{r}_A)$ are homogeneous polynomials of degree l. Since the standard and transformed solid harmonics are homogeneous polynomials with the same leading terms, they must be identical. We may therefore write the solid harmonics in the form

$$S_{lm}(\boldsymbol{r}_A) = \sum_{i+j+k=l} S_{ijk}^{lm} \mathscr{H}_i(x_A) \mathscr{H}_j(y_A) \mathscr{H}_k(z_A) \tag{33}$$

which differs from the standard expression eqn (10) in the replacement of x_A^i, y_A^j, and z_A^k by $\mathscr{H}_i(x_A)$, $\mathscr{H}_j(y_A)$, and $\mathscr{H}_k(z_A)$, respectively.

Combining eqns (1) and (33), we may now express the solid-harmonic Gaussians as

$$\begin{aligned} G_{lm}(\boldsymbol{r}, a, \boldsymbol{A}) =& a^{-l/2} \sum_{i+j+k=l} S_{ijk}^{lm} \mathscr{H}_i(\sqrt{a} x_A) \mathscr{H}_j(\sqrt{a} y_A) \mathscr{H}_k(\sqrt{a} z_A) \\ &\times \exp(-a r_A^2) \end{aligned} \tag{34}$$

where the homogeneity of the solid harmonics ensures that the coordinate scaling is canceled by the prefactor $a^{-l/2} = a^{(-i-j-k)/2}$. Substituting x_A by $\sqrt{a} x_A$ in eqn (25) and multiplying the resulting equation from the left by $\exp(-a x_A^2)$, we obtain:[2]

$$\mathscr{H}_i(\sqrt{a} x_A) \exp(-a x_A^2) = (2\sqrt{a})^{-i} \left(\frac{\mathrm{d}}{\mathrm{d}A_x}\right)^i \exp(-a x_A^2) \tag{35}$$

Using this result in eqn (34), we arrive at eqn (14), where we have introduced the Hermite Gaussians eqn (13). Thus, we may globally replace the Cartesian Gaussians eqn (11) by the Hermite Gaussians eqn (13) in the expansion of the solid harmonics eqn (12).

3. Molecular integrals over solid-harmonic Gaussians

In this section, we demonstrate how integrals over solid-harmonic Gaussians can be expanded in integrals over Hermite Gaussians, expressed as scaled derivatives of integrals over spherical Gaussians. The evaluation of these Hermite integrals is discussed in Section 4.

3.1 Overlap and multipole-moment integrals

Consider the multipole-moment integrals about $\boldsymbol{M}$ between two solid-harmonic Gaussians at $\boldsymbol{A}$ and $\boldsymbol{B}$ expanded in Hermite Gaussians eqn (14):

$$\begin{aligned} M_{ab}^{\boldsymbol{k}} &= \int G_{l_a m_a}(\boldsymbol{r}, a, \boldsymbol{A})\, G_{l_b m_b}(\boldsymbol{r}, b, \boldsymbol{B})\, x_M^{k_x} y_M^{k_y} z_M^{k_z}\, \mathrm{d}\boldsymbol{r} \\ &= \sum_{\boldsymbol{ij}} S_{\boldsymbol{i}}^{l_a m_a} S_{\boldsymbol{j}}^{l_b m_b} m_{\boldsymbol{ijk}}^{ab} \end{aligned} \tag{36}$$

An important special case is the overlap integral $S_{ab} = M_{ab}^{\boldsymbol{0}}$. The Hermite multipole-moment integrals $m_{\boldsymbol{ijk}}^{ab}$ are given by

$$\begin{aligned} m_{\boldsymbol{ijk}}^{ab} &= \int H_{\boldsymbol{i}}(\boldsymbol{r}, a, \boldsymbol{A})\, H_{\boldsymbol{j}}(\boldsymbol{r}, b, \boldsymbol{B})\, x_M^{k_x} y_M^{k_y} z_M^{k_z}\, \mathrm{d}\boldsymbol{r} \\ &= \frac{\partial^{i+j} \int \exp(-a r_A^2)\, \exp(-b r_B^2)\, x_M^{k_x} y_M^{k_y} z_M^{k_z}\, \mathrm{d}\boldsymbol{r}}{(2a\partial A_x)^{i_x} \cdots (2b\partial B_z)^{j_z}} \end{aligned} \tag{37}$$

We have here inserted the Hermite Gaussians eqn (13) and used the Leibniz integral rule to take the differential operators outside the integration sign, noting that the integration limits are independent of the Gaussian coordinates.[2,7] For brevity, we have introduced the notation $\boldsymbol{i} = (i_x, i_y, i_z)$ and $i = i_x + i_y + i_z$ (and likewise for $\boldsymbol{j}$ and $\boldsymbol{k}$); we also adopt the convention of denoting integrals over solid-harmonic Gaussians eqn (36) by uppercase letters $M_{ab}^{\boldsymbol{k}}$ and the corresponding Hermite integrals eqn (37) by lowercase letters $m_{\boldsymbol{ijk}}^{ab}$. Invoking the Gaussian product rule[13]

$$\exp(-a r_A^2)\, \exp(-b r_B^2) = \exp(-\mu R_{AB}^2)\, \exp(-p r_P^2) \tag{38}$$

with

$$\begin{aligned} p &= a + b, \quad \mu = \frac{ab}{p}, \quad \boldsymbol{R}_{AB} = \boldsymbol{A} - \boldsymbol{B}, \\ \boldsymbol{P} &= \frac{a\boldsymbol{A} + b\boldsymbol{B}}{p} \end{aligned} \tag{39}$$

we find that the integral over the product of spherical Gaussians eqn (38) is given by

$$\int \exp(-a r_A^2)\, \exp(-b r_B^2)\, x_M^{k_x} y_M^{k_y} z_M^{k_z}\, \mathrm{d}\boldsymbol{r} = \exp(-\mu R_{AB}^2)\, M_{\boldsymbol{k}}(p, \boldsymbol{R}_{PM}) \tag{40}$$

We have here introduced the multipole-moment integral

$$M_{\boldsymbol{k}}(p, \boldsymbol{R}_{PM}) = \int x_M^{k_x} y_M^{k_y} z_M^{k_z} \exp(-p r_P^2)\, \mathrm{d}\boldsymbol{r} \tag{41}$$

which depends on p and $\boldsymbol{R}_{PM} = \boldsymbol{P} - \boldsymbol{M}$, with the special value $M_{\boldsymbol{0}}(p, \boldsymbol{R}_{PM}) = (\pi/p)^{3/2}$. Inserting this result into eqn (37), we obtain the following Hermite multipole-moment integral

$$m_{\boldsymbol{ijk}}^{ab} = \frac{\partial^{i+j} \exp(-\mu R_{AB}^2)\, M_{\boldsymbol{k}}(p, \boldsymbol{R}_{PM})}{(2a\partial A_x)^{i_x} \cdots (2b\partial B_z)^{j_z}} \tag{42}$$

whose recursive evaluation is discussed in Section 4. However, we note here that the overlap integral may be expressed as a scaled Hermite Gaussian eqn (13) in $\boldsymbol{R}_{AB}$ with exponent μ:[7]

$$s_{\boldsymbol{ij}}^{ab} = m_{\boldsymbol{ij0}}^{ab} = (-1)^i \left(\frac{\pi}{p}\right)^{3/2} \left(\frac{b}{p}\right)^i \left(\frac{a}{p}\right)^j H_{\boldsymbol{i}+\boldsymbol{j}}(\boldsymbol{R}_{AB}, \mu, \boldsymbol{0}) \tag{43}$$

Since odd-order Hermite Gaussians vanish at the origin, the overlap integrals vanish for odd $i_x + j_x$ if $A_x = B_x$ (and likewise for the y and z directions).

The integrals discussed above were evaluated over a two-center overlap distribution, generated by a product of two Gaussians. Sometimes, we are interested in integrals over one-center overlap distributions—in particular, in density-fitting methods. Let us therefore consider the one-center overlap integrals

$$S_p = \int G_{l_p m_p}(\boldsymbol{r}, p, \boldsymbol{P})\, \mathrm{d}\boldsymbol{r} = \sum_{\boldsymbol{t}} S_{\boldsymbol{t}}^{l_p m_p} s_{\boldsymbol{t}}^{p} \tag{44}$$

where, by convention, we use p and $\boldsymbol{P}$ for one-center overlap distributions, with the Hermite quantum numbers $\boldsymbol{t} = (t_x, t_y, t_z)$. Proceeding as for two-center overlap distributions eqn (37), we find that

$$s_{\boldsymbol{t}}^{p} = \left(\frac{\pi}{p}\right)^{3/2} \delta_{t0} \tag{45}$$

where $t = t_x + t_y + t_z$, in agreement with the fact that integration over a single solid-harmonic Gaussian gives zero except in the totally symmetric case $l_p = m_p = 0$.

3.2 Integrals over differential operators

In the Hermite scheme, one-electron integrals over differential operators are easily obtained from the overlap integrals:

$$\begin{aligned} D_{ab}^{\boldsymbol{k}} &= \int G_{l_a m_a}(\boldsymbol{r}, a, \boldsymbol{A}) \left(\tfrac{\partial^{k_x}}{\partial x^{k_x}}\right) \left(\tfrac{\partial^{k_y}}{\partial y^{k_y}}\right) \left(\tfrac{\partial^{k_z}}{\partial z^{k_z}}\right) G_{l_b m_b}(\boldsymbol{r}, b, \boldsymbol{B})\, \mathrm{d}\boldsymbol{r} \\ &= \sum_{\boldsymbol{ij}} S_{\boldsymbol{i}}^{l_a m_a} S_{\boldsymbol{j}}^{l_b m_b} d_{\boldsymbol{ijk}}^{ab} \end{aligned} \tag{46}$$

where the Hermite integrals may be calculated in a variety of equivalent ways, such as

$$d_{\boldsymbol{ijk}}^{ab} = (2a)^k s_{\boldsymbol{i}+\boldsymbol{k}\,\boldsymbol{j}}^{ab} = (-2b)^k s_{\boldsymbol{i}\,\boldsymbol{j}+\boldsymbol{k}}^{ab} \tag{47}$$

As an important special case, the kinetic-energy integral is given by

$$T_{ab} = -\tfrac{1}{2}(D_{ab}^{200} + D_{ab}^{020} + D_{ab}^{002}) \tag{48}$$

and is thus easily obtained from overlap integrals with incremented quantum numbers.

3.3 One-electron Coulomb integrals

For the one-electron Coulomb integrals, we follow the same approach as for the multipole-moment integrals in Subsection

3.1, expanding in Hermite integrals

$$V_{ab} = \iint \frac{G_{l_a m_a}(\boldsymbol{r}, a, \boldsymbol{A})\, G_{l_b m_b}(\boldsymbol{r}, b, \boldsymbol{B})}{r_C}\, \mathrm{d}\boldsymbol{r} = \sum_{\boldsymbol{ij}} S_{\boldsymbol{i}}^{l_a m_a} S_{\boldsymbol{j}}^{l_b m_b} v_{\boldsymbol{ij}}^{ab} \tag{49}$$

where the Leibniz rule gives

$$v_{\boldsymbol{ij}}^{ab} = \iint \frac{H_{\boldsymbol{i}}(\boldsymbol{r}, a, \boldsymbol{A}) H_{\boldsymbol{j}}(\boldsymbol{r}, b, \boldsymbol{B})}{r_C}\, \mathrm{d}\boldsymbol{r} = \frac{\partial^{i+j} \int \exp(-a r_A^2) \exp(-b r_B^2)\, r_C^{-1}\, \mathrm{d}\boldsymbol{r}}{(2a\partial A_x)^{i_x} \cdots (2b\partial B_z)^{j_z}} \tag{50}$$

Following Boys, we next invoke the Gaussian product rule eqn (38), obtaining

$$\int \frac{\exp(-a r_A^2) \exp(-b r_B^2)}{r_C}\, \mathrm{d}\boldsymbol{r} = \frac{2\pi}{p} \exp(-\mu R_{AB}^2)\, F_0(p R_{PC}^2) \tag{51}$$

where we have introduced the Boys function[13]

$$F_n(x) = \int_0^1 \exp(-x t^2) t^{2n}\, \mathrm{d}t \tag{52}$$

Inserting eqn (51) in eqn (50), we find that the two-center one-electron Coulomb integrals are scaled derivatives of $\exp(-\mu R_{AB}^2) F_0(p R_{PC}^2)$ with respect to $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$v_{\boldsymbol{ij}}^{ab} = \frac{2\pi}{p} \frac{\partial^{i+j} \exp(-\mu R_{AB}^2)\, F_0(p R_{PC}^2)}{(2a\partial A_x)^{i_x} \cdots (2b\partial B_z)^{j_z}} \tag{53}$$

Unlike the multipole-moment integral eqn (42), it cannot be factorized in the Cartesian directions. For the corresponding one-center Coulomb integrals, we obtain

$$V_p = \int \frac{G_{l_p m_p}(\boldsymbol{r}, p, \boldsymbol{P})}{r_C}\, \mathrm{d}\boldsymbol{r} = \sum_{\boldsymbol{t}} S_{\boldsymbol{t}}^{l_p m_p} v_{\boldsymbol{t}}^{p} \tag{54}$$

$$v_{\boldsymbol{t}}^{p} = \int \frac{H_{\boldsymbol{t}}(\boldsymbol{r}, p, \boldsymbol{P})}{r_C}\, \mathrm{d}\boldsymbol{r} = \frac{2\pi}{p} \frac{\partial^t F_0(p R_{PC}^2)}{(2p\partial P_x)^{t_x} \cdots (2p\partial P_z)^{t_z}} \tag{55}$$

Comparing with the two-center case eqn (53), we note the expected absence of the exponential $\exp(-\mu R_{AB}^2)$, greatly simplifying its evaluation, as discussed in Section 4.

3.4 Two-electron Coulomb integrals

The four-center two-electron repulsion integrals over solid-harmonic Gaussians

$$G_{abcd} = \iint \frac{G_{l_a m_a}(\boldsymbol{r}_1, a, \boldsymbol{A})\, G_{l_b m_b}(\boldsymbol{r}_1, b, \boldsymbol{B})\, G_{l_c m_c}(\boldsymbol{r}_2, c, \boldsymbol{C})\, G_{l_d m_d}(\boldsymbol{r}_2, d, \boldsymbol{D})}{r_{12}}\, \mathrm{d}\boldsymbol{r}_1\, \mathrm{d}\boldsymbol{r}_2 \tag{56}$$

and the corresponding two- and three-center integrals G_{pq}, $G_{ab,q}$ and $G_{p,cd}$ may be treated in the same way as the one-electron integrals in Subsection 3.3. Substituting the expansions of the solid-harmonic functions eqn (14) in Hermite Gaussians, we obtain

$$G_{abcd} = \sum_{\boldsymbol{ijkl}} S_{\boldsymbol{i}}^{l_a m_a} S_{\boldsymbol{j}}^{l_b m_b} S_{\boldsymbol{k}}^{l_c m_c} S_{\boldsymbol{l}}^{l_d m_d} g_{\boldsymbol{ijkl}}^{abcd} \tag{57}$$

$$G_{ab,q} = \sum_{\boldsymbol{iju}} S_{\boldsymbol{i}}^{l_a m_a} S_{\boldsymbol{j}}^{l_b m_b} S_{\boldsymbol{u}}^{l_q m_q} g_{\boldsymbol{ij},\boldsymbol{u}}^{ab,q} \tag{58}$$

$$G_{p,cd} = \sum_{\boldsymbol{tkl}} S_{\boldsymbol{t}}^{l_p m_p} S_{\boldsymbol{k}}^{l_c m_c} S_{\boldsymbol{l}}^{l_d m_d} g_{\boldsymbol{t},\boldsymbol{kl}}^{p,cd} \tag{59}$$

$$G_{pq} = \sum_{\boldsymbol{tu}} S_{\boldsymbol{t}}^{l_p m_p} S_{\boldsymbol{u}}^{l_q m_q} g_{\boldsymbol{tu}}^{pq} \tag{60}$$

where the Hermite integrals are denoted by $g_{\boldsymbol{ijkl}}^{abcd}$, $g_{\boldsymbol{ij},\boldsymbol{u}}^{ab,q}$, $g_{\boldsymbol{t},\boldsymbol{kl}}^{p,cd}$, and $g_{\boldsymbol{tu}}^{pq}$. As for the one-electron integrals, we substitute the Hermite Gaussians eqn (13) in the Hermite integrals, invoke the Leibniz rule and apply the Gaussian product rule eqn (38), introducing the exponents and coordinates eqn (39) for the first electron and

$$q = c + d, \quad \nu = \frac{cd}{q}, \quad \boldsymbol{R}_{CD} = \boldsymbol{C} - \boldsymbol{D}, \quad \boldsymbol{Q} = \frac{c\boldsymbol{C} + d\boldsymbol{D}}{q} \tag{61}$$

for the second one. Finally, using the result of Boys[13]

$$\iint \frac{\exp(-p r_{1P}^2) \exp(-q r_{2Q}^2)}{r_{12}}\, \mathrm{d}\boldsymbol{r}_1\, \mathrm{d}\boldsymbol{r}_2 = \frac{2\pi^{5/2}}{pq\sqrt{p+q}} F_0(\alpha R_{PQ}^2) \tag{62}$$

with

$$\alpha = \frac{pq}{p+q}, \quad \boldsymbol{R}_{PQ} = \boldsymbol{P} - \boldsymbol{Q} \tag{63}$$

we find that the two-electron Hermite integrals are given by

$$g_{\boldsymbol{ijkl}}^{abcd} = \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \frac{\partial^{i+j+k+l} \exp(-\mu R_{AB}^2) \exp(-\nu R_{CD}^2)\, F_0(\alpha R_{PQ}^2)}{(2a\partial A_x)^{i_x} \cdots (2d\partial D_z)^{l_z}} \tag{64}$$

$$g_{\boldsymbol{ij},\boldsymbol{u}}^{ab,q} = \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \frac{\partial^{i+j+u} \exp(-\mu R_{AB}^2)\, F_0(\alpha R_{PQ}^2)}{(2a\partial A_x)^{i_x} \cdots (2q\partial Q_z)^{u_z}} \tag{65}$$

$$g_{\boldsymbol{t},\boldsymbol{kl}}^{p,cd} = \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \frac{\partial^{t+k+l} \exp(-\nu R_{CD}^2)\, F_0(\alpha R_{PQ}^2)}{(2p\partial P_x)^{t_x} \cdots (2d\partial D_z)^{l_z}} \tag{66}$$

$$g_{\boldsymbol{tu}}^{pq} = \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \frac{\partial^{t+u} F_0(\alpha R_{PQ}^2)}{(2p\partial P_x)^{t_x} \cdots (2q\partial Q_z)^{u_z}} \tag{67}$$

to be compared with the multipole-moment integrals eqn (42) and the one-electron Coulomb integrals eqns (53) and (55). The two- and three-center two-electron integrals are easier to calculate than the four-center integrals, not only because there are fewer differentiations to be carried out but also since the functions to be differentiated are simpler. In Section 4, we shall consider the evaluation of the Hermite Coulomb integrals.

3.5 Differentiated molecular integrals

Let us consider the evaluation of integrals over the differentiated solid-harmonic Gaussians

$$G^{I_x I_y I_z}_{l_a m_a}(\boldsymbol{r}, a, \boldsymbol{A}) = \frac{\partial^{I_x+I_y+I_z} G_{l_a m_a}(\boldsymbol{r}, a, \boldsymbol{A})}{\partial A_x^{I_x}\, \partial A_y^{I_y}\, \partial A_z^{I_z}} \tag{68}$$

Substituting here the expansion of solid-harmonic Gaussians in Hermite Gaussians eqn (14) and noting that the expansion coefficients are independent of $\boldsymbol{A}$, we obtain

$$G^{\boldsymbol{I}}_{l_a m_a}(\boldsymbol{r}, a, \boldsymbol{A}) = (2a)^I \sum_{\boldsymbol{i}} S^{lm}_{\boldsymbol{i}}\, H_{\boldsymbol{i}+\boldsymbol{I}}(\boldsymbol{r}, a, \boldsymbol{A}) \tag{69}$$

where $\boldsymbol{I} = (I_x, I_y, I_z)$ and $I = I_x + I_y + I_z$. Letting $G^{\boldsymbol{IJKL}}_{abcd}$ denote the two-electron integral evaluated over such functions, we obtain

$$\begin{aligned} G^{\boldsymbol{IJKL}}_{abcd} =& (2a)^I (2b)^J (2c)^K (2d)^L \\ & \times \sum_{\boldsymbol{ijkl}} S^{l_a m_a}_{\boldsymbol{i}} S^{l_b m_b}_{\boldsymbol{j}} S^{l_c m_c}_{\boldsymbol{k}} S^{l_d m_d}_{\boldsymbol{l}} g^{abcd}_{\boldsymbol{i}+\boldsymbol{I},\boldsymbol{j}+\boldsymbol{J},\boldsymbol{k}+\boldsymbol{K},\boldsymbol{l}+\boldsymbol{L}} \end{aligned} \tag{70}$$

as a generalization of eqn (57). The same result applies to all other integrals evaluated in terms of Hermite Gaussians. Molecular integrals over differentiated solid-harmonic Gaussians are thus obtained in the same manner as undifferentiated integrals, by incrementing the quantum numbers of the Hermite Gaussians eqn (69). A code written for general angular momentum is therefore also a code for general geometrical derivatives.

Assume that we wish to calculate the nth-order Cartesian derivatives arising from of an orbital shell of angular momentum l_0 (which we may take to be the first of four orbital shells in a two-electron integral). There are $(n + 1)(n + 2)/2$ independent Cartesian derivatives of each of the $2l_0 + 1$ solid-harmonic Gaussians in this shell. In the Hermite scheme, we begin by calculating all integrals arising from this shell with the angular momentum increased from l_0 to $l_0 + n$. Next, the resulting $(l_0 + n + 1)(l_0 + n + 2)/2$ Hermite components are combined to differentiated solid harmonics, using eqn (69). By contrast, in the Cartesian scheme, we first calculate all integrals with angular momentum $\max(0, l_0 - n) \le l \le l_0 + n$ on the first orbital. The number of such components is proportional to $l_0^2 n + n^3/3$. Subsequently, these Cartesian integrals are transformed to the derivative solid-harmonic basis, each transformation of which is more expensive than that from the Hermite basis.[14] Clearly, in this case, it is advantageous to use Hermite rather than Cartesian integrals as intermediates, as their number depends quadratically rather than cubically on n.

The advantages of the Hermite scheme become less pronounced when all derivatives in a given range $0 \le n \le n_{\max}$ are needed—we then need Hermite integrals with $l_0 \le l \le l_0 + n_{\max}$, compared with $\max(0, l_0 - n_{\max}) \le l \le l_0 + n_{\max}$ in the Cartesian case. The Hermite scheme is still preferable, however, since the subsequent transformation to the derivative solid harmonics is simpler, the same number of Hermite Gaussians contributing to each solid-harmonic function, for all orders of differentiation n.

4. Evaluation of Hermite integrals

In Section 3, we expanded molecular integrals over solid-harmonic Gaussians in integrals over Hermite Gaussians, expressed as derivatives of a generating function—see eqn (42) for multipole-moment integrals, eqns (53) and (55) for one-electron Coulomb integrals, and eqns (64)–(67) for two-electron Coulomb integrals. In the present section, we consider the evaluation of these Hermite integrals, using the Obara–Saika scheme[3] in Subsection 4.1 and the McMurchie–Davidson scheme[2] in Subsection 4.2. Živković and Maksić have given expressions for integrals over Hermite functions, without the use of recurrence relations.[7]

4.1 The Obara–Saika scheme for Hermite integrals

In the Obara–Saika scheme, the integrals are calculated from recurrence relations between integrals of different Hermite quantum numbers. Consider first the multipole-moment integrals eqn (42), which may be factorized in the three Cartesian directions, yielding the x component:

$$m^{ab}_{ijk} = \frac{\partial^{i+j} \exp(-\mu X_{AB}^2)\, M_k(p, X_{PM})}{(2a\partial A_x)^i (2b\partial B_x)^j} \tag{71}$$

in the short-hand notation i, j and k for i_x, j_x and k_x, respectively. The function $M_k(p, X_{PC})$, which is the x factor of eqn (41), satisfies the relations

$$M_0(p, X_{PM}) = \sqrt{\frac{\pi}{p}} \tag{72}$$

$$M_{k+1}(p, X_{PM}) = X_{PM} M_k(p, X_{PM}) + \frac{k}{2p} M_{k-1}(p, X_{PM}) \tag{73}$$

$$\frac{\partial M_k(p, X_{PM})}{\partial P_x} = k M_{k-1}(p, X_{PM}) \tag{74}$$

Incrementing the three indices, we obtain after a little algebra

$$m^{ab}_{i+1,j,k} = X_{PA} m^{ab}_{ijk} + \frac{1}{2p}(i w_{ab} m^{ab}_{i-1,j,k} + j m^{ab}_{i,j-1,k} + k m^{ab}_{i,j,k-1}) \tag{75}$$

$$m^{ab}_{i,j+1,k} = X_{PB} m^{ab}_{ijk} + \frac{1}{2p}(i m^{ab}_{i-1,j,k} + j w_{ba} m^{ab}_{i,j-1,k} + k m^{ab}_{i,j,k-1}) \tag{76}$$

$$m^{ab}_{i,j,k+1} = X_{PM} m^{ab}_{ijk} + \frac{1}{2p}(i m^{ab}_{i-1,j,k} + j m^{ab}_{i,j-1,k} + k m^{ab}_{i,j,k-1}) \tag{77}$$

where we have introduced the factor

$$w_{ab} = -\frac{b}{a} \tag{78}$$

From these "vertical" recurrence relations (which increment the highest $i + j + k$), we may generate the full set of integrals, beginning with m^{ab}_{000}. For example, we may first generate the overlap integrals using the first two recurrences eqns (75) and (76), followed by the generation of the multipole-moment integrals from the overlap integrals using eqn (77). We also note the "horizontal" recurrence relation

$$2a m^{ab}_{i+1,j,k} + 2b m^{ab}_{i,j+1,k} = k m^{ab}_{i,j,k-1} \tag{79}$$

which follows from translational invariance and conserves the highest $i + j + k$.

The same approach can be applied to the Coulomb integrals. Because of the presence of the Boys function, the integrals can no longer be factorized into Cartesian factors, although the recurrence relations in the three Cartesian directions are still independent. In the following, we therefore consider only increments in the x direction. Introducing the auxiliary functions

$$\Theta^n_{ij} = \frac{2\pi}{p}\frac{\partial^{i+j}\exp(-\mu R^2_{AB})\ (-2p)^n F_n(pR^2_{PC})}{(2a\partial A_x)^i(2b\partial B_x)^j} \tag{80}$$

$$\Theta^n_{ijkl} = \frac{2\pi^{5/2}}{pq\sqrt{p+q}}\ \times \frac{\partial^{i+j+k+l}\exp(-\mu R^2_{AB})\exp(-\nu R^2_{CD})(-2\alpha)^n F_n(\alpha R^2_{PQ})}{(2a\partial A_x)^i(2b\partial B_x)^j(2c\partial C_x)^k(2d\partial D_x)^l} \tag{81}$$

and using $F_n'(x) = -F_{n+1}(x)$, we obtain the following vertical recurrence relations

$$\Theta^n_{i+1,j} = X_{PA}\Theta^n_{ij} + \frac{1}{2p}(iw_{ab}\Theta^n_{i-1,j} + j\Theta^n_{i,j-1}) + \frac{1}{2p}X_{PC}\Theta^{n+1}_{ij} + \frac{1}{4p^2}(i\Theta^{n+1}_{i-1,j} + j\Theta^{n+1}_{i,j-1}) \tag{82}$$

$$\Theta^n_{i+1,j,k,l} = X_{PA}\Theta^n_{ijkl} + \frac{1}{2p}(iw_{ab}\Theta^n_{i-1,j,k,l} + j\Theta^n_{i,j-1,k,l}) + \frac{1}{2p}X_{PQ}\Theta^{n+1}_{ijkl} + \frac{1}{4p^2}(i\Theta^{n+1}_{i-1,j,k,l} + j\Theta^{n+1}_{i,j-1,k,l}) - \frac{1}{4pq}(k\Theta^{n+1}_{i,j,k-1,l} + l\Theta^{n+1}_{i,j,k,l-1}) \tag{83}$$

and likewise for increments in j, k and l. Beginning from Θ^n_{00} with $0 \le n \le \max(i+j)$ or Θ^n_{0000} with $0 \le n \le \max(i+j+k+l)$, we may thus generate the full set of Hermite Coulomb integrals $\nu^{ab}_{i00,j00} = \Theta^0_{ij}$ and $g^{abcd}_{i00,j00,k00,l00} = \Theta^0_{ijkl}$. Note that the one-electron Coulomb recurrence relations eqn (82) are identical to the two-electron relations eqn (83) except for the replacement of X_{PQ} by X_{PC} and the absence of the last two terms; also, the first three terms in eqns (82) and (83) are the same as for the multipole-moment integral eqn (75).

The Obara–Saika recurrence relations for Hermite Coulomb integrals eqns (82) and (83) resemble closely those for Cartesian integrals,[3,8] to which they reduce if we arbitrarily set $w_{ab} = 1$. As in the Cartesian case, it may be advantageous to use additional recurrence relations. From eqn (21), we obtain the horizontal recurrence relations

$$\Theta^n_{i,j+1,k,l} = \Theta^n_{i+1,j,k,l} + X_{AB}\Theta^n_{ijkl} + \frac{i}{2a}\Theta^n_{i-1,j,k,l} - \frac{j}{2b}\Theta^n_{i,j-1,k,l} \tag{84}$$

which differ from the Cartesian recurrences[4] by the presence of the two last terms. Finally, from the translational invariance of the integrals, we obtain the electron-transfer relations

$$\Theta^n_{i,0,k+1,0} = -\frac{bX_{AB} + \mathrm{d}X_{CD}}{q}\Theta^n_{i0k0} + \frac{iw_{ab}}{2q}\Theta^n_{i-1,0,k,0} + \frac{kw_{cd}}{2q}\Theta^n_{i,0,k-1,0} - \frac{p}{q}\Theta^n_{i+1,0,k,0} \tag{85}$$

which differ from to the corresponding relations for Cartesian Gaussians[15,16] only in the presence of the w_{ab} and w_{cd} factors. Using these relations, we may simplify the evaluation of integrals by using a reduced Obara–Saika recurrence eqn (83) to generate all $\Theta^n_{i+j+k+l,0,0,0}$, followed by use of the electron-transfer relation eqn (85) to generate all $\Theta_{i+j,0,k+l,0}$. In the final step, we use the horizontal recurrences eqn (84) to generate the final integrals Θ_{ijkl}.

The Coulomb recurrence relations given above eqns (82) and (83) are for two-center overlap distributions. The corresponding relations for integrals with one-center distributions eqn (55) and eqns (65)–(67) are obtained by setting $a = b = p$, $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{P}$, and $i + j = t$ for the first electron and $c = d = q$, $\boldsymbol{C} = \boldsymbol{D} = \boldsymbol{Q}$, and $k + l = u$ for the second electron. For example, for two-center two-electron Coulomb integrals, we obtain from eqn (83) the recursion

$$\Theta^n_{t+1,u} = \frac{1}{2p}X_{PQ}\Theta^{n+1}_{tu} + \frac{t}{4p^2}\Theta^{n+1}_{t-1,u} - \frac{u}{4pq}\Theta^{n+1}_{t,u-1} \tag{86}$$

where the first three terms of eqn (83) vanish and the pairs of terms in parentheses collapse into single terms, reducing the total number of terms from eight to three. By further rewriting the intermediate integrals in the form

$$\Theta^n_{tu} = (2p)^{-t}(-2q)^{-u}\theta^n_{t+u} \tag{87}$$

$$\theta^n_t = (-2\alpha)^n\frac{2\pi^{5/2}}{pq\sqrt{p+q}}\frac{\partial^t F_n(\alpha R^2_{PQ})}{\partial P^t_x} \tag{88}$$

we obtain the even simpler McMurchie–Davidson recursion:[2]

$$\theta^n_{t+1} = X_{PQ}\theta^{n+1}_t + t\theta^{n+1}_{t-1} \tag{89}$$

and similarly for the y and z directions. We have thus reduced the Obara–Saika recurrence relations from eight to two terms. No such reduction is possible in the Cartesian basis.

Finally, we note that it is possible to set up a scheme for the Hermite integrals eqn (64), where we first calculate the derivatives of the three factors $F^{ab}_n = \partial^n\exp(-\mu R^2_{AB})/\partial A^n_x$, $F^{cd}_n = \partial^n\exp(-\nu R^2_{CD})/\partial C^n_x$, and $R^{pq}_n = \partial^n F_0(\alpha R^2_{PQ})/\partial P^n_x$, after which the Hermite integrals are assembled by binomial expansion, as done by Živković and Maksić.[7]

4.2 The McMurchie–Davidson scheme for Hermite integrals

In the Obara–Saika scheme, molecular integrals are generated recursively, using different sets of recurrence relations for one- and two-center overlap distributions, as discussed in Subsection 4.1. In the McMurchie–Davidson scheme, we take a different approach, expanding two-center overlap distributions in one-center overlap distributions. In this way, all four-center two-electron Coulomb integrals are reduced to two-center Coulomb integrals, which are evaluated using the same recurrence relations as in the Obara–Saika scheme.

Consider the overlap distribution generated by the Hermite Gaussians $H_i(\boldsymbol{r}_1, a, \boldsymbol{A})$ and $H_j(\boldsymbol{r}_1, b, \boldsymbol{B})$:

$$\Omega_{ij}(\boldsymbol{r}, a, \boldsymbol{A}, b, \boldsymbol{B}) = H_i(\boldsymbol{r}, a, \boldsymbol{A})\, H_j(\boldsymbol{r}, b, \boldsymbol{B}) \tag{90}$$

In the McMurchie–Davidson scheme, this two-center distribution is expanded in Hermite Gaussians about the product

center $\boldsymbol{P}$:

$$\Omega_{\boldsymbol{ij}}(\boldsymbol{r}, a, \boldsymbol{A}, b, \boldsymbol{B}) = \sum_{\boldsymbol{t}=\boldsymbol{0}}^{\boldsymbol{i+j}} F_{\boldsymbol{t}}^{\boldsymbol{ij}}(a, b, \boldsymbol{R}_{AB}) H_{\boldsymbol{t}}(\boldsymbol{r}, p, \boldsymbol{P}) \tag{91}$$

noting that the expansion coefficients depend only on the relative positions of the Gaussians. Substituting eqn (91) in eqn (37) and using eqn (45), we obtain for the overlap integrals

$$s_{\boldsymbol{ij}}^{ab} = \sum_{\boldsymbol{t}} F_{\boldsymbol{t}}^{\boldsymbol{ij}} s_{\boldsymbol{t}}^{p} = F_{\boldsymbol{0}}^{\boldsymbol{ij}} s_{\boldsymbol{0}}^{p} \tag{92}$$

For the two-center one-electron Coulomb integrals eqn (50), substitution of eqn (91) yields

$$v_{\boldsymbol{ij}}^{ab} = \sum_{\boldsymbol{t}} F_{\boldsymbol{t}}^{\boldsymbol{ij}} v_{\boldsymbol{t}}^{p} \tag{93}$$

where the one-center integral is given by eqn (55). Finally, for the two-electron Coulomb integrals of Subsection 3.4, eqn (91) and a similar expansion of $\Omega_{\boldsymbol{kl}}(\boldsymbol{r}, c, \boldsymbol{C}, d, \boldsymbol{D})$ yield

$$g_{\boldsymbol{ij},\boldsymbol{u}}^{ab,q} = \sum_{\boldsymbol{t}} F_{\boldsymbol{t}}^{\boldsymbol{ij}} g_{\boldsymbol{tu}}^{pq} \tag{94}$$

$$g_{\boldsymbol{t},\boldsymbol{kl}}^{p,cd} = \sum_{\boldsymbol{u}} g_{\boldsymbol{tu}}^{pq} F_{\boldsymbol{u}}^{\boldsymbol{kl}} \tag{95}$$

$$g_{\boldsymbol{ijkl}}^{abcd} = \sum_{\boldsymbol{tu}} F_{\boldsymbol{t}}^{\boldsymbol{ij}} g_{\boldsymbol{tu}}^{pq} F_{\boldsymbol{u}}^{\boldsymbol{kl}} \tag{96}$$

where the basic two-center integrals $g_{\boldsymbol{tu}}^{pq}$ are given by eqn (67). In the original Cartesian-based McMurchie–Davidson scheme,[2] also one-center overlap distributions are expanded in Hermite orbitals according to eqn (96). In the purely Hermite scheme presented here, only two-center distributions are expanded, greatly simplifying the evaluation of few-center integrals. The Hermite integrals $\nu_{\boldsymbol{t}}^{p}$ and $g_{\boldsymbol{tu}}^{pq}$ are evaluated using eqns (87)–(89).[2]

It only remains to discuss the evaluation of the expansion coefficients of eqn (91). Factorizing the expansion in the Cartesian directions and introducing the following short-hand notation for the x direction

$$\Omega_{ij} = \sum_{t} F_{t}^{ij} H_{t} \tag{97}$$

we obtain

$$2a\Omega_{i+1,j} + 2b\Omega_{i,j+1} = 2p \sum_{t} F_{t}^{ij} H_{t+1} \tag{98}$$

where we have applied $\partial/\partial A_x + \partial/\partial B_x = \partial/\partial P_x$ and then eqn (22) on both sides of the equation. Inserting eqn (97) on the left-hand side and collecting terms, we arrive at the horizontal recurrence relation

$$aF_{t}^{i+1,j} + bF_{t}^{i,j+1} = pF_{t-1}^{ij} \tag{99}$$

Next, we rewrite each term in eqn (98) using eqn (21), yielding

$$2px_P\Omega_{ij} - i\Omega_{i-1,j} - j\Omega_{i,j-1} = 2px_P \sum_{t} F_{t}^{ij} H_{t} - \sum_{t} F_{t}^{ij} t H_{t-1} \tag{100}$$

where we have used the relation $ax_A + ax_B = px_P$ on the left-hand side. Canceling the first term on each side and rearranging, we obtain the vertical recurrence relation

$$iF_{t}^{i-1,j} + jF_{t}^{i,j-1} = (t+1)F_{t+1}^{ij} \tag{101}$$

However, this relation can only be used to increment the upper indices ij for F_{t}^{ij} with $t > 0$. Different vertical recurrences are obtained by incrementing the first index of Ω_{ij}, yielding

$$\begin{aligned}\Omega_{i+1,j} &= x_A\Omega_{ij} - \frac{i}{2a}\Omega_{i-1,j} \\ &= X_{PA}\Omega_{ij} + \sum_{t} F_{t}^{ij} x_P H_{t} - \frac{i}{2a}\Omega_{i-1,j} \\ &= X_{PA}\Omega_{ij} + \sum_{t} F_{t}^{ij} H_{t+1} + \sum_{t} F_{t}^{ij} \frac{t}{2p} H_{t-1} - \frac{i}{2a}\Omega_{i-1,j}\end{aligned} \tag{102}$$

where we have first used eqn (21) for orbital a, then expanded $x_A = x_P + X_{PA}$, followed by the application of eqn (21) for orbital p. Collecting terms, we obtain

$$F_{t}^{i+1,j} = F_{t-1}^{ij} + X_{PA}F_{t}^{ij} + \frac{t+1}{2p}F_{t+1}^{ij} - \frac{i}{2a}F_{t}^{i-1,j} \tag{103}$$

A similar set of recurrence relations may be derived for increments in the second index j. However, the most efficient scheme for the evaluation of the full set of expansion coefficients is to use the following combination of eqns (99), (101) and (103):

$$F_{0}^{i+1,0} = X_{PA}F_{0}^{i0} + \frac{1}{2p}F_{1}^{i0} - \frac{i}{2a}F_{0}^{i-1,0} \tag{104}$$

$$F_{0}^{i,j+1} = -\frac{a}{b}F_{0}^{i+1,j} \tag{105}$$

$$F_{t}^{ij} = \frac{i}{t}F_{t-1}^{i-1,j} + \frac{j}{t}F_{t-1}^{i,j-1}, \quad t > 0 \tag{106}$$

beginning with $F_{0}^{00} = \exp(-\mu X_{AB}^2)$. The recurrence relations of the Hermite Gaussians are therefore no more complicated than those for the Cartesian Gaussians.[2,8] We note, however, that our definition of Hermite Gaussians eqn (13) differs from that of McMurchie and Davidson,[2] who use $\Lambda_{\boldsymbol{t}}(\boldsymbol{r}, p, \boldsymbol{P}) = (2p)^{t_x+t_y+t_z} H_{\boldsymbol{t}}(\boldsymbol{r}, p, \boldsymbol{P})$. We have therefore here denoted the expansion coefficients in eqn (91) by $F_{\boldsymbol{t}}^{\boldsymbol{ij}}$ rather than by $E_{\boldsymbol{t}}^{\boldsymbol{ij}}$ as in ref. 2.

5. Conclusions

We have presented a scheme for the evaluation of molecular integrals over solid-harmonic Gaussians, in which the integration is carried out over Hermite Gaussians rather than over Cartesian Gaussians, based on the observation that solid-harmonic Hermite Gaussians are identical to the corresponding solid-harmonic Cartesian Gaussians. The presented scheme simplifies the evaluation of derivative integrals (needed for energy derivatives and in relativistic theory) since differentiation with respect to nuclear coordinates merely increments the quantum numbers of the Hermite integrals, in the same way as when the angular momentum is increased. Consequently, the differentiation can be carried out to arbitrary order using the same code as for undifferentiated

integrals. Moreover, the presented Hermite-based scheme simplifies the evaluation of two- and three-center two-electron integrals, of importance in density-fitting schemes, bypassing the traditional time-consuming transformation to Cartesian basis.

Note added in proof

After this article had been accepted for publication, we became aware of the general theory of spherical tensor gradient operators, as reviewed by Weniger.[17] From this theory, the equivalence of the expansions of solid-harmonic Gaussians in Cartesian and Hermite Gaussians follows as a special result. For further details and references, we refer the reader to this work. However, it is proper here to mention the work of Dunlap[18–20] and of Ishida,[21] who have used spherical tensor gradient theory to develop integration techniques based on angular-momentum recoupling, bypassing the evaluation of integrals over Cartesian and Hermite Gaussians entirely. We would also like to draw attention to the work of Fortunelli and Salvetti,[22] where two-electron integrals over Hermite functions are considered within the Obara–Saika scheme.

Acknowledgements

This work has been supported by the Norwegian Research Council through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420) and by the European Research and Training Network NANOQUANT, Understanding Nanomaterials from the Quantum Perspective, contract No. MRTN-CT-2003-506842. We would like to thank Peter R. Taylor and Filip Pawłowski for comments.

References

1 M. Dupuis, J. Rys and H. F. King, *J. Chem. Phys*., 1976, **65**, 111.
2 L. E. McMurchie and E. R. Davidson, *J. Comput. Phys*., 1978, **26**, 218.
3 S. Obara and A. Saika, *J. Chem. Phys*., 1986, **84**, 3963.
4 M. Head-Gordon and J. A. Pople, *J. Chem. Phys*., 1988, **89**, 5777.
5 P. M. W. Gill, M. Head-Gordon and J. A. Pople, *J. Phys. Chem*., 1990, **94**, 5564.
6 A. M. Köster, *J. Chem. Phys*., 2003, **118**, 9943.
7 T. Živković and Z. B. Maksić, *J. Chem. Phys*., 1968, **49**, 3083.
8 T. Helgaker, P. Jørgensen and J. Olsen, *Molecular Electronic–Structure Theory*, Wiley, Chichester, 2000.
9 K. Ishida, *J. Chem. Phys*., 1998, **109**, 881.
10 J. L. Whitten, *J. Chem. Phys*., 1973, **58**, 4496.
11 B. I. Dunlap, J. W. D. Connolly and J. R. Sabin, *J. Chem. Phys*., 1979, **71**, 3396.
12 R. Ahlrichs, *Phys. Chem. Chem. Phys*., 2004, **6**, 5119.
13 S. F. Boys, *Proc. R. Soc. London, Ser. A*, 1950, **200**, 542.
14 K. Hald, A. Halkier, P. Jørgensen, S. Coriani, C. Hättig and T. Helgaker, *J. Chem. Phys*., 2003, **118**, 2985.
15 T. P. Hamilton and H. F. Schaefer III, *Chem. Phys*., 1991, **150**, 163.
16 R. Lindh, U. Ryu and B. Liu, *J. Chem. Phys*., 1991, **95**, 5889.
17 E. J. Weniger, *Collect. Czech. Chem. Commun*., 2005, **70**, 1225.
18 B. I. Dunlap, *Phys. Rev. A*, 1990, **42**, 1127.
19 B. I. Dunlap, *Phys. Rev. A*, 2002, **66**, 032502.
20 B. I. Dunlap, *J. Chem. Phys*., 2003, **118**, 1036.
21 K. Ishida, *J. Comput. Chem*., 2002, **23**, 378.
22 A. Fortunelli and O. Salvetti, *Int. J. Quantum Chem*., 1993, **48**, 257.

Paper II

Linear-scaling implementation of molecular electronic self-consistent field theory
P. Sałek, S. Høst, L. Thøgersen, P. Jørgensen, P. Manninen, J. Olsen, B. Jansik, **S. Reine**, F. Pawłowski, E. Tellgren, T. Helgaker and S. Coriani
The Journal of Chemical Physics, **126**, 114110 (2007)

Linear-scaling implementation of molecular electronic self-consistent field theory

Paweł Sałek
Department of Theoretical Chemistry, The Royal Institute of Technology, SE-10691 Stockholm, Sweden

Stinne Høst,[a] Lea Thøgersen, Poul Jørgensen, Pekka Manninen,[b]
Jeppe Olsen, and Branislav Jansík
The Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Aarhus, DK-8000 Århus C, Denmark

Simen Reine,[c] Filip Pawłowski,[d] Erik Tellgren,[e] and Trygve Helgaker[e]
Department of Chemistry, University of Oslo, P.O. Box 1033, Blindern N-0315, Norway

Sonia Coriani
Dipartimento di Scienze Chimiche, Università degli Studi di Trieste, Via Licio Giorgieri 1, I-34127 Trieste, Italy

(Received 4 October 2006; accepted 9 January 2007; published online 21 March 2007)

A linear-scaling implementation of Hartree-Fock and Kohn-Sham self-consistent field (SCF) theories is presented and illustrated with applications to molecules consisting of more than 1000 atoms. The diagonalization bottleneck of traditional SCF methods is avoided by carrying out a minimization of the Roothaan-Hall (RH) energy function and solving the Newton equations using the preconditioned conjugate-gradient (PCG) method. For rapid PCG convergence, the Löwdin orthogonal atomic orbital basis is used. The resulting linear-scaling trust-region Roothaan-Hall (LS-TRRH) method works by the introduction of a level-shift parameter in the RH Newton equations. A great advantage of the LS-TRRH method is that the optimal level shift can be determined at no extra cost, ensuring fast and robust convergence of both the SCF iterations and the level-shifted Newton equations. For density averaging, the authors use the trust-region density-subspace minimization (TRDSM) method, which, unlike the traditional direct inversion in the iterative subspace (DIIS) scheme, is firmly based on the principle of energy minimization. When combined with a linear-scaling evaluation of the Fock/Kohn-Sham matrix (including a boxed fitting of the electron density), LS-TRRH and TRDSM methods constitute the linear-scaling trust-region SCF (LS-TRSCF) method. The LS-TRSCF method compares favorably with the traditional SCF/DIIS scheme, converging smoothly and reliably in cases where the latter method fails. In one case where the LS-TRSCF method converges smoothly to a minimum, the SCF/DIIS method converges to a saddle point. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2464111]

I. INTRODUCTION

During the last decade, much effort has been directed towards the development and implementation of Hartree-Fock (HF) and Kohn-Sham (KS) self-consistent field (SCF) theories in such a manner that, for sufficiently large systems, the cost of the calculations scales linearly with system size $O(N)$, where N may be taken as the number of atoms in the molecule. To achieve linear scaling, two bottlenecks must be overcome: first, the construction of the Fock/KS matrix $\mathbf{F}$ in the atomic-orbital (AO) basis, which conventionally scales as $O(N^2)$; second, the generation of a new density matrix from a given Fock/KS matrix, which is conventionally achieved by an $O(N^3)$ diagonalization step $\mathbf{FC}=\mathbf{SC}\boldsymbol{\epsilon}$, where $\mathbf{S}$ is the AO overlap matrix. Over the years, many strategies have been proposed to make the cost of these key steps scale linearly with system size.

To remove the diagonalization bottleneck, many methods have been suggested—see Refs. 1 and 2 for an overview. We focus here on the density-matrix methods,[1] which may be subdivided into two categories:[2] the Fermi-operator expansion (FOE) methods and the density-matrix minimization (DMM) methods. The FOE methods include rational-function, polynomial, or recursive-polynomial expansions to compute the density matrix, of which the canonical-purification method of Palser and Manolopoulos,[3] the purification[4,5] of McWeeny and the Chebyshev expansion[6,7] of Baer and Head-Gordon serve as examples. Alternatively, the DMM methods use the fact that the density matrix obtained from a Fock/KS matrix diagonalization represents the

[a]Author to whom correspondence should be addressed. Fax: +45 8619 6199. Electronic mail: stinne@chem.au.dk
[b]Present address: Helsinki University of Technology, P.O. Box 1100 (Otakaari 1 M), FI-02015 Hut, Finland.
[c]Present address: The Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Aarhus, DK-8000 Århus C, Denmark.
[d]Present address: Institute of Physics, Kazimierz Wielki University, Plac Weyssenhoffa 11, 85-072 Bydgoszcz, Poland.
[e]Present address: Department of Chemistry, University of Durham, South Road, Durham DH1 3LE, United Kingdom.

global minimum of the Roothaan-Hall (RH) energy function $E^{\mathrm{RH}}=\mathrm{Tr}\,\mathbf{DF}$ (with $\mathbf{F}$ fixed),[8,9] thereby replacing the diagonalization by a minimization, suitably constrained so as to satisfy the idempotency condition $\mathbf{DSD}=\mathbf{D}$. Li *et al.* proposed to deal with the idempotency constraint by replacing the density matrix in the optimization by its McWeeny-purified counterpart, noting that the variations are then idempotent to first order;[10] their approach was further developed by Millam and Scuseria[11] and by Challacombe.[12] Alternatively, the idempotency condition may be incorporated into the parametrization of the density matrix $\mathbf{D}(\mathbf{X})=\exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX})$ with $\mathbf{X}$ antisymmetric, as described by Helgaker and co-workers.[8,13,14] The first attempts to use this parametrization to minimize E^{RH} employed a sequence of Newton iterations but encountered difficulties in the solution of the Newton linear equations.[14] Subsequently, these difficulties were solved by Shao *et al.*[15] in their curvy-step method by transforming the Newton equations to the Cholesky basis, where the Hessian has a smaller condition number and is more diagonally dominant than in the AO basis. We discuss here improvements to the algorithm of Shao *et al.*[15] using the Löwdin or principal square-root basis rather than the Cholesky basis for the following reasons. First, the convergence of the Newton equations is marginally better in the Löwdin basis than in the Cholesky basis; second, the Löwdin basis is the orthogonal AO basis that resembles most closely the original AO basis,[16] ensuring that locality is preserved to the greatest possible extent; and third, the transformation to the Löwdin basis can be performed straightforwardly within a linear-scaling framework.[17]

In setting up the SCF iterations, we note that the RH energy $E^{\mathrm{RH}}(\mathbf{X})=\mathrm{Tr}\,\mathbf{D}(\mathbf{X})\mathbf{F}$ constitutes a rather crude model of the true SCF energy E_{SCF}. In particular, at the expansion point $\mathbf{X}=\mathbf{0}$, E^{RH} has the same gradient as E_{SCF} but only an approximate Hessian. A global minimization of E^{RH} (as traditionally accomplished by diagonalization of the Fock/KS matrix) may therefore lead to steps that are too long and therefore unreliable. To avoid such problems, we impose in the trust-region RH (TRRH) method the condition that steps should not be taken outside the trust region, that is, outside the region where E^{RH} is a good approximation to E_{SCF}. We have previously implemented the TRRH method in conjunction with the diagonalization of the Fock/KS matrix.[18,19] In the present paper, we describe how the TRRH method may be implemented without diagonalization, making it suitable for linear-scaling SCF calculations. We denote the obtained algorithm the linear-scaling TRRH (LS-TRRH) method.

The information in the density and Fock/KS matrices (gradients) $\mathbf{D}_i$ and $\mathbf{F}_i$ that have been generated during an SCF optimization may be used to accelerate the convergence of the SCF iterations. Traditionally, this is accomplished by Pulay's method of direct inversion in the iterative subspace (DIIS),[20] where an improved density matrix is obtained in the subspace of the previous density matrices by minimizing the norm of the gradient. As an alternative to DIIS, we recently introduced the trust-region density-subspace minimization (TRDSM) algorithm,[18,19] where a local energy model E^{DSM} is set up in the subspace of the previous density matrices $\mathbf{D}_i$. Disregarding the idempotency conditions, the TRDSM algorithm reduces to the energy-DIIS (EDIIS) algorithm of Kudin *et al.*[21] A disadvantage of the EDIIS algorithm is that, even at the expansion point, the EDIIS gradient is not equal to the SCF gradient. By contrast, the E^{DSM} energy of the TRDSM algorithm constitutes an accurate representation of the true energy E_{SCF} in the subspace of previous density matrices $\mathbf{D}_i$; consequently, a trust-region optimization may be safely performed on E^{DSM} to obtain the improved density matrix.

Combining the LS-TRRH and TRDSM algorithms, we obtain the linear-scaling trust-region SCF (LS-TRSCF) method. In the LS-TRSCF calculations, sparse-matrix algebra is used both in the LS-TRRH part and in the TRDSM part of the optimization to achieve linear scaling. Sample calculations are reported on polyalanine peptides containing up to 119 alanine residues to demonstrate linear scaling. The LS-TRSCF convergence is also examined and compared with the convergence of conventional SCF/DIIS calculations, that is, diagonalization without level shifting, improved by the DIIS algorithm. The calculations demonstrate that the LS-TRSCF algorithm constitutes an efficient and robust algorithm for optimizing SCF wave functions.

For the Fock/KS matrix evaluation to scale linearly, a number of techniques have been introduced for the different contributions to $\mathbf{F}$: the fast multipole method (FMM) for the Coulomb contribution;[22–26] the order-N exchange method and the linear exchange K (LinK) method for the exact HF exchange contribution,[27–32] and efficient numerical-quadrature methods for the exchange-correlation (XC) contribution.[33–35] Our SCF code uses FMM combined with boxed density fitting for the Coulomb contribution, LinK for the exact-exchange contribution, and linear-scaling numerical quadrature for the XC contribution.

The remainder of this paper contains three sections. We begin by discussing the optimization of the RH energy in Sec. II. Section III contains some illustrative calculations, whereas Sec. IV contains conclusions.

II. OPTIMIZATION OF THE ROOTHAAN-HALL ENERGY

A. Parametrization of the density matrix

Let $\mathbf{D}$ be a valid KS density matrix of an N-electron system, which together with the AO overlap matrix $\mathbf{S}$ satisfies the following relations:

$$\mathbf{D}^{\mathrm{T}}=\mathbf{D}, \tag{1}$$

$$\mathrm{Tr}\,\mathbf{DS}=N, \tag{2}$$

$$\mathbf{DSD}=\mathbf{D}. \tag{3}$$

Introducing the projectors $\mathbf{P}_{\mathrm{o}}$ and $\mathbf{P}_{\mathrm{v}}$ onto the occupied and virtual orbital spaces

$$\mathbf{P}_{\mathrm{o}}=\mathbf{DS}, \tag{4}$$

$$\mathbf{P}_{\mathrm{v}}=\mathbf{I}-\mathbf{DS}, \tag{5}$$

we may, from the reference density matrix $\mathbf{D}$, generate any other valid density matrix by the transformation[8]

$$\mathbf{D}(\mathbf{X}) = \exp[-\mathcal{P}(\mathbf{X})\mathbf{S}]\mathbf{D}\exp[\mathbf{S}\mathcal{P}(\mathbf{X})], \tag{6}$$

where $\mathbf{X}$ is an antisymmetric matrix and where we have introduced the notation

$$\mathcal{P}(\mathbf{X}) = \mathbf{P}_o\mathbf{X}\mathbf{P}_v^T + \mathbf{P}_v\mathbf{X}\mathbf{P}_o^T. \tag{7}$$

The matrix exponential is evaluated as

$$\exp(\mathbf{XS}) = \sum_{n=0}^{\infty} \frac{(\mathbf{XS})^n}{n!}. \tag{8}$$

In an orthonormalized AO basis, such as will be discussed in Sec. II E, simplifications and a typically faster convergence follow from the fact that $\mathbf{S}=\mathbf{I}$.

The density matrix $\mathbf{D}(\mathbf{X})$ may be expanded in orders of $\mathbf{X}$ as

$$\mathbf{D}(\mathbf{X}) = \mathbf{D} + [\mathbf{D},\mathcal{P}(\mathbf{X})]_S + \tfrac{1}{2}[[\mathbf{D},\mathcal{P}(\mathbf{X})]_S,\mathcal{P}(\mathbf{X})]_S + \cdots, \tag{9}$$

where we have introduced the S commutator

$$[\mathbf{A},\mathbf{B}]_S = \mathbf{ASB} - \mathbf{BSA}. \tag{10}$$

We shall here in particular be concerned with expansions of the type $\mathrm{Tr}[\mathbf{MD}(\mathbf{X})]$, where $\mathbf{M}$ is symmetric. Inserting the S commutator expansion of the density matrix $\mathbf{D}(\mathbf{X})$, we obtain

$$\begin{aligned}\mathrm{Tr}[\mathbf{MD}(\mathbf{X})] = {} & \mathrm{Tr}(\mathbf{MD}) + \mathrm{Tr}(\mathbf{MP}_o\mathbf{X}\mathbf{P}_v^T - \mathbf{MP}_v\mathbf{X}\mathbf{P}_o^T) \\ & + \tfrac{1}{2}\mathrm{Tr}(\mathbf{MP}_o\mathbf{X}\mathbf{P}_v^T\mathbf{SP}_v\mathbf{X}\mathbf{P}_o^T \\ & - 2\mathbf{MP}_v\mathbf{X}\mathbf{P}_o^T\mathbf{SP}_o\mathbf{X}\mathbf{P}_v^T \\ & + \mathbf{MP}_o\mathbf{X}\mathbf{P}_v^T\mathbf{SP}_v\mathbf{X}\mathbf{P}_o^T) + \cdots,\end{aligned} \tag{11}$$

where we have made repeated use of the idempotency relations $\mathbf{P}_o^2=\mathbf{P}_o$ and $\mathbf{P}_v^2=\mathbf{P}_v$ and of the orthogonality relations $\mathbf{P}_o\mathbf{P}_v=\mathbf{P}_v\mathbf{P}_o=\mathbf{0}$ and $\mathbf{P}_o^T\mathbf{SP}_v=\mathbf{P}_v^T\mathbf{SP}_o=\mathbf{0}$. Introducing the shorthand notation

$$\mathbf{M}^{ab} = \mathbf{P}_a^T\mathbf{MP}_b, \tag{12}$$

this result may be written compactly as

$$\begin{aligned}\mathrm{Tr}[\mathbf{MD}(\mathbf{X})] = {} & \mathrm{Tr}(\mathbf{MD}) + \mathrm{Tr}(\mathbf{M}^{vo}\mathbf{X} - \mathbf{M}^{ov}\mathbf{X}) \\ & + \mathrm{Tr}(\mathbf{M}^{oo}\mathbf{XS}^{vv}\mathbf{X} - \mathbf{M}^{vv}\mathbf{XS}^{oo}\mathbf{X}) + \cdots.\end{aligned} \tag{13}$$

Note that, whereas the off-diagonal blocks $\mathbf{M}^{ov}$ and $\mathbf{M}^{vo}$ of $\mathbf{M}$ contribute to the terms linear in $\mathbf{X}$, the diagonal blocks $\mathbf{M}^{oo}$ and $\mathbf{M}^{vv}$ contribute to the quadratic terms.

B. The Roothaan-Hall Newton equations

In an SCF optimization, diagonalization of the Fock/KS matrix $\mathbf{F}$ is equivalent to minimization of the RH energy[8,9]

$$E^{RH}(\mathbf{X}) = \mathrm{Tr}[\mathbf{FD}(\mathbf{X})] \tag{14}$$

in the sense that both approaches yield the same density matrix. However, E^{RH} is only a crude model of the true SCF energy function E_{SCF}, having the correct gradient but an approximate Hessian at the point of expansion; this can be understood from the observation that, whereas E^{RH} depends linearly on $\mathbf{D}(\mathbf{X})$, the true energy E_{SCF} depends nonlinearly on $\mathbf{D}(\mathbf{X})$. Therefore, a complete minimization of E^{RH} (as achieved, for example, by diagonalization of the Fock/KS matrix) may give steps that are too long to be trusted, increasing, for example, rather than decreasing the total SCF energy. We therefore impose on the minimization the condition that the new occupied space does not differ appreciably from the old occupied space. Noting that $\mathbf{D}(\mathbf{X})\mathbf{S}$ and $\mathbf{DS}$ are projectors onto the new and old occupied spaces, respectively, we require that

$$\begin{aligned}\|\mathbf{D}(\mathbf{X}) - \mathbf{D}\|_S^2 &= \mathrm{Tr}\{[\mathbf{D}(\mathbf{X}) - \mathbf{D}]\mathbf{S}[\mathbf{D}(\mathbf{X}) - \mathbf{D}]\mathbf{S}\} \\ &= 2N - 2\,\mathrm{Tr}[\mathbf{DSD}(\mathbf{X})\mathbf{S}]\end{aligned} \tag{15}$$

is equal to some real parameter δ that characterizes the trust region of E^{RH}.

When the trust-region algorithm[36] is used for E^{RH}, the Newton step is taken only if the Hessian is positive definite and the Newton step is inside the trust region; otherwise, the minimum is determined on the boundary of the trust region of the second-order Taylor expansion of $E^{RH}(\mathbf{X})$. This is achieved by setting up the Lagrangian where the step-size constraint in Eq. (15) is added, multiplied by an undetermined multiplier μ:

$$L^{RH}(\mathbf{X}) = \mathrm{Tr}[\mathbf{FD}(\mathbf{X})] - 2\mu\{N - \mathrm{Tr}[\mathbf{DSD}(\mathbf{X})\mathbf{S}] - \delta\}. \tag{16}$$

Expanding this Lagrangian in orders of $\mathbf{X}$, we obtain

$$\begin{aligned}L^{RH}(\mathbf{X}) = {} & \mathrm{Tr}(\mathbf{FD}) + \mathrm{Tr}(\mathbf{F}^{vo}\mathbf{X} - \mathbf{F}^{ov}\mathbf{X}) \\ & + \mathrm{Tr}(\mathbf{F}^{oo}\mathbf{XS}^{vv}\mathbf{X} - \mathbf{F}^{vv}\mathbf{XS}^{oo}\mathbf{X}) \\ & + 2\mu[\mathrm{Tr}(\mathbf{S}^{oo}\mathbf{XS}^{vv}\mathbf{X}) - \delta] + \mathcal{O}(\mathbf{X}^3).\end{aligned} \tag{17}$$

To obtain Eq. (17), we have used Eq. (13) where $\mathbf{M}$ is replaced by $\mathbf{F}$ and $\mathbf{SDS}$, respectively, for the first and second terms of Eq. (16), recognizing that the only nonzero component of $\mathbf{SDS}$ is $\mathbf{P}_o^T\mathbf{SDSP}_o=\mathbf{S}^{oo}$. Differentiating this Lagrangian with respect to the elements $\mathbf{X}$, we obtain

$$\begin{aligned}\frac{\partial L^{RH}(\mathbf{X})}{\partial \mathbf{X}} = {} & \mathbf{F}^{ov} - \mathbf{F}^{vo} - \mathbf{S}^{vv}\mathbf{XF}^{oo} - \mathbf{F}^{oo}\mathbf{XS}^{vv} + \mathbf{F}^{vv}\mathbf{XS}^{oo} \\ & + \mathbf{S}^{oo}\mathbf{XF}^{vv} - 2\mu(\mathbf{S}^{vv}\mathbf{XS}^{oo} + \mathbf{S}^{oo}\mathbf{XS}^{vv}) + \cdots,\end{aligned} \tag{18}$$

where we have used the relation

$$\frac{\partial \mathrm{Tr}(\mathbf{AX})}{\partial \mathbf{X}} = \mathbf{A}^T. \tag{19}$$

Since $\mathbf{X}^T=-\mathbf{X}$, the right-hand side of Eq. (18) is antisymmetric. Finally, setting the right-hand side equal to zero and ignoring higher-order contributions, we obtain the matrix equation

$$\begin{aligned}\mathbf{F}^{vv}\mathbf{XS}^{oo} - \mathbf{F}^{oo}\mathbf{XS}^{vv} + \mathbf{S}^{oo}\mathbf{XF}^{vv} - \mathbf{S}^{vv}\mathbf{XF}^{oo} - 2\mu(\mathbf{S}^{vv}\mathbf{XS}^{oo} \\ + \mathbf{S}^{oo}\mathbf{XS}^{vv}) = \mathbf{F}^{vo} - \mathbf{F}^{ov}\end{aligned} \tag{20}$$

for the stationary point of the RH energy function.

We note that for each nonredundant solution $\mathbf{X}=\mathcal{P}(\mathbf{X})$, Eq. (20) has redundant solutions $\mathbf{X}+\mathbf{X}_R$, where $\mathbf{X}_R$ contains only redundant elements, that is, $\mathcal{P}(\mathbf{X}_R)=\mathbf{0}$. Restricting ourselves to the nonredundant solutions and introducing the notation

$$\mathbf{G} = \mathbf{F}^{\mathrm{ov}} - \mathbf{F}^{\mathrm{vo}}, \tag{21}$$

$$\mathbf{H}(\mu) = \mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}} - \mu\mathbf{S} \tag{22}$$

for the RH gradient and level-shifted Hessian, we may write these matrix equations more compactly as

$$\mathbf{H}(\mu)\tilde{\mathbf{X}}\mathbf{S} + \mathbf{S}\tilde{\mathbf{X}}\mathbf{H}(\mu) = -\mathbf{G}, \tag{23}$$

where it is assumed that $\tilde{\mathbf{X}}$ is a pure matrix in the sense that $\tilde{\mathbf{X}} = \mathcal{P}(\tilde{\mathbf{X}})$. These equations are solved iteratively, in a manner to be discussed shortly, so as to minimize the RH energy [Eq. (14)] subject to the constraint $\|\mathbf{D}(\mathbf{X}) - \mathbf{D}\|_{\mathrm{S}} = \delta$. In passing, we note that the RH Newton equations [Eq. (23)] may be viewed as a special case of the generalized Lyapunov equation of control theory $\mathbf{AXB} + \mathbf{BXA} = \mathbf{Q}$, where $\mathbf{X}$ is (anti)symmetric for (anti)symmetric $\mathbf{Q}$.

C. Vectorization transformation of the Roothaan-Hall Newton equations

In discussing the solution of the RH Newton matrix equations [Eq. (23)], it is instructive to rewrite it in a different form. For this purpose, we introduce the vec operator, which vectorizes a matrix by stacking its columns, for example,

$$\mathrm{vec}\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{pmatrix}. \tag{24}$$

For three arbitrary, conformable matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, we note the relationship

$$\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathrm{vec}\,\mathbf{B}. \tag{25}$$

Applying the vec operator to both sides of Eq. (23), we arrive at the RH Newton linear equations

$$\boldsymbol{\mathcal{H}}(\mu)\mathrm{vec}\,\tilde{\mathbf{X}} = -\mathrm{vec}\,\mathbf{G}, \tag{26}$$

with a level-shifted Hessian matrix given by

$$\boldsymbol{\mathcal{H}}(\mu) = \mathbf{H}(\mu) \otimes \mathbf{S} + \mathbf{S} \otimes \mathbf{H}(\mu). \tag{27}$$

The Newton matrix equations [Eq. (23)] for $\tilde{\mathbf{X}}$ are thus equivalent to the Newton linear equations for vec $\tilde{\mathbf{X}}$. We emphasize, however, that in practice the more compact matrix form [Eq. (23)] is used rather than the linear equations [Eq. (26)].

D. The transformed preconditioned conjugate-gradient method

For large dimensions, linear equations such as Eq. (26) are typically solved iteratively using the conjugate-gradient (CG) method, the convergence depending critically on the condition number of the level-shifted Hessian $\kappa[\boldsymbol{\mathcal{H}}(\mu)]$, where $\kappa(\boldsymbol{A})$ is the condition number of $\boldsymbol{A}$. To accelerate convergence, the preconditioned CG (PCG) method is used, replacing the linear equations [Eq. (26)] by the preconditioned equations

$$\boldsymbol{\mathcal{W}}^{-1}\boldsymbol{\mathcal{H}}(\mu)\mathrm{vec}\,\tilde{\mathbf{X}} = -\boldsymbol{\mathcal{W}}^{-1}\,\mathrm{vec}\,\mathbf{G}, \tag{28}$$

where $\boldsymbol{\mathcal{W}}$ is a symmetric, positive-definite matrix that approximates $\boldsymbol{\mathcal{H}}(\mu)$ but is easy to invert. We can now solve the linear equations more quickly with the CG method provided that $\kappa[\boldsymbol{\mathcal{W}}^{-1}\boldsymbol{\mathcal{H}}(\mu)] \ll \kappa[\boldsymbol{\mathcal{H}}(\mu)]$. A disadvantage of this approach is that $\boldsymbol{\mathcal{W}}^{-1}\boldsymbol{\mathcal{H}}(\mu)$ is, in general, neither symmetric nor positive definite, even for symmetric and positive-definite $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{H}}(\mu)$. To avoid this problem, we factorize the preconditioner

$$\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{V}}^T\boldsymbol{\mathcal{V}}, \tag{29}$$

where the positive-definite matrix $\boldsymbol{\mathcal{V}}$ may or may not be symmetric. Inserting Eq. (29) into Eq. (28) and rearranging, we obtain the similarity-transformed linear equation

$$[\boldsymbol{\mathcal{V}}^{-\mathrm{T}}\boldsymbol{\mathcal{H}}(\mu)\boldsymbol{\mathcal{V}}^{-1}][\boldsymbol{\mathcal{V}}\,\mathrm{vec}\,\tilde{\mathbf{X}}] = -\boldsymbol{\mathcal{V}}^{-\mathrm{T}}\,\mathrm{vec}\,\mathbf{G}, \tag{30}$$

which constitutes the basis for the transformed PCG method.

Returning to the matrix equations [Eq. (23)], we write the preconditioner factor $\boldsymbol{\mathcal{V}}$ in Eq. (29) as a Kronecker product

$$\boldsymbol{\mathcal{V}} = \mathbf{V} \otimes \mathbf{V} \tag{31}$$

and we find

$$\boldsymbol{\mathcal{V}}^{-\mathrm{T}}(\mathbf{A} \otimes \mathbf{B})\boldsymbol{\mathcal{V}}^{-1} = \mathbf{A}_{\mathrm{V}} \otimes \mathbf{B}_{\mathrm{V}}, \tag{32}$$

$$\boldsymbol{\mathcal{V}}\,\mathrm{vec}\,\mathbf{A} = \mathrm{vec}\,\mathbf{A}^{\mathrm{V}}, \tag{33}$$

$$\boldsymbol{\mathcal{V}}^{-\mathrm{T}}\,\mathrm{vec}\,\mathbf{A} = \mathrm{vec}\,\mathbf{A}_{\mathrm{V}}, \tag{34}$$

where we have used Eq. (25) and introduced the notation

$$\mathbf{A}_{\mathrm{V}} = \mathbf{V}^{-\mathrm{T}}\mathbf{A}\mathbf{V}^{-1}, \tag{35}$$

$$\mathbf{A}^{\mathrm{V}} = \mathbf{V}\mathbf{A}\mathbf{V}^{\mathrm{T}}. \tag{36}$$

We may therefore write the preconditioned RH Newton matrix equations as

$$\mathbf{H}_{\mathrm{V}}(\mu)\tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{S}_{\mathrm{V}} + \mathbf{S}_{\mathrm{V}}\tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{H}_{\mathrm{V}}(\mu) = -\mathbf{G}_{\mathrm{V}}, \tag{37}$$

where

$$\mathbf{G}_{\mathrm{V}} = \mathbf{F}_{\mathrm{V}}^{\mathrm{ov}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{vo}}, \tag{38}$$

$$\mathbf{H}_{\mathrm{V}}(\mu) = \mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}} - \mu\mathbf{S}_{\mathrm{V}}. \tag{39}$$

The application of the transformed PCG method for the Newton equations is thus equivalent to carrying out similarity transformations of the Fock/KS and overlap matrices with $\mathbf{V}^{-1}$. Our task now is to identify a useful preconditioner $\mathbf{V}$.

E. Choice of preconditioner

For large values of the level-shift parameter μ, the matrix Newton equations [Eq. (37)] take the form $\mu\mathbf{S}_{\mathrm{V}}\tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{S}_{\mathrm{V}} = \mathbf{G}_{\mathrm{V}}$, suggesting that a suitable preconditioner $\mathbf{V}$ is obtained by factorizing the (positive-definite) overlap matrix

$$\mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{V}, \tag{40}$$

since then $\mathbf{S}_{\mathrm{V}}=\mathbf{I}$ in Eq. (37). Such a factorization may be accomplished in infinitely many ways, for example, by introducing a Cholesky factorization[37] ($\mathbf{V}_{\mathrm{C}}$) or the Löwdin decomposition[38] ($\mathbf{V}_{\mathrm{s}}$, also called the principal square root)

$$\mathbf{V}_{\mathrm{C}} = \mathbf{U}, \tag{41}$$

$$\mathbf{V}_{\mathrm{s}} = \mathbf{S}^{1/2}, \tag{42}$$

where $\mathbf{U}$ is an upper triangular nonsingular matrix and where $\mathbf{S}^{1/2}$ is a positive-definite symmetric matrix. With these preconditioners, the RH Newton equations [Eq. (37)] take the form

$$\mathbf{H}_{\mathrm{V}}(\mu)\tilde{\mathbf{X}}^{\mathrm{V}} + \tilde{\mathbf{X}}^{\mathrm{V}}\mathbf{H}_{\mathrm{V}}(\mu) = -\mathbf{G}_{\mathrm{V}}, \quad \mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{V}, \tag{43}$$

where

$$\mathbf{H}_{\mathrm{V}}(\mu) = \mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}} - \mu\mathbf{I}. \tag{44}$$

These matrix equations, which are a special case of the continuous Lyapunov equation $\mathbf{AX}+\mathbf{XA}^{\mathrm{T}}=\mathbf{Q}$, are equivalent to the following Newton linear equations:

$$\boldsymbol{\mathcal{H}}_{\mathrm{V}}(\mu)\mathrm{vec}\,\tilde{\mathbf{X}}^{\mathrm{V}} = -\mathrm{vec}\,\mathbf{G}_{\mathrm{V}}, \tag{45}$$

$$\boldsymbol{\mathcal{H}}_{\mathrm{V}}(\mu) = \mathbf{H}_{\mathrm{V}}(\mu)\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{H}_{\mathrm{V}}(\mu), \tag{46}$$

which are the orthonormal counterpart of Eq. (26). A further improvement is possible by extracting the diagonal part of the similarity-transformed RH Hessian:

$$\mathbf{V}_{\mathrm{H}} = \mathrm{diag}([\mathbf{H}_{\mathrm{V}}(\mu)]_{11}^{1/2},[\mathbf{H}_{\mathrm{V}}(\mu)]_{22}^{1/2},\ldots)\mathbf{V}, \tag{47}$$

which is trivially set up, requiring only the extraction of the diagonal elements of the Hessian

$$[\boldsymbol{\mathcal{H}}_{\mathrm{V}}]_{\alpha\beta,\alpha\beta} = [\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}}]_{\alpha\alpha} + [\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}}]_{\beta\beta} - \mu, \tag{48}$$

where we have assumed an orthonormal basis.

The Cholesky and symmetric (square-root) preconditioners are equivalent in the sense that they yield the same condition number $\kappa[\boldsymbol{\mathcal{W}}^{-1}\boldsymbol{\mathcal{H}}(\mu)]$. Indeed, since the structures of $\mathbf{F}$ and $\mathbf{S}$ are broadly similar (with similar eigenvalues), these preconditioners typically reduce the condition number by several orders of magnitude, greatly improving CG convergence and reducing the overall computational effort. In passing, we note that, in any orthonormalized AO basis, the condition number of the RH Newton equations is the same as the condition number in the canonical orbital basis, to which it is related by an (condition-number conserving) orthonormal transformation.

An advantage of the Löwdin preconditioner over the Cholesky preconditioner is that it is often more diagonally dominant, as we shall see in some of the examples in Sec. III. Moreover, among all possible orthogonal bases, the Löwdin basis is the one that most closely resembles the original (local) AO basis, ensuring that locality is preserved to the greatest possible extent.[16] A possible misgiving about the Löwdin preconditioner is the practicality of generating $\mathbf{S}^{1/2}$ and $\mathbf{S}^{-1/2}$ in linear time. However, in Ref. 17, we demonstrate that $\mathbf{S}^{1/2}$ and $\mathbf{S}^{-1/2}$ can always be calculated at linear cost, in an iterative manner. Unless otherwise specified, we use the Löwdin basis in our calculations.

We conclude this section by noting that the use of an orthonormal Löwdin or Cholesky AO basis also simplifies the evaluation of the matrix exponential [Eq. (8)] to

$$\exp(\mathbf{X}) = \sum_{n=0}^{\infty}\frac{\mathbf{X}^n}{n!}. \tag{49}$$

However, this series converges rapidly only for small $\mathbf{X}$. To accelerate convergence for large arguments, we can use the relation

$$\exp(\mathbf{X}) = [\exp(2^{-k}\mathbf{X})]^{2^k}, \tag{50}$$

where on the right-hand side $\mathbf{X}$ is scaled by some suitably small parameter 2^{-k} such that the Frobenius norm of $\mathbf{X}$ is small enough for Eq. (49) to be rapidly convergent. In this way, the transformed density matrix can be evaluated in about ten matrix multiplications, regardless of the magnitude of $\mathbf{X}$. Furthermore, since $\mathbf{X}$ is antisymmetric, $\exp(-\mathbf{X})$ is given by $[\exp(\mathbf{X})]^{\mathrm{T}}$.

F. The level-shifted Newton equations in the canonical molecular-orbital basis

To gain insight into the convergence of the PCG algorithm and, in particular, to understand how the level-shift parameter should be chosen, it is instructive to express Eq. (37) in the unoptimized canonical molecular-orbital (MO) basis. In this basis, the Fock/KS matrix has diagonal occupied-occupied and virtual-virtual blocks with the pseudo-orbital energies ϵ_P on the diagonal, whereas the occupied-virtual and virtual-occupied blocks are nonzero. The level-shifted Hessian elements are then given by (using indices A, B, C, and D for virtual MOs and I, J, K, and L for occupied MOs)

$$H_{AIBJ}(\mu) = \delta_{AB}\delta_{IJ}(\epsilon_A - \epsilon_I - \mu), \tag{51}$$

and the virtual-occupied elements of Eq. (37) become

$$(\epsilon_A - \epsilon_I - \mu)X_{AI} = F_{AI}, \tag{52}$$

where X_{AI} is the solution vector in the canonical MO basis. The step-length function

$$\|\mathbf{X}\|_{\mathrm{S}}^2 = \sum_{AI}\frac{F_{AI}^2}{(\epsilon_A - \epsilon_I - \mu)^2} \tag{53}$$

has $k+1$ branches, where k is the number of eigenvalues $\epsilon_A-\epsilon_I$ of the (unshifted) Hessian (see Fig. 1). The function is positive for all μ with asymptotes at the eigenvalues. For $\mu<\min(\epsilon_A-\epsilon_I)$, the RH energy is lowered to both first and second orders.[8,36] In the trust-region formalism, the step length is taken to be the stationary point that corresponds to the minimum on the boundary of the trust region. The stationary point is therefore given by the intersection marked by a cross in Fig. 1.

In the canonical MO basis, the Hessian is diagonal and the solution to the level-shifted Newton equations is trivial. In the AO basis, by contrast, the Hessian is not diagonal and

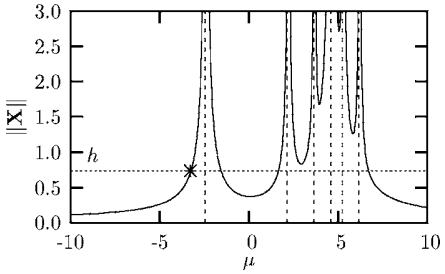

FIG. 1. The step length in Eq. (52) as a function of the level-shift parameter. The asymptotes indicated by vertical dashed lines occur at the eigenvalues of the Hessian for the RH energy. The trust region (h) is marked by the horizontal dotted line. The crossing between the dotted line and the step length function marked with a cross represents the chosen level shift.

the equations must be solved iteratively. The PCG convergence is then critically dependent on the judicious choice of preconditioner.

G. The level-shifted Newton equation as an eigenvalue problem

The solution to the level-shifted Newton equations [Eq. (45)] may alternatively be found by solving the eigenvalue problem[39–41]

$$\boldsymbol{\mathcal{A}}(\alpha)\begin{pmatrix}1\\ \tilde{\mathbf{x}}\end{pmatrix}=\mu\begin{pmatrix}1\\ \tilde{\mathbf{x}}\end{pmatrix}, \tag{54}$$

where we have introduced the short-hand notation

$$\tilde{\mathbf{x}}=\operatorname{vec}\tilde{\mathbf{X}}, \tag{55}$$

$$\mathbf{g}_{\mathrm{V}}=\operatorname{vec}\mathbf{G}_{\mathrm{V}}, \tag{56}$$

and where the dimension of the augmented Hessian

$$\boldsymbol{\mathcal{A}}(\alpha)=\begin{pmatrix}0 & \alpha\mathbf{g}_{\mathrm{V}}^{\mathrm{T}}\\ \alpha\mathbf{g}_{\mathrm{V}} & \boldsymbol{\mathcal{H}}_{\mathrm{V}}(0)\end{pmatrix} \tag{57}$$

is one larger than that of the Hessian $\boldsymbol{\mathcal{H}}_{\mathrm{V}}(\mu)$. To see that the solution of Eq. (54) determines the solution to the level-shifted Newton equations, we write the second component of Eq. (54) as

$$\boldsymbol{\mathcal{H}}_{\mathrm{V}}(0)\tilde{\mathbf{x}}+\alpha\mathbf{g}_{\mathrm{V}}=\mu\tilde{\mathbf{x}}, \tag{58}$$

or equivalently,

$$\boldsymbol{\mathcal{H}}_{\mathrm{V}}(\mu)\alpha^{-1}\tilde{\mathbf{x}}=-\mathbf{g}_{\mathrm{V}}. \tag{59}$$

Thus, the solution to the Newton equations [Eq. (45)] with the level-shift parameter μ is given by $\alpha^{-1}\tilde{\mathbf{x}}$, where $(1,\tilde{\mathbf{x}})^{\mathrm{T}}$ is the eigenvector that belongs to the eigenvalue μ of the augmented-Hessian eigenvalue problem [Eq. (54)]. Since the dimension of the augmented Hessian $\boldsymbol{\mathcal{A}}(\alpha)$ in Eq. (57) is one larger than that of $\boldsymbol{\mathcal{H}}_{\mathrm{V}}(\mu)$, the Hylleraas-Undheim theorem[42] predicts that the lowest eigenvalue of $\boldsymbol{\mathcal{A}}(\alpha)$ is lower than the lowest eigenvalue of $\boldsymbol{\mathcal{H}}_{\mathrm{V}}(\mu)$. Therefore, by selecting the lowest eigenvalue of Eq. (54), we generate a step in the left-hand branch of Fig. 1. Moreover, by adjusting α so that $\|\alpha^{-1}\tilde{\mathbf{x}}\|^2\approx h^2$, we generate a step to the minimum on the boundary of the trust region with trust radius h.

The augmented-Hessian eigenvalue problem [Eq. (54)] may be solved iteratively, updating α in the course of the iterations to give a step of length h. Assume that, during the iterative procedure, we have obtained a set of $n+1$ trial vectors

$$\begin{pmatrix}1\\ \mathbf{0}\end{pmatrix},\begin{pmatrix}0\\ \mathbf{b}_1\end{pmatrix},\begin{pmatrix}0\\ \mathbf{b}_2\end{pmatrix},\ldots,\begin{pmatrix}0\\ \mathbf{b}_n\end{pmatrix}, \tag{60}$$

where the $\mathbf{b}_i=\operatorname{vec}\mathbf{B}_i$ are orthonormal

$$\mathbf{b}_i^{\mathrm{T}}\mathbf{b}_j=\operatorname{Tr}(\mathbf{B}_i\mathbf{B}_j)=\delta_{ij}, \tag{61}$$

and where $\mathbf{b}_1$ is the normalized gradient vector

$$\mathbf{b}_1=\|\mathbf{g}_{\mathrm{V}}\|^{-1}\mathbf{g}_{\mathrm{V}}. \tag{62}$$

The augmented-Hessian eigenvalue problem [Eq. (54)] for the lowest eigenvalue may be set up in the basis of the $n+1$ trial vectors

$$\boldsymbol{\mathcal{A}}^{\mathrm{R}}(\alpha)\begin{pmatrix}1\\ \tilde{\mathbf{x}}^{\mathrm{R}}(\alpha)\end{pmatrix}=\mu^{\mathrm{R}}\begin{pmatrix}\mathbf{1}\\ \tilde{\mathbf{x}}^{\mathrm{R}}(\alpha)\end{pmatrix}, \tag{63}$$

where

$$\boldsymbol{\mathcal{A}}_{00}^{\mathrm{R}}(\alpha)=0, \tag{64}$$

$$\boldsymbol{\mathcal{A}}_{10}^{\mathrm{R}}(\alpha)=\boldsymbol{\mathcal{A}}_{01}^{\mathrm{R}}(\alpha)=\alpha\mathbf{b}_1^{\mathrm{T}}\mathbf{g}_{\mathrm{V}}=\alpha\|\mathbf{g}_{\mathrm{V}}\|, \tag{65}$$

$$\boldsymbol{\mathcal{A}}_{0i}^{\mathrm{R}}(\alpha)=\boldsymbol{\mathcal{A}}_{i0}^{\mathrm{R}}(\alpha)=0\quad(i>1), \tag{66}$$

$$\boldsymbol{\mathcal{A}}_{ij}^{\mathrm{R}}(\alpha)=\mathbf{b}_i^{\mathrm{T}}\boldsymbol{\sigma}_j, \tag{67}$$

and $\boldsymbol{\sigma}_j$ is the linearly transformed vector

$$\boldsymbol{\sigma}_j=\boldsymbol{\mathcal{H}}_{\mathrm{V}}(0)\mathbf{b}_j. \tag{68}$$

The first component in the reduced eigenvalue problem [Eq. (63)] spans the augmented dimension and is normalized to 1 according to Eq. (54). The solution to the level-shifted Newton equations [Eq. (45)] with $\mu=\mu^{\mathrm{R}}$ is given by $\alpha^{-1}\tilde{\mathbf{x}}^{\mathrm{R}}$ expanded in the basis of the trial vectors. By adjusting α so as to satisfy

$$\|\alpha^{-1}\tilde{\mathbf{x}}^{\mathrm{R}}\|^2=h^2, \tag{69}$$

we obtain a step of length h in the reduced space. When the lowest eigenvalue of Eq. (54) is determined iteratively, we may straightforwardly adjust α until it satisfies Eq. (69). Storing $\mathcal{A}^{\mathrm{R}}(1)$ with $\alpha=1$, we obtain $\mathcal{A}^{\mathrm{R}}(\alpha)$ for $\alpha\neq 1$ by a simple scaling of $\mathcal{A}_{10}^{\mathrm{R}}(1)$ and $\mathcal{A}_{01}^{\mathrm{R}}(1)$ according to Eq. (65).

To solve the augmented-Hessian eigenvalue problem, we may use the Davidson algorithm.[43] When the lowest eigenvalue is determined in the reduced space, α may be dynamically updated. In this manner, the minimum on the boundary of the trust region may be determined in the same number of iterations as required for solving the eigenvalue equation with a fixed α parameter.

To determine the lowest eigenvalue of the augmented Hessian efficiently, a good initial guess is required. However, since the augmented Hessian is not strongly diagonally dominant, such a guess is usually not readily available. In practice, therefore, we use the augmented-Hessian eigenvalue equation only to update α, so as to ensure that the level

shift is in the proper interval and of the correct size. The improved trial vectors are themselves obtained by solving the level-shifted Newton equations in the same reduced space $\{\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_n\}$ as for the eigenvalue equation but with an updated level-shift parameter. Essentially, we perform a sequence of PCG iterations, dynamically updating the level-shift parameter in the subspace generated by the PCG iterations.

In the PCG minimization, we first determine a solution with the step-size constraint $\|\mathbf{X}^{\mathrm{V}}\|=0.6$, where $\|\mathbf{X}^{\mathrm{V}}\|$ is the Frobenius norm. Next, the subspace generated during this minimization is utilized as the starting point for a subsequent minimization, now with the step-size constraint $X^{\mathrm{V}}_{\mathrm{max}}=0.35$, where $X^{\mathrm{V}}_{\mathrm{max}}$ is the largest component of $\mathbf{X}^{\mathrm{V}}$. Unlike the constraint on $\|\mathbf{X}^{\mathrm{V}}\|$, the constraint on $X^{\mathrm{V}}_{\mathrm{max}}$ is size-intensive. The algorithm is not sensitive to the choice of the $\|\mathbf{X}^{\mathrm{V}}\|$ parameter, whereas $X^{\mathrm{V}}_{\mathrm{max}}$ should be chosen carefully. We have found $\|\mathbf{X}^{\mathrm{V}}\|=0.6$ and $X^{\mathrm{V}}_{\mathrm{max}}=0.35$ to be suitable parameters.

The first level shift is obtained by solving the augmented-Hessian eigenvalue problem in a two-dimensional subspace, corresponding to a reduced space containing only one trial vector, namely, the normalized gradient in Eq. (62). The PCG iterations are terminated when the level shift has converged and when the residual has been reduced by a factor of 100 relative to the residual in the two-dimensional reduced space.

The RH SCF iterations are continued until the gradient norm $\|\mathbf{g}_{\mathrm{V}}\|$ is smaller than some preset threshold. However, just like $\|\mathbf{X}^{\mathrm{V}}\|$, the norm $\|\mathbf{g}_{\mathrm{V}}\|$ is an extensive property. Indeed for two noninteracting, identical systems, the total squared norm is equal to twice the norm of each subsystem:

$$\|\mathbf{g}^{\mathrm{A+B}}\|^2=\sum_i (g_i^{\mathrm{A}})^2+\sum_i (g_i^{\mathrm{B}})^2=\|\mathbf{g}^{\mathrm{A}}\|^2+\|\mathbf{g}^{\mathrm{B}}\|^2. \tag{70}$$

A size-intensive requirement on the SCF convergence is thus to use the gradient norm divided by the square root of the number of electrons $\|\mathbf{g}^{\mathrm{V}}\|/\sqrt{N}$.

H. Diagonalization of the level-shifted Fock/KS matrix by Newton's method

The minimum of the RH energy subject to the step-size constraint [Eq. (15)] may alternatively be determined by using the MO coefficients as variational parameters. In this parametrization, the density matrix may be expressed as

$$\mathbf{D}(\mathbf{X})=\mathbf{C}_{\mathrm{occ}}\mathbf{C}_{\mathrm{occ}}^{\mathrm{T}}, \tag{71}$$

where the coefficients of the occupied MOs $\mathbf{C}_{\mathrm{occ}}$ satisfy the orthonormality constraint

$$\mathbf{C}_{\mathrm{occ}}^{\mathrm{T}}\mathbf{S}\mathbf{C}_{\mathrm{occ}}=\mathbf{I}. \tag{72}$$

Imposing this orthonormality constraint simultaneously with the step-size constraint [Eq. (15)] on the energy E^{RH}, we obtain the Lagrangian

$$L^{\mathrm{RH}}(\mathbf{C}_{\mathrm{occ}})=\mathrm{Tr}[\mathbf{F}\mathbf{D}(\mathbf{X})]-\mu\{2N-2\,\mathrm{Tr}[\mathbf{D}\mathbf{S}\mathbf{D}(\mathbf{X})\mathbf{S}]-\delta\}-\mathrm{Tr}\,\boldsymbol{\lambda}(\mathbf{C}_{\mathrm{occ}}^{\mathrm{T}}\mathbf{S}\mathbf{C}_{\mathrm{occ}}-\mathbf{I}). \tag{73}$$

Differentiation of this Lagrangian with respect to the MO coefficients gives

$$(\mathbf{F}-\mu\mathbf{S}\mathbf{D}\mathbf{S})\mathbf{C}_{\mathrm{occ}}(\mu)=\mathbf{S}\mathbf{C}_{\mathrm{occ}}(\mu)\boldsymbol{\epsilon}(\mu), \tag{74}$$

where $\boldsymbol{\lambda}(\mu)$ is chosen to be diagonal $[\boldsymbol{\epsilon}(\mu)]$ since the energy is invariant with respect to rotations among the occupied MOs. The density matrix for the new RH iteration becomes

$$\mathbf{D}(\mu)=\mathbf{C}_{\mathrm{occ}}(\mu)\mathbf{C}_{\mathrm{occ}}^{\mathrm{T}}(\mu), \tag{75}$$

where $\mathbf{C}_{\mathrm{occ}}(\mu)$ are the eigenvectors of the generalized eigenvalue problem [Eq. (74)] with the level-shifted Fock/KS matrix $\mathbf{F}-\mu\mathbf{S}\mathbf{D}\mathbf{S}$.

In the local part of the RH SCF optimization, where $\mu=0$ and $\mathbf{X}$ is small, the solution of the Newton matrix equations [Eq. (23)] and the diagonalization of the Fock matrix [Eq. (74)] give essentially the same step and the same density matrix. To first order in $\mathbf{X}$, the solution of the Newton equations then corresponds to a diagonalization of the Fock/KS matrix. By contrast, in the global part of the RH SCF optimization, where $\mathbf{X}$ is larger, the steps obtained by diagonalizing the Fock/KS matrix [Eq. (74)] and by solving the Newton equations [Eq. (23)] differ.

In our implementation, the Newton step is always taken in the local region, where $\mu=0$. In the global region, each SCF iteration begins by solving the Newton eigenvalue equation [Eq. (54)] to determine the level-shift parameter μ_{max} by requiring that the largest step-length component is equal to $X^{\mathrm{V}}_{\mathrm{max}}$. The minimization of

$$E^{\mathrm{RH}}_{\mu_{\mathrm{max}}}=\mathrm{Tr}(\mathbf{F}-\mu_{\mathrm{max}}\mathbf{S}\mathbf{D}\mathbf{S})\mathbf{D}(\mathbf{X}) \tag{76}$$

is represented by the solution of the Fock/KS eigenvalue equation [Eq. (74)] with $\mu=\mu_{\mathrm{max}}$. The solution of the level-shifted Newton equations with level-shift parameter μ_{max} then represents a first-order diagonalization of the level-shifted Fock/KS matrix in Eq. (74), whereas the full minimization of $E^{\mathrm{RH}}_{\mu_{\mathrm{max}}}$ requires a complete diagonalization and may be accomplished by a sequence of level-shifted Newton iterations with $\mu=\mu_{\mathrm{max}}$. In practice, a partial rather than exact minimization of $E^{\mathrm{RH}}_{\mu_{\mathrm{max}}}$ is sufficient in the global region. Thus, in our implementation, no more than one or two level-shifted Newton iterations [Eq. (23)] with $\mu=\mu_{\mathrm{max}}$ are taken since, after two iterations, the Newton steps have become so small that they no longer affect the global SCF convergence. Indeed, our standard procedure is to take only one level-shifted Newton step although we also report some calculations where two level-shifted Newton steps are taken at each SCF iteration.

I. Evaluation of the Coulomb contribution

The Coulomb contributions to the Fock/KS matrix and the energy are given by

$$J_{ab}=(ab|\rho), \tag{77}$$

$$J=\tfrac{1}{2}(\rho|\rho), \tag{78}$$

in terms of the one-electron density

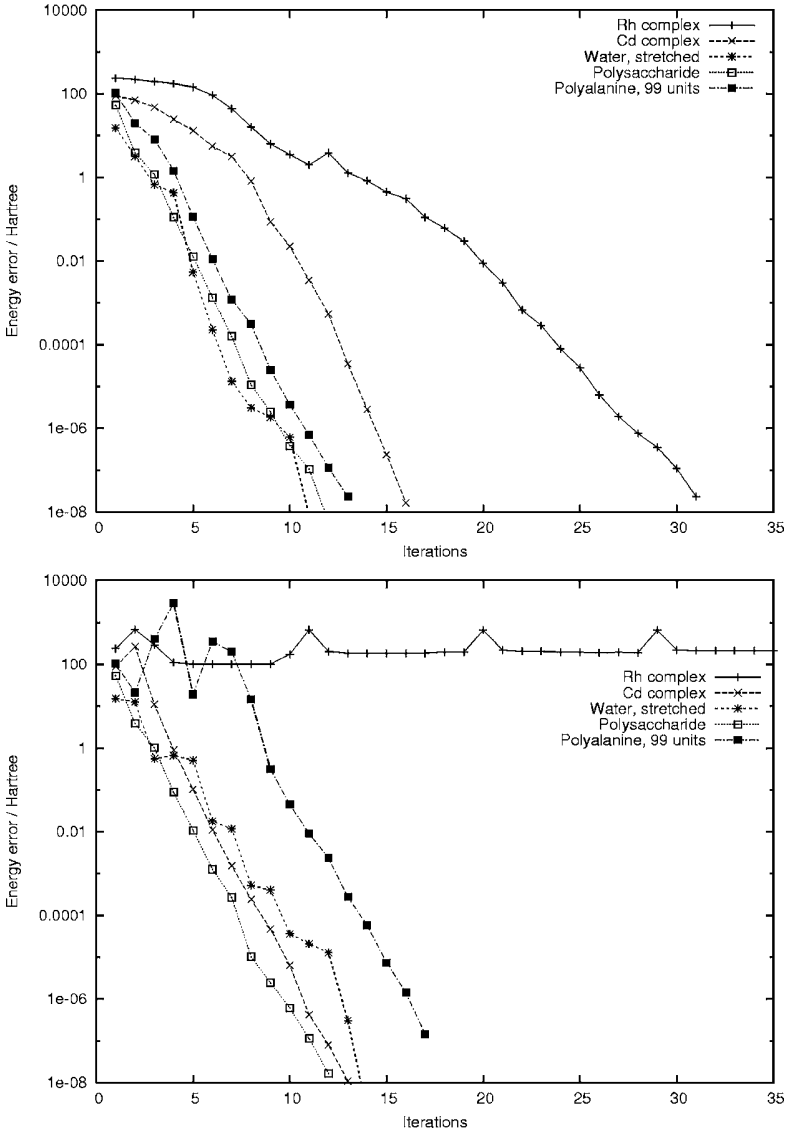

FIG. 2. The error in the energy in HF LS-TRSCF (top) and SCF/DIIS (bottom) optimizations.

$$\rho(\mathbf{r}) = \sum_{cd} \chi_c(\mathbf{r})\chi_d(\mathbf{r})D_{cd}. \tag{79}$$

In density fitting,[44,45] the computational cost is significantly reduced by evaluating these contributions as

$$\tilde{J}_{ab} = (ab|\tilde{\rho}), \tag{80}$$

$$\tilde{J} = (\rho|\tilde{\rho}) - \tfrac{1}{2}(\tilde{\rho}|\tilde{\rho}) = J - \tfrac{1}{2}(\rho - \tilde{\rho}|\rho - \tilde{\rho}) \tag{81}$$

from an approximate density $\tilde{\rho}$ expanded in an atom-centered auxiliary basis:

$$\tilde{\rho}(\mathbf{r}) = \sum_{\alpha} \xi_\alpha(\mathbf{r})c_\alpha. \tag{82}$$

We determine the c_α by minimizing the fitting error $(\rho - \tilde{\rho}|w|\rho - \tilde{\rho})$ with metric w subject to the charge-conserving constraint $\int\tilde{\rho}(\mathbf{r})\mathrm{d}\mathbf{r} = N_\mathrm{e}$, leading to the linear equation

$$\sum_{\beta} (\alpha|w|\beta)c_\beta = (\alpha|w|\rho) + \lambda(\alpha), \tag{83}$$

where the one-center overlaps are given by $(\alpha) = \int\xi_\alpha(\mathbf{r})\mathrm{d}\mathbf{r}$, and with

$$\lambda = \frac{N_\mathrm{e} - \Sigma_{\alpha\beta}(\alpha)(\alpha|w|\beta)^{-1}(\beta|\rho)}{\Sigma_{\alpha\beta}(\alpha)(\alpha|w|\beta)^{-1}(\beta)}. \tag{84}$$

From Eq. (81), we see that the fitted Coulomb repulsion energy is always lower than the regular repulsion energy. The smallest fitting error is obtained by an unconstrained minimization $\lambda = 0$ in the Coulomb metric $w = r_{12}^{-1}$ in Eq. (83). The use of constraints or of a non-Coulomb metric increases the error, lowering the Coulomb energy.

With density fitting, large speedups are observed, but scaling becomes a problem for large systems—the inversion [Eq. (83)] scales cubically in time, whereas the memory requirements for the $(\alpha|\beta)$ matrix scale quadratically. To achieve linear scaling, there are two main strategies. One is to fit the density in a metric different from the long-range Coulomb metric, so that $(\alpha|w|\beta)$ of Eq. (83) becomes

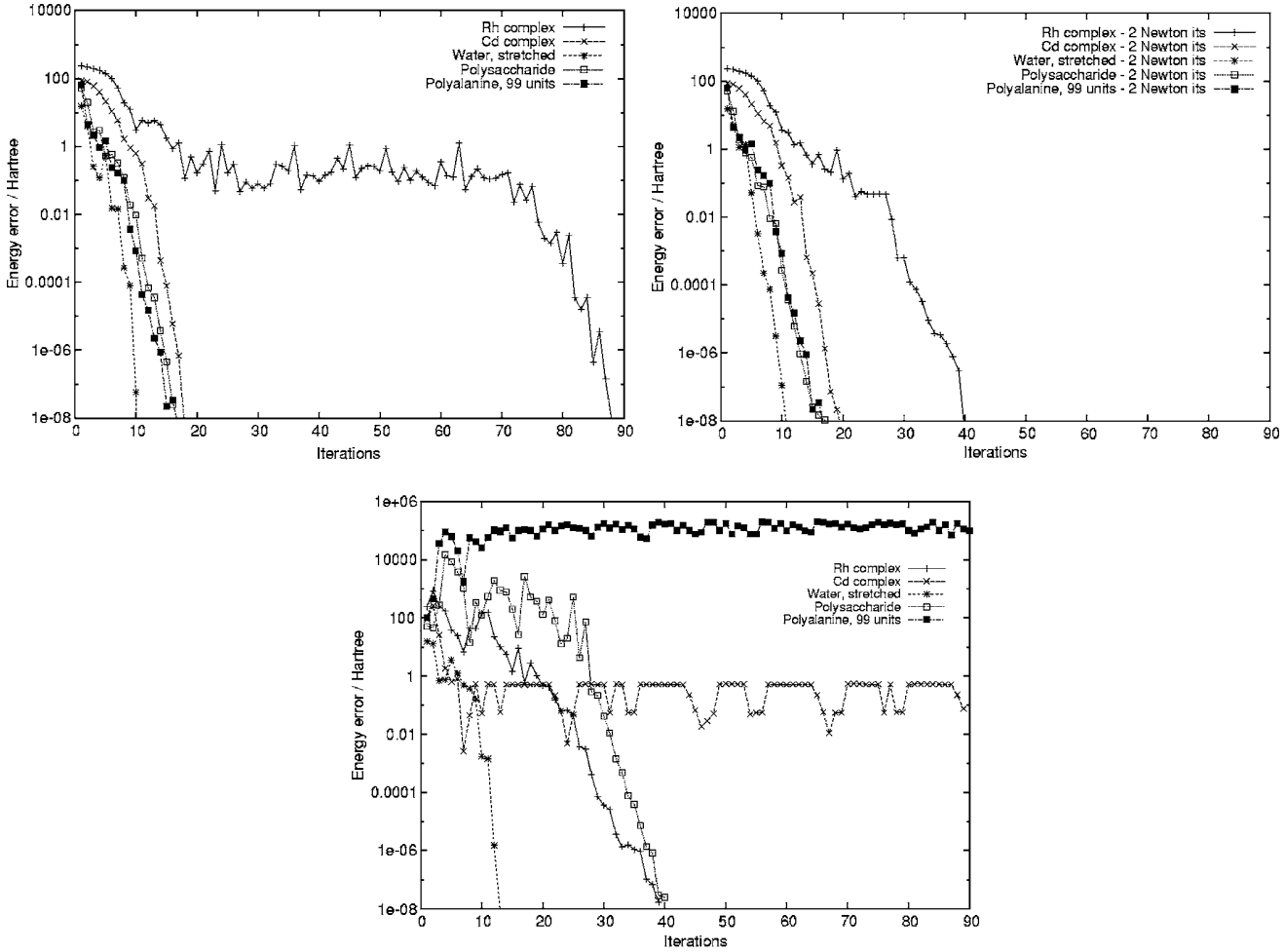

FIG. 3. The error in the energy in LDA KS optimizations using the one-step Newton LS-TRSCF (top left), two-step Newton LS-TRSCF (top right), and SCF/DIIS (bottom) methods.

sparse.[45–47] Alternatively, the density is partitioned into localized parts, which are fitted separately.[48–50] We use an approach similar to that of Ref. 48. The system is divided into localized parts i using the density partitioning[51] of Yang and Lee

$$\rho(\mathbf{r}) = \sum_i \rho^{(i)}(\mathbf{r}) = \sum_i \sum_{ab} \chi_a(\mathbf{r})\chi_b(\mathbf{r})D_{ab}x_{ab}^{(i)}, \tag{85}$$

where $x_{ab}^{(i)}=1$ for both a and b in i, $x_{ab}^{(i)}=1/2$ for either a or b in i, and $x_{ab}^{(i)}=0$ otherwise. With this decomposition, some of the overlap distributions belonging to subsystem i may in fact be centered outside this subsystem (by the Gaussian product rule), but these decay exponentially with the square of the separation between the two Gaussian functions.

Each subsystem density $\rho^{(i)}$ is fitted using auxiliary functions located within an extended subsystem i, comprising the original subsystem i padded with a buffer zone δ_i around the subsystem

$$\sum_{\beta\in i+\delta_i} (\alpha|\beta)c_\beta^{(i)} = \sum_{cd} (\alpha|cd)D_{cd}x_{cd}^{(i)} - \lambda^{(i)}(\alpha). \tag{86}$$

The multipliers are given by

$$\lambda^{(i)} = \frac{Q^{(i)} - \Sigma_{\alpha\beta\in i+\delta_i}(\alpha)(\alpha|\beta)^{-1}(\beta|\rho^{(i)})}{\Sigma_{\alpha\beta\in i+\delta_i}(\alpha)(\alpha|\beta)^{-1}(\beta)}, \tag{87}$$

with the subsystem charge

$$Q^{(i)} = \int \rho^{(i)}(\mathbf{r})\mathrm{d}\mathbf{r} = \sum_{ab} S_{ab}D_{ab}x_{ab}^{(i)}. \tag{88}$$

The cost of solving Eq. (86) depends on the size of the subsystem rather than on the size of the full system. Given that the number of subsystems increases linearly with system size, the full density is fitted in linear time. In our calculations, the total system is put in a rectangular box, which is recursively bisected until no subbox contains more than 5000 auxiliary basis functions. In the fitting, a buffer zone of width 5 Bohr is used. In the applications presented here, we used the optimized basis set developed by Eichkorn *et al.*[52,53] with no charge constraints imposed on the fitted density.

III. SAMPLE CALCULATIONS

In the HF and KS calculations reported here, we use the LS-TRSCF method, combining the LS-TRRH algorithm for the RH iterations of Sec. II B with the TRDSM algorithm for density averaging [implemented in a local version of DALTON

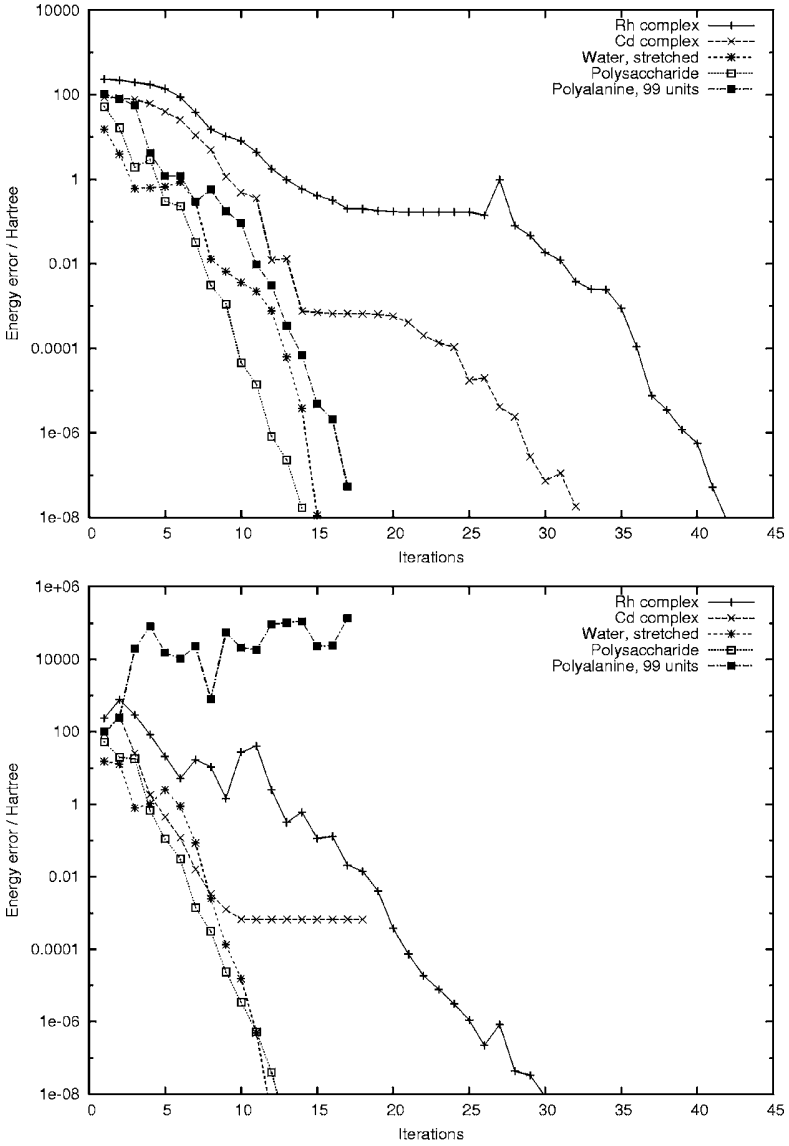

FIG. 4. The error in the energy in B3LYP KS LS-TRSCF (top) and SCF/DIIS (bottom) optimizations.

(Ref. 54)]. First, in Sec. III A, we compare the LS-TRSCF scheme with the traditional SCF/DIIS scheme. Next, in Sec. III B, we examine the CG solution of the RH Newton equations [Eq. (45)]. Finally, in Sec. III C, we consider the cost of the LS-TRRH optimization, demonstrating that linear scaling may be obtained within this framework.

A. Convergence of the LS-TRSCF method

To compare the LS-TRSCF and SCF/DIIS methods, we use the following five molecules which represent a variety of bonding situations: the water molecule in the d-aug-pVTZ basis with the bonds stretched to twice their equilibrium value; the rhodium complex of Ref. 19, with the STO-3G basis for rhodium and the Ahlrichs VDZ basis for the other atoms;[55] the cadmium-imidazole complex of Ref. 18 in the 3-21G basis; a 438-atom polysaccharide in the 6-31G basis; and a 992-atom polypeptide of 99 alanine residues in the 6-31G basis. For all systems, we have carried out calculations at the HF and KS levels of theory, using the local density approximation (LDA) and B3LYP functionals. As initial guesses for the optimization, we used the one-electron core Hamiltonian for water and for the metal complexes, while the Hückel model was used for the large polysaccharide and polypeptide molecules. Unless otherwise indicated, only one Newton step is taken in each TRRH iteration. In the DIIS and TRDSM density-averaging steps, a maximum of eight density matrices and Fock/KS matrices are used. In Figs. 2, 3, and 4, we have plotted the error in the energy (on a logarithmic scale) at each SCF iteration for the HF model, the LDA model, and the B3LYP model, respectively.

In the LS-TRSCF calculations, we observe a smooth convergence to an error of 10^{-8} a.u. in 10-30 iterations. The only exceptions are the the KS rhodium calculations, where a significant energy lowering is observed in the first few iterations, followed by a large number of iterations with no apparent progress, in particular, for the LDA functional. Eventually, the local region is reached and fast convergence is established. In Fig. 3, we have also included plots for LDA calculations that take two Newton steps in each SCF iteration. In general, there is little difference between the one- and two-step calculations, but a striking improvement is ob-

TABLE I. H_2O stretched, LDA/d-aug-cc-pVTZ. SCF convergence and number of microiterations needed to get the new density in the Löwdin and Cholesky bases, respectively.

Iteration	Energy	Gradient norm	n_{it} Löwdin	n_{it} Cholesky
1	−60.173 329 477 53	17.278 869		
2	−71.778 669 286 89	12.938 274	10	9
3	−75.418 451 510 91	3.170 448	16	15
4	−74.234 639 836 42	11.774 826	22	18
5	−75.549 995 553 04	2.278 892	21	14
6	−75.539 701 417 42	2.994 470	17	19
7	−75.562 908 067 61	2.070 197	13	17
8	−75.579 986 850 34	0.903 419	16	18
9	−75.583 777 493 19	0.093 106	17	21
10	−75.583 817 470 68	0.004 338	18	17
11	−75.583 817 561 34	0.000 448	14	20
12	−75.583 817 562 31	0.000 051	18	19
13	−75.583 817 562 33	0.000 008	13	18

served for the rhodium complex. This improvement is not yet understood (it may be accidental), and we continue to use one Newton iteration as the default in our optimizations.

A comparison of the SCF/DIIS plots with the LS-TRSCF plots in Figs. 2–4 clearly reveals the poorer SCF/DIIS performance, in particular, for the KS calculations. However, some differences are also observed in the HF calculations—unlike the LS-TRSCF method, the SCF/DIIS method diverges for the rhodium complex and performs erratically in the global part of the polyalanine optimization. In general, we note that the SCF/DIIS and LS-TRSCF differences are largest in the global region, where the SCF/DIIS scheme suffers from the fact that it is not based on the principles of energy minimization and step-size control, sometimes leading to an erratic behavior. In the local region, size constraints become irrelevant, since both methods use the quasi-Newton condition to speed up the local convergence, which becomes very similar for the two methods.

The SCF/DIIS LDA calculations in Fig. 3 show a strikingly erratic behavior. The cadmium and polyalanine calculations both diverge; for the polysaccharide, no convergence is observed until iteration 25. Surprisingly, the LDA calculation on the rhodium complex converges, unlike the Hartree-Fock calculation. Finally, concerning the B3LYP functional in Fig. 4, we note that the SCF/DIIS polyalanine optimization diverges. Interestingly, for the cadmium complex, the horizontal line that begins at iteration 10 indicates convergence to a stationary point of higher energy than that obtained with the LS-TRSCF algorithm. Indeed, a closer examination of this stationary point reveals that it is a saddle point with the lowest Hessian eigenvalue of −0.0147 a.u. From an inspection of the corresponding LS-TRSCF curve, it appears that the LS-TRSCF approaches the same saddle point in iterations 15-20. At iteration 17, however, TRDSM detects a negative Hessian eigenvalue, and from iteration 20, convergence is established towards the minimum (lowest Hessian eigenvalue of 0.0275 a.u.), which is reached in 32 iterations.

B. The solution of the Roothaan-Hall Newton equations

To examine the convergence of the RH Newton equations, we consider the stretched water molecule of Sec. III A at the LDA/d-aug-cc-pVTZ level of theory, solving Eq. (45) in the Cholesky and Löwdin bases with and without the diagonal preconditioner [Eq. (48)]. Although small, this example is representative of the present calculations. We begin by noting that the SCF convergence illustrated in Fig. 3 is the same in the Cholesky and Löwdin bases. In both cases, the electronic gradient is reduced to less than 10^{-5} a.u. after 13 SCF iterations—see Table I, where we have also listed the number of PCG iterations required to solve a set of RH Newton equations at each SCF iteration. Typically, 10-20 PCG iterations are needed to solve the Newton equations, with an average number of 17 iterations needed in the Cholesky basis and 16 iterations in the Löwdin basis.

To understand better the performance of the CG method in the Cholesky and Löwdin bases, we have selected for closer examination one level-shifted SCF iteration in the glo-

TABLE II. Global H_2O LDA/d-aug-cc-pVTZ convergence, second SCF iteration. Convergence of the RH Newton equations [Eq. (43)] in the Cholesky basis with and without a diagonal preconditioner. The constrained step-size parameter is marked with an asterisk.

	No preconditioner				Diagonal preconditioner			
Iteration	$\|\mathbf{R}\|$	μ	X^v_{max}	$\|\mathbf{X}^v\|$	$\|\mathbf{R}\|$	μ	X^v_{max}	$\|\mathbf{X}^v\|$
1	3.31	−9.51	0.190	0.592*	3.31	−9.51	0.190	0.592*
2	1.09	−11.78	0.174	0.562*	0.34	−11.85	0.163	0.573*
3	0.28	−11.47	0.175	0.590*	0.17	−11.48	0.179	0.592*
4	0.06	−12.23	0.171	0.557*	0.05	−12.23	0.170	0.558*
5	0.02	−11.83	0.176	0.576*	0.01	−11.83	0.176	0.576*
6	0.13	−6.68	0.315*	1.048	0.18	−6.05	0.346*	1.193
7	0.08	−6.48	0.323*	1.091	0.06	−6.42	0.323*	1.104
8	0.06	−6.28	0.332*	1.137	0.03	−6.22	0.336*	1.151
9	0.06	−6.09	0.344*	1.186	0.02	−6.60	0.314*	1.065
10	0.02	−6.46	0.322*	1.096				
11	0.02	−6.26	0.334*	1.142				
12	0.02	−6.07	0.347*	1.192				
13	0.01	−6.44	0.324*	1.100				

TABLE III. Local H_2O LDA/d-aug-cc-pVTZ convergence, seventh SCF iteration. Convergence of the RH Newton equations [Eq. (43)] in the Cholesky basis with and without a diagonal preconditioner.

	No preconditioner			Diagonal preconditioner		
Iteration	$\|\mathbf{R}\|$	$X^v_{\max}$	$\|\mathbf{X}^v\|$	$\|\mathbf{R}\|$	$X^v_{\max}$	$\|\mathbf{X}^v\|$
1	0.163	0.014	0.076	0.089	0.006	0.036
2	0.138	0.025	0.132	0.081	0.020	0.103
3	0.138	0.043	0.214	0.049	0.034	0.160
4	0.127	0.053	0.259	0.040	0.041	0.181
5	0.120	0.065	0.309	0.029	0.044	0.194
6	0.090	0.072	0.338	0.023	0.046	0.204
7	0.078	0.080	0.367	0.019	0.047	0.212
8	0.080	0.085	0.384	0.015	0.050	0.217
9	0.047	0.092	0.407	0.015	0.052	0.220
10	0.055	0.096	0.416	0.011	0.055	0.224
11	0.028	0.103	0.430	0.007	0.058	0.226
12	0.040	0.106	0.435	0.007	0.059	0.228
13	0.027	0.111	0.441	0.005	0.060	0.228
14	0.031	0.115	0.447	0.003	0.061	0.229
15	0.020	0.117	0.449	0.002	0.061	0.229
16	0.023	0.121	0.454	0.001	0.061	0.230
17	0.018	0.123	0.455	0.001	0.061	0.230
18	0.019	0.124	0.457			
19	0.018	0.126	0.459			
20	0.013	0.127	0.461			
21	0.019	0.129	0.463			
22	0.011	0.130	0.465			
23	0.016	0.131	0.466			
24	0.009	0.132	0.468			
25	0.010	0.133	0.468			
26	0.005	0.133	0.469			
27	0.007	0.134	0.469			
28	0.004	0.134	0.470			
29	0.005	0.134	0.470			
30	0.003	0.134	0.470			
31	0.003	0.134	0.470			
32	0.002	0.134	0.470			

bal region (iteration 2) and one unshifted SCF iteration in the local region (iteration 7); see Tables II–V, each of which contains the following information on each (P)CG iteration needed for the solution of the RH Newton equations: the residual $\|\mathbf{R}\|$, the level-shift value μ, the largest component $X^{\mathrm{V}}_{\max}$, and the norm $\|\mathbf{X}^{\mathrm{V}}\|$ of the current solution vector $\mathbf{X}^{\mathrm{V}}$.

In the global SCF iteration of Table II, we first solve the RH Newton equations [Eq. (45)] in the Cholesky basis with the constraint $\|\mathbf{X}^{\mathrm{V}}\|=0.6$ imposed, followed by solution with the new constraint $X^{\mathrm{V}}_{\max}=0.35$. The level shift that gives a total step length of about 0.6 ($\mu=-11.8$) is quickly established, as is subsequently the shift that gives the final step $\mathbf{X}^{\mathrm{V}}$ with the largest component of about 0.35 ($\mu=-6.6$). Note how the step size increases as we change the constraint from $\|\mathbf{X}^{\mathrm{V}}\|$ to $X^{\mathrm{V}}_{\max}$. The reason that we determine a step of total length 0.6 before attempting a step with the largest component 0.35 is that it gives a more robust algorithm. In a small subspace, the individual components of $\mathbf{X}^{\mathrm{V}}$ may change strongly in the first few iterations, making the identification of μ difficult; after a few iterations where $\|\mathbf{X}^{\mathrm{V}}\|$ is determined to be equal 0.6, the individual components become more stable and the application of the constraint on the individual components more straightforward.

The CG iterations are terminated when the residual has been reduced by a factor of 100 in the $\|\mathbf{X}^{\mathrm{V}}\|$-constrained search and by a factor of 50 in the $X^{\mathrm{V}}_{\max}$-constrained search. The overall SCF convergence is not sensitive to the choice of these convergence thresholds. At each iteration, only one matrix multiplication is required to carry out the linear transformation [Eq. (45)]. When using a diagonal preconditioner, two additional multiplications are needed for projection of each trial vector, giving a total of three matrix multiplications in each iteration.

From Table II, we see that the use of a diagonal preconditioner improves the convergence in the global SCF iteration slightly, reducing the number of iterations from 13 to 10. In the local iteration of the same SCF optimization in Table III, the preconditioner is even more effective, almost halving the number of iterations. Clearly, the best strategy for solving the RH equations is to always apply a diagonal preconditioner, giving a more robust CG algorithm at the modest cost of a single projection. In passing, we note that the optimization in the left-hand column of Table III corresponds to the curvy-step method of Shao *et al.*[15] where the unshifted Newton equations are solved in the Cholesky basis without a diagonal preconditioner.

TABLE IV. Global H_2O LDA/d-aug-cc-pVTZ convergence, second SCF iteration. Convergence of the RH Newton equations [Eq. (43)] in the Löwdin basis with and without a diagonal preconditioner. The constrained step-size parameter is marked with an asterisk.

	No preconditioner				Diagonal preconditioner			
Iteration	$\|\mathbf{R}\|$	μ	$X^v_{\max}$	$\|\mathbf{X}^v\|$	$\|\mathbf{R}\|$	μ	$X^v_{\max}$	$\|\mathbf{X}^v\|$
1	3.31	−9.51	0.198	0.592*	3.31	−9.51	0.198	0.592*
2	1.09	−11.78	0.188	0.562*	0.49	−11.85	0.189	0.571*
3	0.28	−11.47	0.199	0.590*	0.15	−11.48	0.197	0.593*
4	0.06	−12.23	0.184	0.557*	0.04	−12.23	0.183	0.558*
5	0.02	−11.83	0.190	0.576*	0.00	−11.83	0.190	0.576*
6	0.13	−6.68	0.324*	1.04	0.16	−6.68	0.325*	1.04
7	0.07	−6.48	0.334*	1.09	0.07	−6.48	0.333*	1.09
8	0.06	−6.28	0.343*	1.13	0.03	−6.28	0.344*	1.13
9	0.03	−6.67	0.326*	1.05	0.02	−6.67	0.325*	1.05
10	0.02	−6.46	0.336*	1.09	0.01	−6.46	0.335*	1.09
11	0.02	−6.26	0.345*	1.14				
12	0.01	−6.65	0.326*	1.05				

Tables IV and V contain the same information as do Tables II and III, respectively, but for the Löwdin rather than Cholesky basis. The convergence is similar to that observed in the Cholesky basis. Again, the preconditioned iterations converge faster—especially in the local SCF iteration, where we save more than 50% of the CG iterations by preconditioning. Also, in this SCF iteration, the CG convergence is slightly faster in the Löwdin basis than in the Cholesky basis. In the following, we use the Löwdin basis with a diagonal preconditioner.

C. Linear scaling using the TRSCF algorithm

To demonstrate that linear scaling is obtained with the LS-TRSCF algorithm, we here carry out polyalanine peptide

TABLE V. Local H_2O LDA/d-aug-cc-pVTZ convergence, seventh SCF iteration. Convergence of the RH Newton equations [Eq. (43)] in the Löwdin basis with and without a diagonal preconditioner.

	No preconditioner			Diagonal preconditioner		
Iteration	$\|\mathbf{R}\|$	$X^v_{\max}$	$\|\mathbf{X}^v\|$	$\|\mathbf{R}\|$	$X^v_{\max}$	$\|\mathbf{X}^v\|$
1	0.0381	0.0012	0.007	0.0331	0.001	0.006
2	0.0279	0.0018	0.011	0.0263	0.003	0.018
3	0.0344	0.0029	0.018	0.0227	0.005	0.032
4	0.0206	0.0036	0.023	0.0172	0.009	0.047
5	0.0316	0.0047	0.029	0.0142	0.011	0.057
6	0.0220	0.0068	0.038	0.0088	0.013	0.063
7	0.0204	0.0079	0.043	0.0064	0.014	0.066
8	0.0175	0.0089	0.047	0.0050	0.015	0.068
9	0.0174	0.0111	0.055	0.0023	0.015	0.068
10	0.0171	0.0122	0.059	0.0018	0.015	0.069
11	0.0120	0.0138	0.065	0.0008	0.015	0.069
12	0.0132	0.0144	0.067	0.0004	0.015	0.069
13	0.0072	0.0150	0.069	0.0003	0.015	0.069
14	0.0102	0.0154	0.071			
15	0.0048	0.0159	0.072			
16	0.0054	0.0160	0.073			
17	0.0032	0.0163	0.074			
18	0.0027	0.0164	0.074			
19	0.0018	0.0164	0.074			
20	0.0018	0.0164	0.074			
21	0.0013	0.0164	0.074			
22	0.0008	0.0164	0.074			
23	0.0009	0.0164	0.074			
24	0.0006	0.0164	0.074			
25	0.0009	0.0164	0.074			
26	0.0003	0.0164	0.074			
27	0.0005	0.0164	0.074			
28	0.0003	0.0164	0.074			

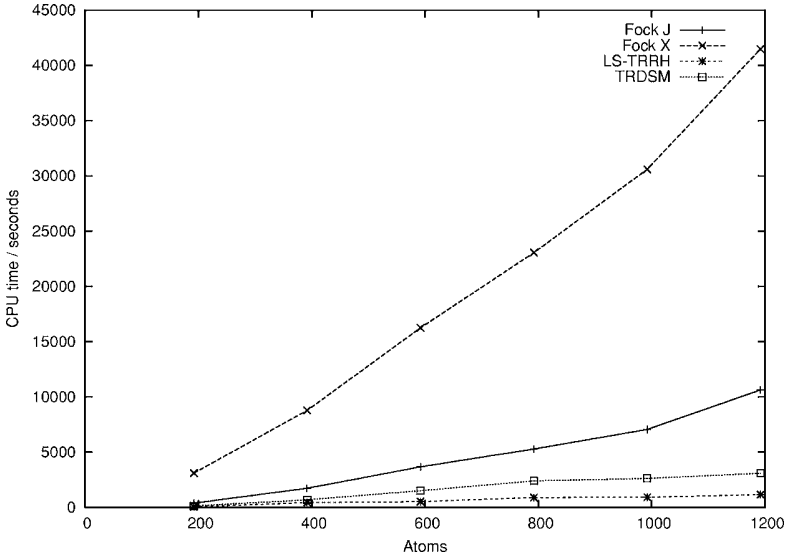

FIG. 5. Timings for the first local iteration of Hartree-Fock calculations on polyalanine peptides. The timings given are for the Coulomb (Fock J) and exchange (Fock X) parts of the Fock matrix and for the LS-TRRH and TRDSM parts (sparse-matrix algebra.)

calculations with up to 119 alanine residues (1192 atoms) using the HF and B3LYP models in the 6-31G basis. Each SCF optimization converges as the 99-residue calculations in Figs. 2 and 4.

In Fig. 5, we have plotted the CPU times spent in the different parts of the LS-TRSCF algorithm in the Hartree-Fock/6-31G calculations using sparse-matrix algebra. The timings are obtained using a single processor on an IBM RS6000 pSeries 690 (1.3 GHz). Except for the DSM step, the timings in this and later plots are for the first local SCF iteration. However, since the time spent in the DSM step depends on the number of density matrices included in the density subspace, the DSM timings are always given for an SCF iteration where the subspace contains the maximum number of density matrices (eight).

The CPU times for the Coulomb and exchange parts of the Fock matrix in Fig. 5 both increase linearly with system size, but with a slight kink as the system increases from 1000 to 1200 atoms. The exchange part is about four times more expensive than the Coulomb part. The LS-TRRH and TRDSM optimization steps are dominated by matrix multiplications. The linearity of the LS-TRRH and TRDSM timings in Fig. 5 therefore indicates that sparsity is exploited efficiently in the matrix multiplications. The importance of efficient sparse-matrix algebra is evident in Fig. 6, where we compare the timings of Fig. 5 with those obtained with dense-matrix algebra. The different behaviors of sparse-matrix algebra (linear scaling) and dense-matrix algebra (cubical scaling) are well illustrated. Some fluctuations are observed in these plots since the LS-TRRH and TRDSM steps both involve iterations, whose number may vary slightly from system to system. The benefits of sparse-matrix algebra are first noticed for TRDSM, since each TRDSM step contains more matrix multiplications than does each LS-TRRH step.

Finally, Fig. 7 shows the CPU timings for B3LYP optimizations with sparse-matrix algebra. This figure differs from Fig. 5 in that it also contains contributions from the KS exchange-correlation potential. In these calculations, the exchange-correlation step is about twice as expensive as the

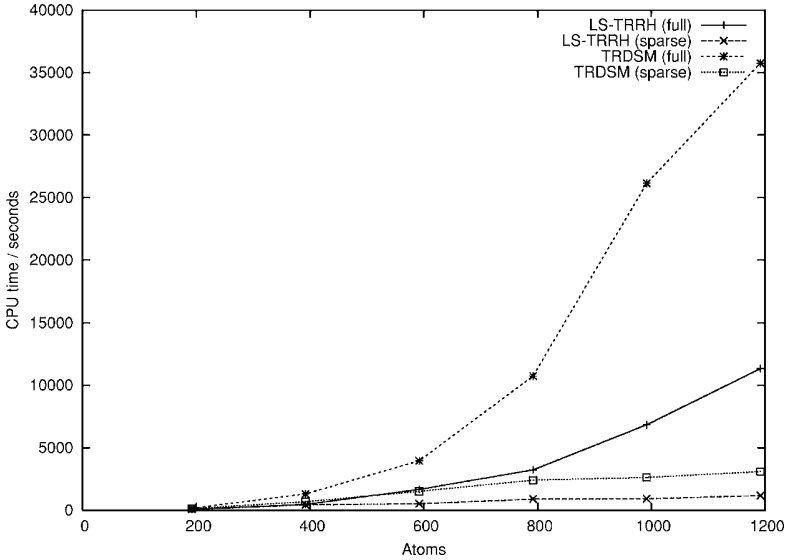

FIG. 6. The timings for the LS-TRRH and TRDSM contributions from Fig. 5 shown with the corresponding timings when full matrices are used.

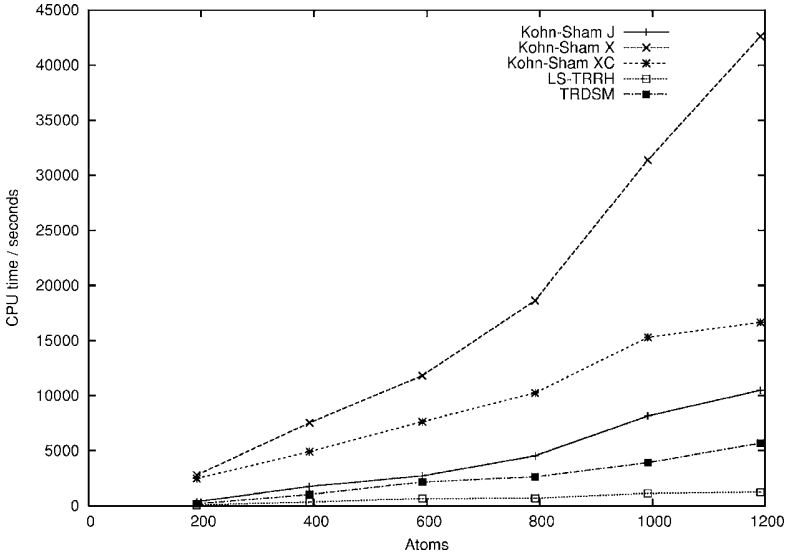

FIG. 7. Timings for the first local iteration of B3LYP calculations on polyalanine peptides. The timings given are for the Coulomb (Kohn-Sham J), exchange (Kohn-Sham X), and exchange-correlation (Kohn-Sham XC) parts of the KS matrix and for the LS-TRRH and TRDSM parts (sparse-matrix algebra.)

Coulomb step and about half as expensive as the exact-exchange step. The cost of the exchange-correlation step clearly scales linearly with system size.

IV. CONCLUSIONS

We have described a linear-scaling implementation of the trust-region self-consistent field (LS-TRSCF) method. In the LS-TRSCF method, each iteration consists of a minimization of the RH energy (equivalent to a minimization of the sum of the orbital energies in canonical HF theory) to generate a new AO density matrix in the trust-region RH (LS-TRRH) step, followed by the determination of an improved averaged density matrix in the subspace of the current and previous density matrices using the trust-region density-subspace minimization (TRDSM) algorithm. A linear-scaling algorithm is obtained by using iterative methods to solve the level-shifted Newton equations and by exploiting the sparsity of the involved matrices.

In the solution of the RH Newton matrix equations, we have shown that the Löwdin and Cholesky orthonormalizations yield similar performances, with a slight preference for the Löwdin orthonormalization since it resembles most closely the original AO basis set (preserving sparsity to the largest possible extent) and since it leads to marginally fewer CG iterations than the Cholesky orthonormalization. We have, moreover, demonstrated that, in the Löwdin and Cholesky bases, use of a diagonal preconditioner significantly improves convergence, typically reducing the number of CG iterations by a factor of 2 in the local SCF iterations. In each LS-TRRH step, a single Newton step is sufficient for the minimization of the RH energy, although we have observed one case where two Newton steps give (perhaps fortuitously) a significantly improved SCF convergence.

When comparing LS-TRRH to the curvy-step method of Shao *et al.*[15] the main differences are the diagonal preconditioning of the CG iterations and the level shifting of the SCF iterations. Without a diagonal CG preconditioner, the convergence of the level-shifted Newton equations is at best much slower than the solution of the preconditioned equations; often, the equations do not converge without preconditioning. Indeed, the latter is almost always the case for molecules with an electronic structure more complicated than those of water clusters or linear alkanes, typically used as test cases. For robust and fast convergence of the SCF and Newton iterations, it is essential to choose a level shift that is neither too small (which will introduce wrong directions and cause divergence) nor too large (which will cause very slow convergence). An important feature of the LS-TRRH algorithm is that the optimal level shift is determined dynamically at no extra cost.

We have demonstrated that the LS-TRSCF method yields a smooth and robust convergence for small and large systems, often converging where the traditional SCF/DIIS scheme fails. For small systems, a TRSCF implementation based on an explicit diagonalization of the Fock/KS matrix may be more efficient. However, since the time spent in the optimization of such systems is insignificant compared with the time spent constructing the Fock/KS matrix, we recommend the LS-TRSCF method as the standard method for systems of all sizes.

ACKNOWLEDGMENTS

This work has been supported by the Lundbeck Foundation and the Danish Natural Research Council and the Norwegian Research Council through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420) and through a grant of computer time from the Program for Supercomputing. The authors acknowledge support from the Danish Center for Scientific Computing (DCSC), the Academy of Finland, and the European Research and Training Network NANOQUANT, Understanding Nanomaterials from the Quantum Perspective, Contract No. MRTN-CT-2003-506842.

[1] S. Goedecker and G. E. Scuseria, Comput. Sci. Eng. **5**, 14 (2003).
[2] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).
[3] A. H. R. Palser and D. E. Manolopoulos, Phys. Rev. B **58**, 12704 (1998).

[4] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).
[5] R. W. Nunes and D. Vanderbilt, Phys. Rev. B **50**, 17611 (1994).
[6] R. Baer and M. Head-Gordon, J. Chem. Phys. **109**, 10159 (1998).
[7] R. Baer and M. Head-Gordon, Phys. Rev. B **58**, 15296 (1998).
[8] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (Wiley, New York, 2000).
[9] R. McWeeny, *Methods of Molecular Quantum Mechanics*, 2nd ed. (Academic, New York, 1992).
[10] X. P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993).
[11] J. M. Millam and G. E. Scuseria, J. Chem. Phys. **106**, 5569 (1997).
[12] M. Challacombe, J. Chem. Phys. **110**, 2332 (1999).
[13] T. Helgaker, H. Larsen, J. Olsen, and P. Jørgensen, Chem. Phys. Lett. **327**, 379 (2000).
[14] H. Larsen, J. Olsen, P. Jørgensen, and T. Helgaker, J. Chem. Phys. **115**, 9685 (2001).
[15] Y. Shao, C. Saravanan, M. Head-Gordon, and C. A. White, J. Chem. Phys. **118**, 6144 (2003).
[16] B. C. Carlson and J. M. Keller, Phys. Rev. **105**, 102 (1957).
[17] B. Jansík, S. Høst, P. Jørgensen, and T. Helgaker, J. Chem. Phys. (accepted for publication).
[18] L. Thøgersen, J. Olsen, A. Köhn, P. Jørgensen, P. Sałek, and T. Helgaker, J. Chem. Phys. **123**, 074103 (2005).
[19] L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek, and T. Helgaker, J. Chem. Phys. **121**, 16 (2004).
[20] P. Pulay, Chem. Phys. Lett. **73**, 393 (1980); J. Comput. Chem. **3**, 556 (1982).
[21] K. N. Kudin, G. E. Scuseria, and E. Cancés, J. Chem. Phys. **116**, 8255 (2002).
[22] C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Lett. **230**, 8 (1994).
[23] C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Lett. **253**, 268 (1996).
[24] M. C. Strain, G. E. Scuseria, and M. J. Frisch, Science **271**, 51 (1996).
[25] M. Challacombe and E. Schwegler, J. Chem. Phys. **106**, 5526 (1997).
[26] Y. Shao and M. Head-Gordon, Chem. Phys. Lett. **323**, 425 (2000).
[27] E. Schwegler and M. Challacombe, J. Chem. Phys. **105**, 2726 (1996).
[28] E. Schwegler, M. Challacombe, and M. Head-Gordon, J. Chem. Phys. **106**, 9708 (1997).
[29] E. Schwegler and M. Challacombe, J. Chem. Phys. **111**, 6223 (1999).
[30] E. Schwegler and M. Challacombe, Theor. Chem. Acc. **104**, 344 (2000).
[31] C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
[32] J. C. Burant, G. E. Scuseria, and M. J. Frisch, J. Chem. Phys. **105**, 8969 (1996).
[33] J. M. Pérez-Jordá and W. Yang, Chem. Phys. Lett. **241**, 469 (1995).
[34] B. G. Johnson, C. A. White, Q. Zang, B. Chen, R. L. Graham, P. M. W. Gill, and M. Head-Gordon, in *Recent Developments in Density Functional Theory*, edited by J. M. Seminario (Elsevier Science, Amsterdam, 1996), Vol. 4.
[35] R. E. Stratman, G. E. Scuseria, and M. J. Frisch, Chem. Phys. Lett. **257**, 213 (1996).
[36] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (Wiley, New York, 1987).
[37] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran*, 2nd ed. (Cambridge University Press, Cambridge, 1992).
[38] P.-O. Löwdin, J. Chem. Phys. **18**, 365 (1950).
[39] H. J. Aa. Jensen and P. Jørgensen, J. Chem. Phys. **80**, 1204 (1984).
[40] B. Lengsfield III, J. Chem. Phys. **73**, 382 (1980).
[41] R. Shepard, I. Shavitt, and J. Simons, J. Chem. Phys. **76**, 543 (1982).
[42] E. A. Hylleraas and B. Undheim, Z. Phys. **65**, 759 (1930); J. K. L. Macdonald, Phys. Rev. **43**, 830 (1933).
[43] E. R. Davidson, J. Comput. Phys. **17**, 87 (1975).
[44] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).
[45] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).
[46] Y. Jung, A. Sodt, P. M. W. Gill, and M. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **102**, 6692 (2005).
[47] O. Vahtras, J. Almlöf, and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).
[48] R. T. Gallant and A. St-Amant, Chem. Phys. Lett. **256**, 569 (1996).
[49] C. F. Guerra, J. G. Snijders, G. te Velde, and E. J. Baerends, Theor. Chem. Acc. **99**, 391 (1998).
[50] A. Sodt, J. E. Subotnik, and M. Head-Gordon, J. Chem. Phys. **125**, 194109 (2006).
[51] W. Yang and T. S. Lee, J. Chem. Phys. **103**, 5674 (1995).
[52] K. Eichkorn, O. Treutler, H. Hm, M. Hser, and R. Ahlrichs, Chem. Phys. Lett. **242**, 652 (1995).
[53] K. Eichkorn, F. Weigend, O. Treutler, and R. Ahlrichs, Theor. Chem. Acc. **97**, 119 (1997).
[54] DALTON, Release 2.0, an *ab initio* electronic structure program, 2005. See http://www.kjemi.uio.no/software/dalton/dalton.html
[55] A. Schäfer, H. Horn, and R. Ahlrichs, J. Chem. Phys. **97**, 2571 (1992).

Paper III

Variational and robust density fitting of four-center two-electron integrals in local metrics

S. Reine, E. Tellgren, A. Krapp, T. Kjærgaard, T. Helgaker, B. Jansik, S. Høst and P. Sałek

The Journal of Chemical Physics, **129**, 104101 (2008)

Variational and robust density fitting of four-center two-electron integrals in local metrics

Simen Reine,[1,a)] Erik Tellgren,[1] Andreas Krapp,[1] Thomas Kjærgaard,[1,b)] Trygve Helgaker,[1] Branislav Jansik,[2] Stinne Høst,[2] and Paweł Salek[3]
[1]*Centre of Theoretical and Computational Chemistry, Department of Chemistry, University of Oslo, P.O. Box 1033 Blindern, N-0315 Oslo, Norway*
[2]*Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Århus, DK-8000 Århus C, Denmark*
[3]*Laboratory of Theoretical Chemistry, The Royal Institute of Technology, Teknikringen 30, Stockholm SE-10044, Sweden*

(Received 20 May 2008; accepted 18 June 2008; published online 8 September 2008)

Density fitting is an important method for speeding up quantum-chemical calculations. Linear-scaling developments in Hartree–Fock and density-functional theories have highlighted the need for linear-scaling density-fitting schemes. In this paper, we present a robust variational density-fitting scheme that allows for solving the fitting equations in local metrics instead of the traditional Coulomb metric, as required for linear scaling. Results of fitting four-center two-electron integrals in the overlap and the attenuated Gaussian damped Coulomb metric are presented, and we conclude that density fitting can be performed in local metrics at little loss of chemical accuracy. We further propose to use this theory in linear-scaling density-fitting developments. © *2008 American Institute of Physics*. [DOI: 10.1063/1.2956507]

I. INTRODUCTION

In molecular electronic-structure theory, an essential step is the evaluation of two-electron integrals over one-electron basis functions, typically taken to be linear combinations of atomic orbitals or other local or semilocal basis functions. Examples of such semilocal basis functions are Gaussian-type orbitals (GTOs) and Slater-type orbitals (STOs). The expansion coefficients are found by applying the variation principle, which ensures that all first-order variations in the energy with respect to the variations in the density are zero. Although a finite basis-set expansion may introduce quite large absolute errors, the variational property leads to high accuracy in the calculated chemical properties.

In a similar fashion, the product of two such basis functions may again be expanded in one-center auxiliary orbitals. Such density-fitting or resolution-of-the-identity (RI) approximations are introduced to speed up calculations involving four-center two-electron integrals, the traditional bottleneck of *ab initio* and density-functional calculations. In effect, the evaluation of four-center two-electron integrals is replaced by the evaluation of two- and three-center two-electron integrals and the solution of a set of linear equations. The speed-up resulting from this approach depends on the system studied and the basis set used; typically, a speed-up by a factor of 3–30 is observed.[1] The auxiliary basis sets introduced for density fitting are about three times larger than the regular basis set, while the errors introduced by the auxiliary basis are about two orders of magnitude smaller than the errors introduced by the regular basis. In this paper, we employ the auxiliary basis sets developed in Refs. 2 and 3.

The next section gives an introduction to density fitting. Next, in Sec. III, we present a robust variational scheme for approximating four-center two-electron integrals in a non-Coulomb metric, demonstrating how it can be used for the two-electron Coulomb and exchange contributions appearing in Hartree–Fock (HF) theory and Kohn–Sham (KS) density-functional theory. Implementational details are given in Sec. IV, whereas results are presented and discussed in Sec. V. Section VI contains some concluding remarks.

II. DENSITY FITTING

Density fitting was introduced independently in the Coulomb metric by Whitten[4] and in the overlap metric by Baerends *et al.*[5] In Ref. 4, Whitten established bounds on individual integrals, and later Jafri and Whitten[6] applied density fitting in self-consistent field (SCF) calculations, where individual integrals are either fitted or calculated directly depending on whether the predicted error in the fit is below a certain threshold or not. In the paper by Baerends *et al.*,[5] the electron density $\rho(\mathbf{r})$ is approximated by an expansion in atom-centered auxiliary basis functions $\xi_\alpha(\mathbf{r})$,

$$\begin{aligned}\rho(\mathbf{r}) &= \sum_{ab} D_{ab}\chi_a(\mathbf{r})\chi_b(\mathbf{r}) \\ &= \sum_{ab} D_{ab}\Omega_{ab}(\mathbf{r}) \approx \tilde{\rho}(\mathbf{r}) \\ &= \sum_{\alpha}^{N_{\text{aux}}} c_\alpha \xi_\alpha(\mathbf{r}). \end{aligned} \tag{1}$$

[a)]Electronic mail: simen.reine@kjemi.uio.no.
[b)]Permanent address: Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Århus, DK-800 Århus C, Denmark.

Here tilde denotes a density-fitted quantity, the D_{ab} are the matrix elements of the electron density expanded in the atomic orbitals (AOs) $\chi_a(\mathbf{r})$, $\Omega_{ab}(\mathbf{r})$ is the product (overlap distribution) between $\chi_a(\mathbf{r})$ and $\chi_b(\mathbf{r})$, and c_α are the fitting coefficients. The fitted density $\tilde{\rho}(\mathbf{r})$ is used to construct an approximate Coulomb potential

$$V_C(\mathbf{r}_1) = \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 \approx \tilde{V}_C(\mathbf{r}_1) = \int \frac{\tilde{\rho}(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2, \tag{2}$$

which is in turn used for the construction of the Coulomb part of the Fock or KS matrix

$$\tilde{J}_{ab} = \int \Omega_{ab}(\mathbf{r}) \tilde{V}_C(\mathbf{r}) d\mathbf{r} = \int \Omega_{ab}(\mathbf{r}_1) \frac{1}{r_{12}} \tilde{\rho}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2. \tag{3}$$

Baerends *et al.* obtained the fitting coefficients c_α by minimizing the fitting error

$$D_w = \langle \rho - \tilde{\rho} | w | \rho - \tilde{\rho} \rangle, \tag{4}$$

in the overlap metric, $w(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$. We here use the notation

$$\langle f | w | g \rangle = \int f(\mathbf{r}_1) w(\mathbf{r}_1, \mathbf{r}_2) g(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2. \tag{5}$$

The fitted density is further constrained to conserve charge,

$$\int \rho(\mathbf{r}) d\mathbf{r} = \int \tilde{\rho}(\mathbf{r}) d\mathbf{r} = N_e, \tag{6}$$

where N_e is the number of electrons, leading to the following set of linear equations for the fitting coefficients

$$\sum_\beta \langle \alpha | w | \beta \rangle c_\beta = \sum_{cd} \langle \alpha | w | cd \rangle D_{cd} + (\alpha)\lambda, \tag{7}$$

with the Lagrange multiplier

$$\lambda = \frac{N_\mathbf{e} - \Sigma_{\alpha\beta}(\alpha)(\alpha|\beta)^{-1}(\beta|\rho)}{\Sigma_{\alpha\beta}(\alpha)(\alpha|\beta)^{-1}(\beta)}. \tag{8}$$

We use the notation $(f|g) \equiv \langle f | 1/r_{12} | g \rangle$,

$$(f|g) = \int f(\mathbf{r}_1) \frac{1}{r_{12}} g(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \tag{9}$$

and

$$(\alpha) = \int \xi_\alpha(\mathbf{r}) d\mathbf{r}. \tag{10}$$

The density-fitting scheme of Baerends *et al.* was further developed by Dunlap *et al.*,[7,8] who replaced the overlap operator, $\delta(\mathbf{r}_1 - \mathbf{r}_2)$, with the Coulomb operator, $1/r_{12}$. In these two papers, Dunlap *et al.* established that the Coulomb metric is superior to the overlap metric, noting that the error in the energy is about one order of magnitude smaller in the Coulomb metric than in the overlap metric.

A. Linear-scaling density fitting

Density fitting offers significant speed-ups for the calculation of four-center integrals at little loss of accuracy. Recent developments toward large systems have highlighted the need for a linear-scaling density-fitting scheme. We note that the fitting equations [Eq. (7)] in the Coulomb metric cannot be solved straightforwardly for large systems, as the computational time scales cubically with system size. In this section, we give a brief overview of different linear-scaling density-fitting schemes presented in the literature. We first discuss methods based on the spatial partitioning of electron density; next, we consider methods based on the use of a local metric.

The partitioning approach has been explored by several authors;[1,9–11] all of them effectively enforce sparsity of the solved fitting equations in different ways. In the paper by Gallant and St-Amant,[9] the density is partitioned using Yang's scheme.[12] Each of these densities is fitted separately by including fitting functions within some predefined vicinity of the density. The fitted density is further constrained to preserve charge. The resulting errors are small, but the fitted density is not variational and the procedure does not provide a continuous potential energy surface. The same is true for the method presented by Salek *et al.* in Ref. 10, although here the energy is correct to second order.

In the partitioning approach by Fonseca Guerra *et al.*,[11] the density is partitioned into diatomic densities, generated by overlaps between basis functions centered on two atoms. The diatomic densities are fitted in the overlap metric subject to charge conservation. The resulting energy is neither variational nor correct to second order. It is worth noting that STOs are used rather than GTOs and that the fitted density is used to build an approximate Coulomb potential that is included in the numerical evaluation together with the exchange-correlation contribution.

In the partitioning approach known as local atomic density fitting (LADF) or atomic resolution of the identity (ARI), Sodt *et al.*[1] partition the density into atomic regions by localizing the individual product overlaps between two basis functions to one of the two atoms that the basis functions originate from. These atomic densities are fitted individually by including fitting functions in some predefined vicinity of the atom and by introducing a bump function on the boundary to turn off smoothly which fitting functions to include. The bump function is important in order for the potential energy surface to be continuous. By including first-order Dunlap corrections,[13] the fitted energy is made variational.

Of the above partitioning schemes, the LADF scheme offers the most elegant and balanced way to obtain the fitted density, although the bump function does represent an artifact. Clearly, in all partitioning schemes, some cutoff scheme must be adopted. A criticism of these partitioning schemes is that impact of the fitting error on the calculated energies is difficult to predict.

We now turn our attention to fitting methods based on the use of a local metric.[5,14,15] In the approach of Baerends *et al.*,[5] the electron density is fitted in the overlap metric, giving errors one order of magnitude greater than in the Coulomb metric.[7] This result was confirmed by Vahtras *et al.*,[14] who compare three different ways of fitting the four-center integrals in the overlap metric to the corresponding fit in the Coulomb metric. In the paper by Jung *et al.*,[15] the expansion coefficients obtained in the Coulomb metric, the overlap

metric, and the complementary error-function attenuated metric $w(\mathbf{r}_1,\mathbf{r}_2)=\mathrm{erfc}(\omega r_{12})/r_{12}$ are compared. The attenuated metric bridges the Coulomb and the overlap metrics by varying the value of the damping parameter ω. The coefficients obtained in the overlap metric decay more or less exponentially with distance, whereas the coefficients obtained in the Coulomb metric decay more slowly at long distances. For a one-dimensional test system studied in that paper, the fitting coefficients decay as $\sim r^{-1.25}$ in the Coulomb metric, with a faster decay observed for two- and three-dimensional systems. The authors further provide statistics on atomization energies for the G2 benchmark set using RI second-order Møller–Plesset perturbation theory in the cc-pVDZ basis, reporting errors six to seven times larger in the overlap metric than in the Coulomb metric.

The scheme of Jung *et al.*[15] is neither robust nor variational. In Sec. II D, we present an extended scheme that is both robust and variational, providing an accurate and reliable linear-scaling density-fitting alternative to the LADF scheme by Sodt *et al.*[1]

B. Density-fitting exchange

Density fitting was originally introduced to approximate the full density $\rho(\mathbf{r})$ and in this way accelerate Coulomb evaluation. Similar methodology can be used in the exchange part and is also here termed density-fitting or RI approximations.

The exact-exchange matrix is given by

$$K_{ab}=\sum_{cd}(ac|bd)D_{cd}=\sum_{i}(ai|bi). \tag{11}$$

Subscript i here denotes an occupied molecular orbital (MO) $\phi_i(\mathbf{r})$, expanded in AOs with coefficients C_{ai},

$$\phi_i(\mathbf{r})=\sum_a C_{ai}\chi_a(\mathbf{r}). \tag{12}$$

The density-matrix elements D_{ab} are related to the MO coefficients according to

$$D_{ab}=\sum_i C_{ai}C^*_{bi}. \tag{13}$$

Density fitting of the exact exchange was introduced by Weigend.[2] In this paper, the exchange matrix of Eq. (11) was approximated as

$$\begin{aligned}\tilde{K}_{ab}&=\sum_{cd}\sum_{\alpha\beta}(ac|\alpha)(\alpha|\beta)^{-1}(\beta|bd)D_{cd}\\&=\sum_{cd}\sum_{\alpha'}(ac|\alpha')(\alpha'|bd)D_{cd},\end{aligned} \tag{14}$$

where the last step constitutes a transformation to an orthogonal auxiliary basis

$$\xi_{\alpha'}(\mathbf{r})=\sum_{\alpha}(\alpha'|\alpha)^{-1/2}\xi_\alpha(\mathbf{r}). \tag{15}$$

C. Linear-scaling density fitting of exchange

Linear-scaling aspects of density fitting in the evaluation of exchange contribution were first considered by Polly *et al.*[16] The fitted exchange matrix,

$$\tilde{K}_{ab}=\sum_i(ai|\alpha)c^{bi}_\alpha, \tag{16}$$

is computed in linear time. This is achieved using localized orbitals $\chi_a(\mathbf{r})$ and $\phi_i(\mathbf{r})$ that interact with auxiliary functions $\xi_\alpha(\mathbf{r})$ only in some local domain. Localization of the MOs is achieved through the Pipek–Mezey localization.[17] The density-fitted exchange energy,

$$\tilde{K}=\sum_{ij}\sum_{\alpha}(ij|\alpha)c^{ij}_\alpha, \tag{17}$$

is computed without use of local fitting domains. It is argued that the fitted exchange energy depends sensitively on the size of the fitting domains, whereas the optimized MO coefficients do not—reported errors are in the microhartree range. In effect the MOs are not optimized variationally, although the energy is corrected through first order. It should be noted that the final step of Eq. (17) does not scale linearly with system size, i.e., without the use of local fitting domains.

The ARI exchange method (ARI-K) of Sodt *et al.*[18] is an extension of the LADF or ARI approach of Ref. 1, applied to the exchange rather than the Coulomb contribution. In this approach, the product overlaps $\Omega_{ai}(\mathbf{r})$ are approximated by auxiliary basis functions $\xi_\alpha(\mathbf{r})$ in the local domain $[A]$ near the parent atom of AO $\chi_a(\mathbf{r})$,

$$\tilde{\Omega}_{ai}(\mathbf{r})=\sum_{\alpha\in[A]}c^{ai}_\alpha\xi_\alpha(\mathbf{r}), \tag{18}$$

with

$$c^{ai}_\alpha=\sum_{\beta\in[A]}(\alpha|\beta)^{-1}_A(\beta|ai). \tag{19}$$

As in the LADF scheme, continuity of the potential energy surface is ensured by the use of individual inverses $(\alpha|\beta)^{-1}_A$ associated with the centers A (see Ref. 18 for details). The exchange matrix of Eq. (11) is further approximated according to

$$\tilde{K}_{ab}=\frac{1}{2}\sum_i\left(\sum_{\alpha\in[A]}c^{ai}_\alpha(\alpha|bi)+\sum_{\beta\in[B]}(ai|\beta)c^{bi}_\beta\right). \tag{20}$$

We note that this approach is nonvariational, which is justified by reporting errors in energies using Eq. (20) that are typically only twice those of regular density fitting of exchange.

D. Local robust and variational fitting of four-center integrals

Let us consider the robust and variational fitting of the two-electron integrals $(ab|cd)$ in a general metric. We denote the fitted overlap distributions and their (negative) errors by

$$\widetilde{\langle ab|}=\sum_\alpha c^{ab}_\alpha\langle\alpha|,\quad \langle\Delta ab|=\langle ab|-\widetilde{\langle ab|}, \tag{21}$$

$$\widetilde{|cd\rangle} = \sum_{\beta} c_{\beta}^{cd}|\beta\rangle, \quad |\Delta cd\rangle = |cd\rangle - \widetilde{|cd\rangle}. \tag{22}$$

Following Dunlap,[13] a robust integral fitting is given by

$$\begin{aligned}(\widetilde{ab|cd}) &= (ab|\widetilde{cd}) + (\widetilde{ab}|cd) - (\widetilde{ab}|\widetilde{cd}) \\ &= (ab|cd) - (\Delta ab|\Delta cd),\end{aligned} \tag{23}$$

which is manifestly quadratic in the fitting errors. The fitting coefficients c_{α}^{ab} are obtained by minimizing the self-interaction energy of the fitting errors,

$$D_{abcd}^{w} = \langle \Delta ab|w|\Delta cd\rangle, \tag{24}$$

in a metric w, possibly different from the Coulomb metric, leading to the linear equations

$$\langle \Delta ab|w|\beta\rangle = 0, \quad \langle \alpha|w|\Delta cd\rangle = 0. \tag{25}$$

These equations are sparse when local metric and basis functions are used, allowing for an iterative solution in time proportional to system size. To make the integral [Eq. (23)] variational in the fitting coefficients, we use Lagrange's method of undetermined multipliers, treating Eq. (25) as constraints on the integral. Multiplying these constraints by $\bar{c}_{\alpha}^{ab}$ and $\bar{c}_{\beta}^{cd}$ and adding the resulting expressions to Eq. (23), we obtain

$$\begin{aligned}(\widetilde{ab|cd}) = (ab|\widetilde{cd}) + (\widetilde{ab}|cd) - (\widetilde{ab}|\widetilde{cd}) - \langle \overline{ab}|w|\Delta cd\rangle \\ - \langle \Delta ab|w|\overline{cd}\rangle,\end{aligned} \tag{26}$$

in the notation

$$\langle \overline{ab}| = \sum_{\alpha} \bar{c}_{\alpha}^{ab}\langle \alpha|, \quad |\overline{cd}\rangle = \sum_{\beta} \bar{c}_{\beta}^{cd}|\beta\rangle. \tag{27}$$

Differentiating Eq. (26) with respect to the fitting coefficients and setting the result equal to zero, we obtain the following linear equations for the multipliers:

$$\langle \overline{ab}|w|\beta\rangle = (\Delta ab|\beta), \quad \langle \alpha|w|\overline{cd}\rangle = (\alpha|\Delta cd), \tag{28}$$

that must be solved to make the integrals variational in all parameters. Because of Eq. (25), these terms do not make a contribution to the unperturbed integrals [Eq. (26)]. However, they do become important for the calculation of molecular properties, as discussed in Sec. III, where we consider the linear-scaling evaluation of the two-electron contributions to molecular gradients.

III. LOCAL FITTING OF COULOMB AND EXACT EXCHANGE

The variational fitting of four-center integrals [Eq. (26)] can be applied to all *ab initio* methods. We here establish explicit expressions for this approach applied to the two-electron Coulomb and exact-exchange contributions in HF and KS theories. The developed theory allows for linear scaling robust variational density fitting of these two contributions in local metrics. We further show how this theory applies to molecular properties.

A. Coulomb energy and the Coulomb matrix

In the notation

$$|\rho\rangle = \sum_{ab} D_{ab}|ab\rangle, \tag{29}$$

the Coulomb repulsion energy is given by

$$J = \frac{1}{2}\sum_{abcd} D_{ab}(ab|cd)D_{cd} = \frac{1}{2}(\rho|\rho). \tag{30}$$

To obtain the corresponding expression with fitted integrals, we replace the integrals $(ab|cd)$ by $(\widetilde{ab|cd})$ of Eq. (26) and obtain

$$\tilde{J} = \frac{1}{2}\sum_{abcd} D_{ab}(\widetilde{ab|cd})D_{cd} = (\rho|\tilde{\rho}) - \frac{1}{2}(\bar{\rho}|\bar{\rho}) - \langle \bar{\rho}|w|\Delta\rho\rangle, \tag{31}$$

where $|\Delta\rho\rangle = |\rho\rangle - |\tilde{\rho}\rangle$ and

$$|\tilde{\rho}\rangle = \sum_{ab} D_{ab}\widetilde{|ab\rangle} = \sum_{\alpha} c_{\alpha}|\alpha\rangle, \quad c_{\alpha} = \sum_{ab} c_{\alpha}^{ab}D_{ab}, \tag{32}$$

$$|\bar{\rho}\rangle = \sum_{ab} D_{ab}|\overline{ab}\rangle = \sum_{\alpha} \bar{c}_{\alpha}|\alpha\rangle, \quad \bar{c}_{\alpha} = \sum_{ab} \bar{c}_{\alpha}^{ab}D_{ab}. \tag{33}$$

In Eq. (31), the last term vanishes by Eq. (25) but it is retained to obtain a variational expression for the fitted two-electron Coulomb repulsion energy, which is important for the calculation of, for example, molecular gradients.

The Fock/KS matrix is the first derivative of the HF/KS energy with respect to the density matrix elements. Therefore, the two-electron Coulomb contribution to the Fock/KS matrix, the Coulomb matrix, is given by

$$J_{ab} = \partial J/\partial D_{ab} = (ab|\rho). \tag{34}$$

We get the fitted Coulomb matrix by differentiation of the approximate Coulomb repulsion energy of Eq. (31) or simply by replacing the four-center integrals of Eq. (34) with the approximate integrals [Eq. (26)], giving

$$\tilde{J}_{ab} = \partial\tilde{J}/\partial D_{ab} = (ab|\tilde{\rho}) + (\widetilde{ab}|\rho) - (\widetilde{ab}|\tilde{\rho}), \tag{35}$$

where we have omitted the w terms, which do not contribute. The fitted Coulomb matrix can be calculated in linear time by standard direct integral evaluation routines, using Cauchy–Schwarz screening and the continuous fast multiple method (see, for example, Ref. 19 and references therein). A linear-scaling implementation also requires that the coefficients c_{β}^{ab} are determined as accurately as possible, with a resource usage proportional to system size, as can be achieved by solving Eq. (25) in a local metric.

B. Exchange energy and exchange contribution to Fock/KS matrix

We now turn our attention to exchange. The exchange energy is given as

$$K = \frac{1}{2}\sum_{abcd} D_{ab}(ac|bd)D_{cd} = \frac{1}{2}\sum_{ij}^{\text{occ}} (ij|ij), \tag{36}$$

where i and j denote the occupied MOs. Proceeding as for the Coulomb energy, we obtain

$$\tilde{K} = \frac{1}{2}\sum_{abcd} D_{ab}(\widetilde{ac|bd})D_{cd}$$

$$= \frac{1}{2}\sum_{ij} [2(ij|\widetilde{ij}) - (\widetilde{ij}|\widetilde{ij}) - 2\langle \overline{ij}|w|\Delta ij\rangle], \quad (37)$$

where we have introduced $|\Delta ij\rangle = |ij\rangle - |\widetilde{ij}\rangle$ and

$$|\widetilde{ij}\rangle = \sum_{\alpha} c_{\alpha}^{ij}|\alpha\rangle, \quad c_{\alpha}^{ij} = \sum_{ab} c_{\alpha}^{ab} C_{ai} C_{bj}, \quad (38)$$

$$|\overline{ij}\rangle = \sum_{\alpha} \bar{c}_{\alpha}^{ij}|\alpha\rangle, \quad \bar{c}_{\alpha}^{ij} = \sum_{ab} \bar{c}_{\alpha}^{ab} C_{ai} C_{bj}. \quad (39)$$

As for the Coulomb energy [Eq. (31)], the last term in Eq. (37) is zero but is retained since it contributes to gradients. The exchange matrix is the derivative of the exchange energy [Eq. (36)] with respect to the density-matrix elements D_{ab},

$$K_{ab} = \partial K/\partial D_{ab} = \sum_{cd} (ac|bd)D_{cd} = \sum_{i}^{\text{occ}} (ai|bi). \quad (40)$$

In the same manner as for the Coulomb energy, we obtain the density-fitted expressions for the exchange energy and matrix,

$$\tilde{K}_{ab} = \partial \tilde{K}/\partial D_{ab}$$

$$= \sum_{cd} (\widetilde{ac|bd})D_{cd}$$

$$= \sum_{i} [(ai|\widetilde{bi}) + (\widetilde{ai}|bi) - (\widetilde{ai}|\widetilde{bi})], \quad (41)$$

where the notation for $|\widetilde{ai}\rangle$ and $|\overline{ai}\rangle$ is analogous to that of Eqs. (38) and (39). When the fitting coefficients c_{α}^{ai} are obtained in the Coulomb metric, the last two terms vanish to give the expression of the fitted exchange matrix of Polly *et al.*[16] given by Eq. (16).

The density-matrix elements couple basis functions on the two electrons. This coupling, together with screening, is exploited for insulators in the order-N exchange[20] and in the linear-scaling exchange[21] (LinK) algorithms to achieve linear scaling with system size in Eq. (40). An alternative approach is to use localized molecular orbitals (LMOs) (see Ref. 22 and references therein). Linear scaling then follows by using these LMOs and Cauchy–Schwarz screening since, provided that the AOs χ_a and χ_b are sufficiently far away from each other, a given LMO will not overlap with both AOs. To see this, we apply the Cauchy–Schwarz inequality twice,

$$|(ai|bi)| \leqslant \sqrt{(ai|ai)}\sqrt{(bi|bi)}$$

$$\leqslant \left[\sum_{c} |C_{ci}|\sqrt{(ac|ac)}\right]\left[\sum_{c} |C_{ci}|\sqrt{(bc|bc)}\right], \quad (42)$$

where we have used

$$(ai|ai) = \sum_{cd} C_{ci}C_{di}(ac|ad)$$

$$\leqslant \sum_{cd} |C_{ci}||C_{di}|\sqrt{(ac|ac)}\sqrt{(ad|ad)}$$

$$= \left[\sum_{c} |C_{ci}|\sqrt{(ac|ac)}\right]^2, \quad (43)$$

and so on.

For insulators, linear-scaling density-fitted exchange-matrix construction can be achieved in a local metric by following the same arguments as for the regular exchange matrix and by pretabulating which three-center Coulomb repulsion integrals $(ab|\alpha)$ [or $(ai|\alpha)$] to calculate. First, we note that, in a local metric, the number of fitting coefficients c_{α}^{ab} scales linearly with system size, as auxiliary basis functions $\xi_{\alpha}(\mathbf{r})$ sufficiently far away from the product overlaps $\Omega_{ab}(\mathbf{r})$ do not contribute to the fitted product overlap $\tilde{\Omega}_{ab}(\mathbf{r})$.[15] Second, since D_{cd} couple basis functions on two different electrons, $\chi_c(\mathbf{r}_1)$ and $\chi_d(\mathbf{r}_2)$, we can neglect all integrals $(ac|bd)$ where the density-matrix elements become sufficiently small for example, using Cauchy–Schwarz screening

$$|(ac|bd)D_{cd}| \leqslant \sqrt{(ac|ac)}\sqrt{(bd|bd)}|D_{cd}|. \quad (44)$$

Therefore, the fitted integrals $(\widetilde{ac|bd})$ of $(ac|bd)$ need only be calculated whenever

$$\sqrt{(ac|ac)}\sqrt{(bd|bd)}|D_{cd}| \geqslant \epsilon, \quad (45)$$

for a given threshold ϵ. For insulators, the density-matrix elements decrease in magnitude with increasing distance, which means, for instance, that $\Omega_{ac}(\mathbf{r}_1)$ only interact with $\tilde{\Omega}_{bd}(\mathbf{r}_2)$ provided that $\chi_c(\mathbf{r}_1)$ and $\chi_d(\mathbf{r}_2)$ are within some finite distance of each other. As a result, $\chi_a(\mathbf{r}_1)$ and $\chi_b(\mathbf{r}_2)$ must also be close to each other. The same argument applies to the fitting functions since $\xi_{\alpha}(\mathbf{r}_2)$, included in $\tilde{\Omega}_{bd}(\mathbf{r}_2)$, have a limited extent from the center of $\Omega_{bd}(\mathbf{r}_2)$, from which $\tilde{\Omega}_{bd}(\mathbf{r}_2)$ originates. The combined effects of locality in the density matrix and locality in the fit imply that the number of contributing three-center integrals $(ac|\alpha)$ scales linearly with system size. The same argument holds for the two-center integrals appearing in the last term of Eq. (41).

C. Contributions to gradient

To conclude this section, we make a note on how to achieve linear scaling for the exchange contribution when calculating properties (such as the molecular gradient) that involve explicit differentiation of the four-center integrals $(ac|bd)$. Let η denote some variable, for example, a nuclear coordinate. Differentiation of the fitted exchange energy of Eq. (37) with respect to η gives

$$\frac{d\tilde{K}}{d\eta} = \sum_{ab} D_{ab}^{\eta}\tilde{K}_{ab} + \sum_{ab} D_{ab}\tilde{K}_{ab}^{\eta} + \sum_{ab} D_{ab}\bar{K}_{ab}^{\eta}, \quad (46)$$

with

Initialization

Non-Coulombic metric w

Construct $G^w_{ab} = \sqrt{\langle ab|w|ab\rangle}$ and $G^w_\alpha = \sqrt{\langle \alpha|w|\alpha\rangle}$

Normalize $\{\xi_\alpha\}$ in metric w (i.e. division with G^w_α)

Construct $\langle\alpha|w|\beta\rangle$ and decompose to $\langle\alpha|w|\beta\rangle^{\pm\frac{1}{2}}$

Construct $\langle ab|w|\alpha\rangle \geq G^w_{ab}G^w_\alpha = G^w_{ab}$

Orthogonalize the auxiliary basis according to $c^{ab}_{\alpha'} = \langle ab|w|\alpha'\rangle = \sum_\alpha \langle ab|w|\alpha\rangle \langle\alpha|w|\alpha'\rangle^{-\frac{1}{2}}$

Coulombic metric

Construct $G_{ab} = \sqrt{(ab|ab)}$ and $G_\alpha = \sqrt{(\alpha|\alpha)}$

Construct $(\alpha|\beta)$ and orthogonalize in metric w, $(\alpha'|\beta') = \langle\alpha'|w|\alpha\rangle^{-\frac{1}{2}}(\alpha|\beta)\langle\beta|w|\beta'\rangle^{-\frac{1}{2}}$

Construct $(ab|\alpha) \geq G_{ab}G_\alpha$

Orthogonalize according to $(ab|\alpha') = (ab|\alpha)\langle\alpha|w|\alpha'\rangle^{-\frac{1}{2}}$

Each iteration

Fitted Coulomb matrix $\tilde{J}_{ab}$

Construct $c_{\alpha'} = \sum_{cd} D_{cd}\langle cd|w|\alpha'\rangle = \sum_{cd} D_{cd}c^{cd}_{\alpha'}$

Compute intermediate $\tilde{J}^I_{ab} = \sum_{\alpha'}(ab|\alpha')c_{\alpha'}$

Construct $\gamma_{\alpha'} = \sum_{cd}(\alpha'|cd)D_{cd} - \sum_{\beta'}(\alpha'|\beta')c_{\beta'}$

Construct $\tilde{J}_{ab} = \tilde{J}^I_{ab} + \sum_{\alpha'} c^{ab}_{\alpha'}\gamma_{\alpha'}$

Fitted exchange matrix $\tilde{K}_{ab}$

Construct Cholesky MO's by Cholesky decomposition of density matrix, $D_{ab} = \sum_i^{\text{occ}} L_{ai}L_{bi}$

MO half-transform according to $c^{ai}_{\alpha'} = \sum_b c^{ab}_{\alpha'}L_{bi}$, and $(ai|\alpha') = \sum_b L_{bi}(ab|\alpha')$

Build intermediate $\tilde{K}^I_{ab} = \sum_{i\alpha'}(ai|\alpha')c^{bi}_{\alpha'}$

Build intermediate $\tilde{K}^{II}_{ab} = \tilde{K}^I_{ab} + \sum_{i\alpha'} c^{ai}_{\alpha'}(bi|\alpha')$

Finalize fitted exchange matrix $\tilde{K}_{ab} = \tilde{K}^{II}_{ab} - \sum_{i\alpha'\beta} c^{ai}_{\alpha'}(\alpha'|\beta')c^{bi}_{\beta'}$

FIG. 1. Outline of the algorithm employed for fitting the Coulomb and exchange matrices in local metrics.

$$D^\eta_{ab} = \frac{dD_{ab}}{d\eta} \tag{47}$$

and

$$\tilde{K}^\eta_{ab} = \sum_{cd}\left(\sum_\beta (\{ac\}^\eta|\beta)c^{bd}_\beta + \sum_\alpha c^{ac}_\alpha(\alpha^\eta|\Delta bd)\right)D_{cd}, \tag{48}$$

and the term including the Lagrangian multipliers

$$\bar{K}^\eta_{ab} = \sum_{cd}\left(\sum_\alpha \lambda^{ac}_\alpha[\langle\alpha^\eta|w|\Delta bd\rangle + \langle\alpha|w|\{\Delta bd\}^\eta\rangle]\right)D_{cd}. \tag{49}$$

In the last two equations the superscript η denotes differentiation with respect to η, so that, for example,

$$\langle\alpha^\eta|w|\Delta bd\rangle = \int \frac{d\xi_\alpha(\mathbf{r}_1)}{d\eta} w(\mathbf{r}_1,\mathbf{r}_2)(\Omega_{bd}(\mathbf{r}_2) - \tilde{\Omega}_{bd}(\mathbf{r}_2))d\mathbf{r}_1 d\mathbf{r}_2. \tag{50}$$

Linear scaling of the two first terms of Eq. (46) follows the same arguments as for the undifferentiated case, whereas linear scaling of the third term requires insertion of the expression for the Lagrangian multipliers of Eq. (28),

$$\bar{K}^\eta_{ab} = \sum_{cd}\left(\sum_{\alpha\beta}(\Delta ac|\alpha)\langle\alpha|w|\beta\rangle^{-1}[\langle\beta^\eta|w|\Delta bd\rangle + \langle\beta|w|\{\Delta bd\}^\eta\rangle]\right)D_{cd}. \tag{51}$$

Linear scaling can be achieved by letting the inverse $\langle\alpha|w|\beta\rangle^{-1}$ matrix work to the right rather than the left—thereby bypassing explicitly solving for the Lagrangian multipliers.

IV. IMPLEMENTATION DETAILS

Figure 1 outlines the algorithm employed in this paper for the construction of the fitted Coulomb and exchange matrices following Eqs. (35) and (41), respectively. To condition the linear set of equations optimally, we orthonormalize the auxiliary basis functions—that is, we normalize in metric w using $g^w_\alpha = \sqrt{\langle\alpha|w|\alpha\rangle}$ and orthogonalize by multiplication of the inverse square root of the auxiliary two-center integrals $\mathbf{V}^w$,

$$V^w_{\alpha,\beta} = \langle\alpha|w|\beta\rangle. \tag{52}$$

The inverse square root $(\mathbf{V}^w)^{-1/2}$ is obtained with the scheme presented in Ref. 23. In the orthogonal basis $\xi_{\alpha'} = \Sigma_\alpha \xi_\alpha \langle\alpha|w|\alpha'\rangle^{-1/2}$, we thus have

$$V^{w}_{\alpha',\beta'} = \langle \alpha' | w | \beta' \rangle = \sum_{\alpha\beta} \langle \alpha' | w | \alpha \rangle^{-1/2} \langle \alpha | w | \beta \rangle \langle \beta | w | \beta' \rangle^{-1/2} = \delta_{\alpha',\beta'}. \quad (53)$$

The three center integrals are calculated using Cauchy–Schwarz screening,

$$|\langle \alpha | w | ab \rangle| \leqslant g^{w}_{ab} g^{w}_{\alpha}, \quad (54)$$

with

$$g^{w}_{f} = \sqrt{\langle f | w | f \rangle}. \quad (55)$$

More specifically, we only calculate the three-center integrals $\langle \alpha | w | ab \rangle$ if $g_{ab} \geqslant \tau / g^{w}_{\alpha}$ for a given threshold τ. Furthermore, the three-center integrals are packed in triangular form to exploit the symmetry of the integrals $\langle a | w | ab \rangle = \langle \alpha | w | ba \rangle$. We do not, in the current implementation, exploit the sparsity obtained in a local metric, although, in Sec. V, we report this sparsity for a selected system.

In the orthogonal basis, the fitting coefficients of Eq. (25) reduce to the three-center integrals,

$$c^{ab}_{\alpha'} = \langle \alpha' | w | ab \rangle. \quad (56)$$

The construction of the fitted Coulomb and exchange matrices, given by Eqs. (35) and (41), respectively, follows straightforwardly by contracting the fitting coefficients with the three-center integrals $(\alpha' | ab)$ and $(\alpha' | \beta')$. However, to speed up the construction of the fitted exchange matrix, we first MO half-transform both the fitting coefficients,

$$c^{ai}_{\alpha'} = \sum_{b} c^{ab}_{\alpha'} L_{bi}, \quad (57)$$

and the three-center Coulomb repulsion integrals,

$$(ai | \alpha') = \sum_{b} (ab | \alpha') L_{bi}, \quad (58)$$

using the localized Cholesky MO coefficients L_{ai} obtained by the incomplete Cholesky decomposition of the density matrix,[22]

$$D_{ab} = \sum_{i} L_{ai} L_{bi}. \quad (59)$$

A. Integral evaluation

In this subsection, we provide a brief overview on how we evaluate the molecular integrals in the different metrics $w(\mathbf{r}_1, \mathbf{r}_2)$ used to determine the fitting coefficients of Eq. (25). Several general integration schemes are available in literature (see, for instance, Ref. 24). The current implementation is part of a development version of DALTON, in which the McMurchie–Davidson scheme forms the basis for integral evaluation.[25] In the McMurchie–Davidson scheme, the product overlap distribution between two (spherical-harmonic) basis functions is expanded in Hermite Gaussian primitives $\Lambda^{\mathbf{P}}_{tuv}(\mathbf{r})$, according to

$$\Omega_{ab}(\mathbf{r}) = \sum_{tuv} E^{ab}_{tuv} \Lambda^{\mathbf{P}}_{tuv}(\mathbf{r}), \quad \Lambda^{\mathbf{P}}_{tuv}(\mathbf{r}) = \frac{\partial^{t+u+v} e^{-p r^2_{\mathbf{P}}}}{\partial P^{t}_{x} \partial P^{u}_{y} \partial P^{v}_{z}}, \quad (60)$$

with $p = a + b$, and $\mathbf{P} = (a\mathbf{A} + b\mathbf{B})/p$ (see Ref. 24 for details).

The two electron integral between two such overlap distributions is, in metric $w(\mathbf{r}_1, \mathbf{r}_2)$, given by

$$\langle ab | w | cd \rangle = \frac{2\pi^{5/2}}{pq\sqrt{p+q}} \sum_{tuv} E^{ab}_{tuv} \times \sum_{\tau\nu\phi} (-1)^{\tau+\nu+\phi} E^{cd}_{\tau\nu\phi} W_{t+\tau,u+\nu,v+\phi}(\gamma, \mathbf{R}_{PQ}), \quad (61)$$

with $\gamma = pq/(p+q)$ and where $\mathbf{R}_{PQ}$ refers to the distance between the two overlap distributions. In this expression, only the Hermite two-electron integral,

$$W_{t,u,v}(\gamma, \mathbf{R}_{PQ}) = \frac{\partial^{t+u+v} W(\gamma, \mathbf{R}_{PQ})}{\partial P^{t}_{x} \partial P^{u}_{y} \partial P^{v}_{z}}, \quad (62)$$

depends on the metric $w(\mathbf{r}_1, \mathbf{r}_2)$ with, for example,

$$W(\gamma, \mathbf{R}_{PQ}) = \begin{cases} F_0(\gamma R^2_{PQ}) & \text{for } w(\mathbf{r}_1, \mathbf{r}_2) = 1/r_{12} \\ \gamma/(2\pi) \exp(-\gamma R^2_{PQ}) & \text{for } w(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2) \\ F_0(\gamma R^2_{PQ}) - \sqrt{\delta} F_0(\gamma \delta R^2_{PQ}) & \text{for } w(\mathbf{r}_1, \mathbf{r}_2) = \operatorname{erfc}(\omega r_{12})/r_{12} \\ \kappa F_0(\kappa\gamma R^2_{PQ}) \exp(-\kappa\omega R^2_{PQ}) & \text{for } w(\mathbf{r}_1, \mathbf{r}_2) = \exp(-\omega r^2_{12})/r_{12}. \end{cases} \quad (63)$$

Here $\delta = \omega^2/(\gamma+\omega)$, $\kappa = \gamma/(\gamma+\omega)$, $F_0(x)$ is the zeroth order Boys function, and ω is the attenuation parameter. The Hermite two-electron integral $W_{t,u,v}(\gamma, \mathbf{R}_{PQ})$ can be found by reccurence.[24,26] Note that for both the attenuated Coulomb metrics of Eq. (63), the complementary error-function Coulomb metric $\operatorname{erfc}(\omega r_{12})/r_{12}$, and the Gaussian damped Coulomb metric $\exp(-\omega r^2_{12})/r_{12}$, we retain the Coulomb integrals in the limit $\omega \to 0$, whereas in the limit $\omega \to \infty$, we get scaled overlap integrals (with prefactors π/ω^2 and $2\pi/\omega$, respectively). Also note that the auxiliary basis functions used for density fitting may, similarly to the expansion of the overlap distributions Eq. (60), be expanded in Hermite

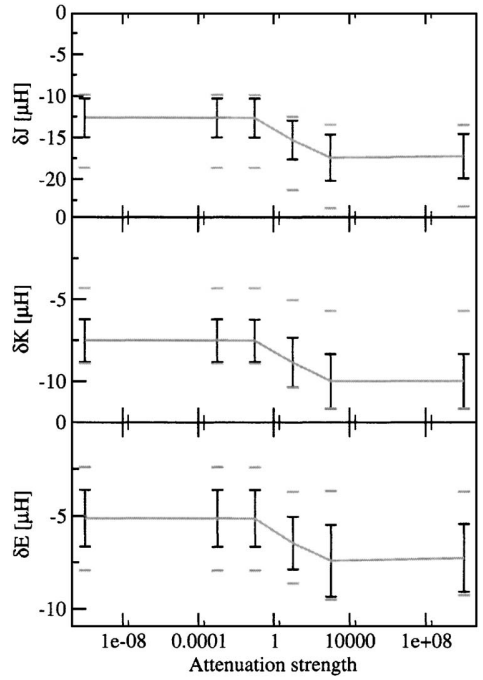

FIG. 2. Average value and standard deviations for the fitting errors in Coulomb δJ, exchange δK, and total energy δE as computed for the benchmark molecule set as a function of varying attenuation strength ω. The basis-set combination cc-pVTZ(df-pVTZ) was used. Errors are computed per nonhydrogen atom, and results were obtained using the Gaussian damped Coulomb metric $\exp(-\omega r_{12}^2)/r_{12}$. Maximal and minimal deviations are marked with bars.

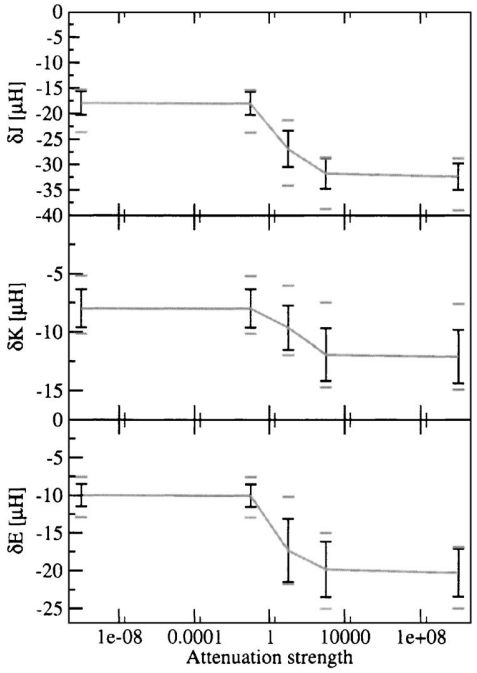

FIG. 3. Average value and standard deviations for the fitting errors in Coulomb δJ, exchange δK, and total energy δE as computed for the benchmark molecule set as a function of varying attenuation strength ω. The basis-set combination 6-31G(df-def2) was used. Errors are computed per nonhydrogen atom, and results were obtained using the Gaussian damped Coulomb metric $\exp(-\omega r_{12}^2)/r_{12}$. Maximal and minimal deviations are marked with bars.

Gaussians. Therefore, the arguments given in this subsection also apply to two-electron integrals involving auxiliary basis functions. Finally note that to speed up the integral evaluation, we use Hermite instead of Cartesian primitive functions for the auxiliary basis functions according to Ref. 27.

V. RESULTS AND DISCUSSION

To assess the presented method with respect to accuracy, we shall now examine the errors introduced in the calculated energies, atomization energies, and reaction enthalpies for a set of test systems. We demonstrate that local density fitting can be applied to molecular energies, at little cost of accuracy. We also take a look at energy differences, presenting results for both atomization energies and reaction enthalpies. The errors in energy differences are more sensitive than the errors in molecular energies when making the transition from Coulomb to overlap density fitting. Although the density-fitting errors are still small, compared to, for example, orbital basis-set errors, the somewhat larger errors for energy differences may constitute a criticism of the presented method. We therefore discuss these issues in greater detail. We finally make a note on computation performance as well as on the sparsity for different screening thresholds.

A. Molecular errors

Figures 2 and 3 display the effect of attenuation on Coulomb, exchange, and total energies at different levels of attenuation ω. Results are for the benchmark set of Ref. 28 at the B3LYP/cc-pVTZ(df-pVTZ) and B3LYP/6-31G(df-def2) levels of theory. The auxiliary basis sets used for density fitting are given in parentheses; df-pVTZ is the triple-zeta valence plus polarization basis set of Ref. 2 and df-def2 is the standard "RI-JK auxiliary basis set" of Ref. 3. Mean errors in the Coulomb δJ, exchange δK, and total energies δE for the full benchmark set are plotted together with the corresponding standard deviations, maximum errors, and minimum errors. The plots were obtained using the Gaussian damped Coulomb metric $\exp(-\omega r_{12}^2)/r_{12}$. The limit of a small ω corresponds to Coulomb fitting, while a large attenuation factor ω approaches overlap fitting. Concerning the choice of a local metric, we note that the performance of the Gaussian and error-function damping is similar with respect to size of the errors. Since Gaussian damping gives rise to more sparsity, it is recommended over error-function damping for large systems.

Inserting Eq. (23) into Eqs. (31) and (37), we obtain

$$\tilde{J} = J + \delta J = J - \tfrac{1}{2}(\Delta\rho|\Delta\rho), \tag{64}$$

$$\tilde{K} = K + \delta K = K - \frac{1}{2}\sum_{ij}(\Delta ij|\Delta ij), \tag{65}$$

and conclude that the density-fitting errors in the Coulomb and exchange energies are both negative. The sign of the total fitting error $\delta E = \delta J - \delta K$ therefore depends on the relation between the Coulomb δJ and the exchange δK errors.

In the B3LYP calculations examined here, the density-fitting error in the Coulomb contribution is larger than the error in the exchange contribution. In Hartree–Fock calcula-

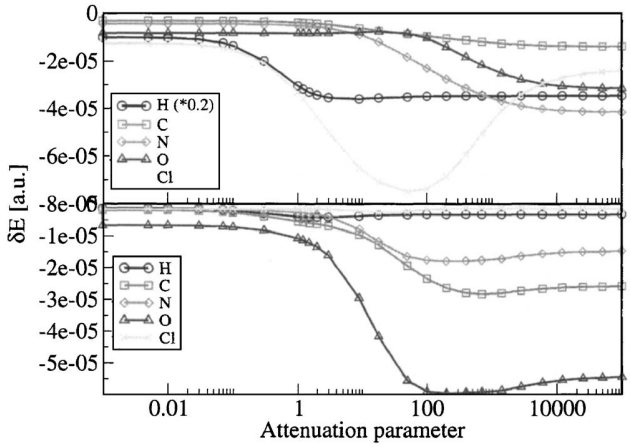

FIG. 4. Total density-fitting error δE obtained for hydrogen, carbon, nitrogen, oxygen, and chlorine atoms at B3LYP density-functional level as a function of Gaussian attenuation strength ω. Top panel displays the fitting error obtained for 6-31G(df-def2) basis-set combination—bottom panel for cc-pVTZ(df-pVTZ) combination. Error for hydrogen and 6-31G(df-def2) combination was scaled down five times.

tions, where the contribution from exact exchange is five times larger than in the B3LYP calculations, the exchange error is two to three times larger than the Coulomb error. Since the Coulomb and exchange errors are both negative, the total error is never larger than the error in one of the contributions. As seen from Figs. 2 and 3, attenuation increases the molecular fitting errors but never by more than 50%. Since the fitting errors are much smaller than the basis-set error, we conclude that attenuated local fitting can be used instead of Coulomb fitting, in molecular calculations, without adversely affecting the resulting total energies.

B. Atomic errors

In this section we address both atomic and atomization density-fitting energies. These two quantities are more sensitive than molecular energies to the transition from Coulomb to overlap density fitting. We attribute this to an auxiliary basis-set superposition error.

In Fig. 4, we have plotted the density-fitting error as a function of attenuation parameter ω for the atoms in the benchmark set.[28] Clearly, the atomic calculations behave differently from the molecular ones. In the atomic calculations, the transition from the Coulomb to the overlap metric increases the fitting error by up to a factor of 8. Moreover, the atomic errors do not increase monotonically as we approach the overlap metric. Instead, the largest fitting error occurs for some intermediate value of ω. Clearly, the attenuation error depends strongly on the auxiliary basis set.

TABLE I. Errors in atomization energies (μhartree) arising from Coulomb and overlap density fitting for N_2 and CO at the B3LYP/cc-pVTZ(df-pVTZ) level of theory, with and without use of the CP correction. The calculations have been carried out at the experimental bond distances of 109.768 pm for N_2 and 112.8323 pm for CO.

	Coulomb		Overlap	
Molecule	No CP	CP	No CP	CP
N_2	21	20	46	26
CO	22	20	52	1

In molecules, auxiliary basis functions on neighboring atoms help to lower the fitting error. These additional functions are unavailable in atomic calculations, giving an unbalanced description of atomic and molecular systems and an associated basis-set superposition error (BSSE) in the energies. The BSSE can, at least to some extent, be corrected for by application of the counterpoise (CP) correction of Ref. 29.

In the case of density fitting such a BSSE effect would be prominent due to limited flexibility of the auxiliary basis set. To examine the BSSE associated with the auxiliary basis set, we have applied the CP correction to atomization-energy calculations on N_2 and CO. The results in Table I show that auxiliary BSSE is more prominent in the overlap metric than in the Coulomb metric. In the Coulomb metric, the CP correction has little effect on the atomization energies, whereas overlap density-fitting errors are reduced from 46 to 26 μhartrees for N_2 and from 52 to 1 μhartree for CO. Clearly, the latter value is an example of a fortuitous cancellation of errors.

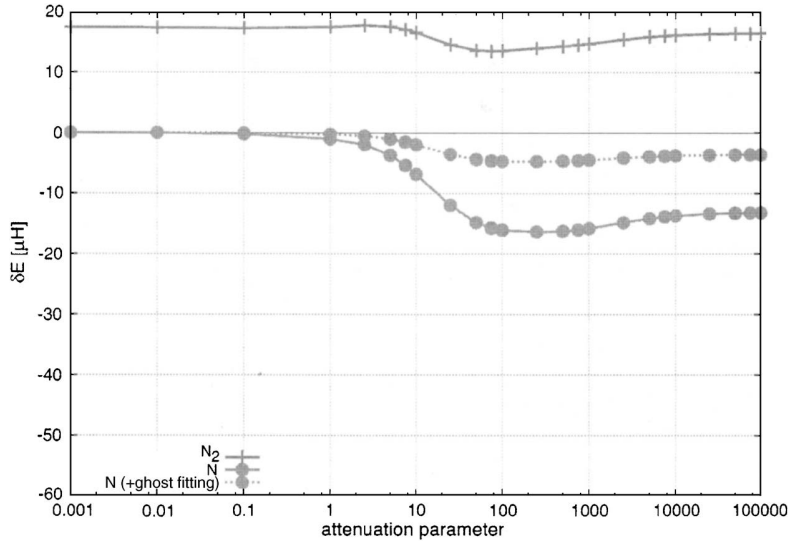

FIG. 5. Variation of the total density-fitting error δE with the attenuation parameter ω for the N_2 molecule and the N atom at B3LYP/cc-pVTZ(df-pVTZ) level of theory. The experimental bond length of 109.768 pm was used for N_2. Results were obtained using the Gaussian damped metric $\exp(-\omega r_{12}^2)/r_{12}$. The atomic calculations labeled "ghost fitting" involve regular (orbital) basis functions for one of the atoms and auxiliary basis functions at the positions of both atoms.

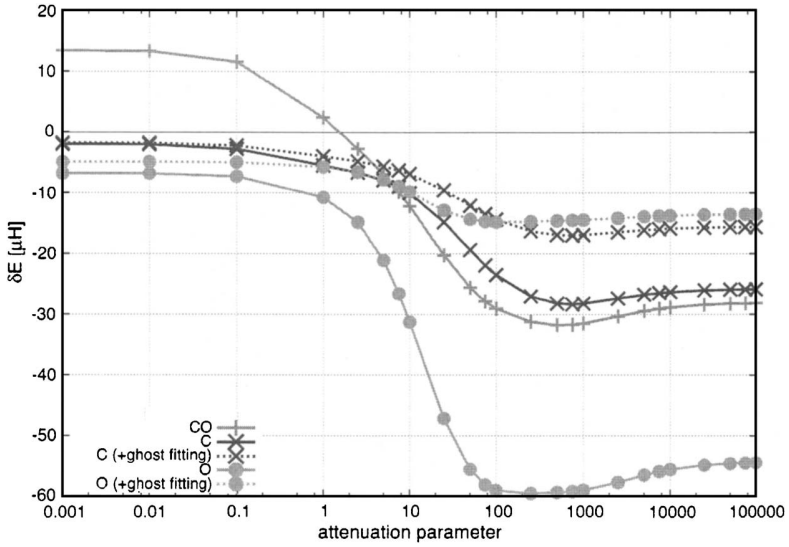

FIG. 6. Variation of the total density-fitting error δE with the attenuation parameter ω for the CO molecule and the C and O atoms at B3LYP/cc-pVTZ(df-pVTZ) level of theory. The experimental bond length of 112.8323 pm was used for CO. Results were obtained using the Gaussian damped metric $\exp(-\omega r_{12}^2)/r_{12}$. The atomic calculations labeled "ghost fitting" involve regular (orbital) basis functions for the atom in question and auxiliary basis functions at the positions of both atoms.

Figures 5 and 6 show the B3LYP/cc-pVTZ(df-pVTZ) density-fitting errors at different levels of attenuation for the two molecules, as well as atomic energies calculated with and without ghost fitting functions at the positions of the bonding partners. The atomic fitting errors are reduced substantially once the flexibility of the auxiliary basis set is enhanced. In particular, for the oxygen atom, the ratio between the overlap and Coulomb density-fitting errors is reduced from 7.7 to 2.7 when the ghost fitting functions are included. Moreover, with ghost functions in the atomic calculations, the dependency of the fitting errors of the attenuation parameter is less pronounced, as seen from the reduced slope of the corresponding curves in Figs. 5 and 6.

In summary, the BSSE is more pronounced in the attenuated and overlap metrics than in the Coulomb metric. However, we note that the auxiliary basis set used in this investigation were optimized with respect to density fitting in the Coulomb metric. The use of auxiliary basis sets optimized in the overlap metric may reduce the BSSE. However, even with the standard auxiliary basis sets used here, the density-fitting error is small compared with the orbital basis-set error.

C. Reaction energies

From a chemical point of view, relative energies are more important than total energies. To obtain reliable reaction energies for a given method, the errors of products and reactants must be balanced. We tested our approach on 11 reaction energies (A–K) at the B3LYP/6-31G(df-def2) level of theory (see Table II). The test set includes isomerization reactions (A–C), bond-breaking reactions leading to closed-shell species (D–G), and bond-breaking reactions leading to open-shell species (H–K). The geometries for all species

TABLE II. Densit-fitting errors δE (μhartree) of reaction energies for reactions A–K in the overlap and Coulomb metris. Also listed are the sum of the density-fitting error of the reactants δE_{reac} and of the products δE_{prod}. Calculations were carried out at the B3LYP/6-31G(df-def2) level of theory.

		δE		δE_{reac}		δE_{prod}	
		Over.	Coul.	Over.	Coul.	Over.	Coul.
A: $C_{12}H_{12}$ (1)	$\rightarrow C_{12}H_{12}$ (2)	66	17	−268	−144	−202	−127
B: $C_{12}H_{12}$ (1)	$\rightarrow C_{12}H_{12}$ (3)	13	25	−268	−144	−254	−119
C: $(CH_3)_3CC(CH_3)_3$	$\rightarrow n-C_8H_{18}$	−33	−21	−113	−103	−146	−124
D: $n-C_6H_{14}+4\ CH_4$	$\rightarrow 5\ C_2H_6$	37	3	−287	−199	−250	−195
E: $n-C_8H_{18}+6\ CH_4$	$\rightarrow 7\ C_2H_6$	59	5	−410	−280	−350	−273
F: adamantane	$\rightarrow 3\ C_2H_4+2\ C_2H_2$	−90	−8	−132	−117	−222	−125
G: bicyclo[2.2.2]octane	$\rightarrow 3\ C_2H_4+C_2H_2$	−63	−8	−133	−101	−195	−109
H: CH_3OCH_3	$\rightarrow CH_3O+CH_3$	29[a]	4[b]	−101[c]	−59[d]	−72[e]	−55[f]
I: $CH_3OCH_2CH_3$	$\rightarrow CH_3O+CH_2CH_3$	18[a]	4[b]	−104[c]	−70[d]	−87[e]	−66[f]
J: $CH_3OCH(CH_3)_2$	$\rightarrow CH_3O+CH(CH_3)_2$	22[a]	4[b]	−121[c]	−83[d]	−99[e]	−78[f]
K: $CH_3OC(CH_3)_3$	$\rightarrow CH_3O+C(CH_3)_3$	34[a]	10[b]	−135[c]	−95[d]	−100[e]	−85[f]

[a] δE using cc-pVTZ(df-pVTZ) for H, I, J, K: 18, 17, −0.4, and −6, respectively.
[b] δE using cc-pVTZ(df-pVTZ) for H, I, J, K: −3, −0.4, 3, and 4, respectively.
[c] δE_{reac} using cc-pVTZ(df-pVTZ) for H, I, J, K: −61, −72, −71, and −71, respectively.
[d] δE_{reac} using cc-pVTZ(df-pVTZ) for H, I, J, K: −31, −39, −47, and −53, respectively.
[e] δE_{prod} using cc-pVTZ(df-pVTZ) for H, I, J, K: −43, −54, −71, and −77, respectively.
[f] δE_{prod} using cc-pVTZ(df-pVTZ) for H, I, J, K: −34, −39, −44, and −49, respectively.

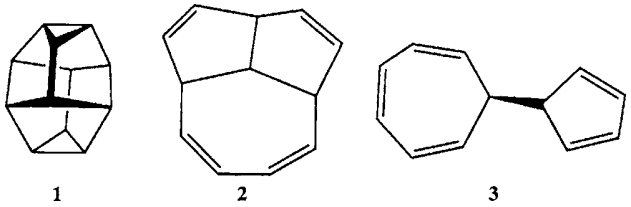

FIG. 7. Structures of three different isomers of $C_{12}H_{12}$: 1, 2, and 3.

were taken from Ref. 30. Figure 7 shows the three isomers 1–3 of $C_{12}H_{12}$ considered in reactions A and B.

The absolute values of the overlap density-fitting errors of the reaction energies are between 13 and 90 μhartrees, to be compared with the Coulomb density-fitting errors between 3 and 25 μhartrees. Moreover, the fitting errors in reaction energies are typically one order of magnitude smaller than the total fitting errors of the products and reactants. A comparison of the fitting errors in the overlap and Coulomb metric reveals that both approaches differ less for isomerization reactions than for bond-breaking reactions, as expected from the interpretation that the fitting error in the overlap metric suffers from BSSE. A reduction in the absolute value of the fitting error is observed when switching from the small 6-31G(df-def2) to the large cc-pVTZ(df-pVTZ) basis set. The very small overlap fitting errors for reaction I (0.4 μhartree) and the Coulomb fitting error for reaction J (−0.4 μhartree) at B3LYP/cc-pVTZ(df-pVTZ) level of theory arise from error cancellation.

To test for auxiliary BSSE, we corrected the interaction energies of H, I, and K at the B3LYP/cc-pVTZ(df-pVTZ) level of theory by applying the CP correction. Interaction energies are defined as the difference between a molecular energy and the sum of the fragment energies with the fragments at the same geometries as in the molecule. Interaction energies thus differ from reaction energies in that the fragment geometries are not relaxed.

For reactions H and I, the fitting errors in the overlap metric of the resulting CP-corrected interaction energies are reduced from 18 to 0.6 μhartree and from 17 to −2 μhartree, respectively. By contrast, for reaction K, the fitting error increases slightly in magnitude, from 11 to −14 μhartree, when the CP correction is included. Thus, when density fitting is performed in the overlap metric, the auxiliary BSSE clearly influences the quality of reaction energies, in agreement with the discussion of the atomization energies. We stress that specifically tuned auxiliary basis sets are expected to reduce the effect of BSSE. Even with the auxiliary basis sets used here, the absolute error in the reaction energies because of density fitting in the overlap metric does not exceed 90 μhartrees.

TABLE III. Timings for a complete HF/cc-pVTZ(df-pVTZ) calculation of the naphthalene molecule. The calculation converged in 14 SCF iterations.

Method	Initialization (s)	Coulomb (s/iter)	exchange (s/iter)
J-engine+LinK		408	1394
Coulomb and exchange fitting	269	1.2	33.1

TABLE IV. Sparsity of fitting integrals in the overlap and Coulomb metrics at different thresholds τ for the acene-5 molecule of Ref. 28 in the cc-pVTZ(df-pVTZ) basis. Also listed is the ratio between the sparsities of the overlap and Coulomb integrals. The prime on the auxiliary basis-set index denote an orthogonal basis.

Integral	$\tau=10^{-10}$	$\tau=10^{-8}$	$\tau=10^{-6}$
$(\alpha\|\beta)$	52%	52%	47%
$\langle\alpha\beta\rangle$	16%	14%	11%
Ratio $\langle\alpha\beta\rangle/(\alpha\|\beta)$	0.31	0.27	0.23
$(ab\|\alpha)$	26%	21%	15%
$\langle ab\alpha\rangle$	7.5%	5.4%	3.2%
$c_\alpha^{ab}=\langle ab\beta\rangle\langle\beta\alpha\rangle^{-1}$	25%	20%	13%
Ratio $\langle ab\alpha\rangle/(ab\|\alpha)$	0.29	0.25	0.22
Ratio $c_\alpha^{ab}/(ab\|\alpha)$	0.95	0.93	0.86
$(ab\|\alpha')$	26%	21%	15%
$c_{\alpha'}^{ab}=\langle ab\alpha'\rangle$	21%	14%	6.4%
Ratio $\langle ab\alpha'\rangle/(ab\|\alpha')$	0.80	0.64	0.42

D. Timings and sparsity

The purpose of this paper is mainly to show that it is possible to perform density fitting in local metrics rather than than in nonlocal Coulomb metric. However, to illustrate that this method is indeed practicable; we present some timings and sparsity results.

As is well known, the application of integral fitting to the calculation of Coulomb and exchange matrices provides a dramatic speed-up of the calculations. Table III contains timings for a B3LYP/cc-pVTZ(df-pVTZ) calculation on naphthalene, with and without density fitting, using a development version of DALTON.[31] The calculation was carried out on a 2200 MHz SUN X2200 AMD Opteron machine, and the code was compiled with ifort 9.0 linked with mkl-libraries (version 8.1). For both Coulomb and exchange matrices, the evaluation is accelerated by almost a factor of 30. The initialization step of 269 s consists of the following main contributions: three-center integral evaluation $(ab|\alpha)$ (89 s), calculation of the inverse square root $(\alpha|\beta)^{-0.5}$ (12 s), and transformation to the orthonormal basis (164 s). The remaining 5 s consists of calculating $g_\alpha^w=\sqrt{(\alpha|\alpha)}$, $g_{ab}^w=\sqrt{(ab|ab)}$, and, after renormalization of the auxiliary basis, the $(\alpha|\beta)$ integrals.

Table IV lists the sparsity of two- and three-center integrals for acene $(n=5)$ of Ref. 28, in the Coulomb and the overlap metrics. Sparsity is listed at different screening thresholds (τ). Also listed are the ratios between the number of integrals in the overlap and Coulomb metrics. The numbers of significant two- and three-center integrals are reduced by a factor of 3–4 in the overlap metric compared to the Coulomb metric. The fitting coefficients c_α^{ab} as obtained in the overlap metric show only a slow onset of sparsity in the nonorthogonal auxiliary basis, reducing the number of significant fitting coefficients in the overlap metric by 5%–14%, for thresholds in range of 10^{-10}–10^{-6} hartree. In the orthogonal basis, the sparsity in the fitting coefficients $c_{\alpha'}^{ab}$ is maintained, reducing the number of significant fitting coefficients by 20%–58%.

VI. CONCLUSIONS AND OUTLOOK

In this paper, we have studied the variational density-fitting technique for the calculation of Coulomb and exchange matrices, with emphasis on the locality of the fitting metric. Such local metrics yield a sparse linear set of equations for the fitting coefficients, allowing for their determination in time proportional to the system size. We have shown that local metrics can be chosen such that the accuracy of the calculation does not suffer as demonstrated by our test implementation. In the derivation of the formulas, we have paid special attention to aspects of density fitting relevant for property calculations (variation principle, etc.).

ACKNOWLEDGMENTS

This work was supported by the Norwegian Research Council through the CeO Centre for Theoretical and Computational Chemistry (Grant No. 179568/V30) and through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420).

[1] A. Sodt, J. E. Subotnik, and M. Head-Gordon, J. Chem. Phys. **125**, 194109 (2006).
[2] F. Weigend, Phys. Chem. Chem. Phys. **4**, 4285 (2002).
[3] F. Weigend, J. Comput. Chem. **29**, 167 (2008).
[4] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).
[5] E. J. Baerends, D. E. Ellis, and P. Ros, Chem. Phys. **2**, 41 (1973).
[6] J. A. Jafri and J. L. Whitten, J. Chem. Phys. **61**, 2116 (1974).
[7] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).
[8] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 4993 (1979).
[9] R. T. Gallant and A. St-Amant, Chem. Phys. Lett. **256**, 569 (1996).
[10] P. Sałek, S. Høst, L. Thøgersen, P. Jørgensen, P. Manninen, J. Olsen, B. Jansik, S. Reine, F. Powlowski, E. Tellgren, T. Helgaker, and S. Coriani, J. Chem. Phys. **126**, 114110 (2007).
[11] C. Fonseca Guerra, J. G. Snijders, G. te Velde, and E. J. Baerends, Theor. Chem. Acc. **99**, 391 (1998).
[12] W. T. Yang and T.-S. Lee, J. Chem. Phys. **103**, 5674 (1995).
[13] B. I. Dunlap, J. Mol. Struct.: THEOCHEM **501**, 221 (2000).
[14] O. Vahtras, J. Almlöf, and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).
[15] Y. Jung, A. Sodt, P. M. W. Gill, and M. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **102**, 6692 (2005).
[16] R. Polly, H.-J. Werner, F. R. Manby, and P. J. Knowles, Mol. Phys. **102**, 2311 (2004).
[17] J. Pipek and P. G. Mezey, J. Chem. Phys. **90**, 4916 (1989).
[18] A. Sodt and M. Head-Gordon, J. Chem. Phys. **128**, 104106 (2008).
[19] M. A. Watson, P. Sałek, P. Macak, and T. Helgaker, J. Chem. Phys. **121**, 2915 (2004).
[20] E. Schwegler and M. Challacombe, J. Chem. Phys. **105**, 2726 (1996).
[21] C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
[22] F. Aquilante, T. B. Pedersen, and R. Lindh, J. Chem. Phys. **126**, 194106 (2007).
[23] B. Jansík, S. Høst, P. Jørgensen, J. Olsen, and T. Helgaker, J. Chem. Phys. **126**, 124104 (2007).
[24] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (Wiley, Chichester, 2000).
[25] L. E. McMurchie and E. R. Davidson, J. Comput. Phys. **26**, 218 (1976).
[26] C. C. M. Samson, W. Klopper, and T. Helgaker, Comput. Phys. Commun. **149**, 1 (2002).
[27] S. Reine, E. Tellgren, and T. Helgaker, Phys. Chem. Chem. Phys. **9**, 4771 (2007).
[28] M. J. G. Peach, P. Benfield, T. Helgaker, and D. J. Tozer, J. Chem. Phys. **128**, 044118 (2008).
[29] S. F. Boys and F. Bernardi, Mol. Phys. **19**, 553 (1970).
[30] Y. Zhao and D. G. Truhlar, Theor. Chem. Acc. **120**, 215 (2008).
[31] DALTON, an ab initio electronic structure program, Release 2.0, 2005 (see http://www.kjemi.uio.no/software/dalton/dalton.html).

Paper IV

An efficient density functional theory force evaluation for large molecular systems
S. Reine, M. F. Iozzi, V. Bakken, A. Krapp, T. Helgaker, F. Pawłowski and P. Sałek
Manuscript

An efficient density-functional-theory force evaluation for large molecular systems

Simen Reine, Maria Francesca Iozzi, Vebjørn Bakken, Andreas Krapp and Trygve Helgaker

Department of Chemistry,
University of Oslo,
P.O. Box 1033 Blindern,
N-0315 Oslo, Norway

Filip Pawłowski

Zakład Fizyki Teoretycznej i Informatyki,
Kazimierz Wielki University,
Plac Weyssenhoffa 11 85-072,
Bydgoszcz, Poland

Pawel Sałek

Laboratory of Theoretical Chemistry,
The Royal Institute of Technology,
Teknikringen 30, Stockholm SE-10044, Sweden

Abstract

In this paper we demonstrate an efficient implementation of molecular forces for systems containing up to 500 atoms. The density-fitted-Coulomb force is obtained in linear time by combining integral screening with the continuous fast multipole method, with the near-field force contributions accelerated using the novel integral evaluation scheme presented earlier (PCCP, **9**, 4771 (2007)). By expanding the solid-harmonic Guassians in Hermite rather than Cartesian Gaussians, the computational cost of differentiated integrals is reduced - and at the same time simplifying the implementation. The computational efficiency of molecular force evaluation is demonstrated by sample calculations, and further applied to the geometry optimization of sample systems.

I. INTRODUCTION

The calculation of molecular energies and their derivatives is a prerequisite for the determination of almost all molecular properties. Through the development of efficient methods, that both reduce the prefactor and computational scaling behavior associated with the different quantum-chemical approaches, and through the on–going revolution in computer technology, quantum-chemical methods are applied to ever large molecular systems. In this context, linear scaling methods based on Kohn-Sham (KS) density functional theory (DFT) are of particular interest, as they give accurate results for large molecules with computational costs proportional to system size. Today, the majority of quantum-chemical calculations are in fact carried out using DFT.

Molecular gradients[1], or the derivatives of the molecular energy with respect to the nuclear coordinates, are needed both for the determination of equilibrium and transition-state structures and for the study of interactions using molecular dynamics. Therefore, efficient evaluation of the forces[2–12], in addition of course to the evaluation of the energies, is an essential component of quantum-chemical softwares. The efficient evaluation of the forces, builds upon the methodology developed for the energy evaluation. Therefore we start by giving a brief overview of efficient methods for the energy evaluation, before we turn the attention to the evaluation of the forces.

In the KS DFT approach the density-matrix is obtained in a self-consistent field (SCF) through a series of SCF iterations. In each iteration a KS-matrix is constructed and a new density-matrix is obtained. The main computational bottlenecks in the evaluation of the KS-matrices, are the Coulomb, exchange-correlation and, for hybrid functionals, also the exact exchange contributions. Almlöf[13] was the first to recognize that splitting the two-electron interactions, into a Coulomb and an exchange part, was beneficial since this allowed the development of efficient algorithms for the two contributions individually. Efficient evaluation of the Coulomb contribution has received significant attention. Linear scaling can be achieved by combining efficient screening techniques[14,15] and the continuous fast multipole method (CFFM)[16]. Integral evaluation may be accelerated using J-engine based integral evaluation[17–20], the Fourier transform Coulomb[21], Cholesky decomposition[22] or density-fitting approximations[23–26]. In the density fitting approximation, the costly four-center two-electron integrals are replaced by two- and three-center

integrals and a set of linear equations, with typical speed-up factor 3-30. Further acceleration of the density fitting approximation is attained through efficient evaluation of the two- and three-center integrals[27,28], by applying the Poisson scheme of Manby and Knowles[29], and through linear scaling density fitting developments[30–34]. The exchange contribution is intrinsically linear scaling for systems with non-vanishing HOMO-LUMO gaps, since the density-matrix elements couple basis-functions belonging to different electrons. Linear-scaling formation of the exchange contribution is achieved by combining integral screening and proper reorganization of the integral loop structure, as implemented first in the order N exchange (ONX)[35] and later in the linear-scaling exchange (LinK)[36] algorithms. Density fitting of the exchange is more involved than for the Coulomb term since three-index fitting coefficients are required, and was introduced as late as 2002 by Weigend[37]. Linear-scaling density-fitting approaches have also been considered for the exchange term[34,38,39]. For the exchange-correlation contribution numerical quadratures are intrinsically linear scaling due to the rapidly decaying nature of the basis functions used[40,41] and the use of linear-scaling grid generation[42]. Finally, the density matrix can be obtained in a linear scaling fashion following Refs.[33,43–46].

In this article we build upon recent developments for energy evaluation, and present a very efficient implementation of molecular gradients for pure DFT functionals within the DALTON program package. The construction of the density-fitted-Coulomb force contributions are accelerated using the McMurchie–Davidson J–engine like integral scheme presented in[20], in combination with integral screening[14] and multipole moment methods[3,4,16]. This allows linear scaling of the different contributions to the gradient, while at the same time avoiding any explicit construction and storage of the differentiated two– and three–center integrals. To the best of our knowledge, linear-scaling density-fitted-Coulomb force evaluations have not yet been reported in the literature, and is reported in this paper. The near-field contributions to the density-fitted-Coulomb force, are accelerated using the novel integral evaluation scheme presented in Ref. [27]. Expanding the solid-harmonic Gaussians in Hermite rather than Cartesian Gaussians as introduced in Ref. [27], has the benefits of reducing the cost of the differentiated integral evaluation and simplifying the implementation. We further outline the details of our exchange-correlation gradient implementation and the techniques applied for geometry optimization.

The paper is organized as follows. In section II we describe the computational details

of our force evaluation, and give a brief report of the geometry-optimization algorithm employed. In section III we report detailed timings for some selected systems. Finally, in section IV we give some concluding remarks.

II. THEORY AND IMPLEMENTATION

In this section we start by introducing the molecular KS DFT gradient in section II A. In section II B we give a detailed account of the density-fitted-Coulomb force evaluation, followed by the exchange-correlation force evaluation in section II C. Finally, in section II D we outline the implementation used for geometry optimization.

A. The molecular gradient

In this paper, we limit the discussion to closed-shell systems. The extension to open-shell systems is straightforward. The closed-shell KS energy is given as

$$E_{\mathrm{KS}}[\rho] = T_s[\rho] + \int v_{\mathrm{ext}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2}\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 + E_{\mathrm{xc}}[\rho] + h_{\mathrm{nuc}}, \tag{1}$$

with $T_s[\rho]$ the non-interacting kinetic energy, v_{ext} the external potential due to the nuclei, $E_{\mathrm{xc}}[\rho]$ the exchange-correlation (XC) energy, $\rho(\mathbf{r})$ the electron density and h_{nuc} the nuclear repulsion energy. In atomic-orbital (AO) representation, the electron density may be written as

$$\rho(\mathbf{r}) = 2N\int \Phi^2(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)\mathrm{d}\mathbf{r}_2 \ldots \mathrm{d}\mathbf{r}_N = 2\sum_{ab}\chi_a(\mathbf{r})\chi_b(\mathbf{r})D_{ab} \tag{2}$$

with $\chi_a(\mathbf{r})$ the AO basis functions and D_{ab} the AO density matrix elements. The molecular gradient elements $G^{\mathbf{e}}$ are the derivatives of the KS energy, Eq. (1), with respect to the nuclear coordinate R_e

$$G^{\mathbf{e}} = \frac{dE_{\mathrm{KS}}[\rho]}{dR_e} = h^{\mathbf{e}} + J^{\mathbf{e}} + E^{\mathbf{e}}_{\mathrm{xc}} + h^{\mathbf{e}}_{\mathrm{nuc}} + \sum_{ab} D^{\mathbf{e}}_{ab} f^{\mathrm{KS}}_{ab}, \tag{3}$$

with $e = x, y, z$, and where we have introduced $\mathbf{e}$ as the first, second or third row of the three by three identity matrix for differentiation with respect to the x, y or z Cartesian directions, respectively. The KS force of Eq. (3) consists of the one-electron $h^{\mathbf{e}}$, which includes the kinetic and the nuclear-electron attraction force, Coulomb $J^{\mathbf{e}}$, XC $E^{\mathbf{e}}_{\mathrm{xc}}$, nuclear

$h^{\mathbf{e}}_{\mathrm{nuc}}$ and the so-called Pulay force[1],

$$G^{\mathbf{e}}_{\mathrm{Pulay}} = \sum_{ab} D^{\mathbf{e}}_{ab} f^{\mathrm{KS}}_{ab}, \tag{4}$$

which arises due to the incompleteness of the AO basis. Here $\mathbf{f}^{\mathrm{KS}}$ is the (converged) KS-matrix and the derivative of the density-matrix is given[47] as $\mathbf{D}^{\mathbf{e}} = -\mathbf{D}\mathbf{S}^{\mathbf{e}}\mathbf{D}$. The main bottlenecks of the molecular KS force contributions are the Coulomb and the XC forces, which will be discussed in the following two sections.

B. The density-fitted-Coulomb contribution to the gradient

In the density-fitting approximation[23–26] the electron density $\rho(\mathbf{r})$ is approximated by a *fitted* density $\tilde{\rho}(\mathbf{r})$, expanded in a set of atom-centered auxiliary basis functions $\chi_\alpha(\mathbf{r})$ according to

$$\tilde{\rho}(\mathbf{r}) = \sum_{\alpha} c_\alpha \chi_\alpha(\mathbf{r}). \tag{5}$$

Following[48] the Coulomb contribution to the KS energy

$$J = \frac{1}{2}\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 = \frac{1}{2}(\rho|\rho), \tag{6}$$

is approximated by

$$\tilde{J} = \frac{1}{2}(\rho|\tilde{\rho}) + \frac{1}{2}(\tilde{\rho}|\rho) - \frac{1}{2}(\tilde{\rho}|\tilde{\rho}), \tag{7}$$

where we use Mulliken notation

$$(f|g) = \int f(\mathbf{r}_1)\frac{1}{r_{12}} g(\mathbf{r}_2) \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2. \tag{8}$$

The expression for the fitted Coulomb energy $\tilde{J}$, Eq. (7), replaces the Coulomb energy J in Eq. (1) yielding the approximated KS energy $\tilde{E}_{\mathrm{KS}}$. $\tilde{E}_{\mathrm{KS}}$ is optimized variationally with respect to both the AO density-matrix elements D_{ab} and the coefficients c_α. The optimization with respect to D_{ab} leads to the *fitted* Coulomb contribution $\tilde{\mathbf{J}}$ to the KS-matrix

$$\tilde{J}_{ab} = (ab|\tilde{\rho}) \tag{9}$$

and the optimization with respect to c_α gives the linear equation set

$$\sum_{\beta} (\alpha|\beta) c_\beta = \sum_{ab} (\alpha|cd) D_{cd}, \tag{10}$$

for the fitting coefficients c_α. Solving the linear set of equations, Eq. (10), for the fitting coefficients c_α, scales cubically with respect to computational time. The scaling becomes a problem only for large systems (typically containing more than 10000 auxiliary basis-functions), due to the very low prefactor using for example the standard mathematical LAPACK library routine DPOSV. For the purpose of this paper we employ Eq. (10), but note that for large systems the fitting coefficients can be obtained in a linear scaling fashion[30–34]. Further note that the above fitting procedure is robust[48] in the sense that the error in the Coulomb repulsion energy is quadratic in the error of the approximated density, according to

$$J - \tilde{J} = \frac{1}{2}(\rho - \tilde{\rho}|\rho - \tilde{\rho}). \tag{11}$$

Differentiation of the fitted electronic Coulomb repulsion energy of Eq. (7) with respect to the nuclear coordinate R_e gives[2]

$$\tilde{J}^{\mathbf{e}} = \frac{d\tilde{J}}{dR_e} = \sum_{ab} D^{\mathbf{e}}_{ab}\tilde{J}_{ab} + \sum_{ab} D_{ab}\tilde{J}^{\mathbf{e}}_{ab} + \sum_{\alpha} c_\alpha(g^{\mathbf{e}}_\alpha - \tilde{g}^{\mathbf{e}}_\alpha) \tag{12}$$

with

$$\begin{aligned} \tilde{J}^{\mathbf{e}}_{ab} &= (\{ab\}^{\mathbf{e}}\,|\tilde{\rho}) \\ g^{\mathbf{e}}_\alpha &= (\alpha^{\mathbf{e}}|\rho) \\ \tilde{g}^{\mathbf{e}}_\alpha &= (\alpha^{\mathbf{e}}|\tilde{\rho}). \end{aligned} \tag{13}$$

Similarly to the undifferentiated Coulomb contribution $\tilde{\mathbf{J}}$ the three differentiated contributions of Eq. (13) are obtained in a linear scaling fashion combining Cauchy-Schwartz screening[14,23],

$$|(f|g)| \leq \sqrt{(f|f)}\sqrt{(g|g)}, \tag{14}$$

(which for the derivative case includes second derivative integrals) and the continuous fast multipole-method (CFMM)[16].

There are two possible ways to obtain the far-field (FF) Coulomb gradient contribution - either by differentiating a given multipole moment expansion of the classical part of the interaction energy[4] $J^{\text{cls.}}$, given by[43]

$$J^{\text{cls.}} = \frac{1}{2}\sum_{p}\sum_{q\in \text{FF}_{\mathbf{P}}} D_p \mathbf{q}_p(\mathbf{P})^{\text{T}}\mathbf{W}(\mathbf{R}_{\bar{\mathbf{P}}\mathbf{P}})^{\text{T}}\mathbf{T}(\mathbf{R}_{\bar{\mathbf{Q}}\bar{\mathbf{P}}})\mathbf{W}(\mathbf{R}_{\bar{\mathbf{Q}}\mathbf{Q}})\mathbf{q}_q(\mathbf{Q})D_q, \tag{15}$$

or by first taking the analytical derivative and then introducing the CFMM approximation[3]. Differentiation of $J^{\text{cls.}}$, gives and exact gradients for a given order of multipole

moment expansion, provided the partitioning of the global system into a hierarchical family of boxes remains the same throughout the optimization, and provided the centers $\mathbf{P}$ of the charge-distributions $\Omega_p(\mathbf{r})$ remain within the same boxes (with centers $\bar{\mathbf{P}}$). During the course of the optimization, however, the charge-distributions can move between different boxes. Furthermore, keeping the boxes fixed throughout the optimization is not a good alternative, since for instance different starting geometries then would converge to different minima. Therefore, both ways of obtaining the CFMM contribution to the gradient are limited by the accuracy of the CFMM expansion. We choose the second approach for obtaining the gradient, namely to first do the analytical derivative of the density-fitted energy according to Eq. (12), and then introduce the CFMM approximation for each of the three terms of Eq. (13) afterwards. This approach is simpler to implement as it only includes the multipole moments $\mathbf{q}_p^{\mathbf{e}}$ of the differentiated charge-distributions $\Omega_p^{\mathbf{e}}(\mathbf{r})$, keeping the FF potential fixed; and thus also leaving the translation matrices $\mathbf{W}(\mathbf{R})$ and the interaction matrices $\mathbf{T}(\mathbf{R})$ unchanged.

Construction of the three contributions of Eq. (13) is accelerated using J-engine based integral evaluation of Ref [20], and we further utilize the fact that solid-harmonic combinations of Cartesian $G_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r})$ and Hermite Gaussian $H_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r})$ atomic orbitals are identical[27]. The Hermite Gaussian,

$$H_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r}) = \frac{\partial^{l_a} \exp(-ar_A^2)}{(2a)^{l_a} \partial A_x^{i_x} \partial A_y^{i_y} \partial A_z^{i_z}} = (2a)^{-l_a} \Lambda_{\mathbf{i}}(a, \mathbf{r}_A), \tag{16}$$

simplifies the evaluation of derivative molecular integrals, since differentiation by nuclear coordinates merely increments the Hermite quantum numbers, whereas differentiation of a Cartesian Gaussian

$$G_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r}) = x_A^{i_x} y_A^{i_y} z_A^{i_z} \exp(-ar_A^2), \tag{17}$$

gives both a higher and a lower order term. Here $l_a = i_x + i_y + i_z$ is the angular momentum quantum number, $\mathbf{A} = (A_x, A_y, A_z)$ is the center of the Gaussian, $r_A = |\mathbf{r} - \mathbf{A}|$, $x_A = x - A_x$ and similarly for y_A and z_A. Following McMurchie and Davidson, the primitive product overlap-distribution $\Omega_{ab}^{\mathbf{ij}}(\mathbf{r})$ between two Hermite Gaussian basis functions is expanded according to[27]

$$\Omega_{ab}^{\mathbf{ij}}(\mathbf{r}) = H_{a,\mathbf{A}}^{\mathbf{i}}(\mathbf{r}) H_{b,\mathbf{B}}^{\mathbf{j}}(\mathbf{r}) = \sum_{|\mathbf{t}|=0}^{\mathbf{i}+\mathbf{j}} E_{\mathbf{t}}^{\mathbf{ij}} \Lambda_{\mathbf{t}}(p, \mathbf{r}_P), \tag{18}$$

with $\mathbf{t} = (t_x, t_y, t_z)$, $p = a + b$ and $\mathbf{P} = (a\mathbf{A} + b\mathbf{B})/p$. The recurrence relations for the expansion coefficients $E_{\mathbf{t}}^{\mathbf{ij}}$ are given in Ref. [27]. In the McMurchie-Davidson based J-engine approaches, the electron density $\rho(\mathbf{r})$ is expanded according to

$$\rho(\mathbf{r}) = \sum_{ab} \Omega_{ab}(\mathbf{r}) D_{ab} = \sum_{p,\mathbf{P}} \sum_{|\mathbf{t}|=0}^{l_a+l_b} (-1)^{|\mathbf{t}|} F_{\mathbf{t}}^{p,\mathbf{P}} \Lambda_{\mathbf{t}}(p, \mathbf{r}_P), \tag{19}$$

with

$$F_{\mathbf{t}}^{p,\mathbf{P}} = (-1)^{|\mathbf{t}|} \sum_{mn \in p,\mathbf{P}} C_m C_n \sum_{\mathbf{ij}} S_{\mathbf{ij}}^{ab} E_{\mathbf{t},mn}^{\mathbf{ij}} D_{ab}, \tag{20}$$

where $mn \in p, \mathbf{P}$ means all primitive pairs mn for which the overlap-distributions $\Omega_{a_m b_n}^{\mathbf{ij}}(\mathbf{r})$ share both exponent p and center $\mathbf{P}$, where C_m and C_n are the contraction coefficients and $S_{\mathbf{i}}^a$ and $S_{\mathbf{j}}^b$ are the solid-harmonic transformation coefficients. Similarly, the fitted density $\tilde{\rho}(\mathbf{r})$ is expanded according to

$$\tilde{\rho}(\mathbf{r}) = \sum_{\alpha} \chi_{\alpha,\mathbf{P}}^{l_\alpha m_\alpha}(\mathbf{r}) c_\alpha = \sum_{p,\mathbf{P}} \sum_{|\mathbf{t}|=l_p} (-1)^{|\mathbf{t}|} \tilde{F}_{\mathbf{t}}^{p,\mathbf{P}} \Lambda_{\mathbf{t}}(p, \mathbf{r}_P), \tag{21}$$

with

$$\tilde{F}_{\mathbf{t}}^{p,\mathbf{P}} = \sum_{m \in p,\mathbf{P}} C_m S_{\mathbf{t}}^{\alpha_m} c_{\alpha_m} (-2\alpha)^{-l_\alpha}. \tag{22}$$

This gives the three differentiated contributions of Eq. (13)

$$\begin{aligned}
\tilde{J}_{ab}^{\mathbf{e}} =& \sum_{\mathbf{ij}} S_{\mathbf{i}}^a S_{\mathbf{j}}^b \sum_{mn} C_m C_n \sum_{|\mathbf{t}|=0}^{l_a+l_b+1} \left(\delta_{\mathbf{EA}} E_{\mathbf{t}}^{\mathbf{i+e,j}} + \delta_{\mathbf{EB}} E_{\mathbf{t}}^{\mathbf{i,j+e}} \right) \\
& \qquad \sum_{q,\mathbf{Q}} \sum_{|\mathbf{u}|=l_q} \tilde{F}_{\mathbf{u}}^{q,\mathbf{Q}} R_{\mathbf{t+e+u}}(\gamma, \mathbf{R}_{PQ}) \\
g_\alpha^{\mathbf{e}} =& \delta_{\mathbf{EP}} \sum_{|\mathbf{t}|=l_\alpha} S_{\mathbf{t}}^\alpha \sum_m C_m (2\alpha_m)^{1-l_\alpha} \sum_{q,\mathbf{Q}} \sum_{|\mathbf{u}|=0}^{l_q} F_{\mathbf{u}}^{q,\mathbf{Q}} R_{\mathbf{t+e+u}}(\gamma, \mathbf{R}_{PQ}) \\
\tilde{g}_\alpha^{\mathbf{e}} =& \delta_{\mathbf{EP}} \sum_{|\mathbf{t}|=l_\alpha} S_{\mathbf{t}}^\alpha \sum_m C_m (2\alpha_m)^{1-l_\alpha} \sum_{q,\mathbf{Q}} \sum_{|\mathbf{u}|=l_q} \tilde{F}_{\mathbf{u}}^{q,\mathbf{Q}} R_{\mathbf{t+e+u}}(\gamma, \mathbf{R}_{PQ}),
\end{aligned} \tag{23}$$

with $\gamma = pq/(p+q)$, $R_{PQ} = |\mathbf{P} - \mathbf{Q}|$, where the Dirac delta function $\delta_{\mathbf{AB}}$ is zero if the centers $\mathbf{A}$ and $\mathbf{B}$ are different, and one if they are identical, and where the primitive Hermite repulsion integrals $R_{tuv}(\gamma, R_{PQ})$ are found by recurrence relations[49]. As can be seen from Eq. (23) the use of Hermite rather than Cartesian Gaussians has two advantages. First, for the one-center auxiliary functions, the number of contractions is reduced, which can be seen for instance from the first term of $\tilde{J}_{ab}^{\mathbf{e}}$ where the innermost summation only

contains terms for which $|\mathbf{u}| = l_q$, rather than stating from $|\mathbf{u}| = 0$. Second, there are no differentiated E-coefficients involved, instead the $E_{\mathbf{t}}^{\mathbf{ij}}$'s are incremented by one order in the quantum numbers.

C. XC contribution

Due to the complex expressions of available exchange-correlation (XC) functionals F_{xc} the integration of F_{xc} is performed numerically in the course of the energy evaluation. For the construction of the numerical quadrature we employ an atomic partitioning scheme[50] which leads to the following energy expression for the XC contribution

$$E_{\mathrm{xc}} = \int F_{\mathrm{xc}}(\mathbf{r})\mathrm{d}\mathbf{r} = \sum_A \sum_i w_i w_A(\mathbf{r}_i) F_{\mathrm{xc}}(\mathbf{r}_i), \tag{24}$$

where the first summation is over all atoms A and the second over the numerical quadrature grid points i of the current atom and w_i is the single center spherical weight and $w_A(\mathbf{r})$ is the atomic weight function which depends on the atom positions. Becke space partitioning with size corrections[50] will be used throughout for the determination of the atomic weights. A detailed description of our implementation of the XC energy evaluation can be found in[20]. The single center integrations in Eq. (24) are separated into radial and angular integrations. For the angular part of the quadrature grid we use Lebedev grids[51–55]. For the radial part we employ a Gauss-Chebychev quadrature of second kind[56] and follow Treutler and Ahlrichs[57] in mapping the integration range from [0,∞) onto [-1,1] via the transformation $r_i = \frac{\xi}{\ln 2}(1+x_i)^{0.6} \ln(\frac{2}{1-x_i})$ using atomic scaling parameters ξ.

The differentiation of the XC-energy expression Eq. (24) leads to the following expression for the XC contribution to the molecular gradient

$$E_{\mathrm{xc}}^{\mathbf{e}} = \sum_i \sum_A w_i (w_A \frac{\partial F_{\mathrm{xc}}}{\partial R_e} + \frac{\partial w_A}{\partial R_e} F_{\mathrm{xc}}), \tag{25}$$

The first term in Eq. (25) requires the derivative of the basis function μ with respect to perturbations of the atoms, and the second term in Eq. (25) involves the derivatives of the atomic grid-weight functions $w_A(\mathbf{r})$ that are calculated following the derivations of Johnson *et* al.[58]. For the calculations presented in this manuscript, the contributions involving the grid-weight derivatives have not been included for the XC forces. This contribution is currently under development.

D. Geometry optimization

To be able to successfully determine for example equilibrium geometries, the implementation of the molecular gradient reported in the previous subsections must be combined with a geometry optimizer. The optimizer makes the system traverse the potential energy surface (PES) by updating the atomic coordinates based on the available information regarding the surface and its derivatives at the current geometry. Also, the optimizer determines when a stationary point on the PES has been reached.

Regular geometry optimizers for molecular systems[59–61] generally contains operations that scale cubically with the number of atoms. The number of atoms will necessarily be far smaller than the number of basis functions, but in order to achieve true linear scaling it is obviously also necessary to address the geometry optimizer. Significant progress has been made in order to reduce the scaling to essentially quadratic in the number of atoms[62–64], and even further to near linear or linear scaling[65–67].

However, constructing a new optimizer was outside the scope of this study, thus it has been necessary to employ the regular geometry optimization routines in the DALTON program package. To quickly summarize, optimizations are run in redundant internal coordinates[59,61] with a quasi-Newton method using the BFGS updating formula[68], and with steps controlled by the level-shifted trust region method[69]. The initial approximate Hessians are constructed according to the model proposed by Lindh *et al.*[70]. It should be briefly noted that within this scheme one may perform constrained geometry optimizations, where one or more internal coordinates are kept frozen. A more detailed description of the DALTON optimizer can be found in Ref. [61].

Two major performance bottlenecks were identified, the first being the determination of the generalized inverse of the Wilson B-matrix (necessary for transforming the calculated Cartesian gradient to internal coordinates), the second was the diagonalization of the updated Hessian in internal coordinates. In both cases previous explicit code was replaced by appropriate calls to LAPACK-routines. For the titin molecule, these modifications caused the time spent in the geometry optimization routines to reduce to less than 3% of the total CPU time, which is very acceptable. Somewhat surprisingly, it is evident that it may in fact not be critical to construct a better scaling optimizer until one gets into the 1500-2000 atoms range.

Convergence criteria determine when an optimization has reached a stationary point on the PES, and it is very important that these criteria are given reasonable thresholds compared to the given accuracy for the energy and particularly the gradient. After each iteration the root-mean-square of the gradient, the maximum element of the gradient, the root-mean-square of the step and the maximum element of the step (all entities in internal coordinates) are determined. We have found the respective convergence criteria ϵ, 5ϵ, 3ϵ and 15ϵ for given threshold ϵ to be reasonable for the above values, and convergence is only declared once all four criteria are met. For constrained optimizations, the *difference* in root-mean-square gradient and maximum gradient element compared to the previous iteration is used, and the two thresholds are tightened by a factor of 10.

During the course of the geometry optimization, the changes in the density-matrix are typically small from one iteration to the next, in particular when the geometry approaches a stationary point. In the current implementation, we exploit this in a simple manner by taking as our starting guess the McWeeny-purified (converged) density-matrix of the previous geometry step. This reduces the number of SCF cycles needed per geometry step, compared to the conventional Hückel guess. The reduction is less in the beginning of the optimization, but become more significant when the geometry steps becomes smaller. On average the number of SCF cycles are reduced by about a factor two. Note that in some cases when the geometry changes significantly this approach does not work, as the McWeeny-purification scheme changes the number of electrons. In such cases, we revert to the regular (extended Hückel) starting guess.

III. RESULTS AND DISCUSSION

To assess the presented method, we here report timings for force and energy evaluations, and the geometry optimization step, for some selected molecular systems, using a development version of DALTON[71]. The calculations were carried out on a Xeon 2.66 Ghz processor, and the code was compiled with ifort 11.0 linked with mkl-libraries (version 10.1). First, we look at the scaling behavior for the energy and force evaluation of linear polyene chains at BP86/6-31G** level of theory. Then we report timings for the geometry optimization of the valinomycine ($C_{54}N_7O_{18}H_{90}$), taxol $(C_{47}N_1)_{14}H_{51}$) and titin ($C_{124}N_{36}O_{37}S_3H_{192}$) molecules. For the presented calculations in this section a threshold

10^{-10} is used for screening in the integral evaluation, the maximum order of the multipoles for the evaluation of the FF is 12 and the 'grid 4' of Treutler and Ahlrichs[57] has been used for the XC numerical quadrature.

At the beginning of the SCF cycle, the screening matrices $\sqrt{(ab|ab)}$ and $\sqrt{(\alpha|\alpha)}$, metric matrix $(\alpha|\beta)$, Cholesky decomposition of the metric matrix (needed in the LAPACK library routine DPOSV), multipole moments and the grid, are constructed in addition to the one-electron contribution to the KS matrix. The density-fitting Coulomb matrix evaluation is composed of three steps in each SCF cycle: 1) the right hand side of the fitting equations, Eq. (10), is constructed, 2) the linear equations are solved using the LAPACK library routine DPOSV, and 3) the fitted Coulomb matrix of Eq. (9) is constructed. In Fig. 1 the computational timings for a single construction of the density-fitted-Coulomb and XC contributions to the KS-matrix are shown to give near linear scaling for the polyene chains. The Coulomb timings are split into the near-field (NF) and the far-field (FF) timings, as well as the time used for the linear equation solver. The presented NF and FF timings are the combined timings for the NF and FF evaluation of steps 1) and 3). The linear equation solver of step 2) scales cubically with system size. As can be seen from Fig. (1) this step exhibits a low prefactor compared to the NF and FF parts of the density-fitted-Coulomb contributions, even for the sparse linear polyenes.

Fig. 2 shows the computational timings of the different force contributions for the polyene chains. The density-fitted-Coulomb contribution shows near linear scaling as for the energy, with the NF and FF contribution about factor two slower than for a single KS-matrix build. For the density-fitted Coulomb force evaluation, the construction of the multipole moments of the differentiated charge-distributions becomes prominent, and further efforts can directed towards improving the computation performance of these moments. Linear scaling evaluation of the one-electron contribution is also clearly needed for these systems. Somewhat surprisingly the XC contribution does not scale linearly, even though the force evaluation closely follows the evaluation of the XC contribution to the KS-matrix.

In table I and table II we report the CPU timings for a single construction of the density-fitted-Coulomb and XC contributions to the KS-matrix and to the molecular force, respectively, for the valinomycine, taxol and titin molecules. In table II we also present the CPU timings for the one-electron and Pulay forces, and for the evaluation of

the multipole moments of the differentiated charge distributions. The calculations were carried out at BP86 level of theory, using the 6-31G and 6-31G** basis sets for valinomycine and taxol, and a combination of the 6-31G and 6-31G* basis sets for the titin molecule. The presented results demonstrate the efficiency of both the density-fitted-Coulomb and XC force evaluations. The NF timings for the density-fitted-Coulomb force evaluation are only factor $2.1 - 2.5$ slower than for the corresponding NF contributions of the density-fitted-Coulomb matrix, and for the FF contribution only factor $1.1 - 1.4$ slower. Compared to the CFFM Coulomb force evaluation of Ref. [4], the FF force contribution exhibits similar scaling behavior compared to the FF Coulomb matrix contribution. Note that it is possible to use the FF potentials of the regular and density-fitted density evaluated for the energy to significantly reduce the FF contribution to the gradient. This approach has not been taken here, but is suggested to further accelerate the FF force contribution.

For the NF force however, the efficiency of the screening is less prominent for the density-fitted-Coulomb force than for the regular Coulomb force. This is due to the double weighting with respect to density-matrix elements for the regular case, which greatly reduces the number of differentiated four-center integrals to be evaluated. Similar computational benefits for the force evaluation are only possible for the first of the three contributions of Eq. (13), since the weighting with the fitting coefficients only gives minor screening benefits. One way to accelerate the NF contribution by up to a factor two is to calculate the two contributions $\tilde{J}^{\mathbf{e}}_{ab}$ and $g^{\mathbf{e}}_{\alpha}$ simultaneously, as they share different intermediates like the Hermite repulsion integrals $R_{\mathbf{t+e+u}}(\gamma, \mathbf{R}_{PQ})$. Such an approach will be most efficient for low angular momentum integrals, whereas less benefit will be gained for the higher angular momentum integrals, as these integrals become dominated by the contraction steps. Another way to accelerate the NF is to use the Poisson scheme[29].

The additional calculation of the multipole moments of the differentiated charge distributions takes less than 23 percent of the total density-fitted-Coulomb force evaluation for the given systems. The evaluation of the XC force is slightly faster, and the one-electron contribution takes a little longer than half the time, compared to the density-fitted-Coulomb force evaluation of the largest system studied here, the titin molecule, and less for the smaller systems. The timings for the Pulay force is small for the considered cases, but for larger systems the linear scaling evaluations of both the Pulay force

and one-electron part are needed (which can be achieved by exploiting the sparsity of the overlap and density matrices and by employing multipole-moment methods for the electron-nuclear attraction part of the one-electron force, respectively).

In table III we report averaged CPU timings of the energy, force and geometry optimization steps for the full geometry optimization of the three systems studied above. Also listed are the number of energy evaluations, the averaged number of SCF cycles per energy evaluation and the number of geometry steps needed for the optimization. The convergence criteria given in section II D were used for the geometry optimization with threshold $\epsilon = 10^{-4}$ Hartree. Note that as outlined in section II D, the McWeeny-purified counterpart of the converged density matrix of the previous geometry step, is used as the starting guess for the next geometry. This typically reduces the averaged number of SCF iterations needed at each geometry step by about a factor two, compared with the number of iterations needed to converge using the extended Hückel guess (from 18 to 9.3 for the titin molecule). Further note that we use the incremental scheme[35] in each SCF cycle, for which the KS-matrix is built using difference densities rather than the full density. The Coulomb matrix (and the exchange matrix) are linear in the density matrix, and the difference Coulomb matrix can be constructed using difference densities. For the calculations presented here the incremental scheme typically gives an average speed-up of about a factor 2-3 for the NF contributions, and $20-30$ percent for the FF contributions. The incremental scheme can also be adapted to the XC contribution by constructing difference density and density-gradient contributions rather than the full density at each grid point, and by using the difference potentials for the final XC matrix build. In the current implementation we have only implemented the former of the two, giving speed-ups of $15-20$ percent for the XC contribution. For the systems presented in table III, the evaluation of the forces takes one third to one half the CPU time compared to the energy evaluation, whereas CPU time for the optimization step is negligible in comparison.

As an example of the energy contribution, we report timings for the titin molecule, the averaged CPU time for a full energy evaluation is distributed as follows: initialization step 812 s (grid partitioning 535 s, metric and screening matrices 158 s, one-electron matrices 49 s, Cholesky decomposition of metric matrix 44 s, and multipole moments 9 s), NF contributions 340 s, FF contributions 1363 s, XC contribution 790 s, RH/DIIS

optimization 279 s and linear solver 12 s. For the energy, the FF contribution dominates the calculation, followed by the initialization step and the XC contribution, whereas the NF contribution takes less than 10 percent of the overall CPU time of the energy evaluation.

As can be seen from these timings the most pressing issue for the density-fitted-Coulomb evaluation is therefore the efficient evaluation of the FF. For the construction of the density-fitted FF Coulomb contribution the FF potentials of both the regular and the fitted electron densities need to be constructed, and is therefore slower than for the regular Coulomb FF contribution. Code optimizations of the current implementation to reduce the prefactor for the FF contribution are possible. Still, the results illustrate the need for a more efficient FF evaluation. One way forward, may be the splitting of the interactions into classical and non-classical interactions, treating the classical contributions by FMM rather than CFFM, as suggested by Ref. [20]. To reduce the density-fitted FF evaluation further, we note that for the construction of the FF potential of the fitted density, the multipole moment of the auxiliary basis functions centered on a single atom can be combined into one multipole moment expansion, expanded exactly in multipole moments of order equal to the highest angular moment function of the auxiliary basis. Similarly, the FF potential of the electron density needs only be constructed to orders equal the highest angular momentum auxiliary basis-functions at each atomic centers. Another way forward is to use the Poisson scheme of Ref. [29], possibly in combination with the above two approaches, to reduce the number of classical interactions to be treated. The Poisson scheme also reduces the NF timings and noted above.

The grid partitioning step takes a significant portion of the energy evaluation. For the evaluation of the forces, without the use of the grid quadrature weight derivatives, we have been forced to use order 11 for the Becke partitioning function, rather than the suggested order 3[57]. As the cost of this step is proportional to the order, this step will be reduced when the weight-derivative contribution is added to the XC force. However, the evaluation of the grid-weight derivatives and their contraction with XC-matrix elements then add to the total cost of the gradient. Clearly, the grid partitioning of the current implementation would benefit from further optimizations. Finally, for the systems studied here the Roothaan-Hall (RH) / direct inversion of the iterative subspace (DIIS) approach is not the time-limiting step. For larger systems, however, linear scaling wave-function

optimization schemes[33,43–46] and linear scaling density-fitting procedures[30–34] will become important.

IV. CONCLUSIONS

In this paper we have demonstrated the efficient implementation of molecular forces for systems containing up to 500 atoms. The forces have been used for the geometry optimization of three selected molecules with up to 400 atoms. We have presented the first linear-scaling density-fitted-Coulomb force evaluation, and have further accelerated the near-field contribution to the Coulomb force evaluation using a novel scheme for the molecular integral evaluation, in which the solid-harmonic Gaussian are expanded in Hermit rather the Cartesian Gaussians - reducing the cost of differentiated integrals and simplifying the implementation. The results presented here clearly demonstrates the efficiency of the presented implementation.

Acknowledgments

We would like to acknowledge the financial support from the Norwegian Research Council through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420), through the Nanomat 'Molecular Modeling in Nanotechnology' program (grant no. 158538/431) and through the CeO Centre for Theoretical and Computational Chemistry (Grant No. 179568/V30). We would further like to acknowledge the NOTUR computing facilities which have been used to conduct the calculation presented in this paper.

---

[1] P. Pulay, *Ab initio* calculation of force constants and equilibrium geometries in polyatomic molecules, Mol. Phys. 17 (1969) 197.

[2] B. I. Dunlap, J. Andzelm, J. W. Mintmire, Local-density-functional total gradients in the linear combination of Gaussian-type orbitals method, Phys. Rev. **A** 42 (1990) 6354–6359.

[3] J. C. Burant, M. C. Strain, G. E. Scuseria, M. J. Frisch, Analytic energy gradients for the gaussian very fast multipole method (gvfmm), Chem. Phys. Lett. 248 (1996) 43–49.

[4] Y. Shao, C. A. White, M. Head-Gordon, Efficient evaluation of the coulomb force in density-functional theory calculation, J. Chem. Phys. 114 (2001) 6572–6577.

[5] D. Rappoport, F. Furche, Analytical time-dependent denisty functional derivative methods within the RI-J approximation, an approach to excited states of large molecules, J. Chem. Phys. 122 (2005) 064105.

[6] C. Ochsenfeld, Linear scaling exchange gradients for hartree-fock and hybrid density functional theory, Chem. Phys. Lett. 327 (2000) 216–223.

[7] R. Fournier, J. Andzelm, D. R. Salahub, Analytical gradient of the linear combination of Gaussian-type orbitals—local spin density energy, J. Chem. Phys. 90 (1989) 6371–6377.

[8] A. S. Amant, D. R. Salahub, Chem. Phys. Lett. 169 (1990) 387.

[9] A. Komornicki, G. Fitzgerald, J. Chem. Phys. 98 (1993) 1399.

[10] N. C. Handy, D. J. Tozer, G. J. Laming, C. W. Murray, R. D. Amos, Irs. J. Chem. 33 (1993) 331.

[11] B. G. Johnson, P. M. W. Gill, J. A. Pople, J. Chem. Phys. 98 (1993) 5612.

[12] L. Versluis, T. Ziegler, J. Chem. Phys. 88 (1988) 322.

[13] J. Almlöf, In Modern Electronic Structure Theory, *Ch. 3*; Yarkony, D. R., Ed., Vol. 1, World Scientific, Singapore, 1995.

[14] M. Häser, R. Ahlrichs, Improvements on the direct scf method, J. Comp. Chem. 10 (1989) 104–111.

[15] D. S. Lambrecht, C. Ochsenfeld, Multipole-based integral estimates for the rigorous description of distance dependence in two-electron integrals, J. Chem. Phys. 123 (2005) 184101.

[16] C. A. White, B. G. Johnson, P. M. W. Gill, M. Head-Gordon, The continuous fast multipole method, Chem. Phys. Lett. 230 (1994) 8–16.

[17] G. R. Ahmadi, J. Almlöf, The coulomb operator in a gaussian product basis, Chem. Phys. Lett. 246 (1995) 364–370.

[18] C. A. White, M. Head-Gordon, A j matrix engine for the density functional theory calculations, J. Chem. Phys. 104 (1996) 2620–2629.

[19] Y. Shao, M. Head-Gordon, An improved j matrix engine for density functional theory calculations, Chem. Phys. Lett. 323 (2000) 425–433.

[20] M. A. Watson, P. Sałek, P. Macak, T. Helgaker, Linear-scaling formation of Kohn-Sham hamiltonian: Application to the calculation of excitation energies an polarizabilities of large

molecular systems, J. Chem. Phys. 121 (2004) 2915–2931.

[21] L. Fsti-Molnr, P. Pulay, The fourier transform coulomb method: Efficient and accurate calculation of the coulomb operator in a gaussian basis, J. Chem. Phys. 117 (2002) 7827–7835.

[22] N. H. F. Beebe, J. Linderberg, Int. J. Quantum Chem. 7 (1977) 683.

[23] J. L. Whitten, Coulombic potential energy integrals and approximations, J. Chem. Phys. 58 (1973) 4496.

[24] E. J. Baerends, D. E. Ellis, R. P., Chem. Phys. 2 (1973) 41–51.

[25] B. I. Dunlap, J. W. D. Connolly, J. R. Sabin, On some approximations in applications of X_α theory, J. Chem. Phys. 71 (1979) 3396–3402.

[26] B. I. Dunlap, J. W. D. Connolly, J. R. Sabin, On first–row diatomic molecules and local density models, J. Chem. Phys. 71 (1979) 4993–4999.

[27] S. Reine, E. Tellgren, T. Helgaker, A unified scheme for the calculation of differentiated and undifferentiated integrals over solid-harmonic Gaussians, Phys. Chem. Chem. Phys. 9 (2007) 4771–4779.

[28] R. Ahlrichs, Efficient evaluation of three-center two-electron integrals over gaussian functions, Phys. Chem. Chem. Phys. 6 (2004) 5119–5121.

[29] F. R. Manby, P. J. Knowles, Poisson equation in the kohn-sham coulomb problem, Phys. Rev. Lett. 87 (2000) 163001.

[30] R. T. Gallant, A. St-Amant, Linear scaling for the charge density fitting procedure of the linear combination of gaussian-type orbitals density functional method, Chem. Phys. Lett. 256 (1996) 569–574.

[31] C. Fonseca Guerra, J. G. Snijders, G. te Velde, E. J. Baerends, Towards an order-N DFT method, Theor. Chem. Acc. 99 (1998) 391–403.

[32] A. Sodt, J. E. Subotnik, M. Head-Gordon, Linear scaling density fitting, J. Chem. Phys. 125 (2006) 194109.

[33] P. Sałek, S. Høst, L. Thøgersen, P. Jørgensen, P. Manninen, J. Olsen, B. Jansik, S. Reine, F. Pawlowski, E. Tellgren, T. Helgaker, S. Coriani, Linear-scaling implementation of molecular electronic self-consistent field theory, J. Chem. Phys. 126 (2007) 114110.

[34] S. Reine, E. Tellgren, A. Krapp, T. Kjærgaard, T. Helgaker, B. Jansik, S. Høst, P. Sałek, Variational and robust density fitting of four-center two-electron integrals in local metric,

J. Chem. Phys. 129 (2008) 104101.

[35] E. Schwegler, M. Challacombe, M. Head-Gordon, Linear scaling computation of the fock matrix. ii. rigorous bounds on exchange integrals and incremental fock build, J. Chem. Phys. 106 (1997) 9708–9717.

[36] C. Ochsenfeld, C. A. White, M. Head-Gordon, Linear and sublinear scaling formation of Hartree–Fock-type exchange matrices, J. Chem. Phys. 109 (1998) 1663–1669.

[37] F. Weigend, A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency, Phys. Chem. Chem. Phys. 4 (2002) 4285–4291.

[38] A. Sodt, M. Head-Gordon, Hartree-fock exchange computed using the atomic resolution of the identity approximation, J. Chem. Phys. 128 (2008) 104106.

[39] R. Polly, H.-J. Werner, F. R. Manby, P. J. Knowles, Fast hartree-fock theory using local density fitting approximations, Mol. Phys. 102 (2004) 2311–2321.

[40] van Wüllen, Chem. Phys. Lett. 219 (1994) 8.

[41] J. M. Perez-Jorda, W. Jang, Chem. Phys. Lett. 241 (1995) 469.

[42] R. E. Stratmann, Chem. Phys. Lett. 257 (1996) 213.

[43] T. Helgaker, P. Jørgensen, J. Olsen, Molecular Electronic–Structure Theory, Wiley, Chichester, 2000.

[44] T. Helgaker, H. Larsen, J. Olsen, P. Jørgensen, Direct optimization of the ao density matrix in hartree-fock and kohn-sham theories, Chem. Phys. Lett. 327 (2000) 397–403.

[45] H. Larsen, J. Olsen, P. Jørgensen, T. Helgaker, Direct optimization of the atomic-orbital density matrix using the conjugate-gradient method with a multilevel preconditioner, J. Chem. Phys. 115 (2001) 9685–9697.

[46] Y. Shao, C. Saravanan, M. Head-Gordon, C. A. White, Curvy steps for the density matrix-based energy minimization: Applications to large-scale self-consistent-field calculations, J. Chem. Phys. 118 (2003) 6144–6151.

[47] H. Larsen, , T. Helgaker, J. Olsen, P. Jørgensen, Geometrical derivatives and magnetic properties in atomic-orbital denisty-based Hartree-Fock theory, J. Chem. Phys. 115 (2001) 10344–10352.

[48] B. I. Dunlap, Robust variational fitting: Gspr's variational exchange can accurately be treated analytically, J. Mol. Struct. 501-502 (2000) 221–228.

[49] L. E. McMurchie, E. R. Davidson, One- and two-electron integrals over cartesian gaussian

functions, J. Comp. Phys. 26 (1978) 218–231.

[50] A. D. Becke, A multicenter numerical integration scheme for polyatomic molecules, J. Chem. Phys. 88 (1988) 2547–2553.

[51] V. I. Lebedev, D. N. Laikov, A quadrature formula for the sphere of the 131st algebraic order of accuracy, Doklady Mathematics 59 (1999) 477–481.

[52] V. I. Lebedev, A quadrature formula for the sphere of the 59th algebraic order of accuracy, Russian Acad. Sci. Dokl. Math. 50 (1995) 283–286.

[53] V. I. Lebedev, A. L. Skorokhodov, Quadrature formulas of orders 41, 47 and 53 for the sphere, Russian Acad. Sci. Dokl. Math. 45 (1992) 587–592.

[54] V. I. Lebedev, Spherical quadrature formulas exact to orders 25-29, Siberian Mathematical Journal 18 (1977) 99–107.

[55] V. I. Lebedev, Values of the nodes and weights of ninth to seventeenth oder gauss-markov quadrature formula invariant under the octahedron group of inversion, Computational Mathematics and Mathmatical Physics 16 (1975) 44–51.

[56] M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions, Dover, New York, 1970.

[57] O. Treutler, R. Ahlrichs, Efficient molecular numerical integration schemes, J. Chem. Phys. 102 (1995) 346–354.

[58] B. G. Johnson, P. M. W. Gill, J. A. Pople, The performance of a family of density functional methods, J. Chem. Phys. 98 (1993) 5612–5626.

[59] C. Peng, P. Y. Ayala, H. B. Schlegel, M. J. Frisch, Using redundant internal coordinates to optimize equilibrium geometries and transition states, J. Comput. Chem. 17 (1996) 49.

[60] F. Eckert, P. Pulay, H.-J. Werner, Ab initio geometry optimization for large molecules, J. Comput. Chem. 18 (1997) 1473.

[61] V. Bakken, T. Helgaker, The efficient optimization of molecular geometries using redundant internal coordinates, J. Chem. Phys. 117 (2002) 9160.

[62] Ö. Farkas, H. B. Schlegel, Methods for geometry optimization of large molecules. I. An $O(n^2)$ algorithm for solving systems of linear equations for the transformation of coordinates and forces, J. Chem. Phys. 109 (1998) 7100.

[63] Ö. Farkas, H. B. Schlegel, Methods for geometry optimization of large molecules. II. Quadratic search, J. Chem. Phys. 111 (1999) 10806.

[64] J. Baker, D. Kinghorn, P. Pulay, Geometry optimization in delocalized internal coordinates: An efficient quadratically scaling algorithm for large molecules, J. Chem. Phys. 110 (1999) 4986.

[65] S. R. Billeter, A. J. Turner, W. Thiel, Linear scaling geometry optimisation and transition state search in hybrid delocalised internal coordinates, Phys. Chem. Chem. Phys. 2 (2000) 2177.

[66] K. Németh, O. Coulaud, G. Monard, J. G. Ángyán, Linear scaling algorithm for the coordinate transformation problem of molecular geometry optimization, J. Chem. Phys. 113 (2000) 5598.

[67] K. Németh, O. Coulaud, G. Monard, J. G. Ángyán, An efficient method for the coordinate transformation problem of massively three-dimensional networks, J. Chem. Phys. 114 (2001) 9747.

[68] R. Fletcher, Practical Methods of Optimization Vol.1 - Unconstrained Optimization, J. Wiley & Sons Ltd., New York, 1981.

[69] H. J. A. Jensen, P. Jørgensen, T. Helgaker, Systematic determination of MCSCF equilibrium and transition structures and reaction paths, J. Chem. Phys. 85 (1986) 3917.

[70] R. Lindh, A. Bernhardsson, G. Karlström, P.-Å. Malmqvist, On the use of a Hessian model function in molecular geometry optimizations, Chem. Phys. Lett. 241 (1995) 423.

[71] DALTON, an *ab initio* electronic structure program, Release 2.0, see http://www.kjemi.uio.no/software/dalton/dalton.html (2005).

[72] K. Eichkorn, F. Weigend, O. Treutler, R. Ahlrichs, Auxiliary basis sets to approximate coulomb potentials, Theor. Chim. Acta 97 (1997) 119–124.

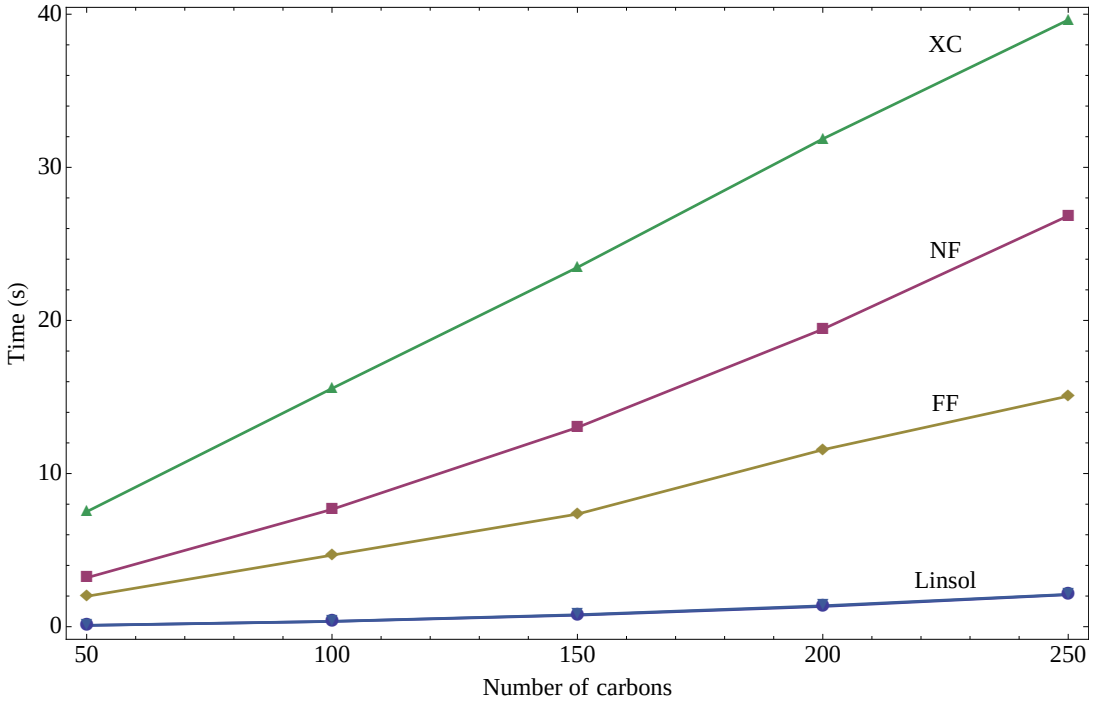

FIG. 1: Timings at the BP86/6-31G** level of theory for a single construction of the exchange-correlation (XC) and the density-fitted-Coulomb contribution to the Kohn-Sham matrix as a function of the number of carbons for linear polyene chains. The Coulomb timings are separated into timings for the near-field (NF), the far-field (FF) and the linear equation solver (Linsol). The auxiliary basis set of Ref. [72] was used as the density-fitting basis.

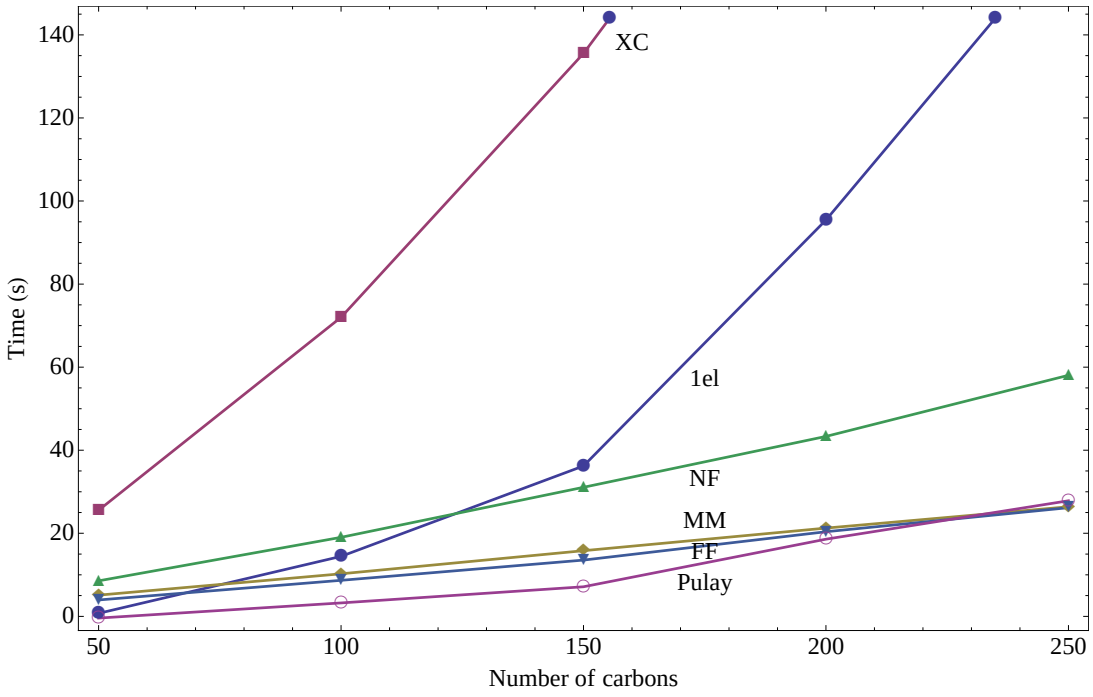

FIG. 2: Timings at the BP86/6-31G** level of theory for the calculation of the molecular forces as a function of the number of carbons for linear polyene chains. Timings are for the one-electron (1el), the exchange-correlation (XC), the near-field (NF) and the far-field (FF) contributions of the density-fitted Coulomb and the so-called Pulay force (see text). Also plotted are the timings for the calculation of the multipole moments of the differentiated charge distributions. The auxiliary basis set of Ref. [72] was used as the density-fitting basis.

System	#bf	NF-J	FF-J	XC	LIN
taxol	660	18.7	19.2	26.9	0.12
taxol*	1123	35.0	31.2	54.2	0.12
valinomycine	882	32.6	44.8	41.0	0.22
valinomycine*	1542	58.5	67.9	81.0	0.23
titin	2221	98.9	171.0	108.5	1.34

* Basis set 6-31G**

TABLE I: Timings in seconds for the single construction of the exchange-correlation (XC) and the density-fitted-Coulomb contribution to the Kohn-Sham matrix for the taxol, the valinomycine and the titin molecule at the BP86 level of theory. The Coulomb timings are given separately for the near-field (NF-J), the far-field (FF-J) and the linear equation solver (LIN). Also listed are the number of contracted basis functions (#bf). Timings for the taxol and valinomycine molecules are reported using both the 6-31G and the 6-31G** basis sets. The titin molecule uses the 6-31G basis set for all atoms, except for the three sulphur atoms and the two carbon atoms bonded to the di-sulphur bridge, which employ the 6-31G* basis. The auxiliary basis set of Ref. [72] was used as the density-fitting basis in all cases.

System	#bf	1-el	MM	NF-J	FF-J	XC	Pulay
taxol	660	11.5	15.8	46.5	26.0	88.5	3.2
taxol*	1123	43.4	40.0	89.8	44.4	153.0	5.7
valinomycine	882	29.3	23.3	72.6	54.7	139.4	6.7
valinomycine*	1542	110.8	57.5	135.6	86.4	236.0	11.7
titin	2221	241.8	61.0	206.0	187.3	419.6	48.4

* Basis set 6-31G**

TABLE II: Timings in seconds for the construction molecular forces for the taxol, the valinomycine and the titin molecule at the BP86 level of theory. Timings for the one-electron (1-el), the density-fitted Coulomb, the exchange-correlation (XC) and the so-called Pulay force (see text) are listed. The Coulomb timings are separated into timings for the near-field (NF-J), the far-field (FF-J) contributions and the construction of the multipole moments (MM) of the differentiated charge distributions. Also listed are the number of contracted basis functions (#bf). Timings for the calculations of taxol and valinomycine molecules are reported using both the 6-31G and the 6-31G** basis sets. For the titin molecule the 6-31G basis set is used for all atoms, except for the three sulphur atoms and the two carbon atoms bonded to the di-sulphur bridge, for which the 6-31G* basis is employed. The auxiliary basis set of Ref. [72] was used as the density-fitting basis in all cases.

System	#bf	#En	#SCF	Energy	#Geo	Force	Optimize
taxol	660	58	9.6	569.6	47	191.5	3.2
taxol*	1123	67	9.2	962.3	44	376.2	3.2
valinomycine	882	78	7.7	879.2	38	326.1	8.5
valinomycine*	1542	54	8.6	1603.0	32	638.1	8.5
titin	2221	31	9.3	3594.0	30	1163.9	105.9

* Basis set 6-31G**

TABLE III: Averaged timings in seconds for the construction of the molecular energies (Energy), the forces (Force) and for the geometry-optimization (Optimize) step for the taxol, the valinomycine and the titin molecule at the BP86 level of theory. Also listed are the number of contracted basis functions (#bf), the number of energy evaluations (#En) the average number of SCF cycles per energy evaluation (#SCF) and the number of geometry steps (#Geo). The results for the calculations of the taxol and valinomycine molecules are reported using both the 6-31G and the 6-31G** basis sets. For the titin molecule the 6-31G basis set is used for all atoms, except for the three sulphur atoms and the two carbon atoms bonded to the di-sulphur bridge, for which the 6-31G* basis is employed. The auxiliary basis set of Ref. [72] was used as the density-fitting basis in all cases.

Paper V

Towards black-box linear scaling optimization in Hartree-Fock and Kohn-Sham theories
S. Høst, J. Olsen, B. Jansik, P. Jørgensen, **S. Reine**, T. Helgaker, P. Sałek and S. Coriani
Lecture Series on Computer and Computational Sciences, **1**, 1-10 (2006)

Brill Academic Publishers
P.O. Box 9000, 2300 PA Leiden,
The Netherlands

Lecture Series on Computer and Computational Sciences
Volume 1, 2006, pp. 1-10

Towards black-box linear scaling optimization in Hartree-Fock and Kohn-Sham theories

Stinne Høst, Jeppe Olsen, Branislav Jansik, Poul Jørgensen[1]

Center for Theoretical Chemistry,
Department of Chemistry,
University of Aarhus,
DK-8000 Århus C, Denmark

Simen Reine, Trygve Helgaker

Department of Chemistry,
University of Oslo,
P. O. Box 1033 Blindern, N-0315 Norway

Paweł Sałek

Laboratory of Theoretical Chemistry,
The Royal Institute of Technology,
Teknikringen 30, Stockholm SE-10044, Sweden

Sonia Coriani

Dipartimento di Scienze Chimiche, Università degli Studi di Trieste, Via Licio Giorgieri 1,
I-34127 Trieste, Italy

Received 1 August, 2006; accepted in revised form ?? , 2006

Abstract: A linear scaling implementation of the trust-region self-consistent field (LS-TRSCF) method for the Hartree-Fock and Kohn-Sham calculations is described. The convergence of the method is smooth and robust and of equal quality for small and large systems. The LS-TRSCF calculations converge in several cases where conventional DIIS calculations diverge. The LS-TRSCF method may be recommended as the standard method for both small and large molecular systems.

Keywords: Linear scaling SCF, Hartree-Fock optimization, Kohn-Sham optimization, trust-region method

Mathematics Subject Classification: 31.15-p

1 Introduction

In Hartree-Fock (HF) and Kohn-Sham (KS) density functional theory (DFT), the electronic energy E_{SCF} is minimized with respect to the density matrix of a single-determinantal wave function. In its original formulation, the minimization was carried out using the self-consistent field (SCF) method consisting of a sequence of Roothaan-Hall iterations. At each iteration, the Fock/KS matrix $\mathbf{F}$ is constructed from the current atomic-orbital (AO) density matrix $\mathbf{D}$; next, the Fock/KS matrix is diagonalized and finally an improved AO density matrix is determined from the molecular orbitals

[1]Corresponding author. E-mail: pou@chem.au.dk

(MOs) obtained by this diagonalization. Unfortunately, this simple SCF scheme converges only in simple cases.

To improve upon the convergence, the optimization is modified by constructing the Fock/KS matrix not directly from the AO density matrix of the last iteration, but rather from an averaged density matrix, obtained as a linear combination of the density matrices of the current and previous iterations. Typically, the averaged density matrix is obtained using the DIIS method of Pulay [1], by minimizing the norm of the linear combination of the gradients. The SCF/DIIS method has been implemented in most electronic-structure programs and has been successfully used to obtain optimized HF/KS energies. However, in some cases the DIIS procedure fails to converge.

During the last decade, much effort has been directed towards developing linear scaling SCF methods. In particular, the computational scaling for the evaluation of the Fock/KS matrix has been successfully reduced by use of the fast multipole method (FMM) for the Coulomb contribution [2]–[6], the order-N exchange (ONX) method and the linear exchange K (LinK) method for the exact (Hartree–Fock) exchange contribution [7]–[12], and efficient numerical quadrature methods for the exchange–correlation (XC) contribution [13]–[15]. Our SCF code uses FMM combined with density fitting for the Coulomb contribution, LinK for the exact exchange contribution, and linear-scaling numerical quadrature for the XC contribution. In the optimization of the SCF energy, the diagonalization of the Fock/KS matrix, which scales cubically with the system size (N^3), may therefore become the time dominating step for large molecules. In this paper, we discuss how the SCF method with DIIS may be improved upon by using an algorithm where the diagonalization of the Fock/KS matrix is avoided in favour of a method of linear complexity.

In the SCF/DIIS method, the minimization of the energy is carried out in two separate steps: the diagonalization of the Fock/KS matrix and the averaging of the density matrix. In neither step an energy lowering is enforced on E_{SCF}. It is simply hoped that, at the end of the SCF iterations, an optimized state is determined. We discuss improvements to both the diagonalization and the density matrix averaging, where a lowering of the energy E_{SCF} is enforced at each iteration. For both steps, we construct a local energy model to E_{SCF} with the current density matrix as the expansion point. At the expansion point, these models have the true gradient, but only an approximate Hessian. They are therefore valid only in a restricted region about the expansion point - the trust region. When these local models are used, it is essential that steps are only generated within the trust region, as otherwise no energy lowering is guaranteed.

Diagonalization may be avoided by recognizing that the density matrix obtained by diagonalizing the Fock/KS matrix represents the global minimum of the Roothaan-Hall energy function $E^{\mathrm{RH}} = \mathrm{Tr}\,\mathbf{FD}$ (with fixed $\mathbf{F}$) [16, 17]. The diagonalization may therefore be replaced by a minimization of E^{RH}. However, since E^{RH} is only a crude model of the true energy E_{SCF}, a complete minimization of E^{RH} (as obtained for example by diagonalization) may give steps that are too long to be trusted. When minimizing E^{RH}, we require the steps to be inside the trust region, solving a set of level shifted Newton equations where the level shift controls the size of the steps. The level shifted Newton equations may be solved using iterative algorithms where the time-dominating step is the multiplication of the Hessian by trialvectors. Linear complexity therefore may be obtained by using sparse matrix algebra. The obtained algorithm will be denoted the linear scaling trust-region Roothaan-Hall (LS-TRRH) method.

To improve on the DIIS scheme we construct an energy function where the expansion coefficients of the averaged density matrix are the variational parameters. Carrying out a second-order expansion of this energy, using the quasi-Newton condition and neglecting terms that require evaluation of new Fock/KS matrices, we arrive at the density subspace minimization (DSM) approximation to the energy E^{DSM} [18, 19]. At the expansion point, E^{DSM} has the same gradient as E_{SCF} and a good approximation to the Hessian. Again, trust-region optimization may be used to determine the optimal expansion coefficients, ensuring also an energy lowering at this step of the iterative procedure. The obtained algorithm is denoted the trust-region density subspace minimization

(TRDSM) method. Combining the LS-TRRH and TRDSM amethods we obtain the LS-TRSCF method.

In the next section we describe the LS-TRRH algorithm while in section 3 the TRDSM algorithm is discussed. Section 4 contains numerical results which demonstrate the convergence of LS-TRSCF calculations and that linear scaling is obtained. The last section contains some concluding remarks.

2 Optimization of the Roothaan–Hall energy

2.1 Parametrization of the density matrix

Let $\mathbf{D}$ be a valid Kohn–Sham density matrix of an N-electron system, which together with the AO overlap matrix $\mathbf{S}$ satisfies the symmetry, trace and idempotency relations:

$$\mathbf{D}^{\mathrm{T}} = \mathbf{D} \tag{1}$$

$$\mathrm{Tr}\,\mathbf{DS} = N \tag{2}$$

$$\mathbf{DSD} = \mathbf{D} \tag{3}$$

Introducing the projectors $\mathbf{P}_{\mathrm{o}}$ and $\mathbf{P}_{\mathrm{v}}$ on the occupied and virtual spaces

$$\mathbf{P}_{\mathrm{o}} = \mathbf{DS} \tag{4}$$

$$\mathbf{P}_{\mathrm{v}} = \mathbf{I} - \mathbf{DS} \tag{5}$$

we may, from the reference density matrix $\mathbf{D}$, generate any other valid density matrix by the transformation [17, 21, 22]

$$\mathbf{D}(\mathbf{X}) = \exp\left[-\mathcal{P}(\mathbf{X})\mathbf{S}\right]\mathbf{D}\exp\left[\mathbf{S}\mathcal{P}(\mathbf{X})\right] \tag{6}$$

where $\mathbf{X}$ is an anti-Hermitian matrix and where

$$\mathcal{P}(\mathbf{X}) = \mathbf{P}_{\mathrm{o}}\mathbf{X}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}} + \mathbf{P}_{\mathrm{v}}\mathbf{X}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}} \tag{7}$$

projects out the redundant occupied-occupied and virtual-virtual components of $\mathbf{X}$.

The density matrix $\mathbf{D}(\mathbf{X})$ may be expanded in orders of $\mathbf{X}$ as

$$\mathbf{D}(\mathbf{X}) = \mathbf{D} + \left[\mathbf{D}, \mathcal{P}(\mathbf{X})\right]_{\mathrm{S}} + \tfrac{1}{2}\left[\left[\mathbf{D}, \mathcal{P}(\mathbf{X})\right]_{\mathrm{S}}, \mathcal{P}(\mathbf{X})\right]_{\mathrm{S}} + \cdots \tag{8}$$

where we have introduced the S commutator

$$\left[\mathbf{A}, \mathbf{B}\right]_{\mathrm{S}} = \mathbf{ASB} - \mathbf{BSA} \tag{9}$$

2.2 The Roothaan–Hall Newton equations in the AO basis

In an SCF optimization, the diagonalization of the Fock/KS matrix $\mathbf{F}$ is equivalent to the minimization of the Roothaan–Hall energy [17]

$$E^{\mathrm{RH}}(\mathbf{X}) = \mathrm{Tr}\left[\mathbf{FD}(\mathbf{X})\right] \tag{10}$$

in the sense that both approaches yield the same density matrix. Inserting the S-commutator expansion of the density matrix $\mathbf{D}(\mathbf{X})$, we obtain

$$\begin{aligned}\mathrm{Tr}\left[\mathbf{FD}(\mathbf{X})\right] = \mathrm{Tr}\left(\mathbf{FD}\right) &+ \mathrm{Tr}\left(\mathbf{F}^{\mathrm{vo}}\mathbf{X} - \mathbf{F}^{\mathrm{ov}}\mathbf{X}\right) \\ &+ \mathrm{Tr}\left(\mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\mathbf{X} - \mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}}\mathbf{X}\right) + \cdots\end{aligned} \tag{11}$$

where we have made repeated use of the idempotency relations $\mathbf{P}_{\mathrm{o}}^2 = \mathbf{P}_{\mathrm{o}}$ and $\mathbf{P}_{\mathrm{v}}^2 = \mathbf{P}_{\mathrm{v}}$ and of the orthogonality relations $\mathbf{P}_{\mathrm{o}}\mathbf{P}_{\mathrm{v}} = \mathbf{P}_{\mathrm{v}}\mathbf{P}_{\mathrm{o}} = \mathbf{0}$ and $\mathbf{P}_{\mathrm{o}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}} = \mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{o}} = \mathbf{0}$ and introduced the short-hand notation

$$\mathbf{F}^{\mathrm{ab}} = \mathbf{P}_{\mathrm{a}}^{\mathrm{T}}\mathbf{F}\mathbf{P}_{\mathrm{b}} \tag{12}$$

Note that, whereas the off-diagonal projections $\mathbf{F}^{\mathrm{ov}}$ and $\mathbf{F}^{\mathrm{vo}}$ of $\mathbf{F}$ contribute to the terms linear in $\mathbf{X}$, the diagonal projections $\mathbf{F}^{\mathrm{oo}}$ and $\mathbf{F}^{\mathrm{vv}}$ contribute to the quadratic terms.

The Roothaan-Hall energy E^{RH} is only a crude model of the true HF/KS energy E_{SCF}, having the correct gradient but an approximate Hessian at the point of expansion; this can be understood from the observation that E^{RH} depends linearly on $\mathbf{D}(\mathbf{X})$, whereas the true energy E_{SCF} depends quadratically on $\mathbf{D}(\mathbf{X})$. Therefore, a complete minimization of E^{RH} (as achieved, for example, by diagonalization of the Fock/KS matrix), may give steps that are too long to be trusted. Such steps may, for example, increase rather than decrease the total SCF energy. We therefore impose on the energy minimization the constraint that the new occupied space does not differ appreciably from the old occupied space. The step must therefore be inside or on the boundary of the trust region of E^{RH}, which we define as a hypersphere with radius h around the density at the current expansion point. In the $\mathbf{S}$ metric norm, the length of the step

$$||\mathcal{P}(\mathbf{X})||_S^2 = \mathrm{Tr}[\mathcal{P}(\mathbf{X})\mathbf{S}\mathcal{P}(\mathbf{X})\mathbf{S}] \tag{13}$$

is thus restricted to h^2. To satisfy this constraint, we introduce an undetermined multiplier μ and set up the Lagrangian

$$L^{\mathrm{RH}}(\mathbf{X}) = \mathrm{Tr}\left[\mathbf{F}\mathbf{D}(\mathbf{X})\right] - \frac{1}{2}\mu\left(\mathrm{Tr}[\mathcal{P}(\mathbf{X})\mathbf{S}\mathbf{X}\mathbf{S}] - h^2\right) \tag{14}$$

Expanding this Lagrangian to second order in $\mathbf{X}$ using Eq. (11), we obtain

$$\begin{aligned} L^{\mathrm{RH}}(\mathbf{X}) = {} & \mathrm{Tr}\left(\mathbf{F}\mathbf{D}\right) + \mathrm{Tr}\left(\mathbf{F}^{\mathrm{vo}}\mathbf{X} - \mathbf{F}^{\mathrm{ov}}\mathbf{X}\right) \\ & + \mathrm{Tr}\left(\mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\mathbf{X} - \mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}}\mathbf{X}\right) + \mu\left[\mathrm{Tr}\left(\mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\mathbf{X}\right) - \frac{1}{2}h^2\right]\cdots \end{aligned} \tag{15}$$

Differentiating this function with respect to the elements of $\mathbf{X}$, we obtain

$$\begin{aligned} \frac{\partial L^{\mathrm{RH}}(\mathbf{X})}{\partial \mathbf{X}} = {} & \mathbf{F}^{\mathrm{ov}} - \mathbf{F}^{\mathrm{vo}} - \mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{F}^{\mathrm{oo}} - \mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}} + \mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{F}^{\mathrm{vv}} \\ & - \mu\left(\mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\right) + \cdots \end{aligned} \tag{16}$$

where we have used the relation

$$\frac{\partial\,\mathrm{Tr}(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}^{\mathrm{T}} \tag{17}$$

Finally, setting the right-hand side equal to zero and ignoring higher-order contributions, we obtain the matrix equation

$$\begin{aligned} & \mathbf{F}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} - \mathbf{F}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{F}^{\mathrm{vv}} - \mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{F}^{\mathrm{oo}} \\ & \qquad - \mu\left(\mathbf{S}^{\mathrm{vv}}\mathbf{X}\mathbf{S}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{X}\mathbf{S}^{\mathrm{vv}}\right) = \mathbf{F}^{\mathrm{vo}} - \mathbf{F}^{\mathrm{ov}} \end{aligned} \tag{18}$$

for the stationary points on the trust sphere of the Roothaan–Hall energy function.

Eq. (18) is equivalent to a level shifted set of Newton equations

$$(\mathbf{H} - \mu\mathbf{M})\mathbf{x} = \mathbf{G} \tag{19}$$

where

$$\mathbf{H} = \mathbf{F}^{\mathrm{vv}} \otimes \mathbf{S}^{\mathrm{oo}} - \mathbf{F}^{\mathrm{oo}} \otimes \mathbf{S}^{\mathrm{vv}} + \mathbf{S}^{\mathrm{oo}} \otimes \mathbf{F}^{\mathrm{vv}} - \mathbf{S}^{\mathrm{vv}} \otimes \mathbf{F}^{\mathrm{oo}} \tag{20}$$

$$\mathbf{M} = \mathbf{S}^{\mathrm{vv}} \otimes \mathbf{S}^{\mathrm{oo}} - \mathbf{S}^{\mathrm{oo}} \otimes \mathbf{S}^{\mathrm{vv}} \tag{21}$$

$$\mathbf{G} = \mathrm{Vec}(\mathbf{F}^{\mathrm{vo}} - \mathbf{F}^{\mathrm{ov}}) \tag{22}$$

$$\mathbf{x} = \mathrm{Vec}\mathbf{X} \tag{23}$$

2.3 The Roothaan–Hall Newton equations in an orthonormal basis

The conditioning number of the level shifted Hessian matrix in Eq. (19) is greatly reduced by transforming the equation to an orthogonal basis. We consider transformations based on the factorization of the overlap in the form

$$\mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{V} \tag{24}$$

Such a factorization may be accomplished in infinitely many ways – for example, by introducing a Cholesky factor (as employed by Shao *et al.* [23] in the curvy step method) or the principal square root

$$\mathbf{V}_{\mathrm{c}} = \mathbf{U} \tag{25}$$

$$\mathbf{V}_{\mathrm{s}} = \mathbf{S}^{1/2} \tag{26}$$

where $\mathbf{U}$ is an nonsingular upper triangular matrix and where $\mathbf{S}^{1/2}$ is a positive-definite symmetric matrix. In the chosen orthonormal basis, the Roothaan–Hall Newton equations Eq. (18) take the form

$$(\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}} - \mu\mathbf{I})\mathbf{X}^{\mathrm{V}} + \mathbf{X}^{\mathrm{V}}(\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}} - \mu\mathbf{I}) = \mathbf{F}_{\mathrm{V}}^{\mathrm{vo}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{ov}} \tag{27}$$

where we have introduced the notation

$$\mathbf{A}_{\mathrm{V}} = \mathbf{V}^{-\mathrm{T}}\mathbf{A}\mathbf{V}^{-1} \tag{28}$$

$$\mathbf{A}^{\mathrm{V}} = \mathbf{V}\mathbf{A}\mathbf{V}^{\mathrm{T}} \tag{29}$$

and where we have further assumed that $\mathbf{X}^{\mathrm{V}}$ contains only non-redundant components.

Eq. (27) represents the solution of a level shifted Newton set of linear equations

$$(\mathbf{H}_{\mathrm{V}} - \mu\mathbf{I})\mathbf{x}^{\mathrm{V}} = \mathbf{G}_{\mathrm{V}} \tag{30}$$

where

$$\mathbf{H}_{\mathrm{V}} = (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}}) \otimes \mathbf{I} + \mathbf{I} \otimes (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}}) \tag{31}$$

$$\mathbf{x}^{\mathrm{V}} = \mathrm{Vec}\mathbf{X}^{\mathrm{V}} \tag{32}$$

$$\mathbf{G}_{\mathrm{V}} = \mathrm{Vec}(\mathbf{F}_{\mathrm{V}}^{\mathrm{vo}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{ov}}) \tag{33}$$

When solving Eq. (30) by the conjugate gradient method, it is advantageous to use a diagonal preconditioner.

In the global region of an SCF optimization, the boundary of the trust region is represented by $X_{\mathrm{V}}^{\max} = k$, where $X_{\mathrm{V}}^{\max}$ is the largest component of $\mathbf{X}^{\mathrm{V}}$ and k is 0.35. Unlike $||\mathbf{X}^{\mathrm{V}}||_S$, $X_{\mathrm{V}}^{\max}$ is size-intensive.

To ensure that the minimum is determined on the boundary of the trust region, the level shift must be restricted to the interval $-\infty < \mu < \epsilon_{\min}$ where $\epsilon_{\min}$ is the lowest eigenvalue of the Hessian Eq. (31). In principle, the lowest Hessian eigenvalue should therefore be determined and a line search carried out in the interval $-\infty < \mu < \epsilon_{\min}$ to find the level shift μ with $X_{\mathrm{V}}^{\max} = 0.35$.

However, a simpler strategy is obtained by recognizing that the solution of the level shifted Newton equations can be determined from the eigenvectors of the augmented Hessian eigenvalue equation [24, 25, 26]. If the solution with the lowest eigenvalue is determined, the level shift is restricted to the interval $-\infty < \mu < \epsilon_{\min}$.

The level shifted Newton equations may be solved using an iterative procedure where the reduced space Hessians and gradients are set up in each iteration. At each iteration, the augmented Hessian may therefore also be set up in the subspace at essentially no cost and the lowest eigenvalue determined. Consequently the level shift may be updated by solving the reduced space augmented Hessian eigenvalue problem at no extra cost. With the updated level shift, a new Newton iteration may be carried out and the iterations continued until convergence is obtained with respect to level shift and the residual of the Newton equations (see Ref. [27]).

When the level shifted Newton equations are solved using iterative algorithms, the time consuming step is the linear transformation of the Hessian matrix on trial vectors. Using sparse matrix algebra, linear scaling may be obtained in these linear transformations.

3 The density subspace minimization (DSM) algorithm

After a sequence of Roothaan–Hall iterations, we have determined a set of density matrices $\mathbf{D}_i$ and a corresponding set of Fock/KS matrices $\mathbf{F}_i = \mathbf{F}(\mathbf{D}_i)$. We now discuss how to make the best use of the information contained in these matrices.

3.1 Parametrization of the DSM density matrix

Using $\mathbf{D}_0$ as the reference density matrix, the improved density matrix may be expressed as a linear combination of the current and previous density matrices [18, 19]

$$\overline{\mathbf{D}} = \mathbf{D}_0 + \sum_{i=0}^{n} c_i \mathbf{D}_i. \tag{34}$$

Ideally $\overline{\mathbf{D}}$ should satisfy the symmetry, trace and idempotency conditions Eqs. (1-3). The symmetry condition Eq. (1) is trivially satisfied while the trace condition Eq. (2) holds only if

$$c_0 = -\sum_{i=1}^{n} c_i. \tag{35}$$

Using c_i with $1 \leq i \leq n$ as independent parameters the density matrix $\overline{\mathbf{D}}$ may be expressed as

$$\overline{\mathbf{D}} = \mathbf{D}_0 + \mathbf{D}_+, \tag{36}$$

where we have introduced the notation

$$\mathbf{D}_+ = \sum_{i=1}^{n} c_i \mathbf{D}_{i0}, \tag{37a}$$

$$\mathbf{D}_{i0} = \mathbf{D}_i - \mathbf{D}_0. \tag{37b}$$

While $\overline{\mathbf{D}}$ satisfies the symmetry and trace conditions Eqs. (1) and (2), the idempotency condition Eq. (3) is not fulfilled. A smaller idempotency error may be obtained using the purified density matrix of McWeeny [20, 28]

$$\widetilde{\mathbf{D}} = 3\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}} - 2\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}}. \tag{38}$$

Emphasizing that $\mathbf{D}_0$ is the reference density matrix, the first-order purified density matrix may be expressed as

$$\widetilde{\mathbf{D}} = \mathbf{D}_0 + \mathbf{D}_+ + \mathbf{D}_\delta. \tag{39}$$

where we have introduced the idempotency correction

$$\mathbf{D}_\delta = \widetilde{\mathbf{D}} - \overline{\mathbf{D}}. \tag{40}$$

3.2 Construction of the DSM energy function

Expanding the energy for the purified averaged density matrix, Eq. (39), around the reference density matrix $\mathbf{D}_0$, we obtain to second order

$$E(\widetilde{\mathbf{D}}) = E(\mathbf{D}_0) + (\mathbf{D}_+ + \mathbf{D}_\delta)^{\mathrm{T}}\, \mathbf{E}_0^{(1)} + \frac{1}{2}\,(\mathbf{D}_+ + \mathbf{D}_\delta)^{\mathrm{T}}\, \mathbf{E}_0^{(2)}\,(\mathbf{D}_+ + \mathbf{D}_\delta) \tag{41}$$

To evaluate the terms containing $\mathbf{E}_0^{(1)}$ and $\mathbf{E}_0^{(2)}$, we first recall that the Fock/KS matrix is defined as

$$\mathbf{E}_0^{(1)} = 2\mathbf{F}_0 \tag{42}$$

Next we carry out an expansion of $\mathbf{E}_i^{(1)}$ with $\mathbf{D}_0$ as expansion point

$$\mathbf{E}_i^{(1)} = \mathbf{E}_0^{(1)} + \mathbf{E}_0^{(2)}(\mathbf{D}_i - \mathbf{D}_0) + \mathcal{O}\big(\mathbf{D}_i - \mathbf{D}_0\big)^2 \tag{43}$$

Neglecting terms of order $\mathcal{O}\big(\mathbf{D}_i - \mathbf{D}_0\big)^2$ we obtain the quasi–Newton condition

$$\mathbf{E}_0^{(2)}(\mathbf{D}_i - \mathbf{D}_0) = 2\mathbf{F}_i - 2\mathbf{F}_0 = 2\mathbf{F}_{i0} \tag{44}$$

which may be used to obtain

$$\mathbf{E}_0^{(2)}\mathbf{D}_+ = 2\mathbf{F}_+ + \mathcal{O}\big(\mathbf{D}_+^2\big), \tag{45}$$

where we have generalized the notation Eq. (37a) to the Fock/KS matrix

$$\mathbf{F}_+ = \sum_{i=1}^{n} c_i \mathbf{F}_{i0} \tag{46}$$

Using Eq. (42) and Eq. (45) and ignoring the terms quadratic in $\mathbf{D}_\delta$ in Eq. (41) and quadratic in $\mathbf{D}_+$ in Eq. (45), we then obtain the DSM energy

$$E^{\mathrm{DSM}}(\mathbf{c}) = E(\mathbf{D}_0) + 2\,\mathrm{Tr}\,\mathbf{D}_+\mathbf{F}_0 + \mathrm{Tr}\,\mathbf{D}_+\mathbf{F}_+ + 2\,\mathrm{Tr}\,\mathbf{D}_\delta\mathbf{F}_0 + 2\,\mathrm{Tr}\,\mathbf{D}_\delta\mathbf{F}_+. \tag{47}$$

Note that $E^{\mathrm{DSM}}(\mathbf{c})$ is expressed solely in terms of the density and Fock/KS matrices of the current and previous iterations. For a more compact notation, we introduce the weighted Fock/KS matrix

$$\overline{\mathbf{F}} = \mathbf{F}_0 + \mathbf{F}_+ = \mathbf{F}_0 + \sum_{i=1}^{n} c_i \mathbf{F}_{i0} \tag{48}$$

and find that the DSM energy may be written in the form

$$E^{\mathrm{DSM}}(\mathbf{c}) = E(\overline{\mathbf{D}}) + 2\,\mathrm{Tr}\,\mathbf{D}_\delta\overline{\mathbf{F}}, \tag{49}$$

where the first term is quadratic in the expansion coefficients c_i

$$E(\overline{\mathbf{D}}) = E(\mathbf{D}_0) + 2\,\mathrm{Tr}\,\mathbf{D}_+\mathbf{F}_0 + \mathrm{Tr}\,\mathbf{D}_+\mathbf{F}_+, \tag{50}$$

and the second (idempotency correction) term is quartic in these coefficients:

$$2\,\mathrm{Tr}\,\mathbf{D}_\delta\overline{\mathbf{F}} = \mathrm{Tr}(6\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}} - 4\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}}\mathbf{S}\overline{\mathbf{D}} - 2\overline{\mathbf{D}})\overline{\mathbf{F}}. \tag{51}$$

The derivatives of $E^{\mathrm{DSM}}(\mathbf{c})$ are straightforwardly obtained by inserting the expansions of $\overline{\mathbf{F}}$ and $\overline{\mathbf{D}}$, using the independent parameter representation and the minimization of $E^{\mathrm{DSM}}(\mathbf{c})$ may straightforwardly be carried out using the trust-region method.

4 Numerical Illustrations

4.1 Convergence of test calculations

We now describe the convergence of test calculations for Hartree-Fock and DFT LDA using the LS-TRSCF algorithm where the level shifted Newton equations are solved in the basis defined by the principal square root in Eq. (26). For comparison, the convergence of the standard SCF/DIIS calculations (diagonalization + DIIS, no level shift) will also be reported. In both DIIS and TRDSM a maximum of eight densities and Fock/KS matrices are stored.

In Fig. 1 we display the convergence (the difference between the energy of a given iteration and the converged energy) of Hartree-Fock calculations using LS-TRSCF (left panel) and SCF/DIIS (right panel) on six molecules representing different types of chemical compounds: 1). Water, stretched: H_2O where the O−H bond is twice its equilibrium value (d-aug-pVTZ basis). 2). Rh complex: The rhodium complex of Ref. [18] (AhlrichsVDZ basis [29], STO-3G on Rh). 3). Cd complex: The cadmium-imidazole complex of Ref. [19] (3-21G basis). 4). Zn complex: The zinc-EDDS complex of Ref. [19] (6-31G basis). 5). Polysaccharide: A polysaccharide containing 438 atoms (6-31G basis). 6). Polyalanine, 24 units: A polypeptide containing 24 alanine residues (6-31G basis). As initial guess we have used H1 core for molecules 1–3 and Hückel for molecules 4–6. Smooth convergence to 10^{-8} a.u. is obtained in all LS-TRSCF calculations. Convergence is obtained in 12-30 iterations. The convergence is very similar for the SCF/DIIS and the LS-TRSCF calculations except for the rhodium complex, where the SCF/DIIS calculation diverges while smooth convergence is obtained using the LS-TRSCF algorithm. The local convergence is very similar for SCF/DIIS and LS-TRSCF reflecting that in both DIIS and DSM, the local convergence is determined by the fact that the quasi-Newton condition is satisfied [19]. In Fig. 2, we report calculations similar to those in Fig. 1 but where the Hartree-Fock model is replaced by LDA. The convergence of the LS-TRSCF Hartree-Fock and LDA calculations is very similar with the exception of the Rh complex where the LDA calculation has a rather erratic behaviour from about iteration 20 to 80 after which fast convergence is obtained. The SCF/DIIS LDA calculations in the left panel in Fig. 2 show a rather erratic convergence behaviour in particular for the Cd complex and polyalanine where the calculations diverge, and for the polysaccharide calculation in the initial 25 iterations. The erratic behaviour which in general is observed in the initial iterations of SCF/DIIS calculations reflects that energy lowering is not an issue in the SCF/DIIS scheme. Surprisingly, the SCF/DIIS LDA calculation on the rhodium complex converges, while the corresponding Hartree-Fock calculation diverges.

To sum it up, similar convergence is seen in Hartree-Fock SCF/DIIS and the LS-TRSCF calculations, whereas for LDA, a much more smooth and robust convergence is obtained by using the LS-TRSCF scheme. Particularly in the initial iterations, a more erratic behaviour is seen with the SCF/DIIS algorithm. In several cases the LS-TRSCF calculations converge, where the SCF/DIIS calculations diverge.

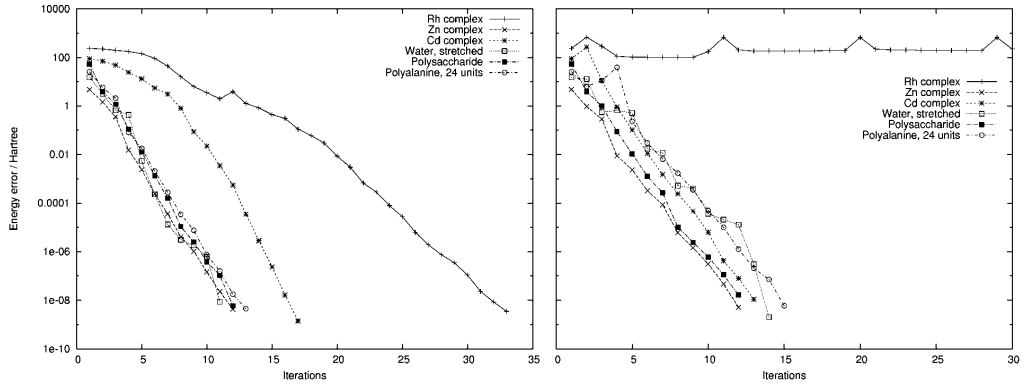

Figure 1: Convergence of the Hartree-Fock LS-TRSCF (left panel) and SCF/DIIS (right panel) calculations for the rhodium complex, the zinc complex, the cadmium complex, the stretched water, the polysaccharide and the polyalanine. The energy error (a.u.) in each iteration is plotted versus number of iterations.

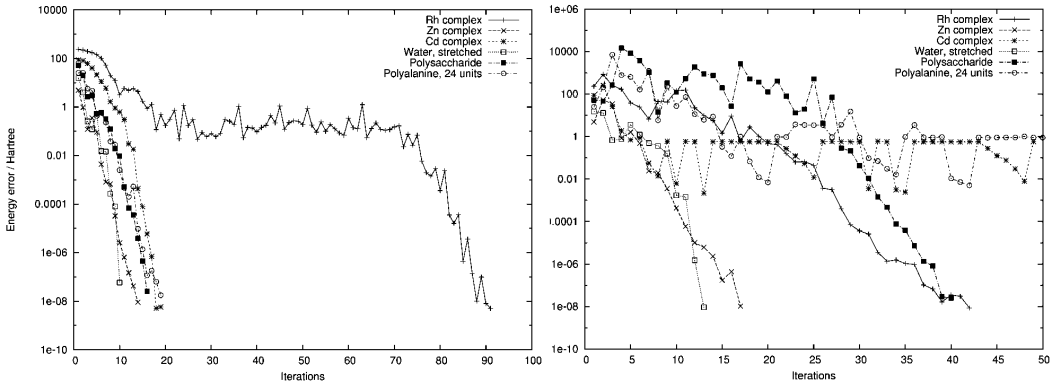

Figure 2: Convergence of the LDA LS-TRSCF (left panel) and SCF/DIIS (right panel) calculations for the rhodium complex, the zinc complex, the cadmium complex, the stretched water, the polysaccharide and the polyalanine. The energy error (a.u.) in each iteration is plotted versus number of iterations.

4.2 Linear scaling using the LS-TRSCF algorithm

In this subsection, we will illustrate that linear scaling is obtained using the LS-TRSCF algorithm. We consider calculations on a polyalanine peptide where we extend the number of alanine residues. We consider both Hartree-Fock and B3LYP calculations in the 6-31G basis. The largest alanine peptide contains 119 alanine residues (a total of 1192 atoms). The convergence of the alanines is similar to the one for the 24 residue peptide given in Figs. 1-2.

In Fig. 3 we have shown the CPU time used in the different parts of the LS-TRSCF algorithm for the Hartree-Fock calculations using sparse matrix algebra In all figures, the timings are for the first iteration in the local region, except for the DSM time, which is dependent on the number of previous densities. Therefore, the DSM time is always given for iteration 8, where we have the maximum number of previous densities involved. The timings are given for the evaluation of the Coulomb (Fock J) and exchange (Fock X) parts of the Fock matrix, respectively, and for the LS-TRRH step and for the TRDSM step. The curve for the most expensive step – the exchange part of the Fock matrix – has a bend probably due to an N^2 scaling sorting routine. For both the LS-TRRH and TRDSM steps, the time consuming part consists of matrix multiplications. Both LS-TRRH and TRDSM scale linearly with system size in the calculations in Fig. 3, showing that sparsity is efficiently exploited in the matrix multiplications. The benefits from exploiting the sparsity of the

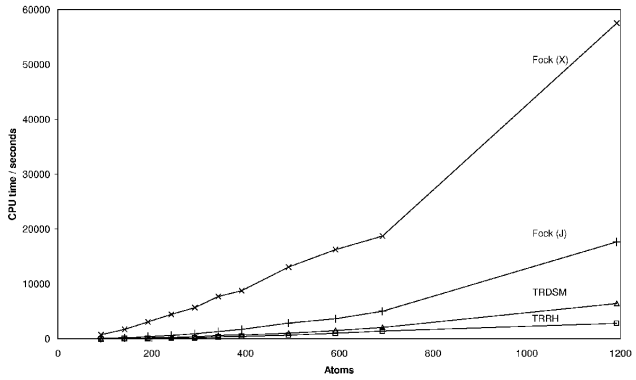

Figure 3: CPU timings for one iteration of a Hartree-Fock calculation using a 6-31G basis plotted as a function of the number of atoms in a polyalanine peptide. The considered contributions are the exchange (X) and Coulomb (J) contributions to Fock matrix in addition to the LS-TRRH and TRDSM optimization steps where sparse matrix algebra is used.

involved matrices are evident from Fig. 4, where we have plotted the CPU times for the LS-TRRH and TRDSM steps from Fig. 3 in combination with timings for calculations where the matrix multiplications involve full (dense) matrices. The timings for full matrix representations increase with system size in accorcance with cubic scaling, but become linear when sparsity is exploited. As seen on the figure, the advantage of going to the sparse matrix representation has an earlier onset for TRDSM than for LS-TRRH, because TRDSM contains more matrix multiplications than LS-TRRH. Fig. 5 shows the CPU timings for the B3LYP calculations in the sparse matrix representation. The timings shown are the same as in Fig. 3, with the addition of the timing for the exchange-correlation (Kohn-Sham XC) contribution. Like the other contributions to the KS matrix (Coulomb and exchange), the exchange-correlation contribution has reached the linear scaling regime. In general, the behaviour of the B3LYP curves is similar to the one observed for the Hartree-Fock curves.

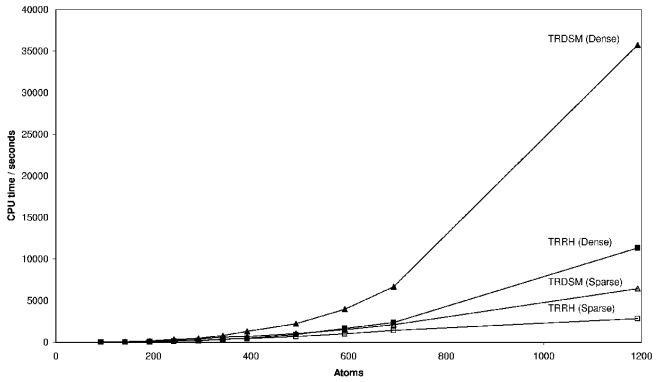

Figure 4: CPU timings for one iteration of a Hartree-Fock calculation using a 6-31G basis for the LS-TRRH and TRDSM steps for sparse and dense matrices plotted as a function of the number of atoms in a polyalanine peptide.

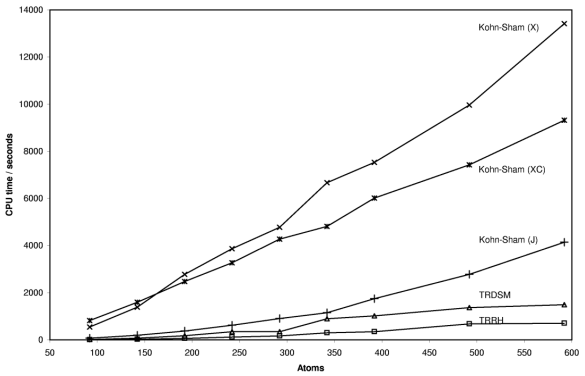

Figure 5: CPU timings for one iteration of a B3LYP calculation using a 6-31G basis plotted as a function of the number of atoms in a polyalanine peptide. The same contributions as in Fig. 3 are considered, in addition to the exchange correlation (XC) contribution.

5 Conclusion

We have described a linear scaling implementation of the trust-region self-consistent field (LS-TRSCF) method. In the LS-TRSCF method, each iteration consists of a incomplete optimization of the Roothaan-Hall energy giving a new density matrix (see Section 2.3) followed by the determination of an improved density matrix in the subspace containing the current and previous density matrices. A linear scaling algorithm is obtained using iterative methods to solve the level shifted Newton equations and sparse matrix algebra.

The convergence of the LS-TRSCF method is examined and for comparison the convergence of conventional SCF/DIIS calculations have been reported. The LS-TRSCF calculations show smooth and robust convergence, and in several cases the LS-TRSCF calculations converge where the SCF/DIIS calculations diverge. The convergence of the LS-TRSCF method is in general equally good for small and large systems. For small systems, a TRSCF implementation based on an explicit diagonalization of the Fock/KS matrix may be more efficient. However, for small systems the computational time for optimizing the density matrix is insignificant compared to the computational time for setting up the Fock/KS matrix. Consequently we recommend using LS-TRSCF as standard method for calculations on both small and large systems.

Acknowledgments

This work has been supported by the Danish Natural Research Council and the Norwegian Research Council. We also acknowledge support from the Danish Center for Scientific Computing (DCSC) and the European Research and Training Network NANOQUANT, Understanding Nanomaterials from the Quantum Perspective, contract No. MRTN-CT-2003-506842.

References

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

[2] C. A. White, B. G. Johnson, P. M. W. Gill and M. Head-Gordon, *Chem. Phys. Lett.* **230**, 8 (1994).

[3] C. A. White,B. G. Johnson, P. M. W. Gill and M. Head-Gordon, *Chem. Phys. Lett.* **253**, 268 (1996).

[4] M. C. Strain, G. E. Scuseria and M. J. Frisch, *Science* **271**, 51 (1996).

[5] M. Challacombe and E. Schwegler, *J. Chem. Phys.* **106**, 5526 (1997).

[6] Y. Shao, and M. Head-Gordon, *Chem. Phys. Lett.* **323**, 425 (2000).

[7] E. Schwegler and M. Challacombe , *J. Chem. Phys.* **105**, 2726 (1996).

[8] E. Schwegler, M. Challacombe and M. Head-Gordon , *J. Chem. Phys.* **106**, 9708 (1997).

[9] E. Schwegler and M. Challacombe , *J. Chem. Phys.* **111**, 6223 (1999).

[10] E. Schwegler and M. Challacombe , *Theor. Chem. Acc.* **104**, 344 (2000).

[11] C. Ochsenfeld, C. A. White and and M. Head-Gordon , *J. Chem. Phys.* **109**, 1663 (1998).

[12] J. C. Burant, G. E. Scuseria and M. J. Frisch, *J. Chem. Phys.* **105**, 8969 (1996).

[13] J. M. Pérez-Jordá and W. Yang, *Chem. Phys. Lett.* **241**, 469 (1995).

[14] B. G. Johnson, C. A. White, Q. Zang, B. Chen, R. L. Graham, P. M. W. Gill and M. Head-Gordon, in *Recent Developments in Density Functional Theory.* edited by J. M. Seminario (Elsevier Science, Amsterdam, 1996), Vol. 4

[15] R. E. Stratman, G. E. Scuseria and M. J. Frisch, *Chem. Phys. Lett.* **257**, 213 (1996).

[16] R. McWeeny. *Methods of Molecular Quantum Mechanics*, 2nd Edition. Academic Press, 1992.

[17] T. Helgaker, P. Jørgensen and J. Olsen. *Molecular Electronic-Structure Theory.* Wiley, New York, 2000.

[18] L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek, T. Helgaker, *J. Chem. Phys.* **121** 16 (2004).

[19] L. Thøgersen, J. Olsen, A. Köhn, P. Jørgensen, P. Sałek, T. Helgaker, *J. Chem. Phys.* **123** 074103 (2005).

[20] R. McWeeny, *Rev. Mod. Phys.* **32**, 325 (1960).

[21] T. Helgaker, H. Larsen, J. Olsen and P. Jørgensen, *Chem. Phys. Lett.* **327**, 379 (2000).

[22] H. Larsen, J. Olsen and P. Jørgensen and T. Helgaker, *J. Chem. Phys.* **115**, 9685 (2001)

[23] Y. Shao, C. Saravanan, M. Head-Gordon and C. A. White, *J. Chem. Phys.* **118**, 6144 (2003).

[24] H. J. Aa. Jensen and P. Jørgensen, *J. Chem. Phys.* **80**, 1204 (1984).

[25] B. Lengsfield III, *J. Chem. Phys.* **73**, 382 (1980).

[26] R. Shepard, I. Shavitt and J. Simons, *J. Chem. Phys.* **76**, 543 (1982).

[27] P. Sałek *et al.*, *J. Chem. Phys.*, to be submitted.

[28] R. W. Nunes and D. Vanderbilt, *Phys. Rev. B* **50**, 17611 (1994).

[29] A. Schafer, H. Horn and R. Ahlrichs, *J. Chem. Phys.* **97**, 2571 (1992).

Paper VI

A ground-state-directed optimization scheme for the Kohn-Sham energy
S. Høst, B. Jansik, J. Olsen, P. Jørgensen, **S. Reine** and T. Helgaker
Physical Chemistry Chemical Physics, **10**, 5344-5348 (2008)

A ground-state-directed optimization scheme for the Kohn–Sham energy†

Stinne Høst,*[a] Branislav Jansík,[a] Jeppe Olsen,[a] Poul Jørgensen,[a] Simen Reine[b] and Trygve Helgaker[b]

Received 6th May 2008, Accepted 27th June 2008
First published as an Advance Article on the web 21st July 2008
DOI: 10.1039/b807639a

Kohn–Sham density-functional calculations are used in many branches of science to obtain information about the electronic structure of molecular systems and materials. Unfortunately, the traditional method for optimizing the Kohn–Sham energy suffers from fundamental problems that may lead to divergence or, even worse, convergence to an energy saddle point rather than to the ground-state minimum—in particular, for the larger and more complicated electronic systems that are often studied by Kohn–Sham theory nowadays. We here present a novel method for Kohn–Sham energy minimization that does not suffer from the flaws of the conventional approach, combining reliability and efficiency with linear complexity. In particular, the proposed method converges by design to a minimum, avoiding the sometimes spurious solutions of the traditional method and bypassing the need to examine the structure of the provided solution.

Nowadays, theoretical and experimental studies of molecular properties and interactions rigorously based on the laws of quantum mechanics are making important contributions to advances in many branches of science. Among the methods developed to describe the electronic structure of molecules and materials, Kohn–Sham (KS) density-functional theory[1] represents the best compromise between cost and accuracy. Indeed, thousands of KS calculations are carried out daily to obtain information about molecular systems of importance not only in chemistry and physics but also in biology and medicine. It is therefore critically important that the KS energy can be determined in a reliable and efficient manner.[2]

In KS theory, the energy is minimized with respect to variations in the one-electron density matrix. The standard method of optimization consists of an iterative procedure, where each iteration consists of two steps. First, in the Roothaan–Hall (RH) step, the KS matrix (the gradient of the KS energy) is diagonalized to generate a new density matrix. Second, in the DIIS (direct inversion in the iterative subspace) step, an improved density matrix is determined by linearly combining this new density matrix with the density matrices of the previous iterations.[3]

Although this two-step RH–DIIS scheme has been very successful, it sometimes fails, either by converging to a saddle point (with an indefinite Hessian) rather than to the minimum (with a positive definite Hessian) or by diverging. Whereas divergence is an obvious failure, convergence to a saddle point is more pernicious in that it typically leaves the user unaware that the solution provided does not properly represent the electronic ground state. A stability analysis that examines the nature of the stationary point is rarely performed since the cost of such an examination is comparable to that of the whole optimization. Convergence to a saddle point may occur since only gradient information is used in the course of the optimization, both in the RH step and in the DIIS step.

We here present a method that does not suffer from the flaws of the two-step RH–DIIS scheme, improving the reliability of the optimization and reducing its cost. The key to its success is the replacement of the two separate steps of each RH–DIIS iteration by a single concerted step that fully exploits the Hessian information available from the previous iterations. At each iteration, we construct a local quadratic model of the KS energy that is exact to second order in the directions of the previous density matrices and a good approximation in the remaining directions. The new density matrix is obtained by applying the trust-region minimization method to this quadratic model,[4] thereby ensuring that the energy is lowered at each iteration. Since the algorithm exploits information about the structure of the Hessian, it converges by design to a minimum. The proposed scheme does not require diagonalization. It is based on matrix multiplications and becomes, for sufficiently large systems, of linear complexity when the sparsity of the matrices is exploited.

Our method differs from previous KS optimization methods in that it does not involve, at each iteration, two separate computational steps such as RH diagonalizaton and DIIS averaging. Rather, each iteration contains a single computational step (the solution of Newton-type equations) that makes a full concerted use of the curvature information available from previous iterations, leading to a dramatic improvement in performance, as demonstrated below. Such an improvement is not possible within the traditional two-step framework, where each step has only limited (and different) curvature information available. Since previous attempts at improving on the RH–DIIS convergence have retained the two-step framework, modifying the RH and DIIS steps separately, they have not constituted a dramatic improvement on the basic RH–DIIS scheme.[5–9] The method is also applicable to Hartree–Fock theory.

Let us assume that we have carried out n iterations in an iterative minimization of the KS energy. In the course of these iterations, we have generated a density-matrix subspace

[a] *Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Aarhus, DK-8000 Århus C, Denmark. E-mail: stinne@chem.au.dk*
[b] *Centre for Theoretical and Computational Chemistry, Department of Chemistry, University of Oslo, P. O. Box 1033, Blindern, N-0315 Oslo, Norway*

† Electronic supplementary information (ESI) available: Molecular geometries of test systems. See DOI: 10.1039/b807639a

consisting of n density matrices $\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_n$, and the corresponding space of n KS matrices $\boldsymbol{F}_1,\boldsymbol{F}_2,\ldots,\boldsymbol{F}_n$, where the KS matrices are the first derivatives of the KS energy E^{KS} with respect to the density matrix

$$\boldsymbol{E}^{[1]}(\boldsymbol{D}_i) = \left.\frac{\partial E^{\mathrm{KS}}(\boldsymbol{D})}{\partial \boldsymbol{D}}\right|_{\boldsymbol{D}=\boldsymbol{D}_i} = 2\boldsymbol{F}(\boldsymbol{D}_i) = 2\boldsymbol{F}_i \qquad (1)$$

for a closed-shell system. The KS energy is now expanded to second order about $\boldsymbol{D}_n$ as

$$\begin{aligned} E^{\mathrm{KS}}(\boldsymbol{D}) &= E(\boldsymbol{D}_n) + \langle \boldsymbol{D}-\boldsymbol{D}_n|\boldsymbol{E}^{[1]}(\boldsymbol{D}_n)\rangle \\ &\quad + \tfrac{1}{2}\langle \boldsymbol{D}-\boldsymbol{D}_n|\boldsymbol{E}^{[2]}(\boldsymbol{D}_n)|\boldsymbol{D}-\boldsymbol{D}_n\rangle \end{aligned} \qquad (2)$$

where $\langle \boldsymbol{A}|\boldsymbol{B}\rangle = \mathrm{Tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B})$. If we ignore the second-order term and minimize the remaining first-order RH energy $\langle \boldsymbol{D}-\boldsymbol{D}_n|\boldsymbol{E}^{[1]}(\boldsymbol{D}_n)\rangle$, we obtain the same density matrix as by diagonalization in RH theory.[10] To make better use of the information present in the density-matrix subspace, we propose to retain the second-order term in an approximate form by invoking the quasi-Newton condition. This approach leads to the quadratic augmented Roothaan–Hall (ARH) energy function, which forms the basis for the proposed ARH optimization method.

The elements of $\boldsymbol{D}$ in eqn (2) cannot be varied freely, as the density matrix must fulfil the symmetry, trace and idempotency relations of a single-determinant density matrix. In an orthonormal basis (such as the orthonormalized atomic-orbital basis used here), the allowed variations in $\boldsymbol{D}_n$ may be parameterized in terms of an anti-symmetric matrix $\boldsymbol{X}$ as[11]

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{X}) &= \exp(-\boldsymbol{X})\boldsymbol{D}_n\exp(\boldsymbol{X}) \\ &= \boldsymbol{D}_n + [\boldsymbol{D}_n,\boldsymbol{X}] \\ &\quad + \tfrac{1}{2}[[\boldsymbol{D}_n,\boldsymbol{X}],\boldsymbol{X}] + \cdots, \end{aligned} \qquad (3)$$

where $\boldsymbol{X}$ contains only non-redundant components, with the redundant occupied–occupied and virtual–virtual components projected out. The second-order energy in eqn (2) contains the density difference $\boldsymbol{D}-\boldsymbol{D}_n$. Expressing this density difference in terms of $\boldsymbol{X}$, we may use eqn (3) to obtain a second-order Taylor expansion of the KS energy in $\boldsymbol{X}$ as

$$\begin{aligned} E^{\mathrm{KS}}(\boldsymbol{X}) &= E(\boldsymbol{D}_n) + \langle [\boldsymbol{D}_n,\boldsymbol{X}]|\boldsymbol{E}^{[1]}(\boldsymbol{D}_n)\rangle \\ &\quad + \tfrac{1}{2}\langle [[\boldsymbol{D}_n,\boldsymbol{X}],\boldsymbol{X}]|\boldsymbol{E}^{[1]}(\boldsymbol{D}_n)\rangle \\ &\quad + \tfrac{1}{2}\langle [\boldsymbol{D}_n,\boldsymbol{X}]|\boldsymbol{E}^{[2]}(\boldsymbol{D}_n)|[\boldsymbol{D}_n,\boldsymbol{X}]\rangle \end{aligned} \qquad (4)$$

In the second-order expansion in eqn (4), the energy derivatives are with respect to variations in the density matrix. The Hessian contribution is in this way separated into a contribution that contains a second-order variation in the density matrix (the third term), and a contribution that contains two first-order variations in the density matrix (the fourth term). This separation of the Hessian contribution is important for the further development.

The Newton equations are obtained by differentiating eqn (4) with respect to $\boldsymbol{X}$, yielding

$$\begin{aligned} &\tfrac{1}{2}\langle [[\boldsymbol{D}_n,\boldsymbol{X}],\boldsymbol{X}]^{\mu}|\boldsymbol{E}^{[1]}(\boldsymbol{D}_n)\rangle + \langle [\boldsymbol{D}_n,\boldsymbol{X}]^{\mu}|\boldsymbol{E}^{[2]}(\boldsymbol{D}_n)|[\boldsymbol{D}_n,\boldsymbol{X}]\rangle \\ &\quad = -\langle [\boldsymbol{D}_n,\boldsymbol{X}]^{\mu}|\boldsymbol{E}^{[1]}(\boldsymbol{D}_n)\rangle, \end{aligned} \qquad (5)$$

where the superscript μ denotes differentiation with respect to element X_μ of $\boldsymbol{X}$, *i.e.* $\boldsymbol{A}^{\mu} = \partial\boldsymbol{A}/\partial X_\mu$. The right-hand side represents the gradient of the energy with respect to $\boldsymbol{X}$ at the point of expansion $\boldsymbol{D}_n$, whereas the left-hand side represents a multiplication of the Hessian by $\boldsymbol{X}$.

When the Newton equations are solved iteratively, each new trial vector is transformed by the Hessian. The first Hessian contribution to the transformation eqn (5) is easily evaluated, as described in ref. 9, from the KS matrix of the nth iteration $\boldsymbol{F}_n$. The second contribution is much more expensive, requiring a new Coulomb and exchange–correlation evaluation for each trial vector. Its full inclusion would lead to a quadratically convergent procedure but at a prohibitively high cost. Fortunately, it is possible to obtain a useful approximation to the second Hessian contribution in eqn (5) from the density and KS matrices of the previous iterations by invoking the quasi-Newton condition:

$$\begin{aligned} \boldsymbol{E}^{[2]}(\boldsymbol{D}_n)(\boldsymbol{D}_i-\boldsymbol{D}_n) &= \boldsymbol{E}^{[1]}(\boldsymbol{D}_i) - \boldsymbol{E}^{[1]}(\boldsymbol{D}_n) \\ &= 2\boldsymbol{F}(\boldsymbol{D}_i) - 2\boldsymbol{F}(\boldsymbol{D}_n) = 2\boldsymbol{F}_{in} \end{aligned} \qquad (6)$$

Introducing the density-matrix subspace projector

$$\mathscr{P}_n = \sum_{i,j=1}^{n-1} |\boldsymbol{D}_{in}\rangle[\boldsymbol{T}^{-1}]_{ij}\langle \boldsymbol{D}_{jn}|,\ T_{ij} = \langle \boldsymbol{D}_{in}|\boldsymbol{D}_{jn}\rangle \qquad (7)$$

where $\boldsymbol{D}_{in} = \boldsymbol{D}_i - \boldsymbol{D}_n$, we may restrict $\boldsymbol{E}^{[2]}(\boldsymbol{D}_n)$ in the second Hessian term in eqn (5) to operate only on the density-matrix subspace:

$$\begin{aligned} &\langle [\boldsymbol{D}_n,\boldsymbol{X}]^{\mu}|\boldsymbol{E}^{[2]}(\boldsymbol{D}_n)\mathscr{P}_n|[\boldsymbol{D}_n,\boldsymbol{X}]\rangle \\ &\quad = 2\sum_{ij}\langle [\boldsymbol{D}_n,\boldsymbol{X}]^{\mu}|\boldsymbol{F}_{in}\rangle[\boldsymbol{T}^{-1}]_{ij}\langle \boldsymbol{D}_{jn}|[\boldsymbol{D}_n,\boldsymbol{X}]\rangle \end{aligned} \qquad (8)$$

which constitutes the ARH approximation to the Hessian trial-vector transformation. Note that the quasi-Newton condition can only be applied straightforwardly for second derivatives with respect to variations in the density matrix, as occurring in the Taylor expansion of the KS energy in eqn (4).

An explicit expression for the ARH quasi-Newton equations in terms of density matrices and KS matrices is given by

$$\begin{aligned} &(\boldsymbol{F}_n^{\mathrm{vv}} - \boldsymbol{F}_n^{\mathrm{oo}})\boldsymbol{X} + \boldsymbol{X}(\boldsymbol{F}_n^{\mathrm{vv}} - \boldsymbol{F}_n^{\mathrm{oo}}) \\ &\quad + \sum_{ij}(\boldsymbol{F}_{in}^{\mathrm{ov}} - \boldsymbol{F}_{in}^{\mathrm{vo}})[\boldsymbol{T}^{-1}]_{ij}\mathrm{Tr}(\boldsymbol{D}_{jn}[\boldsymbol{D}_n,\boldsymbol{X}]) \\ &\quad = \boldsymbol{F}_n^{\mathrm{vo}} - \boldsymbol{F}_n^{\mathrm{ov}} \end{aligned} \qquad (9)$$

where $\boldsymbol{F}^{\mathrm{ab}} = \boldsymbol{P}_{\mathrm{a}}\boldsymbol{F}\boldsymbol{P}_{\mathrm{b}}$ are the projections of the KS matrix onto the occupied and virtual orbital spaces, with $\boldsymbol{P}_{\mathrm{o}} = \boldsymbol{D}_n$ and $\boldsymbol{P}_{\mathrm{v}} = \boldsymbol{I} - \boldsymbol{P}_{\mathrm{o}}$. The left-hand side of eqn (9) describes how the Hessian may be multiplied on a trial matrix $\boldsymbol{X}$ as required in an iterative solution of the Newton equations.

The first Hessian contribution originates from the RH energy and contains the difference between the virtual–virtual and occupied–occupied projections of the KS matrix $\boldsymbol{F}_n^{\mathrm{vv}} - \boldsymbol{F}_n^{\mathrm{oo}}$. The second contribution goes beyond the RH energy, containing an averaged combination of occupied–virtual projections of KS matrices $\boldsymbol{F}_{in}^{\mathrm{vo}} - \boldsymbol{F}_{in}^{\mathrm{vo}}$ determined by invoking the quasi-Newton condition in the density-matrix subspace. In this subspace, the ARH Hessian is equal to the true second-order KS Hessian, to within the finite-difference errors of the quasi-Newton approximation; in the orthogonal complement to this subspace, the ARH Hessian is equal to the RH Hessian, which is in fact by itself quite accurate, except in

the directions that represent excitations (rotations) between orbitals of similar energies. Since the density-matrix subspace spans primarily such directions, the ARH second-order energy function has a Hessian that qualitatively and quantitatively constitutes a good approximation to the true KS Hessian. Moreover, in the limit of a complete density-matrix subspace, the ARH energy expansion is equal to the exact second-order expansion of the KS energy (to within the finite-difference errors of the quasi-Newton approximation).

As the quadratic energy defined by the ARH model accurately resembles the true second-order KS energy of the electronic system, we have sufficient information about the full KS energy function to apply the trust-region method—a general method of non-linear optimization with guaranteed convergence to a minimum.[4] At iteration n, a rotation matrix is determined by minimizing the ARH energy subject to the trust-region condition $\|\boldsymbol{X}\| \leq h_n$, where h_n is the current trust radius. The corresponding minimizer $\boldsymbol{X}^*$ is obtained by solving the Newton equations in eqn (9) with a level-shift parameter added. If $\boldsymbol{D}(\boldsymbol{X}^*)$ determined from eqn (3) lowers the KS energy, then we set $\boldsymbol{D}_{n+1} = \boldsymbol{D}(\boldsymbol{X}^*)$ and the trust radius may be increased $h_{n+1} \geq h_n$, based on a comparison of the observed and predicted decreases in the KS energy. Conversely, if $\boldsymbol{D}(\boldsymbol{X}^*)$ increases rather than decreases the energy, then $\boldsymbol{X}^*$ is rejected, the trust region is reduced $h'_n < h_n$, and a new, smaller step is tried subject to the condition $\|\boldsymbol{X}\| \leq h'_n$.

To illustrate the performance, we report ARH and standard RH–DIIS calculations on five molecules representing various bonding situations (with the model and basis set given in parentheses): a Cd^{2+}–imidazole complex (B3LYP using the VWN3 parameterization[12–15]/3-21G[16,17]), a 29-residue polyalanine peptide (B3LYP(VWN5)/6-31G[18,19]), a model of vitamin B12 (HF/AhlrichsVDZ[20] and BP86[13,21]/AhlrichsVDZ), insulin (HF/6-31G) and a water cluster containing 51 monomers (B3LYP(VWN5)/cc-pVTZ[22]). See the ESI for the molecular geometries.† No level shift is added in the standard RH–DIIS calculations except where specified. For the cadmium–imidazole complex, a bare-nucleus Hamiltonian was used to obtain the initial density matrix. For all other calculations, the Hückel starting guess was used. Convergence is illustrated by plotting the difference between the current energy and the energy of the minimum (ground-state) solution, and convergence is obtained when the gradient norm is smaller than 10^{-4}. The reported calculations are all carried out using a local version of the DALTON program.[23]

During the optimizations with our code, some surprising results were obtained. To verify that these unexpected results do not arise from numerical instabilities or from artefacts or deficiencies of our implementation, additional calculations have been carried out using four widely distributed standard quantum-chemistry program packages. In the following, we shall refer to these collectively as the quantum-chemistry programs (QCPs), using the specific labels QCP1, QCP2, QCP3 and QCP4 when necessary. Unless otherwise specified, we use the standard optimization schemes with these codes.

The vitamin B12 model (HF/AhlrichsVDZ), cadmium–imidazole, and polyalanine calculations in Fig. 1 provide examples where the standard RH–DIIS scheme (Fig. 1b) fails by converging to a saddle point rather than to the minimum,

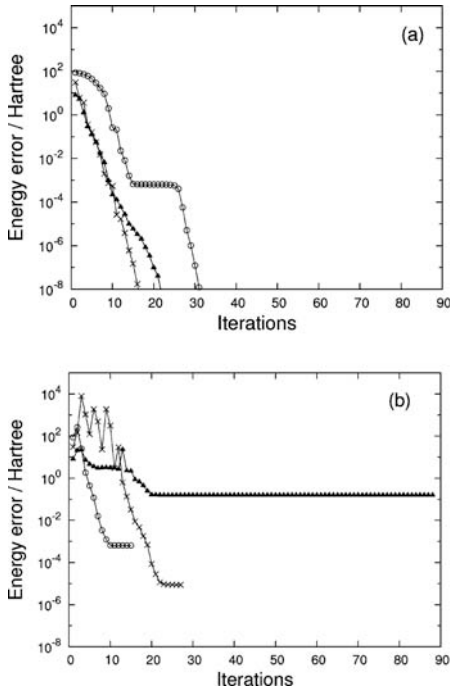

Fig. 1 Convergence of a cadmium complex (circles), polyalanine (crosses) and model B12 vitamin (HF/AhlrichsVDZ, triangles) using (a) the ARH method and (b) the RH–DIIS method. At each iteration, we have plotted the error in the energy relative to the energy of the true minimum (on a logarithmic scale). All calculations have been converged to a gradient norm of 10^{-4}. While the RH–DIIS optimizations all converge to a saddle point, the ARH optimizations correctly identify the minimum.

which is correctly located by the ARH scheme (Fig. 1a). For these molecules, therefore, only the ARH density matrices are valid representations of the electronic ground state. Moreover, all ARH calculations proceed smoothly, while the RH–DIIS convergence is erratic in the initial iterations for polyalanine and slow for the vitamin B12 model, where a gradient norm smaller than 10^{-4} is not obtained until after about 90 iterations. For the cadmium complex, the saddle point is approached in the ARH calculation in iterations 15–25, but the ARH optimization eventually accumulates sufficient information to reject this critical point as a saddle point. Subsequently, the minimum is located in a few iterations. By contrast, even with very tight convergence criteria imposed, the RH–DIIS calculation cannot escape the saddle point, because no information about the Hessian is available.

The erroneous convergence behaviour of the RH–DIIS scheme is seen also with the four QCPs. Thus, for the cadmium complex, all four QCPs converge to the saddle point. With some of the QCPs, it is possible (at a much increased cost) to request a stability analysis followed by a reoptimization if a saddle point has been reached, so as to obtain a minimum. For the cadmium complex, this QCP reoptimization did result in the detection of the minimum. For the B12 model (HF/AhlrichsVDZ), QCP1 correctly locates the minimum, whereas the QCP2–QCP4 optimizations converge to the saddle point. By enforcing a (non-standard) zero-level shift, QCP2 locates

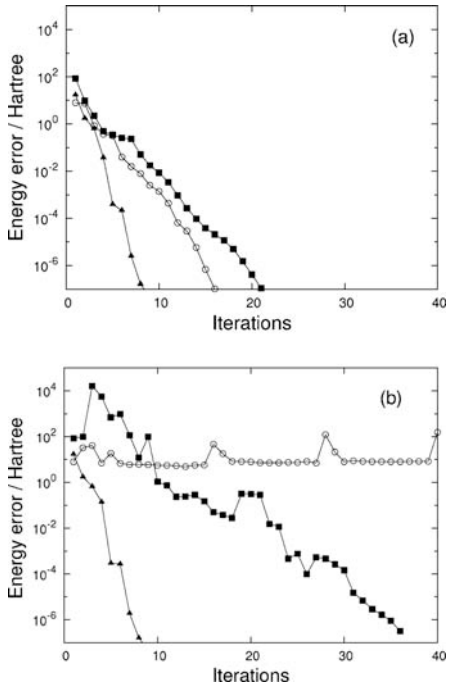

Fig. 2 Convergence of water clusters (full triangles), model B12 vitamin (BP86/AhlrichsVDZ, open circles), and insulin (full squares) using (a) the ARH method, and (b) the RH–DIIS method. At each iteration, we have plotted the error in the energy relative to the energy of the minimum (on a logarithmic scale). All calculations have been converged to a gradient norm of 10^{-4}.

the minimum after more than 200 iterations. The QCP behaviour for polyalanine is discussed below.

We now turn our attention to the vitamin B12 model (BP86/AhlrichsVDZ), water cluster and insulin calculations reported in Fig. 2. For these systems, the ARH scheme (Fig. 2a) converges smoothly, whereas the RH–DIIS scheme (Fig. 2b) struggles with increasing complexity of the system. For the water cluster, the two schemes converge (in our implementation) in the same number of iterations, whereas all four QCPs require less than 15 iterations for convergence. For insulin, our RH–DIIS optimization displays an erratic behavior in the global region but eventually recovers and converges to the correct solution in 36 iterations. By comparison, QCP1, QCP2 and QCP4 required 84, 86 and 46 iterations, respectively, whereas QCP3 failed to converge. For the vitamin B12 model, our RH–DIIS implementation does not locate the local region and fails to converge; QCP1–QCP4 required 36, 37, 51 and 96 iterations for convergence respectively. By contrast, the ARH optimizations converge smoothly, locally as well as globally, in a total of 22 iterations. Note also that no expensive *post factum* examination of the electronic Hessian is necessary in the ARH scheme to confirm that the solution represents the electronic ground state.

To improve the RH–DIIS convergence, level shifts are often applied, in a somewhat haphazard manner, in the solution of the RH eigenvalue equation. While level shifting may indeed stabilize convergence, it may also lead to solutions that violate the Aufbau principle. We here illustrate the effect of level shifting by carrying out calculations on polyalanine peptides

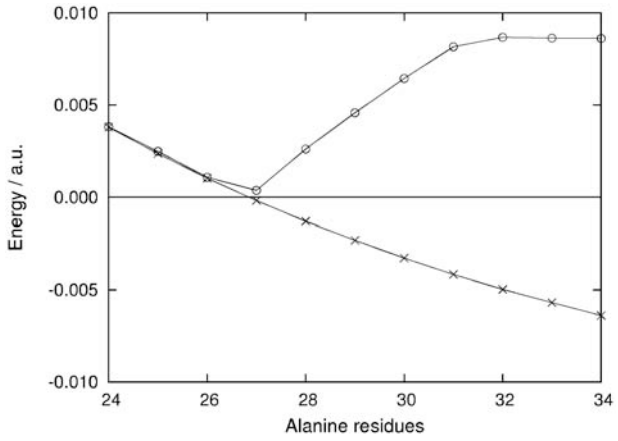

Fig. 3 Lowest Hessian eigenvalue of polyalanines of increasing size in the ARH method (open circles) and RH–DIIS method (crosses).

of increasing size. The standard RH–DIIS calculations with zero-level shift in the RH step converge straightforwardly for peptides containing up to 29 alanine residues—for larger peptides, a level shift of 0.01 is applied to enforce convergence. However, as seen from Fig. 3, the RH–DIIS solutions obtained by level shifting are saddle points and clearly not valid representations of the electronic ground state. By contrast, the ARH calculations converge to a true ground-state minimum, in all cases. For peptides with up to 26 residues, the ARH and RH–DIIS solutions are identical. The pattern with negative Hessian eigenvalues as obtained with the RH–DIIS scheme was reproduced with QCP1 and QCP2.

The calculations presented here demonstrate that the ARH algorithm is both robust and efficient, converging smoothly to a minimum also when the RH–DIIS algorithm fails to converge or when it converges to a saddle point rather than to a minimum. In particular, for the larger and more complicated systems that are nowadays studied by KS theory, the results of RH–DIIS optimizations cannot be trusted without performing a stability analysis. When the two algorithms converge to the same minimum, their performance is comparable in simple cases; in more difficult cases, the RH–DIIS algorithm requires more iterations since it often behaves erratically in the global region of the optimization, while the ARH algorithm converges smoothly. As the time-dominant step in an iteration of both the ARH scheme and the RH–DIIS scheme is the construction of a new KS matrix, the ARH algorithm is in general the most cost-efficient requiring fewer iterations than the RH–DIIS scheme. For the ARH method, it is unnecessary to perform a stability analysis of the stationary point. In this respect, the ARH algorithm has the advantages of the quadratically convergent second-order trust-region algorithm, at a fraction of its cost. We recommend it as the standard algorithm for optimizing KS energies.

Acknowledgements

This work has been supported by the Lundbeck Foundation, the Danish Natural Science Research Council and the

Norwegian Research Council through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420). We also acknowledge support from the Danish Center for Scientific Computing (DCSC).

References

1 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133.
2 W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, Wiley-VCH, Weinheim, 2000.
3 P. Pulay, *Chem. Phys. Lett.*, 1980, **73**, 393.
4 R. Fletcher, *Practical Methods of Optimization*, Wiley, Chichester, 2nd edn, 1987.
5 K. N. Kudin, G. E. Scuseria and E. Cancés, *J. Chem. Phys.*, 2002, **116**, 8255.
6 L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek and T. Helgaker, *J. Chem. Phys.*, 2004, **121**, 16.
7 L. Thøgersen, J. Olsen, A. Köhn, P. Jørgensen, P. Sałek and T. Helgaker, *J. Chem. Phys.*, 2005, **123**, 074103.
8 S. Høst, J. Olsen, B. Jansík, P. Jørgensen, S. Reine, T. Helgaker, P. Sałek and S. Coriani, *Lecture Series on Computer and Computational Sciences*, Brill Academic Publishers, Leiden, The Netherlands, 2006, vol. **6**, p. 177.
9 P. Sałek, S. Høst, L. Thøgersen, P. Jørgensen, P. Manninen, J. Olsen, B. Jansík, S. Reine, F. Pawlowski, E. Tellgren, T. Helgaker and S. Coriani, *J. Chem. Phys.*, 2007, **126**, 114110.
10 R. McWeeny, *Methods of Molecular Quantum Mechanics*, Academic Press, London, 2nd edn, 1992.
11 T. Helgaker, H. Larsen, J. Olsen and P. Jørgensen, *Chem. Phys. Lett.*, 2000, **327**, 397.
12 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648.
13 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098.
14 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785.
15 S. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200.
16 J. Binkley, J. Pople and W. Hehre, *J. Am. Chem. Soc.*, 1980, **102**, 939.
17 K. Dobbs and W. Hehre, *J. Comput. Chem.*, 1987, **8**, 880.
18 W. Hehre, R. Ditchfield and J. Pople, *J. Chem. Phys.*, 1972, **56**, 2257.
19 M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley and M. S. Gordon, *J. Chem. Phys.*, 1982, **77**, 3654.
20 A. Schäfer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571.
21 J. P. Perdew, *Phys. Rev. B*, 1986, **33**, 8822.
22 T. H. Dunning, Jr, *J. Chem. Phys.*, 1989, **90**, 1007.
23 C. Angeli, K. L. Bak, V. Bakken, O. Christiansen, R. Cimiraglia, S. Coriani, P. Dahle, E. K. Dalskov, T. Enevoldsen, B. Fernandez, C. Hättig, K. Hald, A. Halkier, H. Heiberg, T. Helgaker, H. Hettema, H. J. Aa. Jensen, D. Jonsson, P. Jørgensen, S. Kirpekar, W. Klopper, R. Kobayashi, H. Koch, A. Ligabue, O. B. Lutnæs, K. V. Mikkelsen, P. Norman, J. Olsen, M. J. Packer, T. B. Pedersen, Z. Rinkevicius, E. Rudberg, T. A. Ruden, K. Ruud, P. Sałek, A. Sanchez de Meras, T. Saue, S. P. A. Sauer, B. Schimmelpfennig, K. O. Sylvester-Hvid, P. R. Taylor, O. Vahtras, D. J. Wilson and H. Ågren, *DALTON, a molecular electronic structure program, Release 2.0*, 2005, see http://www.kjemi.uio.no/software/dalton/dalton.html.

Paper VII

Linear-scaling implementation of molecular response theory in self-consistent field electronic-structure theory

S. Coriani, S. Høst, B. Jansik, L. Thøgersen, J. Olsen, P. Jørgensen, **S. Reine**, F. Pawłowski, T. Helgaker and P. Sałek

The Journal of Chemical Physics, **126**, 154108 (2007)

Linear-scaling implementation of molecular response theory in self-consistent field electronic-structure theory

Sonia Coriani
Dipartimento di Scienze Chimiche, Università degli Studi di Trieste, Via Licio Giorgieri 1, I-34127 Trieste, Italy

Stinne Høst,[a] Branislav Jansík, Lea Thøgersen, Jeppe Olsen, and Poul Jørgensen
The Lundbeck Foundation Center for Theoretical Chemistry, Department of Chemistry, University of Aarhus, DK-8000 Århus C, Denmark

Simen Reine,[b] Filip Pawłowski,[c] and Trygve Helgaker[d]
Centre of Theoretical and Computational Chemistry, Department of Chemistry, University of Oslo, P.O. Box 1033 Blindern, N-0315 Norway

Paweł Sałek
Department of Theoretical Chemistry, The Royal Institute of Technology, SE-10691 Stockholm, Sweden

(Received 1 November 2006; accepted 15 February 2007; published online 18 April 2007)

A linear-scaling implementation of Hartree-Fock and Kohn-Sham self-consistent field theories for the calculation of frequency-dependent molecular response properties and excitation energies is presented, based on a nonredundant exponential parametrization of the one-electron density matrix in the atomic-orbital basis, avoiding the use of canonical orbitals. The response equations are solved iteratively, by an atomic-orbital subspace method equivalent to that of molecular-orbital theory. Important features of the subspace method are the use of paired trial vectors (to preserve the algebraic structure of the response equations), a nondiagonal preconditioner (for rapid convergence), and the generation of good initial guesses (for robust solution). As a result, the performance of the iterative method is the same as in canonical molecular-orbital theory, with five to ten iterations needed for convergence. As in traditional direct Hartree-Fock and Kohn-Sham theories, the calculations are dominated by the construction of the effective Fock/Kohn-Sham matrix, once in each iteration. Linear complexity is achieved by using sparse-matrix algebra, as illustrated in calculations of excitation energies and frequency-dependent polarizabilities of polyalanine peptides containing up to 1400 atoms. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2715568]

I. INTRODUCTION

Quantum chemistry has evolved in a spectacular fashion during the last two decades. Using quantum-chemical methods, it is nowadays possible to investigate a large number of molecular properties of increasing complexity, from computationally simple energy differences such as reaction enthalpies to more involved high-order frequency-dependent polarizabilities and multiphoton strengths, with control over the accuracy of the results.[1] Molecular properties are fundamental quantities underlying the macroscopic behavior of matter and their determination constitutes one of the most fruitful areas of interplay between experiment and theory.[2]

A difficulty in the application of quantum chemistry to compute molecular properties is the restriction on the size of systems that can be treated by current technology. Even with the recent dramatic improvements in computer technology and introduction of Kohn-Sham theory, the routine study of systems such as myoglobin, containing 150 amino acids, is still beyond our capabilities. This situation is particularly unfortunate in view of the considerable academic and industrial interest in macromolecules containing thousands of atoms such as polymers, proteins, enzymes, and nucleic acids.

The bottleneck for quantum-mechanical methods in their application to large systems is the scaling of the cost—in other words, the increase of CPU usage with increasing system size. Formally, Hartree-Fock and Kohn-Sham self-consistent field (SCF) methods scale as $O(N^4)$, where N refers to the system size. Moreover, wave-function-based correlated methods typically scale as $O(N^5)$ or higher. With such a steep scaling, advances in computer hardware alone will never allow us to treat large systems such as myoglobin. During the last decade, a large effort has been directed towards the development of new algorithms with a better scaling—see, for instance, Refs. 3–5 and references therein. The goal is to develop "linear-scaling" methods—that is, methods where the computational cost scales linearly with the system size, $O(N)$.

In Hartree-Fock and Kohn-Sham theories, the two major obstacles for the optimization of the energy have now been eliminated—namely, the construction of the Fock/Kohn-

[a]Author to whom correspondence should be addressed. Electronic mail: stinne@chem.au.dk
[b]Present address: Department of Chemistry, University of Aarhus, DK-8000 Århus C, Denmark.
[c]Present address: Institute of Physics, Kazimierz Wielki University, Plac Weyssenhoffa 11, 85-072 Bydgoszcz, Poland.
[d]Present address: Department of Chemistry, University of Durham, South Road, Durham DH1 3LE, UK.

Sham (KS) matrix and the generation of a new density matrix from the current Fock/KS matrix, see Ref. 3 for a recent overview. With these obstacles removed, it has become appropriate to address the problem of calculating molecular properties at linear cost.

In this paper, we describe a linear-scaling method for the calculation of molecular properties that may be expressed in terms of frequency-dependent response functions and their poles and residues. In particular, we consider properties calculated from the linear response function such as frequency-dependent polarizabilities, excitation energies, and one-photon transition moments. Molecular properties that are expressed in terms of higher-order response functions[6] can be obtained by a straightforward extension of the presented scheme.

In our linear-scaling response implementation, the expressions for the response functions are derived using a non-redundant exponential parametrization of the density matrix in the atomic-orbital (AO) basis. The formal derivation of the response functions and their residues is given in Refs. 7–9; for perturbation-dependent basis sets (used to calculate geometrical derivatives with atom-fixed AOs and magnetic properties with London AOs), the theory is given in Ref. 8. In this paper, we only discuss computational aspects that exclusively refer to property calculations; the strategy adopted for linear-scaling energy optimizations is described in Ref. 3.

Since all key computational steps of response theory presented here consist of multiplications of density, Fock/KS, and property matrices in the AO basis, matrix sparsity must be explored to achieve linear scaling. First, the response eigenvalue equations and linear sets of equations are solved. Their solution constitutes the major challenge with respect to linear scaling. We describe here how this may be achieved with iterative AO techniques, generalizing the algorithm previously developed to solve the response equations in the molecular-orbital (MO) basis at various levels of theory.[10,11]

An important feature of the response solver is that it maintains the paired structure of the response generalized Hessian and metric matrices. By adding trial vectors in pairs, the solver imposes the paired structure of the full-space response equations on the reduced-space equations, ensuring that complex eigenvalues do not arise during their solution. Furthermore, monotonic convergence is ensured towards the lowest eigenvalues. Another important feature of our algorithm is that we take over, in the AO basis, the preconditioner that has been so successfully employed in the MO basis. However, this preconditioner cannot be applied directly in the AO basis as the generalized AO Hessian has a large condition number and is not diagonally dominant. Rather, it is applied in an orthogonalized AO basis such as those defined by the Cholesky or Löwdin symmetric decomposition of the overlap matrix. In such a basis, the generalized Hessian becomes diagonally dominant and the condition number is significantly reduced. For the optimization of Hartree-Fock and Kohn-Sham energies, the Newton equations have previously been successfully solved when transformed from the AO basis to the Löwdin basis[3] or the Cholesky basis.[12]

The evaluation of static molecular properties within a linear-scaling framework has previously been considered by Ochsenfeld and Head-Gordon,[13] adopting a parametrization of the density matrix where idempotency is taken care of by replacing the density matrix with its McWeeny-purified counterpart,[14] as suggested by Li *et al.*[15] Using this approach, Ochsenfeld *et al.* have reported a linear-scaling implementation of NMR shifts for linear alkanes and presented results for three-dimensional systems with more than 1000 atoms.[16] An alternative strategy for static molecular properties, based on a purification of the density matrix, has recently been proposed by Weber *et al.*[17]

The remainder of this paper is divided into three main sections. In Sec. II, we present the theory and implementation of linear-scaling SCF linear response theory. Section III contains some numerical examples of calculations of frequency-dependent polarizabilities and excitation energies. Finally, Sec. IV contains some concluding remarks.

II. THEORY

The present section consists of four parts. First, in Sec. II A, the basic expressions of AO-based linear response theory are given, in a manner suitable for linear-scaling implementation. In Sec. II B we discuss the iterative algorithm used for solving the response equations. Finally, in Secs. II C and II D, respectively, we describe preconditioning and initial guesses of the iterative algorithm.

A. AO-based SCF linear response theory

In Hartree-Fock and Kohn-Sham theories, response functions may be efficiently calculated in the AO basis, expressing the AO density matrix in the exponential form[1,7,18–21]

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX}), \tag{1}$$

where $\mathbf{S}$ is the AO overlap matrix and $\mathbf{X}$ is an anti-Hermitian matrix that contains the variational parameters, with the redundant parameters projected out:

$$\mathbf{X} = \mathcal{P}(\mathbf{X}). \tag{2}$$

We have here introduced the projection operator on a matrix $\mathbf{M}$,

$$\mathcal{P}(\mathbf{M}) = \mathbf{P}_o\mathbf{M}\mathbf{P}_v^T + \mathbf{P}_v\mathbf{M}\mathbf{P}_o^T, \tag{3}$$

where $\mathbf{P}_o$ and $\mathbf{P}_v$ are projectors onto the occupied and virtual orbital spaces, respectively,

$$\mathbf{P}_o = \mathbf{DS}, \tag{4}$$

$$\mathbf{P}_v = \mathbf{I} - \mathbf{DS}, \tag{5}$$

fulfilling the idempotency ($\mathbf{P}_o^2=\mathbf{P}_o$ and $\mathbf{P}_v^2=\mathbf{P}_v$) and orthogonality relations ($\mathbf{P}_o\mathbf{P}_v=\mathbf{P}_v\mathbf{P}_o=\mathbf{0}$ and $\mathbf{P}_o^T\mathbf{S}\mathbf{P}_v=\mathbf{P}_v^T\mathbf{S}\mathbf{P}_o=\mathbf{0}$). Using the above exponential parametrization of the AO density matrix, the linear response function associated with the time-independent operators $\hat{A}$ and $\hat{B}$ becomes[7]

$$\langle\langle \hat{A};\hat{B}\rangle\rangle_\omega = \mathrm{Tr}[\mathbf{A}^{[1]}\mathbf{N}^B(\omega)], \tag{6}$$

$$(\mathbf{E}^{[2]} - \omega\mathbf{S}^{[2]})\text{vec}\,\mathbf{N}^{B}(\omega) = -\,\text{vec}\,\mathbf{B}^{[1]}, \tag{7}$$

where $\mathbf{M}^{[1]}$ is the property gradient of the operator $\hat{M}$ represented by the AO matrix $\mathbf{M}$:[7]

$$\mathbf{M}^{[1]} = \mathbf{SDM} - \mathbf{MDS} = \mathbf{P}_{\mathrm{o}}^{\mathrm{T}}\mathbf{M} - \mathbf{MP}_{\mathrm{o}}. \tag{8}$$

In Eq. (7), the vec operator transforms a matrix $\mathbf{M}$ into a column vector vec $\mathbf{M}$ by stacking its columns. Since the linear equations are solved iteratively, the generalized Hessian matrix $\mathbf{E}^{[2]}$ and the metric matrix $\mathbf{S}^{[2]}$ are not needed explicitly but may instead be defined in terms of their linear transformations of an arbitrary trial vector vec $\mathbf{b}$. Thus, in the notations

$$\text{vec}\,\mathbf{E}^{[2]}(\mathbf{b}) = \mathbf{E}^{[2]}\,\text{vec}\,\mathbf{b}, \tag{9}$$

$$\text{vec}\,\mathbf{S}^{[2]}(\mathbf{b}) = \mathbf{S}^{[2]}\,\text{vec}\,\mathbf{b}, \tag{10}$$

the Hessian and metric linear transformations are given by[7]

$$\begin{aligned}\boldsymbol{\sigma} &= \mathbf{E}^{[2]}(\mathbf{b}) \\ &= \mathcal{P}_{\mathrm{T}}[\mathbf{FD}_{\mathrm{b}}\mathbf{S} - \mathbf{SD}_{\mathrm{b}}\mathbf{F} + \mathbf{G}(\mathbf{D}_{\mathrm{b}})\mathbf{DS} - \mathbf{SDG}(\mathbf{D}_{\mathrm{b}})],\end{aligned} \tag{11}$$

$$\boldsymbol{\rho} = \mathbf{S}^{[2]}(\mathbf{b}) = -\,\mathcal{P}_{\mathrm{T}}(\mathbf{SD}_{\mathrm{b}}\mathbf{S}). \tag{12}$$

Here the Fock/KS matrix takes the form

$$\mathbf{F} = \mathbf{h} + \mathbf{G}(\mathbf{D}), \tag{13}$$

where $\mathbf{G}(\mathbf{D})$ denotes the Coulomb and exact-exchange contributions. In Kohn-Sham theory, there is an additional contribution from the exchange-correlation potential, not included here. All formulas, however, are equally valid for Kohn-Sham theory. We have furthermore introduced the projector

$$\mathcal{P}_{\mathrm{T}}(\mathbf{M}) = \mathbf{P}_{\mathrm{o}}^{\mathrm{T}}\mathbf{MP}_{\mathrm{v}} + \mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{MP}_{\mathrm{o}} \tag{14}$$

by analogy with Eq. (3) and the transformed density matrix

$$\mathbf{D}_{\mathrm{b}} = [\mathcal{P}(\mathbf{b}),\mathbf{D}]_{\mathrm{S}} = \mathcal{P}([\mathbf{b},\mathbf{D}]_{\mathrm{S}}) = \mathbf{P}_{\mathrm{v}}\mathbf{bP}_{\mathrm{o}}^{\mathrm{T}} - \mathbf{P}_{\mathrm{o}}\mathbf{bP}_{\mathrm{v}}^{\mathrm{T}} \tag{15}$$

in terms of the S commutator $[\mathbf{M},\mathbf{N}]_{\mathrm{S}}=\mathbf{MSN}-\mathbf{NSM}$. Assuming that $\mathcal{P}(\mathbf{b})=\mathbf{b}$, we may also write the linear transformations Eqs. (11) and (12) in the form

$$\begin{aligned}\mathbf{E}^{[2]}(\mathbf{b}) = (\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}})\mathbf{bS} + \mathbf{Sb}(\mathbf{F}^{\mathrm{vv}} - \mathbf{F}^{\mathrm{oo}}) + \mathbf{G}^{\mathrm{vo}}(\mathbf{D}_{\mathrm{b}}) \\ - \mathbf{G}^{\mathrm{ov}}(\mathbf{D}_{\mathrm{b}}),\end{aligned} \tag{16}$$

$$\mathbf{S}^{[2]}(\mathbf{b}) = -\,\mathbf{S}^{\mathrm{vv}}\mathbf{bS}^{\mathrm{oo}} + \mathbf{S}^{\mathrm{oo}}\mathbf{bS}^{\mathrm{vv}}, \tag{17}$$

where we have introduced the notation

$$\mathbf{M}^{\mathrm{mn}} = \mathbf{P}_{\mathrm{m}}^{\mathrm{T}}\mathbf{MP}_{\mathrm{n}}, \tag{18}$$

noting that $\mathbf{S}^{\mathrm{vo}}=\mathbf{S}^{\mathrm{ov}}=\mathbf{0}$.

Excitation energies—that is, the poles of the linear response function in Eq. (6)—are the eigenvalues of the generalized eigenvalue problem

$$(\mathbf{E}^{[2]} - \omega_{n0}\mathbf{S}^{[2]})\text{vec}\,\mathbf{X}_n = \mathbf{0}, \tag{19}$$

where ω_{n0} is the excitation energy from the ground state $|0\rangle$ to the excited state $|n\rangle$. The corresponding transition moment of $\hat{A}$ is obtained from the residue of the linear response function

$$\langle 0|\hat{A}|n\rangle = \text{Tr}[\mathbf{A}^{[1]}\mathbf{X}_n]. \tag{20}$$

In this paper, we describe how linear response functions, excitation energies, and transition moments (one-photon transition strengths) may be evaluated at a cost that, for sufficiently large systems, scales linearly with system size.

In iterative algorithms, which are here used to solve the response equations, the Hessian and metric linear transformations Eqs. (11) and (12) require the AO overlap matrix $\mathbf{S}$, the AO density matrix $\mathbf{D}$, and the AO Fock/KS matrix $\mathbf{F}$, all of which are also needed for the (linear-scaling) AO-based optimization of the energy.[3] The additional contribution from $\mathbf{G}(\mathbf{D}_{\mathrm{b}})$ in Eq. (11), which is not needed for energy optimizations, can also be calculated at linear cost. The transformations Eqs. (11) and (12) consist entirely of sparse-matrix algebra and may for sufficiently large systems be carried out in linear time. We now turn our attention to the linear-scaling iterative solution of the linear set of equations Eq. (7) and the eigenvalue problem Eq. (19). Once their solutions have been found, molecular properties are straightforwardly obtained as the trace of sparse matrices Eqs. (6) and (20).

B. Iterative solution of response equations

Before describing the iterative algorithm, we note the relations

$$[\mathbf{E}^{[2]}(\mathbf{b})]^{\mathrm{T}} = \mathbf{E}^{[2]}(\mathbf{b}^{\mathrm{T}}), \tag{21}$$

$$[\mathbf{S}^{[2]}(\mathbf{b})]^{\mathrm{T}} = -\,\mathbf{S}^{[2]}(\mathbf{b}^{\mathrm{T}}). \tag{22}$$

Therefore, if the transformations Eqs. (11) and (12) are known for a given trial matrix $\mathbf{b}_i$,

$$\boldsymbol{\sigma}_i = \mathbf{E}^{[2]}(\mathbf{b}_i), \tag{23}$$

$$\boldsymbol{\rho}_i = \mathbf{S}^{[2]}(\mathbf{b}_i), \tag{24}$$

they are also known for the transposed trial matrix,

$$\boldsymbol{\sigma}_i^{\mathrm{T}} = \mathbf{E}^{[2]}(\mathbf{b}_i^{\mathrm{T}}), \tag{25}$$

$$-\,\boldsymbol{\rho}_i^{\mathrm{T}} = \mathbf{S}^{[2]}(\mathbf{b}_i^{\mathrm{T}}). \tag{26}$$

Since the transformations of $\mathbf{b}_i$ and $\mathbf{b}_i^{\mathrm{T}}$ are related to each other in such a simple manner, new trial matrices are always added in pairs $\mathbf{b}_i$ and $\mathbf{b}_i^{\mathrm{T}}$.

Let us now assume that we solve the response equations Eq. (7) iteratively and that, in the course of the iterations, n pairs of trial matrices have been generated. These matrices constitute a $2n$-dimensional reduced basis:

$$\mathbf{b}^{2n} = \{\mathbf{b}_1,\mathbf{b}_1^{\mathrm{T}},\mathbf{b}_2,\mathbf{b}_2^{\mathrm{T}},\ldots,\mathbf{b}_n,\mathbf{b}_n^{\mathrm{T}}\}. \tag{27}$$

We assume that the trial matrices are orthonormal,

$$\text{Tr}(\mathbf{b}_i\mathbf{b}_j) = \text{Tr}(\mathbf{b}_i^{\mathrm{T}}\mathbf{b}_j^{\mathrm{T}}) = \delta_{ij}, \tag{28}$$

$$\text{Tr}(\mathbf{b}_i^{\mathrm{T}}\mathbf{b}_j) = \text{Tr}(\mathbf{b}_i\mathbf{b}_j^{\mathrm{T}}) = 0, \tag{29}$$

and that they satisfy the projection relation

$$\mathbf{b}_i = \mathcal{P}(\mathbf{b}_i). \tag{30}$$

The transformed trial matrices $\boldsymbol{\sigma}_i = \mathbf{E}^{[2]}(\mathbf{b}_i)$ and $\boldsymbol{\rho}_i = \mathbf{S}^{[2]}(\mathbf{b}_i)$ are then given by

$$\boldsymbol{\sigma}^{2n} = \{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_1^{\mathrm{T}}, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_2^{\mathrm{T}}, \ldots, \boldsymbol{\sigma}_n, \boldsymbol{\sigma}_n^{\mathrm{T}}\}, \tag{31}$$

$$\boldsymbol{\rho}^{2n} = \{\boldsymbol{\rho}_1, -\boldsymbol{\rho}_1^{\mathrm{T}}, \boldsymbol{\rho}_2, -\boldsymbol{\rho}_2^{\mathrm{T}}, \ldots, \boldsymbol{\rho}_n, -\boldsymbol{\rho}_n^{\mathrm{T}}\}. \tag{32}$$

The basis of trial matrices and their transformed counterparts are then used to set up the response equations in a reduced space of dimension $2n$:

$$(\mathbf{E}_{\mathrm{R}}^{[2]} - \omega \mathbf{S}_{\mathrm{R}}^{[2]})\mathbf{X}_{\mathrm{R}} = -\mathbf{B}_{\mathrm{R}}^{[1]}, \tag{33}$$

where the reduced-space gradient elements are given as

$$(\mathbf{B}_{\mathrm{R}}^{[1]})_i = \mathrm{Tr}[(\mathbf{B}^{[1]})^{\mathrm{T}} \mathbf{b}_i^{2n}], \tag{34}$$

whereas the reduced-space generalized Hessian and metric matrices become

$$(\mathbf{E}_{\mathrm{R}}^{[2]})_{ij} = \mathrm{Tr}[(\mathbf{b}_i^{2n})^{\mathrm{T}} \boldsymbol{\sigma}_j^{2n}], \tag{35}$$

$$(\mathbf{S}_{\mathrm{R}}^{[2]})_{ij} = \mathrm{Tr}[(\mathbf{b}_i^{2n})^{\mathrm{T}} \boldsymbol{\rho}_j^{2n}]. \tag{36}$$

The reduced equations Eq. (33) are easily solved since the dimension $2n$ is small.

From the solution to the reduced problem Eq. (33), we may expand the current optimal solution matrix $\mathbf{X}$ as

$$\mathbf{X} = \sum_{i=1}^{2n} (X_{\mathrm{R}})_i \mathbf{b}_i^{2n}, \tag{37}$$

whereas the residual is evaluated from the transformed matrices Eqs. (31) and (32):

$$\begin{aligned} \mathbf{R} &= \mathbf{E}^{[2]}(\mathbf{X}) - \omega \mathbf{S}^{[2]}(\mathbf{X}) + \mathbf{B}^{[1]} \\ &= \sum_{i=1}^{2n} (X_{\mathrm{R}})_i (\boldsymbol{\sigma}_i^{2n} - \omega \boldsymbol{\rho}_i^{2n}) + \mathbf{B}^{[1]}. \end{aligned} \tag{38}$$

To accelerate convergence, this residual is preconditioned as

$$\mathbf{M}\ \mathrm{vec}\ \mathbf{R}_{\mathrm{p}} = \mathrm{vec}\ \mathbf{R}, \tag{39}$$

where the preconditioner $\mathbf{M}$ is an easily constructed approximation to $\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}$ as discussed in the next section. From the projected preconditioned residual $\mathcal{P}(\mathbf{R}_{\mathrm{p}})$, a new pair of trial matrices $\mathbf{b}_{n+1}$ and $\mathbf{b}_{n+1}^{\mathrm{T}}$ is generated by orthogonalization against the previous basis matrices Eq. (27), ensuring that the new vector pair is normalized and orthogonal,

$$\mathrm{Tr}(\mathbf{b}_{n+1}\mathbf{b}_{n+1}) = 1, \quad \mathrm{Tr}(\mathbf{b}_{n+1}^{\mathrm{T}}\mathbf{b}_{n+1}) = 0, \tag{40}$$

These iterations are continued until the residual is smaller than some preset threshold, using the right-hand side $\mathbf{B}^{[1]}$ of Eq. (7) as an initial guess.

The eigenvalue problem Eq. (19) is solved in the same manner as the response equations, setting up the reduced equations in the space of the $2n$ trial vectors Eq. (27):

$$(\mathbf{E}_{\mathrm{R}}^{[2]} - \omega_{n0}^{\mathrm{R}} \mathbf{S}_{\mathrm{R}}^{[2]})\mathbf{X}_{\mathrm{R},n} = \mathbf{0}. \tag{41}$$

These low-dimensional equations may be solved straightforwardly, yielding an optimal excitation energy ω_{n0}^{R} and eigenvector $\mathbf{X}_{\mathrm{R},n}$. The corresponding residual is given by

$$\mathbf{R} = \mathbf{E}^{[2]}(\mathbf{X}_n) - \omega_{n0}^{\mathrm{R}} \mathbf{S}^{[2]}(\mathbf{X}_n), \tag{42}$$

where $\mathbf{X}_n$ is the expansion of the reduced-space eigenvector $\mathbf{X}_{\mathrm{R},n}$ in the trial vectors Eq. (37). This residual may be preconditioned as in Eq. (39) (with ω_{n0}^{R} replacing ω in $\mathbf{M}$) to generate the new pair of trial vectors $\mathbf{b}_{n+1}$ and $\mathbf{b}_{n+1}^{\mathrm{T}}$, and the iterations are continued until convergence. We discuss below how the initial guess of the excitation vector is obtained.

The strategy of adding trial vectors in conjugate pairs (rather than one at a time) not only accelerates the solution by adding two vectors at the cost of one. More importantly, it imposes the correct paired structure on $\mathbf{E}_{\mathrm{R}}^{[2]}$ and $\mathbf{S}_{\mathrm{R}}^{[2]}$, thereby avoiding complex eigenvalues and ensuring monotonic convergence.

C. Preconditioning

1. The AO basis

The preconditioner $\mathbf{M}$ in Eq. (39) should be a good approximation to the response matrix in the sense that the condition number of $\mathbf{M}^{-1}(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]})$ should be significantly smaller than that of $\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}$. Moreover, the cost of solving the preconditioning equation Eq. (39) should be significantly smaller than that of solving the original response equation Eq. (7). The most expensive step in the solution of Eq. (7) is the evaluation of $\mathbf{G}(\mathbf{D}_{\mathrm{b}})$, which contributes to the two last terms in Eq. (11). Since these terms are small compared with the other terms in Eq. (11), a good preconditioner is given by

$$\mathbf{M} = \mathbf{E}_{\mathrm{F}}^{[2]} - \omega \mathbf{S}^{[2]}, \tag{43}$$

where $\mathbf{E}_{F}^{[2]}$ is an approximation to $\mathbf{E}^{[2]}$ with the last two terms in Eq. (11) neglected:

$$\boldsymbol{\sigma}_{\mathrm{F}} = \mathbf{E}_{\mathrm{F}}^{[2]}(\mathbf{b}) = \mathcal{P}_{\mathrm{T}}(\mathbf{F}\mathbf{D}_{\mathrm{b}}\mathbf{S} - \mathbf{S}\mathbf{D}_{\mathrm{b}}\mathbf{F}). \tag{44}$$

The equations for the preconditioned residual $\mathbf{R}_{\mathrm{p}}$ Eq. (39) may be solved iteratively in the same manner that we solved the response equations Eq. (7).

In solving the response eigenvalue problem Eq. (16), the residual Eq. (42) may be preconditioned as in Eq. (39), using Eq. (43) with ω replaced by ω_{n0}^{R}. However, the solution of the preconditioning equation Eq. (39),

$$(\mathbf{E}_{\mathrm{F}}^{[2]} - \omega_{n0}^{\mathrm{R}} \mathbf{S}^{[2]})\mathrm{vec}\ \mathbf{R}_{\mathrm{p}} = \mathrm{vec}\ \mathbf{R}, \tag{45}$$

in the AO basis is difficult since the condition number of $\mathbf{E}_{\mathrm{F}}^{[2]} - \omega_{n0}^{\mathrm{R}} \mathbf{S}^{[2]}$ is large. For a solution to this problem, we examine in the next section the preconditioner in the MO basis.

2. The MO basis

An iterative algorithm for solving response eigenvalue and linear equations similar to that presented for the AO basis above has been successfully used in the MO basis,[10,11] where $\mathbf{E}^{[2]}$ and $\mathbf{S}^{[2]}$ are diagonally dominant. In the MO basis, the preconditioner Eq. (43) becomes diagonal,

$$\mathbf{M}_{\mathrm{MO}} = (\mathbf{E}_{\mathrm{F}}^{[2]})_{\mathrm{MO}} - \omega(\mathbf{S}^{[2]})_{\mathrm{MO}} = \begin{pmatrix} \boldsymbol{\Delta\epsilon} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta\epsilon} \end{pmatrix} - \omega \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}, \tag{46}$$

where the diagonal matrix $\boldsymbol{\Delta\epsilon}$ contains the differences between virtual and occupied orbital energies,

$$\Delta\epsilon_{AI,AI} = \epsilon_A - \epsilon_I. \tag{47}$$

In the MO basis, therefore, the preconditioning may be carried out in a simple manner, dividing the residual $\mathbf{R}$ by the diagonal elements of Eq. (46).

In the AO basis, by contrast, neither $\mathbf{E}_{\mathrm{F}}^{[2]}$ nor $\mathbf{S}^{[2]}$ are diagonally dominant. Furthermore, the condition number of $\mathbf{E}_{\mathrm{F}}^{[2]}$ is significantly larger in the AO basis than in the MO basis, making the iterative solution of Eq. (39) difficult. Since the condition number of a matrix is unaffected by a similarity transformation, we may dramatically improve the conditioning of the equations (reducing the condition number to that of the MO basis) by transforming them to an orthogonal AO basis (OAO) such as the Cholesky basis or the Löwdin basis. Furthermore, in the OAO basis, the preconditioner $\mathbf{M}_{\mathrm{OAO}}$ is much more diagonally dominant than in the original AO basis. In Sec. II C 3, we consider how the preconditioned equations may be solved in the OAO basis. However, we first discuss here how an initial guess of an excitation vector may be obtained.

In the MO basis, the initial guess of an excitation vector has previously been successfully obtained as the solution to the simplified response eigenvalue equations

$$\left[\begin{pmatrix} \boldsymbol{\Delta\epsilon} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta\epsilon} \end{pmatrix} - \omega \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \right] \begin{pmatrix} \mathbf{Y}_{\mathrm{vo}} \\ \mathbf{Y}_{\mathrm{ov}} \end{pmatrix} = \mathbf{0}, \tag{48}$$

where we recognize the simplified response matrix of Eq. (46). In Eq. (48), $\mathbf{Y}_{\mathrm{vo}}$ and $\mathbf{Y}_{\mathrm{ov}}$ are the virtual-occupied and occupied-virtual blocks, respectively, of the matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{0} & \mathbf{Y}_{\mathrm{vo}} \\ \mathbf{Y}_{\mathrm{ov}} & \mathbf{0} \end{pmatrix}. \tag{49}$$

The solution of Eq. (48) has zero elements in $\mathbf{Y}_{\mathrm{vo}}$ and $\mathbf{Y}_{\mathrm{ov}}$ except for a unit element in $\mathbf{Y}_{\mathrm{vo}}$ corresponding to the considered orbital-energy difference $\epsilon_A - \epsilon_I$. In Sec. II D, we shall discuss how an equivalent initial vector may be set up in the OAO basis.

3. The orthogonal AO basis

In the OAO basis, the AO overlap matrix is factorized as

$$\mathbf{S} = \mathbf{V}^{\mathrm{T}}\mathbf{V}, \tag{50}$$

where $\mathbf{V}$ is either an upper triangular matrix $\mathbf{U}$ (in the Cholesky basis) or the principal square-root matrix $\mathbf{S}^{1/2}$ (in the Löwdin basis):

$$\mathbf{V}_{\mathrm{C}} = \mathbf{U}, \tag{51}$$

$$\mathbf{V}_{\mathrm{L}} = \mathbf{S}^{1/2}. \tag{52}$$

In Ref. 3, we found that both schemes give diagonally dominant Hessians, with a slight preference for the Löwdin basis. An advantage of the Löwdin basis is that, among all possible orthogonal bases, it resembles most closely the original AO basis, ensuring that locality is preserved to the greatest possible extent. Furthermore, the transformation to the Löwdin basis can be performed straightforwardly within a linear-scaling framework.[22] Except as noted, we use the Löwdin basis in our calculations.

In the OAO basis defined by Eq. (50), the linear transformations entering Eq. (39) become

$$(\boldsymbol{\sigma}_{\mathrm{F}})_{\mathrm{V}} = (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}})\mathbf{X}^{\mathrm{V}} + \mathbf{X}^{\mathrm{V}}(\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}}), \tag{53}$$

$$\boldsymbol{\rho}_{\mathrm{V}} = \mathbf{D}^{\mathrm{V}}\mathbf{X}^{\mathrm{V}} - \mathbf{X}^{\mathrm{V}}\mathbf{D}^{\mathrm{V}}, \tag{54}$$

where we have used the notations

$$\mathbf{A}_{\mathrm{V}} = \mathbf{V}^{-\mathrm{T}}\mathbf{A}\mathbf{V}^{-1}, \tag{55}$$

$$\mathbf{A}^{\mathrm{V}} = \mathbf{V}\mathbf{A}\mathbf{V}^{\mathrm{T}}. \tag{56}$$

The preconditioning of the residual for the response equations Eq. (39) is performed in the OAO basis and the conjugate-gradient algorithm may be used with the diagonal preconditioner

$$M_{\alpha\beta,\alpha\beta} = (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}})_{\alpha\alpha} + (\mathbf{F}_{\mathrm{V}}^{\mathrm{vv}} - \mathbf{F}_{\mathrm{V}}^{\mathrm{oo}})_{\beta\beta} - \omega[(\mathbf{D}^{\mathrm{V}})_{\alpha\alpha} - (\mathbf{D}^{\mathrm{V}})_{\beta\beta}]. \tag{57}$$

The preconditioning of the residual of the eigenvalue equations may be carried out in the same manner but with the frequency ω replaced by the excitation energy ω_{n0}^{R}.

D. Initial vectors for the response eigenvalue equation

In the MO basis, the $\mathbf{Y}$ matrix in Eq. (49) has been successfully used to obtain an initial guess of the excitation vector in the iterative solution of the response eigenvalue equations. The $\mathbf{Y}$ matrix is zero except for a unit element Y_{AI} corresponding to the considered orbital-energy difference $\epsilon_A - \epsilon_I$. If the lowest excitation energy is determined, the lowest orbital energy difference [i.e., the highest occupied molecular orbital (HOMO)–lowest unoccupied molecular orbital (LUMO) gap] is considered and similarly for higher excited states. In the OAO basis of Eq. (50), the initial vector becomes

$$\mathbf{Y}_{\mathrm{OAO}} = \mathbf{C}\mathbf{Y}_{\mathrm{MO}}\mathbf{C}^{\mathrm{T}}, \tag{58}$$

where $\mathbf{C}$ contains the eigenvectors of the Fock/KS matrix in the OAO basis,

$$\mathbf{F}_{\mathrm{V}}\mathbf{C} = \boldsymbol{\epsilon}\mathbf{C}. \tag{59}$$

For an initial guess that is represented by a unit element Y_{AI} in the MO basis, the OAO initial vector becomes

$$(\mathbf{Y}_{\mathrm{OAO}})_{\mu\nu} = C_{\mu A}C_{\nu I}, \tag{60}$$

where indices μ and ν refer to OAO basis. In this basis, the projectors onto the occupied and virtual spaces become

$$\mathbf{P}_{\mathrm{o}}^{\mathrm{OAO}} = \mathbf{D}^{\mathrm{V}}, \tag{61}$$

$$\mathbf{P}_{\mathrm{v}}^{\mathrm{OAO}} = \mathbf{1} - \mathbf{D}^{\mathrm{V}}. \tag{62}$$

The Fock/KS eigenvalue equation in Eq. (59) may then be written as

$$(\mathbf{P}_o^{OAO}\mathbf{F}_V\mathbf{P}_o^{OAO} + \mathbf{P}_v^{OAO}\mathbf{F}_V\mathbf{P}_v^{OAO})\mathbf{C} = \boldsymbol{\epsilon}\mathbf{C}, \quad (63)$$

since $\mathbf{P}_o^{OAO}\mathbf{F}_V\mathbf{P}_v^{OAO} = \mathbf{P}_v^{OAO}\mathbf{F}_V\mathbf{P}_o^{OAO} = \mathbf{0}$ for an optimized state. Projecting Eq. (63) onto the occupied and virtual spaces, we obtain

$$\mathbf{P}_o^{OAO}\mathbf{F}_V\mathbf{P}_o^{OAO}\mathbf{C} = \boldsymbol{\epsilon}\mathbf{P}_o^{OAO}\mathbf{C}, \quad (64)$$

$$\mathbf{P}_v^{OAO}\mathbf{F}_V\mathbf{P}_v^{OAO}\mathbf{C} = \boldsymbol{\epsilon}\mathbf{P}_v^{OAO}\mathbf{C}, \quad (65)$$

demonstrating that the orbital energies and eigenvectors of the occupied and virtual spaces can be obtained from Eqs. (64) and (65), respectively. Using iterative techniques, we may thus determine the eigenvectors of the highest occupied orbitals from Eq. (64) and of the lowest virtual orbitals from Eq. (65). Subsequently, Eq. (60) may be used to generate start vectors in the OAO basis.

III. ILLUSTRATIVE RESULTS

In this subsection, we report calculations of excitation energies and frequency-dependent polarizabilities for polyalanine peptides of increasing size. The polyalanines are one-dimensional systems and thus ideal systems for demonstrating that linear scaling is approached. The largest peptide contains 139 alanine residues and 1392 atoms. We use CAM-B3LYP/6-31G to calculate the lowest excitation energy and Hartree-Fock/6-31G to calculate the frequency-dependent polarizability at a frequency of 0.1 a.u. The CAM-B3LYP (Ref. 23) functional is chosen because it gives significantly improved molecular properties compared with the B3LYP functional.[24] For each type of property calculation, we analyze both the scaling with respect to increasing molecular size and the convergence characteristics of the algorithm on one selected peptide—namely, ALA119 (containing 119 alanine residues) for the frequency-dependent polarizability and ALA59 for the excitation energy calculation. The SCF convergence is similar to that described for the ALA99 calculations in Ref. 3. All calculations have been carried out using a local version of DALTON.[25] The timings are obtained using a single processor on a SUN Fire X4600 (Opteron, 2.6 GHz).

A. The frequency-dependent polarizability of a peptide with 119 alanine residues

In this section, we describe a typical frequency-dependent polarizability calculation using ALA119 as an example. First, the response equations Eq. (7) are solved at a frequency of 0.1 a.u., after which the polarizability is obtained as the trace of the property gradient and the solution matrix according to Eq. (6). The linear equations are solved in the AO basis, using the iterative algorithm of Sec. II B. At each iteration, the residual is transformed to the Löwdin basis and preconditioned as described in Sec. II C. As initial trial vector for the response equations, the property gradient is used.

In Table I, we have listed the residual at each iteration in the solution of the linear equations Eq. (7). Convergence to a Frobenius norm of 10^{-2} of the residual is obtained in ten iterations. At each linear-response iteration, the residual is preconditioned by solving the linear equations Eq. (39) with $\mathbf{M}$ given in Eq. (43) in the Löwdin basis, using the linear transformations Eqs. (53) and (54) and the diagonal preconditioner Eq. (57). The iterations are terminated when the residual of Eq. (39) (in the Löwdin basis) has been reduced by a factor of 100 (the overall convergence of the response equations is not sensitive to the choice of this threshold.) As seen from Table I, for all response iterations, the preconditioning equations converge in seven iterations, which is the case in all polarizability calculations presented here.

TABLE I. The residual norm $\|\mathbf{R}\|$ and the number of preconditioning iterations n_{pre} for the calculation of the frequency-dependent polarizability at $\omega=0.1$ a.u. of ALA119 at the Hartree-Fock/6-31G level of theory.

It.	$\|\mathbf{R}\|$	n_{pre}
1	28.10	7
2	12.71	7
3	3.967	7
4	1.776	7
5	0.746	7
6	0.326	7
7	0.125	7
8	0.060	7
9	0.027	7
10	0.011	7

We now consider in more detail the preconditioning of the first trial vector of the ALA119 calculation. In Table II, we have listed the residual of the preconditioning equations Eq. (39), with and without the diagonal preconditioner Eq. (57). Although the diagonal preconditioner dramatically improves convergence, its use requires that the trial vectors are projected, since the preconditioning introduces redundant components. The projection requires four additional matrix

TABLE II. Convergence of the preconditioning equations in the first response iteration of the Hartree-Fock/6-31G calculation of the frequency-dependent polarizability $\omega=0.1$ a.u. of ALA119. The residual norms are given in the Löwdin and Cholesky bases, with and without diagonal preconditioning.

	Löwdin basis		Cholesky basis	
It.	No. prec.	Dia. prec.	No. prec.	Dia. prec.
1	82.20	82.20	82.20	82.20
2	42.15	19.39	41.94	19.27
3	41.16	7.36	41.12	9.11
4	27.82	5.04	27.96	5.56
5	13.47	2.26	13.19	2.68
6	18.05	0.98	18.05	1.24
7	8.44	0.43	8.41	0.59
8	6.59		6.61	
9	7.18		7.16	
10	3.87		3.88	
11	2.90		2.89	
12	2.19		2.19	
13	1.48		1.48	
14	1.29		1.29	
15	0.82		0.81	
16	0.56			

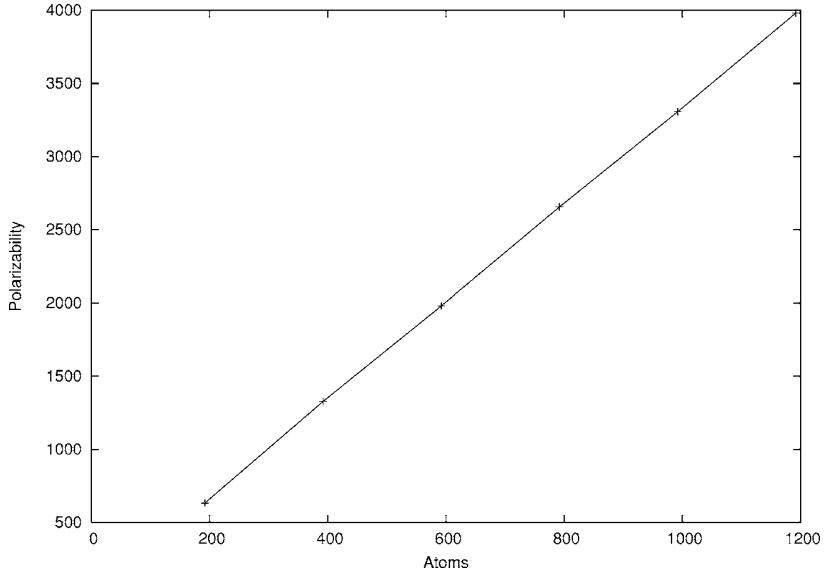

FIG. 1. The xx component of the Hartree-Fock/6-31G frequency-dependent dipole polarizability (ω =0.1 a.u.) for polyalanines (atomic units).

multiplications per iteration, making preconditioning less attractive. We nevertheless recommend its use since otherwise the equations sometimes do not converge.

We use the Löwdin OAO basis by default but have also included in Table II information about convergence in the Cholesky basis. Although the Löwdin basis sometimes gives faster convergence than the Cholesky basis, the situation illustrated in Table II is fairly typical, with a nearly identical behavior in the two bases.

The convergence of the response equations reported here is typical of polarizability calculations and similar to that of the standard MO-based iterative algorithm of Ref. 11 [implemented in DALTON (Ref. 25)]. Any difference in convergence arises because the preconditioning equations are terminated when the residual has been reduced by a factor of 100. If the preconditioning equations were converged to full accuracy, identical results would be obtained in the AO and MO bases.

B. Linear-scaling frequency-dependent polarizability calculations

In Fig. 1, the frequency-dependent dipole longitudinal polarizability $\alpha_{xx}(\omega)$ at ω=0.1 a.u. is plotted as a function of the number of atoms in polyalanine peptides, calculated at the Hartree-Fock/6-31G level of theory. As expected, the longitudinal polarizability depends linearly on the number of atoms. In Fig. 2, we have plotted the CPU times of the different parts of the polarizability calculations, using the block sparse-matrix scheme described by Rubensson and Sałek in Ref. 26. The timings are for the following contributions in the first response iteration: the Coulomb part ("Fock J") and the exchange part ("Fock X") of the $\mathbf{G}(\mathbf{D}_b)$ contribution to the linear transformation Eq. (11), the remainder of the linear transformations Eqs. (11) and (12) ("Lintra"), and the preconditioning of the trial vectors ("Precond").

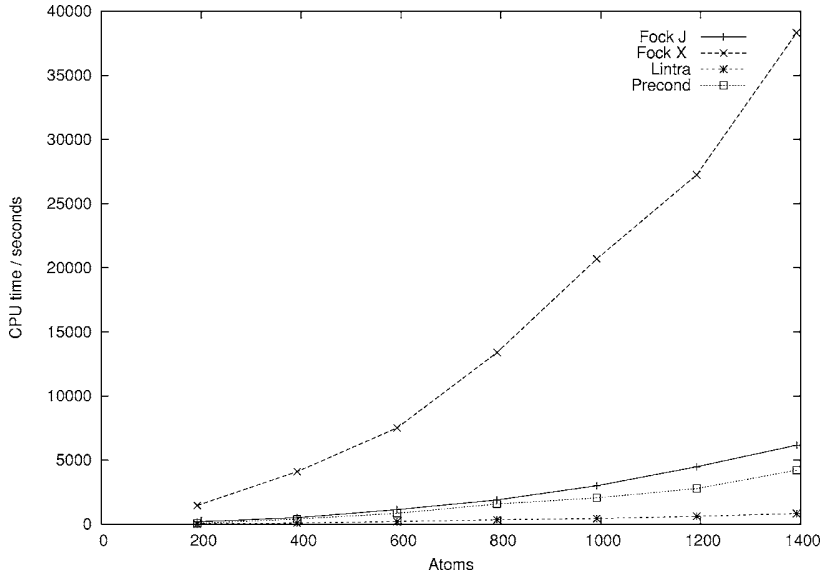

FIG. 2. Timings for the different parts of a Hartree-Fock/6-31G response iteration in a frequency-dependent polarizability calculation (ω=0.1 a.u.) for polyalanines. Sparse-matrix algebra is used.

TABLE III. The residual norm $\|\mathbf{R}\| \times 10^3$ and the number of preconditioning iterations n_{pre} for the calculation of the lowest excitation energy of ALA59 at the CAM-B3LYP/6-31G level of theory (full-matrix algebra).

It.	$\|\mathbf{R}\| \times 10^3$	n_{pre}
1	0.197	20
2	0.057	20
3	0.066	20
4	0.014	20
5	0.008	20

The Hartree-Fock polarizability calculations are dominated by the exchange contribution to the linear transformation, whose calculation approaches linearity for systems containing more than 600 atoms. The Coulomb evaluation (by density fitting) is several times faster than the exchange evaluation. For the remaining two contributions in Fig. 2, the time-consuming parts consist of matrix multiplications and scale linearly with system size, showing that matrix sparsity is efficiently exploited in our calculations.

C. The lowest excitation energy for a peptide with 59 alanine residues

In this section, we consider the calculation of excitation energies for large systems using ALA59 in full-matrix algebra as an example. Excitation energies are eigenvalues of the response eigenvalue problem Eq. (19), which is solved in the AO basis using the iterative algorithm of Sec. II B. At each iteration, the residual is transformed to the Löwdin basis and preconditioned as described in Sec. II C. As an initial eigenvector guess, we use Eq. (60), where the eigenvectors of the Fock/KS matrix in the occupied and virtual spaces are determined from Eqs. (64) and (65), respectively.

In Table III, we have listed the residual at each iteration of the solution of the response eigenvalue problem Eq. (19). The response iterations are terminated when the residual norm has been reduced by a factor of 100, which is obtained in five iterations. At each iteration, the residual is preconditioned by solving the simplified response equations Eq. (45) in the Löwdin basis, using the linear transformations of Eqs. (53) and (54) and the diagonal preconditioner Eq. (57). The preconditioning iterations are also terminated when the residual of Eq. (45) (in the Löwdin basis) has been reduced by a factor of 100 or after a maximum of 20 iterations. As indicated in Table III, the preconditioning equations always terminated at the maximum number of iterations, which is true for all excitation energy calculations presented here.

We now consider in more detail the preconditioning of the first trial vector of the ALA59 calculation. Full-matrix rather than sparse-matrix algebra was used in this calculation, since the residual is very small already in the first response iteration, and the efficiency of the preconditioner becomes blurred by numerical noise when sparse-matrix algebra is used. In Table IV, we have listed the residual of the preconditioning equations Eq. (45), with and without the diagonal preconditioner Eq. (57). Without preconditioning, the equations do not converge. With preconditioning, the residual decreases slowly until, after the maximum number of allowed iterations, it has been reduced by about an order of magnitude. Clearly, it is much more difficult to converge the preconditioning equations for the response eigenvalue equations than for the response linear equations.

TABLE IV. Convergence of the preconditioning equations in the first response iteration of the CAM-B3LYP/6-31G calculation of the lowest excitation energy of ALA59 (full-matrix algebra). The residual norms are given with and without diagonal preconditioning

	$\|\mathbf{R}\| \times 10^3$	
It.	No prec.	Dia. prec.
1	0.309	0.309
2	0.364	0.253
3	0.235	0.349
4	0.413	0.303
5	0.368	0.262
6	0.276	0.249
7	0.429	0.201
8	0.272	0.161
9	0.335	0.135
10	0.321	0.135
11	0.230	0.122
12	0.370	0.122
13	0.258	0.118
14	0.241	0.107
15	0.368	0.091
16	0.201	0.073
17	0.230	0.060
18	0.226	0.047
19	0.153	0.042

To understand this difference between the response linear and eigenvalue equations, consider the carrier matrix Eq. (43) for the preconditioning equations $\mathbf{E}_{\text{F}}^{[2]} - \omega\mathbf{S}^{[2]}$, where ω is either the frequency of the applied field (linear equations) or a reduced-space eigenvalue (eigenvalue equations). The approximate generalized electronic Hessian $\mathbf{E}_{\text{F}}^{[2]}$ is positive definite provided the optimized Hartree-Fock or Kohn-Sham energy is a minimum—it is a well-behaved matrix that (when preconditioned) has a relatively small condition number. Consequently, rapid convergence is observed when the linear equations are solved in the static limit; moreover, since the applied frequency is typically small (0.1 a.u. in Sec. III A), the addition of $-\omega\mathbf{S}^{[2]}$ to $\mathbf{E}_{\text{F}}^{[2]}$ does not affect the convergence of the linear equations.

By contrast, in the solution of the response eigenvalue problem, the addition of $\omega_{n0}^{\text{R}}\mathbf{S}^{[2]}$ changes the structure of the carrier matrix, making $\mathbf{E}_{\text{F}}^{[2]} - \omega_{n0}^{\text{R}}\mathbf{S}^{[2]}$ nearly singular and ill conditioned. As a result, the preconditioning equations are much more difficult to converge for the response eigenvalue equations than for the linear equations. However, as seen from Table III, the eigenvalue equations can nevertheless be converged in a few iterations, because of the good starting guess of Eq. (60). When the Cholesky rather than the Löwdin basis is used for the preconditioning equations, the convergence is similar to that in the Löwdin basis.

D. Linear-scaling calculations of excitation energies

In Fig. 3, the lowest excitation energy, the lowest Hessian eigenvalue, and the HOMO-LUMO gap are plotted as

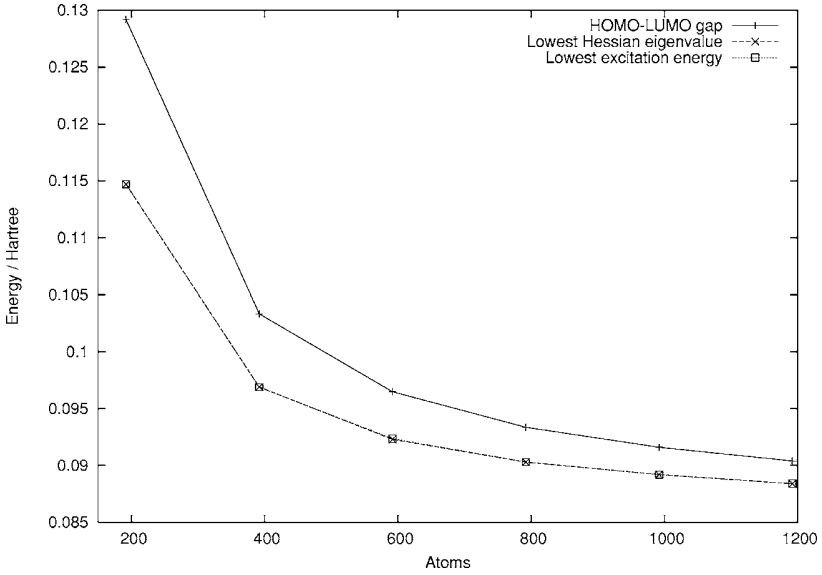

FIG. 3. The HOMO-LUMO gap, the lowest Hessian eigenvalue, and the lowest excitation energy for polyalanines, calculated at the CAM-B3LYP/6-31G level of theory.

functions of the number of atoms in polyalanine peptides at the CAM-B3LYP/6-31G level of theory. As expected, the excitation energy decreases with increasing system size. More surprisingly, the lowest Hessian eigenvalue and the lowest excitation energy are equal to the number of significant digits. To understand this behavior, consider the evaluation of excitation energies in the MO basis, where the response eigenvalue equation in a notation similar to that of Eq. (48) becomes

$$\left[\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B} & \mathbf{A}\end{pmatrix} - \omega\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}\end{pmatrix}\right]\begin{pmatrix}\mathbf{Y}_{\mathrm{vo}}\\ \mathbf{Y}_{\mathrm{ov}}\end{pmatrix} = \mathbf{0}. \tag{66}$$

The **B** matrix contains Hamiltonian matrix elements between the Kohn-Sham determinant and a doubly excited configuration. If it vanishes, the eigenvalues ω of Eq. (66) become equal to the eigenvalues of the electronic Hessian $\mathbf{A}-\mathbf{B}$. The exact exchange in CAM-B3LYP gives a nonzero **B** matrix contribution but is too small to be detected. The HOMO-LUMO gap is slightly above the calculated excitation energies, as expected from the form of the **A** matrix, which contains the orbital-energy difference of Eq. (48) with Coulomb and exchange-correlation contributions subtracted.

In Fig. 4, we have plotted the CPU times of excitation-energy calculations when sparse-matrix algebra is used, for the following contributions to the first response iteration: the Coulomb contribution to the linear transformation Eq. (11) ("Kohn-Sham J"), the exact-exchange contribution ("Kohn-Sham X"), the exchange-correlation contribution ("Kohn-Sham XC"), the remainder of the linear transformations Eqs. (11) and (12) ("Lintra"), and the preconditioning of the trial vectors ("Precond").

The excitation energy calculations are dominated by the exchange-correlation contribution. Comparing with the Hartree-Fock polarizability calculations in Fig. 2, we note that the evaluation of the Coulomb and exact-exchange contributions is much faster in the Kohn-Sham excitation energy calculations in Fig. 4. The difference arises since the timings are given for the first response iteration, for which the start-

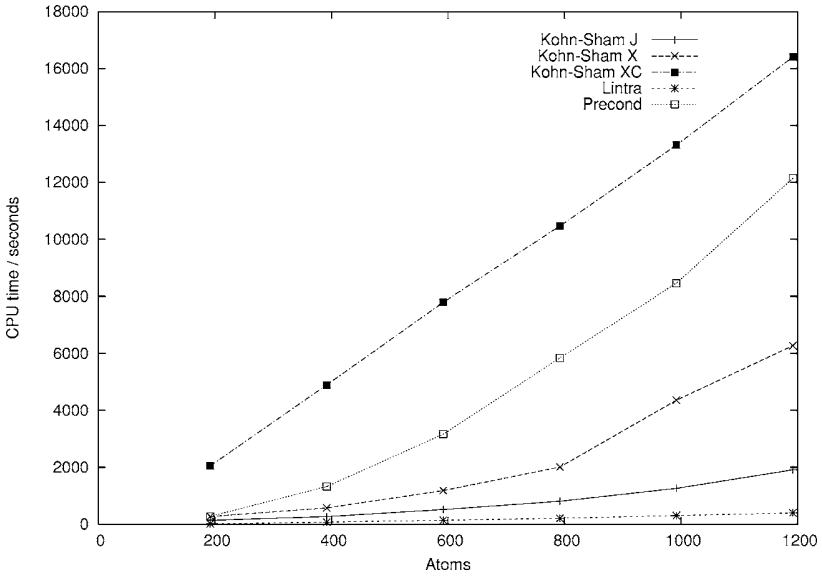

FIG. 4. Timings for different parts of a response iteration in a CAM-B3LYP/6-31G excitation energy calculation for polyalanines. Sparse-matrix algebra is used.

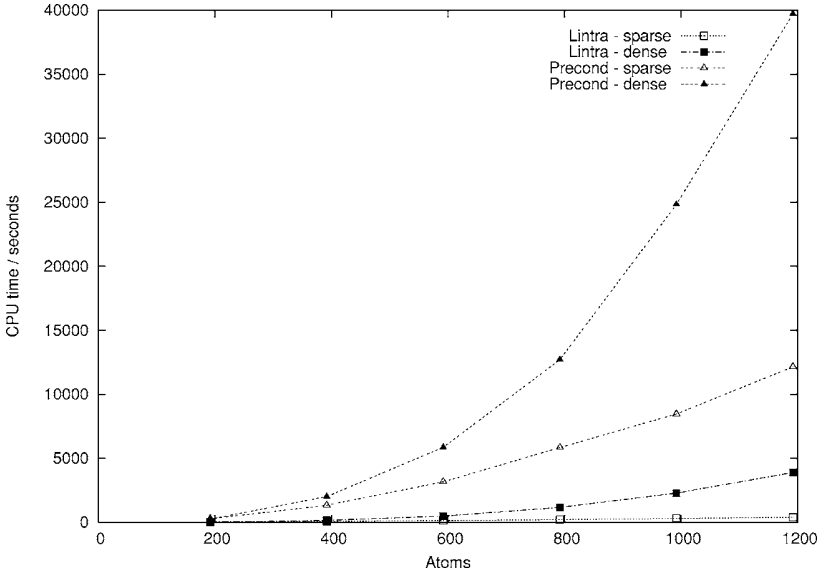

FIG. 5. Comparison of sparse- and dense-matrix timings for selected parts of a CAM-B3LYP/6-31G response iteration in an excitation energy calculation for polyalanines.

ing guess for excitation energies is more sparse than the one for polarizabilities. Also, preconditioning is more expensive for excitation energies than for polarizabilities since about three times more iterations are needed. As a result, preconditioning becomes more expensive than the evaluation of the Coulomb and exact-exchange contributions.

As seen from Fig. 4, the cost of the exchange-correlation contribution to the linear transformation scales linearly with system size. The same is true for the Coulomb contribution, while the evaluation of exact exchange is nonlinear, at least for small systems. For the Lintra and Precond contributions to the response equations, the time-consuming parts consist of matrix-matrix multiplications. The scaling of these contributions shows that sparsity is efficiently exploited, although the Precond contribution shows signs of nonlinear scaling. Investigation of the time spent in the preconditioning shows that the matrices involved in the linear transformation have not yet reached the regime of linear scaling in the number of nonzero elements. The benefits of sparse-matrix algebra are nevertheless evident from Fig. 5, where we compare the Lintra and Precond timings of Fig. 4 with those obtained using full-matrix algebra. Whereas the cost increases cubically when full-matrix algebra is used, linear scaling is approached with sparse-matrix algebra.

IV. CONCLUSIONS

Using the nonredundant exponential parametrization of the density matrix introduced in Refs. 1 and 18, we have presented a linear-scaling implementation of excitation energies and frequency-dependent second-order molecular properties. The response eigenvalue and linear equations are solved using an iterative subspace method equivalent to the one that has been successfully used in the MO basis. Important features of the subspace method are the use of paired trial vectors (to preserve the structure of the full equations in the reduced space), a nondiagonal preconditioning (for rapid convergence), and good start vectors (for robust and fast solution). The performance is similar to that in the MO basis, with five to ten iterations needed for convergence. The preconditioning is carried out in the Löwdin basis, solving a simplified version of the response equations with an iterative method similar to the one used for the full response equations. To reduce the residual of the preconditioning equations by a factor of 100, less than ten iterations are typically needed for the response linear equations. For the response eigenvalue equations, the preconditioning equations are more difficult to converge, easily requiring 20 iterations.

As for the optimization of the Hartree-Fock and KS density matrices, the solution of the response equations is dominated by the construction of the Fock/KS matrix, once at each iteration of the subspace algorithm. The solution of the preconditioning equations is dominated by matrix-matrix multiplications, for which linear scaling is approached by using sparse-matrix algebra. Calculations of the frequency-dependent polarizability at $\omega=0.1$ a.u. and of the lowest excitation energy have been presented for polyalanine peptides containing up to 1400 atoms, demonstrating the efficiency and robustness of the presented algorithm and that linear scaling can be obtained in such calculations.

ACKNOWLEDGMENTS

This work has been supported by the Lundbeck Foundation, the Danish Natural Research Council, and the Norwegian Research Council through a Strategic University Program in Quantum Chemistry (Grant No. 154011/420) and through a grant of computer time from the Program for Supercomputing. The authors also acknowledge support from the Danish Center for Scientific Computing (DCSC) and the European Research and Training Network "NANOQUANT, Understanding Nanomaterials from the Quantum Perspective," Contract No. MRTN-CT-2003-506842. One of the authors (S.C.) acknowledges support from the Italian Consiglio Nazionale delle Ricerche through a Short Term Mobility Grant.

[1] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (Wiley, Chichester, 2000).
[2] *Advances in Quantum Chemistry 50 - Response Theory and Molecular Properties (A Tribute to Jan Linderberg and Poul Jørgensen)*, edited by

H. J. Aa. Jensen (Elsevier, New York, 2005).
[3] P. Sałek, S. Høst, L. Thøgersen *et al.*, J. Chem. Phys. **126**, 114110 (2007).
[4] S. Goedecker and G. E. Scuseria, Comput. Sci. Eng. **5**, 14 (2003).
[5] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).
[6] J. Olsen and P. Jørgensen, J. Chem. Phys. **82**, 3235 (1985).
[7] H. Larsen, P. Jørgensen, J. Olsen, and T. Helgaker, J. Chem. Phys. **113**, 8908 (2000).
[8] J. Olsen (unpublished).
[9] L. Thøgersen, Ph.D. thesis, Aarhus University, 2005.
[10] J. Olsen and P. Jørgensen, in *Modern Electronic Structure Theory*, edited by D. R. Yarkony (World Scientific, Singapore, 1995), pt. II.
[11] J. Olsen, H. J. A. Jensen, and P. Jørgensen, J. Comput. Phys. **74**, 265 (1988).
[12] Y. Shao, C. Saravanan, M. Head-Gordon, and C. A. White, J. Chem. Phys. **118**, 6144 (2003).
[13] C. Ochsenfeld and M. Head-Gordon, Chem. Phys. Lett. **270**, 399 (1997).
[14] R. McWeeny, Rev. Mod. Phys. **32**, 325 (1960).
[15] X. P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993).
[16] C. Ochsenfeld, J. Kussmann, and F. Koziol, Angew. Chem., Int. Ed. **43**, 4485 (2004).
[17] V. Weber, A. M. N. Niklasson, and M. Challacombe, J. Chem. Phys. **123**, 044107 (2005).
[18] T. Helgaker, H. Larsen, J. Olsen, and P. Jørgensen, Chem. Phys. Lett. **327**, 397 (2000).
[19] H. Larsen, J. Olsen, P. Jørgensen, and T. Helgaker, J. Chem. Phys. **115**, 9685 (2001).
[20] H. Larsen, T. Helgaker, P. Jørgensen, and J. Olsen, J. Chem. Phys. **115**, 10344 (2001).
[21] H. Larsen, Ph.D. thesis, Aarhus University, 2001.
[22] B. Jansík, S. Høst, P. Jørgensen, and T. Helgaker, J. Chem. Phys. **126**, 124104 (2007).
[23] T. Yanai, D. P. Tew, and N. C. Handy, Chem. Phys. Lett. **393**, 51 (2004).
[24] M. J. G. Peach, T. Helgaker, P. Sałek, T. W. Keal, O. B. Lutnæs, D. J. Tozer, and N. C. Handy, Phys. Chem. Chem. Phys. **558**, 8 (2006).
[25] DALTON, an *ab initio* electronic structure program, Release 2.0, 2005 (see http://www.kjemi.uio.no/software/dalton/dalton.html).
[26] E. H. Rubensson and P. Sałek, J. Comput. Chem. **26**, 1628 (2005).