Fock Matrix Construction for Large Systems
Licentiate thesis
© Elias Rudberg, 2006
ISBN 91-7178-535-3
ISBN 978-91-7178-535-0
Printed by Universitetsservice US AB,
Stockholm, Sweden, 2006
Typeset in \LaTeX{} by the author
Abstract

This licentiate thesis deals with quantum chemistry methods for large systems. In particular, the thesis focuses on the efficient construction of the Coulomb and exchange matrices which are important parts of the Fock matrix in Hartree–Fock calculations. The methods described are also applicable in Kohn–Sham Density Functional Theory calculations, where the Coulomb and exchange matrices are parts of the Kohn–Sham matrix. Screening techniques for reducing the computational complexity of both Coulomb and exchange computations are discussed, as well as the fast multipole method, used for efficient computation of the Coulomb matrix.

The thesis also discusses how sparsity in the matrices occurring in Hartree–Fock and Kohn–Sham Density Functional Theory calculations can be used to achieve more efficient storage of matrices as well as more efficient operations on them.

As an example of a possible type of application, the thesis includes a theoretical study of Heisenberg exchange constants, using unrestricted Kohn–Sham Density Functional Theory calculations.
Preface

The work presented in this thesis has been carried out at the Department of Theoretical Chemistry, Royal Institute of Technology, Stockholm, Sweden.

List of papers included in the thesis


**Paper 2**  *A hierarchic sparse matrix data structure for large-scale Hartree-Fock/Kohn-Sham calculations*, Emanuel H. Rubensson, Elias Rudberg, and Paweł Salek, *(Submitted Manuscript)*.


List of papers not included in the thesis


- *Sparse Matrix Algebra for Quantum Modeling of Large Systems*, Emanuel H. Rubensson, Elias Rudberg, and Paweł Salek, *(Submitted Manuscript)*.

- *Near-Idempotent Matrices*, Emanuel H. Rubensson and Elias Rudberg, *(In Preparation)*.
Comments on my contribution to the papers included

• I was responsible for the calculations and for the writing of Paper 1.
• I was responsible for the calculations and for part of the writing of Paper 2.
• I was responsible for the calculations in Paper 3.
Acknowledgments

Many thanks to the following people:

Pawel Salek for being an excellent supervisor. Hans Ågren for making it possible for me to work with this. Emanuel Rubensson and Zilvinas Rinkevicius for working with me. Peter Taylor for his support. Everyone at the Theoretical Chemistry group in Stockholm, for making it a good place to work at.

This research has been supported by the Sixth Framework Programme Marie Curie Research Training Network under contract number MRTN-CT-2003-506842.
## Contents

<table>
<thead>
<tr>
<th>I Introductory Chapters</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>3</td>
</tr>
<tr>
<td>2 Hartree–Fock Theory</td>
<td>5</td>
</tr>
<tr>
<td>2.1 The Schrödinger Equation</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Slater Determinants</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Basis Sets</td>
<td>7</td>
</tr>
<tr>
<td>2.4 The Hartree–Fock Method</td>
<td>8</td>
</tr>
<tr>
<td>2.4.1 Restricted Hartree–Fock</td>
<td>8</td>
</tr>
<tr>
<td>2.4.2 The Self-Consistent Field Procedure</td>
<td>10</td>
</tr>
<tr>
<td>2.4.3 Unrestricted Hartree–Fock</td>
<td>10</td>
</tr>
<tr>
<td>3 Density Functional Theory</td>
<td>13</td>
</tr>
<tr>
<td>4 Integral Evaluation in HF and DFT</td>
<td>15</td>
</tr>
<tr>
<td>4.1 Primitive Gaussian Integrals</td>
<td>15</td>
</tr>
<tr>
<td>4.2 Computational Scaling</td>
<td>16</td>
</tr>
<tr>
<td>4.3 Screening</td>
<td>16</td>
</tr>
<tr>
<td>4.4 Cauchy–Schwartz Screening</td>
<td>17</td>
</tr>
<tr>
<td>4.5 The Coulomb and Exchange Matrices</td>
<td>17</td>
</tr>
<tr>
<td>5 Coulomb Matrix Construction</td>
<td>19</td>
</tr>
</tbody>
</table>
5.1 The Fast Multipole Method ........................................... 20

6 Storage and Manipulation of Matrices ......................... 21
6.1 Sparsity ................................................................... 21
6.2 Hierarchic Matrix Library ........................................... 22

7 Results ........................................................................... 23
7.1 Benchmarks ................................................................. 23
7.2 Application to Computation of Heisenberg Exchange Constants ...... 25

Bibliography .................................................................... 26

II Included Papers ............................................................. 29

1 Efficient implementation of the fast multipole method .......... 31

2 A hierarchic sparse matrix data structure for large-scale Hartree-Fock/Kohn-Sham calculations .................. 41

3 Heisenberg Exchange in Dinuclear Manganese Complexes: A Density Functional Theory Study ........... 57
Part I

Introductory Chapters
Most chemists today agree that chemistry is in principle well described by the theory of Quantum Mechanics. Quantum Mechanics was developed during the 1920’s. In 1929, P. A. M. Dirac wrote a paper including the following famous statement:

\[1\]

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

P. A. M. Dirac

This is a quite bold statement, which might seem provocative to a theoretical chemist if misinterpreted as “the whole of chemistry is known”. That is not, of course, what is meant: the statement says something about the underlying physical laws. To solve real problems in theoretical chemistry, two things are necessary. Firstly, knowledge of the underlying physical laws. Secondly, ways to get meaningful results from the mathematical equations dictated by those laws. It is interesting that Dirac used the word “only” when referring to the task of solving the equations. This wording does not really add any information to the statement; it merely reflects the author’s opinion that knowing the equations is more important than solving them.

Now, nearly 80 years later, the science of theoretical chemistry is still struggling with the solution of these equations. Therefore, one might argue that the word “only” should be omitted in a modern version of the above statement.

The mathematical equations in question can be “solved” in many different ways depending on which level of approximation is adopted. This thesis deals with one type of “solutions”: 
those given by the Hartree–Fock and Kohn–Sham Density Functional Theory methods. With these methods, large molecular systems can be treated quantum mechanically with the use of rather crude approximations.
Chapter 2

Hartree–Fock Theory

This chapter gives a brief description of the Hartree–Fock method from a computational point of view. A more thorough treatment of Hartree–Fock theory, including a complete derivation, can be found in the book by Szabo and Ostlund [2].

2.1 The Schrödinger Equation

The algorithms discussed in this thesis are useful when quantum mechanics is applied to the physical problem of \( N \) identical charged particles (the electrons) in the field of \( M \) fixed classical point charges (the nuclei). In the most common formulation of quantum mechanics, solving this problem means solving the \( N \)-electron Schrödinger equation:

\[
\hat{H} \psi = E \psi
\]  

The Schrödinger equation (2.1) is an eigenvalue equation involving the Hamiltonian operator \( \hat{H} \), the wave function \( \psi \), and the eigenvalue \( E \). Here, we will focus on finding the ground state of the system; that is, the eigenfunction \( \psi \) that gives the lowest energy \( E \). The Hamiltonian operator \( \hat{H} \) is given by

\[
\hat{H} = -\sum_{i=1}^{N} \frac{1}{2} \nabla_i^2 - \sum_{i=1}^{N} \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{1}{r_{ij}}
\]  

The wave function \( \psi \) is a function of the spatial positions and spins of the electrons,

\[
\psi = \psi(r_1, \omega_1, r_2, \omega_2, \ldots, r_N, \omega_N)
\]  

with the vectors \( r_i = (x_i, y_i, z_i) \) being the position of electron \( i \) and \( \omega_i \) being the spin of electron \( i \) : \( \psi \) is a function of \( 4N \) variables. In the definition of the Hamiltonian operator
(2.2), \( \nabla_i^2 \) is the Laplacian operator

\[
\nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}
\]

(2.4)

\( Z_A \) is the charge of nucleus \( A \), \( r_{iA} = |r_i - r_A| \) is the distance between electron \( i \) and nucleus \( A \), and \( r_{ij} = |r_i - r_j| \) is the distance between electron \( i \) and electron \( j \). The spin variables \( \omega_i \) do not appear in the Hamiltonian operator, but become important because of the Pauli exclusion principle, which states that the wave function \( \psi \) must be antisymmetric with respect to the interchange of any two electrons:

\[
\psi(x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_N) = -\psi(x_1, x_2, \ldots, x_j, \ldots, x_i, \ldots, x_N)
\]

(2.5)

Here, the notation \( x_i = (r_i, \omega_i) \) is used for the combination of spatial and spin variables of electron \( i \).

To summarize, the wave function \( \psi \) must satisfy two requirements: firstly, it must fulfill the Schrödinger equation (2.1), and secondly it must fulfill the antisymmetry condition (2.5).

### 2.2 Slater Determinants

As described in section 2.1, the wave function must satisfy two distinct conditions. It is then natural to try to satisfy one condition first, and then see what can be done about the other condition. In the approach described in the following, one first makes sure the antisymmetry condition is fulfilled.

Now, how can one construct a function \( \psi \) that satisfies the antisymmetry condition (2.5)? For a case with just two electrons, a possible choice of \( \psi \) is

\[
\psi(x_1, x_2) = \phi_1(x_1)\phi_2(x_2) - \phi_1(x_2)\phi_2(x_1)
\]

(2.6)

where \( \phi_i(x) \) are one-electron functions, i.e. functions of the space and spin variables of just one electron. More generally, an antisymmetric function \( \psi \) can be achieved by taking linear combinations of products of one-electron functions using the combinatorics of a matrix determinant:

\[
\psi(x_1, x_2, \ldots, x_N) = \begin{vmatrix}
\phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_N(x_1) \\
\phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_N(x_2) \\
\vdots & \vdots & \ddots & \vdots \\
\phi_1(x_N) & \phi_2(x_N) & \cdots & \phi_N(x_N)
\end{vmatrix}
\]

(2.7)

A function \( \psi \) constructed according to (2.7) is called a Slater determinant. A Slater determinant will always satisfy the antisymmetry condition (2.5), for any choice of the one-electron
functions $\phi_i(x)$. Next, we note that any linear combination of Slater determinants will also satisfy (2.5). For this reason, Slater determinants are very useful as basic building blocks when one tries to construct a solution to the Schrödinger equation (2.1) that at the same time satisfies the antisymmetry condition (2.5). This thesis focuses on the computationally simplest option: to use a single Slater determinant, varying the one-electron functions $\phi_i(x)$ to get as near as possible to a solution of (2.1). This is known as the Hartree–Fock approximation.

2.3 Basis Sets

When working with Slater determinants, everything is based on the one-electron functions $\phi_i(x)$, usually taken to be products of a spatial part and a spin part:

$$\phi_i(x) = \varphi_i(r)\gamma_i(\omega)$$  \hfill (2.8)

The spin part $\gamma(\omega)$ is completely described using just two functions $\alpha(\omega)$ and $\beta(\omega)$. The spatial part $\varphi_i(r)$, on the other hand, could in principle be any function of the three spatial coordinates. One therefore constructs $\varphi_i(r)$ as a linear combination of basis functions:

$$\varphi_i(r) = \sum_{j=1}^{n} c_{ij}b_i(r)$$  \hfill (2.9)

The set of basis functions $\{b_i(r)\}$ is often called a basis set. In principle, allowing $\varphi_i(r)$ complete flexibility to take any functional form would require an infinite number of basis functions. In practice, however, one has to settle for a finite number of basis functions. This thesis will focus on one particular type of basis functions: the Gaussian Type Linear Combination of Atomic Orbital (GT–LCAO) type of functions. The primitive functional form used is:

$$h_j(r) = e^{-\alpha_j(r-r_0)^2}$$  \hfill (2.10)

where $\alpha$ is referred to as the exponent of the primitive function, and $r_0$ is the spatial point around which the function is centered. Given primitive functions of the type (2.10), the basis functions $b_i(r)$ are constructed as linear combinations of the primitive functions, multiplied by some polynomials in the displacement coordinates $r - r_0$:

$$b(r) = P(r - r_0)\sum_{j=1}^{k} g_jh_j(r)$$  \hfill (2.11)

The center points $r_0$ of the basis functions are usually set to the nuclear positions and the polynomials $P$ are chosen as the real solid harmonics[5]. The exponents $\alpha$ and expansion
coefficients $g_j$ are chosen to allow as good a description as possible of the one-electron functions $\phi_i(x)$ using only a limited number of basis functions.

## 2.4 The Hartree–Fock Method

As was mentioned in Section 2.2, the use of a single Slater determinant as an approximate wave function is known as the *Hartree–Fock* approximation. There are a few different variants of this approximation depending on how one treats the spins of the electrons that make up the Slater determinant. The most widely used variant is called *Restricted Hartree–Fock* (RHF). In RHF, one makes use of the fact that electrons tend to stay in pairs; a “spin-up” electron together with a “spin-down” electron. A molecular system with an even number of electrons that are all paired in this way is called a *closed shell* system. Many systems studied with quantum chemistry methods are quite well described by a closed shell Slater determinant, and therefore the RHF method is commonly used. The computational details of the RHF method are given in Section 2.4.1. In Section 2.4.3, an alternative known as *Unrestricted Hartree–Fock* (UHF) is described, which is useful in cases when all electrons are not sitting in pairs.

### 2.4.1 Restricted Hartree–Fock

This section describes the RHF computational procedure for a molecular system with $N$ electrons and $n$ basis functions. The following three $n \times n$ matrices are important: the overlap matrix $S$, the density matrix $D$, and the Fock matrix $F$.

The overlap matrix is given by

$$ S_{pq} = \int b_p(r)b_q(r)dr $$

Before discussing the density matrix, consider the single Slater determinant used to approximate the wave function. In RHF, all electrons are paired; this means there are only $\frac{N}{2}$ different spatial functions for the $N$ electrons. These spatial functions are known as *orbitals*. Each orbital $\varphi_i$ is defined as a linear combination of basis functions, as stated in equation (2.9). The density matrix $D$ is related to the coefficients $c_{ij}$ of equation (2.9):

$$ D_{pq} = 2 \sum_{a=1}^{\frac{N}{2}} c_{ap}c_{aq} $$
Hartree–Fock Theory

Using the density matrix $D$, the total electron density $\rho(r)$ can be written as

$$\rho(r) = \sum_{pq} D_{pq} b_p(r) b_q(r)$$  \hfill (2.14)

hence the name “density matrix”. The Fock matrix $F$ is given by

$$F = H_{\text{core}} + G$$  \hfill (2.15)

where

$$H_{\text{core}} = T + V$$  \hfill (2.16)

with the kinetic energy term $T$ and the electron-nuclear attraction term $V$ given by

$$T_{pq} = -\frac{1}{2} \int b_p(r) \nabla^2 b_q(r) dr$$  \hfill (2.17)

$$V_{pq} = -\int b_p(r) \left[ \sum_{A=1}^{M} \frac{Z_A}{|r - r_A|} \right] b_q(r) dr$$  \hfill (2.18)

The matrix $G$ is the two-electron part of the Fock matrix, defined using the density matrix as

$$G_{pq} = \sum_{rs} D_{rs} [(pq|rs) - \frac{1}{2} (pr|sq)]$$  \hfill (2.19)

The numbers $(pq|rs)$ and $(pr|sq)$ are known as two-electron integrals:

$$(pq|rs) = \int \frac{b_p(r_1)b_q(r_1)b_r(r_2)b_s(r_2)}{|r_1 - r_2|} dr_1 dr_2$$  \hfill (2.20)

The three different contributions $T$, $V$, and $G$ to the Fock matrix $F$ each correspond to one of the terms in the Hamiltonian operator of equation (2.2). The matrices $T$ and $V$ depend only on basis set and nuclear positions; therefore $H_{\text{core}}$ is independent of $D$. Given a set of orbitals and a corresponding density matrix $D$, a Fock matrix $F$ can be computed. Then the energy corresponding to that particular set of orbitals can be computed as

$$E_{\text{RHF}} = \text{Tr}(DH_{\text{core}}) + \frac{1}{2}\text{Tr}(DG)$$  \hfill (2.21)

In a way, this is all we need to know in order to find the RHF electron density and corresponding energy; if we could try all possible choices of $D$, the one with lowest energy is the RHF solution. In practice, however, it is not possible to try all possible choices of $D$. Some kind of optimization scheme is needed. Section 2.4.2 describes such a scheme, known as the Self-Consistent Field procedure (SCF).
2.4.2 The Self-Consistent Field Procedure

The Self-Consistent Field procedure (SCF) is an iterative method for finding the density matrix that gives the lowest energy. It consists of two main steps, here labeled \((\mathbf{D} \rightarrow \mathbf{F})\) and \((\mathbf{F} \rightarrow \mathbf{D})\). In the \((\mathbf{D} \rightarrow \mathbf{F})\) step, a new Fock matrix \(\mathbf{F}\) is constructed from \(\mathbf{D}\) using equation (2.15). The elements of \(\mathbf{D}\) enters through equation (2.19). In the \((\mathbf{F} \rightarrow \mathbf{D})\) step, a new density matrix is constructed for a given Fock matrix, in the following way: first, the generalized eigenvalue problem

\[
\mathbf{FC} = \mathbf{SCA}
\] (2.22)

is solved, giving the matrix of eigenvectors \(\mathbf{C}\) and corresponding eigenvalues in the diagonal matrix \(\mathbf{\Lambda}\). Then the new density matrix is created as

\[
D_{pq} = 2 \sum_{a=1}^{\frac{N}{2}} C_{pa} C_{qa}
\] (2.23)

Note that equation (2.23) is essentially the same as equation (2.13): the eigenvectors of \(\mathbf{F}\) are interpreted as orbital coefficients.

Now that we know how to carry out the two main steps \((\mathbf{D} \rightarrow \mathbf{F})\) and \((\mathbf{F} \rightarrow \mathbf{D})\), the SCF procedure is really simple: first get a starting guess for \(\mathbf{D}\). This guess could in principle be anything, but it helps if it is close to the final solution. Then compute a new \(\mathbf{F}\) from \(\mathbf{D}\), a new \(\mathbf{D}\) from that \(\mathbf{F}\), and so on, until \(\mathbf{D}\) does not change any more. At that point one says that a self-consistent field has been found, hence the name SCF.

2.4.3 Unrestricted Hartree–Fock

In the Unrestricted Hartree–Fock (UHF) method, one does not force the electrons to stay in pairs. Instead, each electron has its own spatial orbital \(\varphi_i\); this means that there are now \(N\) different spatial orbitals \(\varphi_i\) instead of \(\frac{N}{2}\) as in the RHF case. We will use the symbols \(N_\alpha\) and \(N_\beta\) for the number of electrons of \(\alpha\) and \(\beta\) spin, with \(N_\alpha + N_\beta = N\). There are now separate density matrices \(\mathbf{D}^\alpha\) and \(\mathbf{D}^\beta\) for the \(\alpha\) and \(\beta\) electrons, respectively:

\[
D_{pq}^\alpha = \sum_{a=1}^{N_\alpha} C_{ap}^{\alpha} C_{aq}^{\alpha}
\] (2.24)

\[
D_{pq}^\beta = \sum_{a=1}^{N_\beta} C_{ap}^{\beta} C_{aq}^{\beta}
\] (2.25)

It is also useful to define the total density matrix \(\mathbf{D}^{\text{tot}}\) as

\[
\mathbf{D}^{\text{tot}} = \mathbf{D}^\alpha + \mathbf{D}^\beta
\] (2.26)
Now, UHF Fock matrices $F^\alpha$ and $F^\beta$ are defined as

$$F^\alpha = H_{\text{core}} + G^\alpha$$  \hspace{1cm} (2.27)

$$F^\beta = H_{\text{core}} + G^\beta$$  \hspace{1cm} (2.28)

where the two-electron parts $G^\alpha$ and $G^\beta$ are given by

$$G^\alpha_{pq} = \sum_{rs} \left[ D_{rs}^{\text{tot}} (pq|rs) - D_{rs}^\alpha (pr|sq) \right]$$  \hspace{1cm} (2.29)

$$G^\beta_{pq} = \sum_{rs} \left[ D_{rs}^{\text{tot}} (pq|rs) - D_{rs}^\beta (pr|sq) \right]$$  \hspace{1cm} (2.30)

The matrix $H_{\text{core}}$ is the same as in the RHF case. Given $D^\alpha$ and $D^\beta$ and the resulting Fock matrices $F^\alpha$ and $F^\beta$, the energy is given by

$$E_{\text{UHF}} = \text{Tr}(D^{\text{tot}} H_{\text{core}}) + \frac{1}{2} \text{Tr}(D^\alpha G^\alpha) + \frac{1}{2} \text{Tr}(D^\beta G^\beta)$$  \hspace{1cm} (2.31)

The RHF method can be seen as a special case of UHF: if the number of electrons is even, with $N_\alpha = N_\beta = \frac{N}{2}$, and each $\alpha$ orbital has an identical $\beta$ orbital corresponding to it, then $D^\alpha = D^\beta = \frac{1}{2} D$, and consequently $F^\alpha = F^\beta = F$. In this case the UHF energy expression (2.31) reduces to the RHF variant (2.21).

The self-consistent field procedure for UHF is analogous to the RHF case; the two main steps are now $(D^\alpha, D^\beta \rightarrow F^\alpha, F^\beta)$ and $(F^\alpha, F^\beta \rightarrow D^\alpha, D^\beta)$. The $(F^\alpha, F^\beta \rightarrow D^\alpha, D^\beta)$ step is carried out by solving generalized eigenvalue problems analogous to (2.22) for $\alpha$ and $\beta$ separately:

$$F^\alpha C^\alpha = SC^\alpha \Lambda^\alpha$$  \hspace{1cm} (2.32)

$$F^\beta C^\beta = SC^\beta \Lambda^\beta$$  \hspace{1cm} (2.33)

and then using the resulting eigenvectors $C^\alpha$ and $C^\beta$ as orbital coefficients when forming new density matrices according to (2.24) and (2.25).
Density Functional Theory

This chapter deals with a computationally cheap way of improving upon the Hartree–Fock approximation: the Kohn–Sham formulation of Density Functional Theory (KS–DFT) [3]. Density Functional Theory is based on the Hohenberg–Kohn theorems [4], which essentially state that the electronic density alone contains enough information to serve as the basic quantity describing a N-electron molecular system. When the Kohn–Sham formulation of DFT is used, the electronic density is assumed to have the form of a single Slater determinant, just as in Hartree–Fock theory. This approach allows for a good description of the kinetic energy of the electrons, something that is otherwise difficult to achieve using DFT.

The KS–DFT computational scheme is the same as the HF scheme, except for two things: there is a DFT exchange-correlation contribution to the energy ($E_{XC}$), and the Fock matrix $F$ is replaced by the Kohn–Sham matrix $F^{KS}$. Just as for HF, there are restricted and unrestricted versions of KS–DFT; in the unrestricted case there are two Kohn–Sham matrices $F^{KS}_\alpha$ and $F^{KS}_\beta$. In the following we will focus on the restricted case.

Before defining $F^{KS}$, we need to introduce the Coulomb matrix $J$ and the exchange matrix $K$ as the two parts of the Hartree–Fock two-electron matrix $G$:

$$G = J + K$$

(3.1)

where

$$J_{pq} = \sum_{rs} D_{rs}(pq|rs)$$

(3.2)

$$K_{pq} = -\frac{1}{2} \sum_{rs} D_{rs}(pr|sq)$$

(3.3)

With the definitions (3.2) and (3.3), (3.1) is equivalent with the previous definition of $G$. 

---

Chapter 3
Now, the DFT Kohn–Sham matrix $F_{KS}$ is defined as

$$ F_{KS} = H_{\text{core}} + J + \theta K + F_{XC} \quad (3.4) $$

where $\theta$ is an empirically determined parameter and $F_{XC}$ is the DFT exchange-correlation matrix. The KS–DFT energy is given by

$$ E = \text{Tr} (DH_{\text{core}}) + \frac{1}{2} \text{Tr} \left( D \left[ J + \theta K \right] \right) + E_{XC} \quad (3.5) $$

The matrix $F_{XC}$ and the exchange-correlation contribution to the energy ($E_{XC}$) are determined by the chosen exchange-correlation functional together with the density matrix $D$. The chosen functional also determines the parameter $\theta$ of equation (3.4). Here, we will assume that the Generalized Gradient Approximation (GGA) is used. In this approximation, the functional is specified as an analytic function of the density $\rho$ and the gradient of the density $\nabla \rho$. We will write this function as $F(\rho, \nabla \rho)$. Given the functional, as described by $F(\rho, \nabla \rho)$, $E_{XC}$ and $F_{XC}$ are given by

$$ E_{XC} = \int F(\rho(r), \nabla \rho(r)) \, dr \quad (3.6) $$

$$ F_{XC}^{pq} = \int \frac{\partial F}{\partial \rho} \bigg|_{\rho(r), \nabla \rho(r)} b_p(r)b_q(r) \, dr + \int \frac{\partial F}{\partial \nabla \rho} \bigg|_{\rho(r), \nabla \rho(r)} \nabla (b_p(r)b_q(r)) \, dr \quad (3.7) $$

The function $F(\rho, \nabla \rho)$ is given as an analytical expression, from which the functions $\frac{\partial F}{\partial \rho}$ and $\frac{\partial F}{\partial \nabla \rho}$ can also be derived analytically. The integrations (3.6) and (3.7) are then evaluated numerically. In the integration, the elements of $D$ enter through the evaluation of $\rho(r)$ and $\nabla \rho(r)$. Because the function $F(\rho, \nabla \rho)$ is usually quite complicated, no complete examples of such analytic expressions will be given here. In any case, it is an analytic expression containing terms like $\rho^{\frac{3}{2}}$ and $\log(\rho^{-\frac{1}{2}})$.

The different DFT exchange-correlation functionals can be divided into two main groups: pure DFT functionals and hybrid DFT functionals. The pure functionals have no $K$ contribution to $F_{KS}$; this corresponds to the parameter $\theta$ being zero. The hybrid functionals all have nonzero $\theta$. The most widely used hybrid functional, Becke’s three-parameter functional with Lee-Young-Parr correlation part (B3LYP), uses $\theta = 0.2$. 
Integral Evaluation in HF and DFT

In a HF/DFT calculation, much computer time is usually spent evaluating two-electron integrals of the type

\[(pq|rs) = \int \frac{b_p(r_1)b_q(r_1)b_r(r_2)b_s(r_2)}{|r_1 - r_2|} dr_1 dr_2\]  \hspace{1cm} (4.1)

The reason is that there are very many such integrals; if none of them are neglected, the number of integrals is \(n^4\), where \(n\) is the number of basis functions. There is some symmetry to take advantage of: from (4.1) it follows that \((pq|rs) = (qp|rs), (pq|rs) = (pq|sr), (pq|rs) = (rs|pq),\) etc., so that if all four indexes are different, there are eight integrals having the same value. This means that the number of unique integrals is about \(\frac{1}{8}n^4\). For large systems, it is possible to reduce the number of integrals that need to be computed even more, as long as one does not request infinite accuracy.

4.1 Primitive Gaussian Integrals

When the GT–LCAO type of basis functions are used, each two-electron integral \((pq|rs)\) is a sum of primitive integrals of the type

\[
\int \frac{\Psi_A(r_1)\Psi_B(r_2)}{|r_{12}|} dr_1 dr_2 \hspace{1cm} (4.2)
\]

where

\[
\Psi_A(r) = (x-x_A)^i_A(y-y_A)^j_A(z-z_A)^k_A e^{-\alpha_A(r-r_A)^2} \hspace{1cm} (4.3)
\]

\[
\Psi_B(r) = (x-x_B)^i_B(y-y_B)^j_B(z-z_B)^k_B e^{-\alpha_B(r-r_B)^2} \hspace{1cm} (4.4)
\]
Chapter 4

At this level, each primitive distribution $\Psi_A$ is determined by its exponent $\alpha_A$, its center coordinates $\mathbf{r}_A = (x_A, y_A, z_A)$, and the three integers $i_A, j_A$ and $k_A$.

Primitive integrals of the type (4.2) can be efficiently computed using the recurrence relations of the McMurchie–Davidson scheme. A thorough explanation of this scheme is found in the book by Helgaker et al [5].

4.2 Computational Scaling

When discussing integral evaluation in HF and DFT, one often talks about the computational scaling or complexity of different algorithms. The idea is that if one considers systems of different size but otherwise similar properties, using the same type of basis set (so that the number of basis functions grows linearly with the number of atoms), the performance of a particular algorithm will usually be rather predictable; the computational time $T$ is often assumed to be of the form

$$T(n) = \text{const} \times n^p$$

(4.5)

where $p$ is some number defining the scaling behavior. One then says that the algorithm has $O(n^p)$ scaling. For example, a naive Fock matrix construction routine, where all integrals $(pq|rs)$ are computed explicitly, will scale as $O(n^4)$. Here, the number of basis functions $n$ was chosen as the variable used to express the scaling behavior; one could equivalently use any other quantity that increases linearly with the system size, such as the number of atoms or the number of electrons.

4.3 Screening

In this section, the concept of screening is introduced in the context of integral evaluation. What is meant by screening in this context is something like “neglect of small contributions”: as was seen in chapter 2, the elements of the two-electron matrix $\mathbf{G}$ are sums in which each term is of the form

$$D_{ab}(pq|rs)$$

(4.6)

Now, if it is known that $|D_{ab}| < \epsilon_1$ and that $|(pq|rs)| < \epsilon_2$, then the contribution to $\mathbf{G}$ is surely smaller than $\epsilon_1\epsilon_2$. The idea of screening is to systematically estimate bounds on $|D_{ab}|$ and $|(pq|rs)|$ and use that information to skip evaluation of $|(pq|rs)|$ whenever it is known that the resulting contribution to $\mathbf{G}$ would be smaller than some threshold value $\tau$. 
4.4 Cauchy–Schwartz Screening

The Cauchy–Schwartz inequality is very useful for providing bounds to two-electron integrals:

$$|\langle pq|rs \rangle| \leq \sqrt{\langle pq|pq \rangle} \sqrt{\langle rs|rs \rangle}$$

(4.7)

The quantity $\sqrt{\langle pq|pq \rangle}$ associated with a pair of basis functions $b_p(r)$ and $b_q(r)$ is therefore of interest. Let $D_{\text{max}}$ be the largest absolute density matrix element, and let $\text{CSMAX}$ be the largest Cauchy-Schwartz factor $\sqrt{\langle pq|pq \rangle}$ among all pairs of basis functions. Then a limit for the largest contribution to $G$ that a particular basis function pair $pq$ can give is given by $D_{\text{max}} \text{CSMAX} \sqrt{\langle pq|pq \rangle}$. Therefore, once $D_{\text{max}}$ and $\text{CSMAX}$ are known, many basis function pairs can be neglected; it makes sense to create a list of all non-negligible basis function pairs. When the GT–LCAO type of basis functions are used, the number of non-negligible basis function pairs will scale as $O(n)$, thanks to locality of basis functions. Using such a list of non-negligible basis function pairs as a starting point, all non-negligible integrals $\langle pq|rs \rangle$ can easily be computed with $O(n^2)$ complexity.

4.5 The Coulomb and Exchange Matrices

In the previous section, it was discussed how the formal $O(n^4)$ scaling can be improved to $O(n^2)$ while still considering all contributions to $G$ on an equal footing. Further improvements can be made if one takes into account the way that density matrix elements are combined with the two-electron integrals. This differs between the Coulomb contribution $J$ and the exchange contribution $K$. Therefore, when attempting to improve the scaling beyond $O(n^2)$, it is helpful to use separate algorithms for the two contributions $J$ and $K$. In the case of $J$, one can exploit the fact that pairs of basis functions together with the corresponding density matrix elements can be seen as a charge distribution for which a simplified description can be generated. In the case of $K$, improved scaling can be achieved if density matrix elements decay with the distance between the corresponding basis functions, as is normally the case for insulating systems. Because Paper 1 of this thesis is devoted to the computation of the Coulomb matrix, some more details about Coulomb matrix construction will be given in chapter 5.
Chapter 5

Coulomb Matrix Construction

The Coulomb matrix is essential for all SCF–type quantum chemistry calculations: this matrix is needed in HF as well as in KS–DFT, for both pure and hybrid functionals. The Coulomb matrix $J$ is given by

$$J_{pq} = \sum_{rs} D_{rs} (pq|rs)$$

(5.1)

Taking into account the expression (2.20) for the two-electron integrals $(pq|rs)$, this can be rewritten as

$$J_{pq} = \int \frac{b_p(r_1)b_q(r_1)\sum_{rs} D_{rs} b_r(r_2)b_s(r_2)}{|r_1 - r_2|} dr_1 dr_2 = \int \frac{b_p(r_1)b_q(r_1)\rho(r_2)}{|r_1 - r_2|} dr_1 dr_2$$

(5.2)

where $\rho(r)$ is the electronic density

$$\rho(r) = \sum_{rs} D_{rs} b_r(r)b_s(r)$$

(5.3)

This means that the Coulomb matrix element $J_{pq}$ can be interpreted as follows: $J_{pq}$ is the Coulombic repulsion energy of an electron whose spatial distribution is given by $b_p(r)b_q(r)$, due to the total charge distribution $\rho(r)$ of all electrons.

To efficiently evaluate Coulomb matrix elements $J_{pq}$ it is useful to have two things precomputed. Firstly, a list of non-negligible basis function products $b_p(r)b_q(r)$, as described in section 4.4. Secondly, a description of the density $\rho(r)$ that allows one to evaluate repulsion from a group of charges that are far away, without considering interactions of individual pairs of charges. Such a description of $\rho(r)$ is used in the fast multipole method, discussed in the next section.
5.1 The Fast Multipole Method

The Fast Multipole Method (FMM) was originally developed for calculation of the Coulomb interaction between classical point charges\[6, 7\]. Later, much research has been devoted to the application of multipole methods in quantum chemistry \[8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\]. The basic idea of FMM is that a group of charges that are close to each other can be described by a multipole expansion, which includes enough information to accurately evaluate the Coulomb repulsion between that group of charges and some other charge that is well separated from the group. This allows Coulomb interaction to be calculated without taking into account each individual charge separately, thus reducing the computational effort. A hierarchy of groups of charges is used, so that larger groups of charges can be used at long distances. This gives improved scaling.

When FMM is applied to quantum chemistry using GT–LCAO basis functions, the “charges” are no longer point charges but instead Gaussian charge distributions arising from products of basis functions. Groups of such charge distributions are then described by multipole expansions.

So what is a “multipole expansion”? It is a list of numbers called “multipole moments”, each multipole moment giving the weight of one particular contribution to the expansion. The simplest part is the monopole contribution. Then there are three dipole contributions, one for each coordinate direction. Higher order contributions are more complicated multipoles; quadrupoles, octopoles, etc. The expansion must be truncated at some expansion order, chosen so that the desired accuracy is maintained.

\[
\begin{align*}
\text{Monopole} & : + \\
\text{Dipole} & : + - \\
\text{Quadrupole} & : - + +
\end{align*}
\]

Paper 1 of this thesis discusses various aspects of FMM applied to the computation of the Coulomb matrix.
Chapter 6

Storage and Manipulation of Matrices

This chapter contains a brief discussion about how matrices occurring in SCF calculations are stored and handled. A more thorough discussion on this subject can be found in Emanuel Rubensson’s licentiate thesis[21].

All matrices involved in Hartree–Fock or Kohn–Sham calculations need to be stored in a way suitable for the kind of matrix operations one needs to perform on them. Typical matrix operations needed are matrix–matrix multiply, addition of matrices, and computation of the trace of a matrix. Some more complicated operations are also of interest, such as diagonalization and inverse Cholesky decomposition.

When treating small systems, full matrix storage is the natural choice; a $n \times n$ matrix $A$ is then stored as a list of its elements, and a matrix element $A_{ij}$ is accessed as $A[i \times n + j]$ or as $A[j \times n + i]$. Full matrix storage is convenient because most matrix operations needed can be done using highly optimized linear algebra software libraries. For large systems, however, full matrix storage is not optimal because many matrix elements are negligibly small, so that they do not need to be stored. Much can then be gained by exploiting sparsity in the matrices.

6.1 Sparsity

The matrices occurring in HF or KS-DFT calculations with GT–LCAO basis functions are such that each matrix element $A_{ij}$ corresponds to a pair of basis functions $b_i(r)$ and $b_j(r)$. If the two basis functions are non–overlapping and separated by a large distance, the corresponding matrix element is usually of small magnitude. For this reason, it is advantageous to order the basis functions in such a way that basis functions that are near each other in space are also near each other in the ordering. Then, negligible matrix elements
will often be located together, so that a block–sparse matrix storage can be used. Block–sparse matrix storage is much preferable to element–wise sparse matrix storage because highly optimized linear algebra software libraries can be used at the level of submatrices.

6.2 Hierarchic Matrix Library

Different types of block–sparse matrix storage are possible. One commonly used option is to use the Compressed Sparse Row (CSR) format, in which only the non–zero submatrices on each row of the matrix are stored. The column indexes of the non–zero submatrices are kept in a separate vector. Another option is to use a hierarchic matrix data structure, in which the simple full–matrix storage is used at each level. This approach has certain advantages over the CSR format. Paper 2 of this thesis describes our implementation of a hierarchic matrix data structure, which we refer to as the Hierarchic Matrix Library (HML). As an example of what is gained compared to the CSR format, Paper 2 of this thesis includes a description of how the inverse Cholesky decomposition can be performed using HML.
The methods, algorithms, and computational approaches discussed in this thesis have been implemented in the quantum chemistry program \textit{ergo}, developed at the Theoretical Chemistry group at KTH in Stockholm. The program is written in the C and C++ programming languages. All matrices used by the program are computed directly in sparse form, and stored in HML format. The Coulomb matrix is computed using the fast multipole method, and the exchange matrix is computed using the screening techniques described in chapter 4, using an algorithm similar to the LinK algorithm\cite{19, 20}. This chapter gives some examples of what the \textit{ergo} program can be used for: section 7.1 includes some benchmark calculations to illustrate scaling behavior, and section 7.2 discusses the application of unrestricted KS–DFT calculations to Heisenberg exchange constants.

7.1 Benchmarks

In this section, benchmark Hartree–Fock calculations on water cluster systems are presented. The benchmark calculations were performed on an Intel Xeon EM64T 3.4 GHz CPU with 8 GB of memory.

The benchmark systems were generated from a snapshot of a molecular dynamics simulation at standard temperature and pressure. Test systems of different sizes were generated by including all water molecules within different radii from an arbitrarily chosen center point. Examples of what these benchmark systems look like are shown in figure 7.1.

Figure 7.2 shows timings for construction of the Coulomb matrix $J$ and the exchange matrix $K$ and memory usage, plotted against the number of basis functions $n$. The basis set used was 3-21G. Both timings and memory usage are heavily dependent on the chosen accuracy (threshold value $\tau$ in section 4.3); higher accuracy would mean longer computational time.
Figure 7.1: Water clusters of different sizes.

Figure 7.2: Timings and memory usage for exchange and Coulomb matrix construction for Hartree–Fock calculations on water clusters of varying size. Basis set: 3-21G.

and less favorable scaling behavior. In the calculations presented here, the threshold value $\tau = 10^{-5}$ was used, which corresponds to a rather low accuracy. The relationship between requested accuracy and computational time is discussed in Paper 1 of this thesis.

The largest test system here contained 1583 water molecules, corresponding to 20579 basis functions in the 3-21G basis set. If full matrix storage were used, a single $20579 \times 20579$ matrix would require 3.4 GB of memory, and the memory usage scaling would be $O(n^2)$. The memory usage plot in Figure 7.2 shows that the use of sparse matrices gives a much better scaling behavior, making it possible to treat even larger systems on a computer with 8 GB of memory.
7.2 Application to Computation of Heisenberg Exchange Constants

This section describes one possible application of unrestricted KS–DFT calculations: theoretical computation of Heisenberg exchange constants using the broken symmetry approach. Such calculations are presented in Paper 3 of this thesis.

Heisenberg exchange constants are of interest when magnetic properties of chemical compounds are studied. In the broken symmetry approach, the Heisenberg exchange constant $J_{AB}$ is computed as

$$J_{AB} = \text{const} \times (E_{BS} - E_{HS})$$ (7.1)

where $E_{BS}$ and $E_{HS}$ are the energies of the broken symmetry (BS) and high-spin (HS) electronic states, computed using unrestricted KS–DFT. The systems studied in Paper 3 are dinuclear Mn$^{IV}$ – Mn$^{IV}$ complexes, in which all electrons are expected to be paired except six. The six unpaired electrons are located three on each of the Mn centers. Two different spin configurations are considered: all six spins up (high-spin state HS), and three up / three down (broken symmetry state BS). The computational procedure is as follows: Two unrestricted KS–DFT calculations are performed, one for the HS state and one for the BS state. The resulting energies $E_{HS}$ and $E_{BS}$ are then used to compute the Heisenberg exchange constant $J_{AB}$ according to equation (7.1).

The calculations reported in Paper 3 are not perfect examples of applications of the computational methods for large systems that this thesis focuses on, because the systems studied are not large enough to make it necessary to use such methods. However, theoretical computation of Heisenberg exchange constants using the broken symmetry approach is of interest also for much larger systems, such as the so–called Mn$_{12}$ magnets [22].
Bibliography


