Chapter 1

FOUR-COMPONENT ELECTRONIC STRUCTURE METHODS FOR MOLECULES

T. Saue

CNRS, UMR 7551, Laboratoire de Chimie Quantique, 4 Rue Blaise Pascal, F-67000 Strasbourg, France

saue@quantix.u-strasbg.fr

L. Visscher

Department of Theoretical Chemistry, Faculty of Sciences, Vrije Universiteit Amsterdam, De Boelaan 1083, 1081 HV Amsterdam, The Netherlands

visscher@chem.vu.nl

Abstract In this chapter we consider the extension of 4-component relativistic methods from atomic to molecular systems, in particular the challenges arising from the introduction of the algebraic approximation. In order to analyze the variational stability of the relativistic many-electron Hamiltonian we derive a variational theory of QED in the semiclassical limit using the second quantization formalism and exponential parametrization. In QED the negative-energy orbitals are filled leading to a true minimization principle for the electronic ground state, whereas in the standard 4-component approach these orbitals are empty and treated as an orthogonal complement, thus leading to a minimax principle. We emphasize the non-uniqueness of the resulting no-pair Hamiltonian of the standard approach. 4-component methods allow the continuous update of the Hamiltonian and thereby complete relaxation of the electronic wave function. We also discuss more practical aspects of the implementation of 4-component relativistic methods. We carefully analyze their computational cost and conclude that the difference with respect to non-relativistic methods constitute a prefactor and not a difference in order. We furthermore discuss how computational cost may be reduced while staying at the 4-component level, e.g. by exploiting the atomic nature of the small component density.

Keywords: Relativistic Quantum Chemistry, Electronic Structure Theory, Computational Chemistry, 4-component methods, variational collapse, quantum electrodynamics, basis sets calculations, algebraic approximation, Dirac equation, molecules

Introduction

4-component electronic structure methods for atoms were introduced in chapter six. In the present chapter we consider the extension of these methods to molecular systems (see [116] and references therein for recent reviews). It is therefore natural in this chapter to focus on the additional complexities added when going from single atoms to polyatomic systems. An obvious difference is the lack of spherical symmetry. In the atomic case the high symmetry allows the separation of radial and angular degrees of freedom. The angular part can be solved completely by symmetry, particularly facilitated by the introduction of Racah algebra [38], whereas radial equations can be solved by finite difference methods. In molecular calculations one generally has to resort to the *algebraic approximation*, that is the use of finite basis set expansions. The first basis set calculations led to rather disastrous results (see [88] for references). Schwarz and Wechsel-Trakowski [89] identified two problems connected with 4-component relativistic calculations in the algebraic approach:

1 The coupling of the large and small components of the Dirac equation requires separate basis set expansions for each component.

2 The relativistic many-electron Hamiltonian is not bounded from below and, according to an argument given Brown and Ravenhall [10], gives only continuum solutions. This has been referred to as the "Brown-Ravenhall disease".

In the atomic case the first problem is avoided by the use of finite difference methods, and the second problem is solved by imposing the boundary conditions at $r = 0$ and $r \to \infty$ for bound solutions [39]. In practice the above two problems have also been solved in the algebraic approximations, and 4-component relativistic molecular calculations are thereby routinely carried out today. On the theoretical side things have apparently not been completely straightened out, as witnessed by a number of misunderstandings in the literature. After an initial overview of the relativistic many-electron Hamiltonian in section 1, we therefore give a variational formulation of QED in the semiclassical limit, that is with continuous electromagnetic fields, in section 2. At this level of theory a true minimization principle is assumed to exist [30], and this puts the discussion of the variational stability of the relativistic many-electron Hamiltonian on firm grounds.

The separate basis set expansion of the large and small components needed to solve the second problem above leads to increased computational cost. 4-component calculations are therefore considered limited to benchmark calculations, and more cost-efficient methods are sought by reducing the 4-component Hamiltonian to 1- or 2-component approximate forms. There is, however, another approach which consists of staying within the 4-component form and instead seek reduction of computational cost by reduction or elimination of intermediate quantities (e.g. two-electron integrals) appearing in the actual calculations. The current status and perspectives of this approach are discussed in section 3.

Notation and units: Unless otherwise stated, all formulas appearing in this chapter are in atomic units. In these units the electron mass m and the elementary charge e are both unity. We have, however, chosen to retain their symbols so that the reader can see how the fundamental characteristics of the electrons enter the equations. This is particular important for the electron charge since it provides the coupling to external fields as will be discussed in section 1.1.1. In this manner one can distinguish equations describing electrons from those describing its antiparticle, the positron. Operators generally come with operator hats, and vectors are written in bold characters. We shall also make extensive use of the Einstein summation convention or implicit summation, which means that a repeated index signals free summation over this index. A dot product is accordingly written $\mathbf{u}\cdot\mathbf{v} = u_i v_i$. This convention allows formulas to be written in a more compact manner, which is particularly useful in the section on QED.

1. The Hamiltonian

The general form of the many-electron Hamiltonian is

$$\hat{H} = \sum_i \hat{h}(i) + \frac{1}{2}\sum_{i\neq j}\hat{g}(i,j) + \hat{V}_{nn}; \quad \hat{V}_{nn} = \frac{1}{2}\sum_{A\neq B}\frac{Z_A Z_B e^2}{|\mathbf{R}_A - \mathbf{R}_B|} \tag{1.1}$$

where $\hat{h}(i)$ are one-electron operators, $\hat{g}(i,j)$ represents the two-electron interaction and $\hat{V}_{nn}$ is the classical repulsion of fixed nuclei. This form is valid in both the relativistic and the non-relativistic domain to the extent that three-particle and higher interactions can be ignored. For the development of the various methods of quantum chemistry it is rarely necessary to be more specific regarding the form of the many-electron Hamiltonian. This holds particularly true if the operator is recast in second quantized form, as will be discussed in section 1.3. This observation signals that one should carefully distinguish Hamiltonians from methods. At the 4-component relativistic molecular level of theory one now finds an almost complete set of methods analogous to the arsenal developed in the non-relativistic domain, such as Hartree-Fock (HF) (see e.g. [95, 61, 3, 64, 102, 23, 81, 73, 117] and references therein), second-order Møller-Plesset perturbation theory (MP2) [25, 57], Multi-Configuration Self Consistent Field (MCSCF) [47, 98], Restricted Active Space Configuration Interaction (RASCI) [101], Coupled Cluster (CC) [106, 109] and Density Functional Theory (DFT) (see e.g. [99, 60, 118, 84]). We therefore find it unnecessary to elaborate on the general principles of these methods since this information is available in a number of textbooks [96, 66, 76, 43]. We shall rather try to point out the specific adaptions and considerations needed to carry these methods over into the 4-component relativistic realm. We would also like to point out that in view of the distinction between Hamiltonians and methods emphasized above, we advise against the use of method names such as Dirac-Hartree-Fock and the shorter version Dirac-Fock (which is not very fair to Douglas Hartree who first suggested to Bertha Swirles the extension of the SCF method to the Dirac equation [33]) and instead recommend "the Hartree-Fock method at the 4-component relativistic level" or, more specifically, DC-HF (the Hartree-Fock method based on the Dirac-Coulomb Hamiltonian).

1.1 The one-electron part

1.1.1 The Dirac equation in an electromagnetic field.

The starting point for 4-component molecular calculations is Dirac's celebrated relativistic wave equation [18, 19]. In covariant form (that is the form in which the equation "looks the same" in all Lorentz frames [77]) it is given by

$$(i\gamma_\mu \partial_\mu - mc)\,\psi = 0; \qquad \left\{ \begin{array}{rcl} \partial_\mu & = & \left(\boldsymbol{\nabla}, -\dfrac{i}{c}\dfrac{\partial}{\partial t}\right) \\ \\ \gamma_\mu & = & \beta\,(\boldsymbol{\alpha}, iI_4) \end{array} \right. \tag{1.2}$$

in which appears the 4-gradient ∂_μ. Note that we do not distinguish between covariant and contravariant 4-vectors, as this is not necessary at the level of special relativity [77, p.6]. The quantity γ_μ is given in terms of the 4×4 identity matrix I_4 and the Dirac matrices

$$\boldsymbol{\alpha} = \left[\begin{array}{cc} 0_2 & \boldsymbol{\sigma} \\ \boldsymbol{\sigma} & 0_2 \end{array} \right] \quad \text{and} \quad \beta = \left[\begin{array}{cc} I_2 & 0_2 \\ 0_2 & -I_2 \end{array} \right] \tag{1.3}$$

and generates the C_4 Clifford algebra [72], the algebra of gamma matrices. The Pauli spin matrices

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}; \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{1.4}$$

are themselves generators of the C_3 Clifford algebra, that is the algebra of complex quaternions, a property that we shall make use of in section 1.1.2. Note also that the Dirac $\boldsymbol{\alpha}$ matrices can be expressed in terms of the 4×4 spin matrices

$$\boldsymbol{\alpha} = \zeta\boldsymbol{\Sigma}; \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma} & 0_2 \\ 0_2 & \boldsymbol{\sigma} \end{bmatrix}; \quad \zeta = \begin{bmatrix} 0_2 & I_2 \\ I_2 & 0_2 \end{bmatrix} \tag{1.5}$$

upon the introduction of the auxiliary matrix ζ[68]. In order to align the relativistic and non-relativistic energy scales one usually performs the substitution

$$\beta \quad \rightarrow \quad \beta' = \beta - I_4 \tag{1.6}$$

but we will for the moment retain the original form of the Dirac equation.

The free-particle Dirac equation in its more familiar form is obtained by multiplication by βc from the left

$$\left[\hat{h}_{D;0} - i\frac{\partial}{\partial t}\right]\psi = 0; \quad \hat{h}_{D;0} = \beta mc^2 + c\left(\boldsymbol{\alpha}\cdot\mathbf{p}\right) \tag{1.7}$$

External fields are introduced through *the principle of minimal electromagnetic interaction* [37]

$$p_\mu \rightarrow \pi_\mu = p_\mu - qA_\mu \tag{1.8}$$

in which appears 4-momentum $p_\mu = -i\partial_\mu$ and the 4-potential $A_\mu = (\mathbf{A}, \frac{i}{c}\phi)$. We are interested in electrons and therefore *choose* the charge $q = -e$. The Dirac equation then attains the form

$$\hat{D}\psi = \left[\hat{h}_{D;A_\mu} - i\frac{\partial}{\partial t}\right]\psi = 0; \quad \hat{h}_{D;A\mu} = \beta mc^2 + c\left(\boldsymbol{\alpha}\cdot\boldsymbol{\pi}\right) - e\phi \tag{1.9}$$

It is important to note that the minimal substitution (1.8) follows from the term

$$\mathcal{L}_{\text{int}} = j_\mu A_\mu \tag{1.10}$$

in the Lagrangian describing the interaction between the particle and the electromagnetic field as the product of the 4-current $j_\mu = (\mathbf{j}, ic\rho)$ and the 4-potential. This term was first proposed by Schwarzschild [90] to satisfy Lorentz covariance. It is employed in *ad hoc* basis in the non-relativistic domain, even though it does not represent the proper non-relativistic limit, which is electrostatics (see [78] and references therein). Comparing (1.9) and (1.10) we can identify the 4-current as

$$j_\mu = -ec\psi^\dagger\beta\gamma_\mu\psi; \quad \rho = -e\psi^\dagger I_4\psi; \quad \mathbf{j} = -ec\psi^\dagger\boldsymbol{\alpha}\psi \tag{1.11}$$

Classically the current density is given as charge multiplied by velocity and indeed one finds from Heisenbergs equation of motion that the velocity operator in the relativistic domain is $c\boldsymbol{\alpha}$. This may appear somewhat surprising, but may be interpreted as the electron oscillating at the speed of light c about its mean position ("Zitterbewegung" [86, 68]). Note that just as we may identify $-ec\boldsymbol{\alpha}$ as the operator of current density, we may identify $-eI_4$ as the operator of charge density. This identification will prove useful in section 1.2.

External electric $\mathbf{E}$ and magnetic $\mathbf{B}$ fields appear in the Hamiltonian only indirectly through the scalar ϕ and vector $\mathbf{A}$ potentials

$$\mathbf{E} = -\boldsymbol{\nabla}\phi - \frac{\partial\mathbf{A}}{\partial t}; \quad \mathbf{B} = \boldsymbol{\nabla}\times\mathbf{A} \tag{1.12}$$

This can be considered an advantage as the freedom of gauge [46] leaves room to choose the most convenient form of the potentials. For instance, a uniform electric field may be represented by $A_\mu = \left(0, -\frac{i}{c}\mathbf{E}\cdot\mathbf{r}\right)$ as well as $A_\mu = (-\mathbf{E}t, 0)$, but the former choice is usually preferred since it can be handled by time-independent theory [7]. Quantum chemistry, be it relativistic or not, is usually expressed in Coulomb gauge, that is through the transversality condition $\boldsymbol{\nabla}\cdot\mathbf{A} = 0$. From Maxwell's equations and the definitions (1.12) the field equations in the presence of a density ρ and a current $\mathbf{j}$ can then be expressed as [41]

$$\nabla^2\phi = -4\pi\rho \tag{1.13}$$

$$\left(\nabla^2\mathbf{A} - \alpha^2\frac{\partial^2\mathbf{A}}{\partial t^2}\right) - \boldsymbol{\nabla}\alpha^2\frac{\partial\phi}{\partial t} = -4\pi\alpha^2\mathbf{j} \tag{1.14}$$

where α is the fine-structure constant. The equation for the scalar potential is simply the Poisson equation with solution

$$\phi(\mathbf{r}, t) = \int \frac{\rho(\mathbf{r}', t)}{|\mathbf{r}-\mathbf{r}'|}d\tau' \tag{1.15}$$

At first sight, this result appears to be in contradiction with the theory of special relativity since the scalar potential is given by the *instantaneous* charge density. However, one must bear in mind that the scalar potential itself is not an observable. The effects of retardation, as well as magnetic interactions, enter through the vector potential. The Coulomb gauge bears its name because it singles out the *instantaneous* Coulomb interaction which is the proper non-relativistic limit of classical electrodynamics and which is the dominant interaction of chemistry. All retardation and magnetic interactions enter as higher-order terms in a perturbation expansion of the total interaction in terms of the fine-structure constant α [14]. For instance, to first order the interaction between two point charges q_1 and q_2 is in Coulomb gauge given by [42, 15, 46]

$$L_{\text{int}} = \frac{q_1 q_2}{r_{12}}\left[-1+\frac{1}{2c^2}\left\{(\mathbf{v}_1\cdot\mathbf{v}_2)+\frac{1}{r_{12}^2}(\mathbf{r}_{12}\cdot\mathbf{v}_1)(\mathbf{r}_{12}\cdot\mathbf{v}_2)\right\}\right] \quad (1.16)$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are their velocities. The first term can be identified as a charge-charge interaction, whereas the second term is a current-current interaction.

In molecular theory we shall employ $\hat{h}_{D;V}$, the Dirac operator in the *molecular field*. It corresponds to the introduction of the 4-potential

$$\phi(\mathbf{r}_i) = \sum_A \frac{Z_A e}{|\mathbf{r}_i - \mathbf{R}_A|}; \quad \mathbf{A}(\mathbf{r}_i) = 0. \quad (1.17)$$

where $Z_A e$ and $\mathbf{R}_A$ is charge and position, respectively, of nucleus A. The nuclei are accordingly treated as sources of external scalar potentials and nuclear spins are ignored. This "clamped nucleus" approximation is essentially the same as the one introduced by Born and Oppenheimer [8] in non-relativistic theory in which the main assumption is that electrons follow the slower movements of nuclei adiabatically. However, in the relativistic domain two consequences of this approximation should be kept in mind. The first is that the restriction to a particular frame in which the nuclei are at rest invalidates Lorentz covariance. In reality, though, nuclei will always move relative to each other so that no frame can be found in which all nuclei are stationary. Second, from (1.16) it is clear that all charge-charge interactions involving nuclei will be neglected in the "clamped nucleus" approximation.

The operators appearing in the Dirac equation (1.9) are 4 × 4 matrix operators and the corresponding wave function is therefore a 4-component vector function

$$\psi = \begin{bmatrix} \psi^L \\ \psi^S \end{bmatrix}; \quad \psi^X = \begin{bmatrix} \psi^{X\alpha} \\ \psi^{X\beta} \end{bmatrix}, \quad X = L, S. \quad (1.18)$$

The four degrees of freedom come from the fact that the Dirac equation describes both electrons and positrons and explicitly includes spin. For a given potential the positive-energy solutions correspond to electronic solutions, and in the non-relativistic limit (to be discussed in section 1.1.4) the lower two components of these solutions go to zero, whereas the upper two components reduce to a spin orbital in which the spatial part becomes a solution of the corresponding non-relativistic Schrödinger equation. The upper and lower two components are therefore generally referred to as the large and small components, respectively. For the same potential the negative-energy branch of the spectrum gives the positronic solutions indirectly, that is by charge conjugation (see section 1.1.3) and in the non-relativistic limit it is now the large components that go to zero. For this reason it has been suggested [56] that one should rather speak of upper and lower components than large and small ones. However, since focus in chemistry is on electronic solutions we will retain the common terminology. On the other hand one should not forget that the large components of positive-energy solutions of the Dirac equation are large by a factor c^{-1} only in an averaged sense; there may be regions in space where the small components dominate.

It is important to realize that the four degrees of freedom in the Dirac spinor (1.18), as compared to the scalar eigenfunctions of the Schrödinger equation, are not to be associated with specific components. It is a common mistake to associate the small components with the positronic degrees of freedom. For instance, one cannot simply delete the small components of a given electronic solution. One can also see from (1.18) that spin (α or β) is associated with two components and not one.

1.1.2 Time reversal symmetry.

In this and the following section we shall analyze two features of the Dirac equation (1.9) that are related to the four degrees of freedom in the relativistic wave equation. The first feature is charge conjugation symmetry, which reflects that the Dirac equation describes both electrons and positrons, but generally, as we shall see, not of the same system. The second feature is time reversal symmetry which to some extent can recover the lack of spin symmetry in the relativistic domain. It is interesting to note that both features involve a pairing of eigenfunctions of the Dirac equation. As we shall see, in the case of time reversal symmetry this pairing is broken by the introduction of an external vector potential. In the case of charge conjugation symmetry the pairing is broken by the introduction of *any* 4-potential.

Both features are represented by antiunitary operators $\widehat{\mathcal{K}}$ defined by

$$\widehat{\mathcal{K}}\left\langle \psi_1 \mid \psi_2 \right\rangle = \left\langle \widehat{\mathcal{K}}\psi_1 \mid \widehat{\mathcal{K}}\psi_2 \right\rangle = \left\langle \psi_2 \mid \psi_1 \right\rangle = \left\langle \psi_1 \mid \psi_2 \right\rangle^* \tag{1.19}$$

An example of an antiunitary operator is provided by the complex conjugation operator $\widehat{\mathcal{K}}_0$. In the non-relativistic domain this operator commutes with the Hamiltonian in the absence of external magnetic fields. It's effect on the non-relativistic Schrödinger equation in the molecular field (1.17)is thereby

$$\begin{aligned}\widehat{\mathcal{K}}_0\left[\hat{h}_{NR} - i\frac{\partial}{\partial t}\right]\widehat{\mathcal{K}}_0^{-1}\widehat{\mathcal{K}}_0\psi_{NR}(\mathbf{r},t) &= \left[\hat{h}_{NR} + i\frac{\partial}{\partial t}\right]\overline{\psi}_{NR}(\mathbf{r},t) \qquad (1.20)\\ &= \left[\hat{h}_{NR} - i\frac{\partial}{\partial t}\right]\overline{\psi}_{NR}(\mathbf{r},-t) = 0\end{aligned}$$

where $\overline{\psi}_{NR} = \widehat{\mathcal{K}}_0\psi_{NR}$. Through the substitution $t \rightarrow -t$ one recovers the form of the original equation. The complex conjugation operator is therefore identified as the time-reversal operator in non-relativistic systems [115]. Consider now the effect of $\widehat{\mathcal{K}}_0$ on the Dirac equation (1.9). Proceeding as in (1.20) we obtain

$$\left[\beta mc^2 + c\zeta\mathbf{\Sigma}^*\cdot(-\mathbf{p}+e\mathbf{A}) - e\phi + i\frac{\partial}{\partial t}\right]\widehat{\mathcal{K}}_0\psi(\mathbf{r},t) = 0 \qquad (1.21)$$

In this case the substitution $t \rightarrow -t$ is not enough to recover the original equation and so the complex conjugation operator can not be identified as the time-reversal operator in the relativistic domain. To find the proper form we may note that the product of a unitary and an antiunitary operator is an antiunitary operator. We may therefore take as our starting point the generic form

$$\widehat{\mathcal{K}} = U\widehat{\mathcal{K}}_0 \qquad (1.22)$$

where U is a 4×4 unitary matrix. Next, we note that the problem with (1.21) is that Σ_y changes sign under complex conjugation, whereas the other 4×4 spin matrices (1.5) do not. If one wants an antiunitary transformation under which the individual terms of the Dirac equation are either symmetric or antisymmetric the unitary operator U (1.22) must therefore contain Σ_y. We then finally arrive at the choice $U_T = -i\Sigma_y$ which gives the time reversal operator $\widehat{\mathcal{K}}$ for relativistic systems. Its application gives

$$\left[\underbrace{\left\{\beta mc^2 + c(\boldsymbol{\alpha}\cdot\mathbf{p}) - e\phi\right\}}_{(+)} - \underbrace{\left\{ec(\boldsymbol{\alpha}\cdot\mathbf{A}) - i\frac{\partial}{\partial t}\right\}}_{(-)}\right]\overline{\psi}(\mathbf{r},t) = 0 \qquad (1.23)$$

where $\overline{\psi} = \widehat{\mathcal{K}}\psi$ and one can see that in the absence of external vector potentials, one recovers the correct form of the Dirac equation through the substitution $t \rightarrow -t$.

Equation (1.23) shows that under time reversal the Dirac equation splits into a symmetric (+) and an antisymmetric (−) part. Further insight is obtained by reordering the Dirac equation

$$\begin{bmatrix} \psi^L \\ \psi^S \end{bmatrix} = \begin{bmatrix} \psi^{L\alpha} \\ \psi^{L\beta} \\ \psi^{S\alpha} \\ \psi^{S\beta} \end{bmatrix} \rightarrow \begin{bmatrix} \psi^{L\alpha} \\ \psi^{S\alpha} \\ \psi^{L\beta} \\ \psi^{S\beta} \end{bmatrix} = \begin{bmatrix} \psi^{\alpha} \\ \psi^{\beta} \end{bmatrix} \tag{1.24}$$

The time-symmetric part of the Dirac equation can then be written as

$$\mathcal{D}_+ = \begin{bmatrix} mc^2 - e\phi & -icd_z & 0 & -icd_- \\ -icd_z & -mc^2 - e\phi & -icd_- & 0 \\ 0 & -icd_+ & mc^2 - e\phi & icd_z \\ -icd_+ & 0 & -icd_z & -mc^2 - e\phi \end{bmatrix} \tag{1.25}$$

where we have introduced the notation

$$d_z = \frac{\partial}{\partial z}; \quad d_\pm = \frac{\partial}{\partial x} \pm i\frac{\partial}{\partial y}. \tag{1.26}$$

In the same manner we find that the time-antisymmetric part can be written as

$$\mathcal{D}_- = \begin{bmatrix} -i\frac{\partial}{\partial t} & ecA_z & 0 & ecA_- \\ ecA_z & -i\frac{\partial}{\partial t} & ecA_- & 0 \\ 0 & ecA_+ & -i\frac{\partial}{\partial t} & -ecA_z \\ ecA_+ & 0 & -ecA_z & -i\frac{\partial}{\partial t} \end{bmatrix} \tag{1.27}$$

where $A_\pm = A_x \pm iA_y$. The above two forms can be summarized by the matrix structure

$$\mathcal{D}_t = \begin{bmatrix} A & B \\ -tB^* & tA^* \end{bmatrix}; \quad t = \pm 1. \tag{1.28}$$

It is then a simple exercise to show that if

$$\begin{bmatrix} A & B \\ -tB^* & tA^* \end{bmatrix} \begin{bmatrix} \psi^{\alpha} \\ \psi^{\beta} \end{bmatrix} = \epsilon \begin{bmatrix} \psi^{\alpha} \\ \psi^{\beta} \end{bmatrix} = \epsilon\psi \tag{1.29}$$

then

$$\begin{bmatrix} A & B \\ -tB^* & tA^* \end{bmatrix} \begin{bmatrix} -\psi^{\beta *} \\ \psi^{\alpha *} \end{bmatrix} = t\epsilon \begin{bmatrix} -\psi^{\beta *} \\ \psi^{\alpha *} \end{bmatrix} = t\epsilon\overline{\psi}. \tag{1.30}$$

This shows that the time symmetric part of the Dirac equation has doubly degenerate eigensolutions. In the reordered equation (1.24) the time reversal operator has the form

$$\mathcal{K} = \begin{bmatrix} 0_2 & -I_2 \\ I_2 & 0_2 \end{bmatrix} \mathcal{K}_0 \tag{1.31}$$

and one therefore easily sees that the eigenvectors ψ and $\overline{\psi}$ are related by time reversal symmetry. They will therefore be referred to as *Kramers partners.* A orthonormal *Kramers paired basis* can be constructed from a spinor set $\{\psi_i\}$ and the corresponding Kramers partners $\left\{\overline{\psi}_i\right\}$. With the introduction of an external vector potential the double degeneracy is lifted. One may use degenerate perturbation theory to obtain the splitting. To zeroth order one obtains the eigenvalues via diagonalization of the 2×2 Hamiltonian matrix in the space of the two Kramers partners as

$$E_{\pm} = \epsilon \pm ec \left| \left\langle \psi \left| (\boldsymbol{\alpha} \cdot \mathbf{A}) \right| \overline{\psi} \right\rangle \right| .$$

The double degeneracy of the time symmetric Dirac equation suggests that a block diagonalization of the matrix operator is possible. This is indeed true, but at the expense of going from complex to quaternion algebra. First, let us recall the definition of a (real) quaternion number

$$q = \sum_{\Lambda=0}^{3} v_{\Lambda} e_{\Lambda} = v_0 + \check{\mathrm{i}} v_1 + \check{\mathrm{j}} v_2 + \check{\mathrm{k}} v_3; \quad v_{\Lambda} \in \mathcal{R} \tag{1.32}$$

The quaternion units $\check{\mathrm{i}}$, $\check{\mathrm{j}}$ and $\check{\mathrm{k}}$ are equivalent in the sense that they may be interchanged by cyclic permutation $\check{\mathrm{i}} \rightarrow \check{\mathrm{j}} \rightarrow \check{\mathrm{k}} \rightarrow \check{\mathrm{i}}$. It was observed already by Jordan [69] that the algebra of imaginary i times the Pauli spin matrices is that of the quaternion units, that is

$$\check{\mathrm{i}} \leftrightarrow i\sigma_z, \quad \check{\mathrm{j}} \leftrightarrow i\sigma_y, \quad \check{\mathrm{k}} \leftrightarrow i\sigma_x \tag{1.33}$$

The link between time reversal symmetry can then be established by noting that the time-symmetric form of (1.28) can be written in terms of Pauli spin matrices

$$\mathcal{D}_{+} = [I_2 \otimes A_R] + [(i\sigma_z) \otimes A_I] + [(i\sigma_y) \otimes B_R] + [(i\sigma_x) \otimes B_I] \tag{1.34}$$

clearly showing the quaternion structure of the matrix operator. The block diagonalization is achieved through the unitary quaternion transformation

$$U^{\dagger} \mathcal{D}_{+} U = \begin{bmatrix} A + B\check{\mathrm{j}} & 0 \\ 0 & -\check{\mathrm{k}} \left(A + B\check{\mathrm{j}} \right) \check{\mathrm{k}} \end{bmatrix} . \quad U = \frac{1}{\sqrt{2}} \begin{bmatrix} I_2 & \check{\mathrm{j}} I_2 \\ \check{\mathrm{j}} I_2 & I_2 \end{bmatrix} \tag{1.35}$$

For the upper block of the quaternion Dirac operator we find the structure

$${}^{Q}\mathcal{D}_{+} = \begin{bmatrix} mc^2 - e\phi & 0 \\ 0 & -mc^2 - e\phi \end{bmatrix} - \check{\mathrm{i}} c \begin{bmatrix} 0 & d_z \\ d_z & 0 \end{bmatrix} - \check{\mathrm{j}} c \begin{bmatrix} 0 & d_y \\ d_y & 0 \end{bmatrix} - \check{\mathrm{k}} c \begin{bmatrix} 0 & d_x \\ d_x & 0 \end{bmatrix} .$$

One observes that the rest mass term and the scalar potential enter the real part of the operator whereas the kinetic energy is represented by the quaternion imaginary part. The quaternion Dirac operator has no preferential axis, in contrast to the conventional form (1.25) where the Pauli spin matrices in their standard form correspond to spin quantization along the z-axis.

Time reversal symmetry provides only a partial compensation for the loss of spin symmetry in the relativistic domain. The coupling of the spin and spatial degrees of freedom by the spin-orbit interaction changes the structure of the equations relative to those of non-relativistic theory. The extra price to be paid due to this coupling can be directly related to the algebra needed to solve the Dirac equation by matrix diagonalization using a finite real basis expansion. In the general case the matrix representation of the symmetric Dirac operator can be block diagonalized through the quaternion unitary transformation (1.35) and one then needs to diagonalize the quaternion subblock $A + B\check{\jmath}$. In the absence of spin-orbit coupling, as in the non-relativistic limit, this subblock becomes real. However, in the relativistic domain symmetry reduction in terms of going from quaternion to complex or real algebra can be achieved by combining time reversal and spatial symmetry [82, 83]. One further thing to note is that Kramers partners do not map directly on to spin α and β orbitals. For instance the Kramers partner of a $p_1\alpha$ orbital is $p_{-1}\beta$ and not $p_1\beta$.

1.1.3 Charge conjugation symmetry.

The choice $U_C = i\beta\alpha_y$ in (1.22) gives the *charge conjugation operator* $\widehat{\mathcal{C}}$ (usually this term is reserved for the unitary part U_C only). Its application gives

$$\left[-\underbrace{\left\{\beta mc^2 + c\left(\boldsymbol{\alpha}\cdot\mathbf{p}\right) - i\frac{\partial}{\partial t}\right\}}_{(-)} + \underbrace{e\left\{c\left(\boldsymbol{\alpha}\cdot\mathbf{A}\right) - \phi\right\}}_{(+)}\right]\widehat{\mathcal{C}}\psi\left(\mathbf{r},t\right) = 0 \quad (1.36)$$

It can be seen that the term containing the 4-potential is symmetric (+) under charge conjugation, whereas the free-particle Dirac equation (1.7) is antisymmetric (−). It follows that if ψ is a solution of the Dirac equation for an electron (thus with charge $-e$) in the 4-potential A_μ, then $\widehat{\mathcal{C}}\psi$ is a solution of the Dirac equation for a particle with charge $+e$ in the *same* potential. In the stationary case the eigenvalues of ψ and $\widehat{\mathcal{C}}\psi$ have the same magnitude, but opposite sign. After an initial false identification with the proton [20], Dirac boldly predicted the existence of the positron [17], confirmed experimentally by Anderson in 1932 [2].

Note, however, that for a given 4-potential the Dirac operator describing an electron is *not* identical to the one describing a positron since the particles couple to the 4-potential through their charge (1.8). Only in the absence of external fields do the two equations become identical.

To see charge conjugation "at work" let us consider a stationary electronic solution ψ of the free-particle Dirac equation (1.7) with eigenvalue ϵ (positive continuum) and it's charge conjugated partner $\widehat{\mathcal{C}}\psi$ with eigenvalue $-\epsilon$ (negative continuum). We now introduce an external 4-potential through minimal coupling (1.8) with the electron charge $-e$, thus adding the term $\hat{V} = -ec\beta\gamma_\mu A_\mu$ to the free-particle Hamiltonian $\hat{h}_{D;0}$. The eigenvalues of the 2×2 Hamiltonian matrix in the space of the two charge conjugated partners is then

$$E_\pm = \Delta \pm \sqrt{\epsilon^2 + \Omega^2}; \quad \left\{ \begin{array}{lcl} \Delta & = & \left\langle \psi \left| \hat{V} \right| \psi \right\rangle \\ & & \\ \Omega & = & \left| \left\langle \psi \left| \hat{V} \right| \widehat{\mathcal{C}}\psi \right\rangle \right| \end{array} \right. . \tag{1.37}$$

The positive sign gives the electronic solution which can be expanded as

$$E_e = E_+ = \epsilon + \Delta + \frac{1}{2}\eta^1 + O(\eta^2) > \epsilon + \Delta; \quad \eta = \left(\frac{\Omega}{\epsilon}\right)^2. \tag{1.38}$$

The corresponding positronic solution can be obtained by introducing the 4-potential through minimal coupling with the positronic charge $+e$. Alternatively, it can be obtained through charge conjugation of the solution corresponding to E_- of (1.37). The corresponding energy is

$$E_p = -E_- = \epsilon - \Delta + \frac{1}{2}\eta + O(\eta^2) \tag{1.39}$$

One can easily see that if the electron is attracted by the 4-potential, that is Δ is negative, then the positron is repulsed, due to it's opposite charge. Unless the coupling term Ω dominates Δ the electronic solution descends below $+mc^2$ and the negative energy solution descends further down the negative continuum. However, this does not imply, as is often stated, that for systems with bound electrons all negative energy orbitals can be identified as having energies below $-mc^2$. Potentials are not always purely attractive or repulsive, e.g. in an anion one may observe negative-energy orbitals entering the gap as such solutions far from the nucleus see a negative and thus attractive potential.

1.1.4 Towards the non-relativistic limit. 4-component operators and wave functions are usually taken to imply relativistic theory.

In this section we shall, however, show that also non-relativistic calculations may be carried out at the 4-component level. In fact, as has been discussed extensively by Visscher and Saue [104], a multitude of Hamiltonians, with and without spin-orbit interaction included, may be formulated at the 4-component level of theory. These Hamiltonians are obtained from non-unitary transformations W of the Dirac equation [31]

$$\psi = W\psi' \quad \Rightarrow \quad W^{\dagger}\left[\hat{h}_{D;A_{\mu}} - i\frac{\partial}{\partial t}\right]W\psi' = 0, \tag{1.40}$$

possibly followed by the deletion of certain parts of the transformed operators. As starting point for these Hamiltonians we choose the time-independent Dirac equation in the molecular field (1.17)

$$\left[\begin{array}{cc} \hat{V} & c\,(\boldsymbol{\sigma}\cdot\mathbf{p}) \\ c\,(\boldsymbol{\sigma}\cdot\mathbf{p}) & \hat{V}-2mc^2 \end{array}\right]\left[\begin{array}{c} \psi^L \\ \psi^S \end{array}\right] = \left[\begin{array}{cc} I_2 & 0_2 \\ 0_2 & I_2 \end{array}\right]\left[\begin{array}{c} \psi^L \\ \psi^S \end{array}\right]E; \quad \hat{V} = -e\phi \tag{1.41}$$

where the metric is explicitly included since non-unitary transformations will be performed. We have also performed the substitution (1.6) to align relativistic and non-relativistic energy scales.

Consider first the non-relativistic limit, generally obtained as $c \to \infty$. The Dirac operator $\hat{h}_{D;V}$ has terms linear and even quadratic in the speed of light and one can therefore not apply this limit directly. Instead, one first performs the non-unitary transformation

$$\left[\begin{array}{c} \psi^L \\ \psi^S \end{array}\right] = \left[\begin{array}{cc} I_2 & 0_2 \\ 0_2 & c^{-1}I_2 \end{array}\right]\left[\begin{array}{c} \psi^L \\ \widetilde{\psi}^S \end{array}\right] \tag{1.42}$$

which gives the transformed equation

$$\left[\begin{array}{cc} \hat{V} & (\boldsymbol{\sigma}\cdot\mathbf{p}) \\ (\boldsymbol{\sigma}\cdot\mathbf{p}) & -2m\left(1-\dfrac{\hat{V}}{2mc^2}\right) \end{array}\right]\left[\begin{array}{c} \psi^L \\ \widetilde{\psi}^S \end{array}\right] = \left[\begin{array}{cc} I_2 & 0_2 \\ 0_2 & c^{-2}I_2 \end{array}\right]\left[\begin{array}{c} \psi^L \\ \widetilde{\psi}^S \end{array}\right]E \tag{1.43}$$

The speed of light now appears only in inverse powers and the proper non-relativistic limit may be obtained, but with the following restrictions [55]:

1 $|E| \ll c^2$, that is we restrict attention to the positive-energy solutions. This also means that a separate non-relativistic limit exists for the negative-energy solutions.

2 The potential ϕ in $\hat{V}$ must be non-singular. This does not hold for the potential of point charges, but does hold for the extended nuclei commonly used in 4-component relativistic calculations.

With the above restrictions we arrive in the non-relativistic limit at the Levy-Leblond equation [58]

$$\begin{bmatrix} \hat{V} & (\boldsymbol{\sigma}\cdot\mathbf{p}) \\ (\boldsymbol{\sigma}\cdot\mathbf{p}) & -2m \end{bmatrix}\begin{bmatrix} \psi^L \\ \widetilde{\psi}^S \end{bmatrix} = \begin{bmatrix} I_2 & 0_2 \\ 0_2 & 0_2 \end{bmatrix}\begin{bmatrix} \psi^L \\ \widetilde{\psi}^S \end{bmatrix} E \tag{1.44}$$

which forms the starting point of the direct perturbation theory of Kutzelnigg [55]. By elimination of the modified small component $\widetilde{\psi}^S$we obtain the more familiar non-relativistic Schrödinger equation in the molecular field using the identity $(\boldsymbol{\sigma}\cdot\mathbf{p})\,(\boldsymbol{\sigma}\cdot\mathbf{p}) = p^2$. However, in some cases the 4-component non-relativistic form (1.44) may be more advantageous than the conventional 1-component form, in particular upon the introduction of external magnetic fields [44].

Another route to the non-relativistic limit, with an interesting stop on the way, is provided by the non-unitary transformation

$$\begin{bmatrix} \psi^L \\ \psi^S \end{bmatrix} = \begin{bmatrix} I_2 & 0_2 \\ 0_2 & \dfrac{1}{2mc}(\boldsymbol{\sigma}\cdot\mathbf{p}) \end{bmatrix}\begin{bmatrix} \psi^L \\ \widetilde{\psi}^S \end{bmatrix} \tag{1.45}$$

which leads to what has been called *the modified Dirac equation* [24]

$$\begin{bmatrix} \hat{V} & \widehat{T} \\ \widehat{T} & \dfrac{1}{4m^2c^2}(\boldsymbol{\sigma}\cdot\mathbf{p})\,\hat{V}\,(\boldsymbol{\sigma}\cdot\mathbf{p}) - \widehat{T} \end{bmatrix}\begin{bmatrix} \psi^L \\ \widetilde{\psi}^S \end{bmatrix} = \begin{bmatrix} I_2 & 0_2 \\ 0_2 & \dfrac{1}{2mc^2}\widehat{T} \end{bmatrix}\begin{bmatrix} \psi^L \\ \widetilde{\psi}^S \end{bmatrix} E \tag{1.46}$$

where $\hat{T}$ is the kinetic energy operator. The speed of light again appears only in inverse powers which facilitates taking the non-relativistic limit. Another possibility is to use the identity

$$(\boldsymbol{\sigma}\cdot\mathbf{p})\,\hat{V}\,(\boldsymbol{\sigma}\cdot\mathbf{p}) = \mathbf{p}\hat{V}\cdot\mathbf{p} + i\sigma\cdot\left(\mathbf{p}\hat{V}\times\mathbf{p}\right) \tag{1.47}$$

By dropping the spin-dependent term on the right hand side one obtains the spin-free form of the modified Dirac equation. It allows scalar relativistic calculations within a 4-component framework, although the uniqueness of the distinction between scalar and spin-orbit relativistic effects has been questioned [103].

1.2 The two-electron part

The extension from one-electron systems to fully interacting many-electron systems is more complicated in the relativistic domain than in the non-relativistic one. From our discussion of Coulomb gauge in section 1.1.1 this can be understood since in the relativistic framework we have to add the effects of retardation and magnetic interaction to the non-relativistic limit represented by the instantaneous Coulomb interaction. The necessary correction terms can be obtained rigorously by invoking the full machinery of QED where the interaction is described in terms of the exchange of virtual photons. We shall however restrict ourselves to the semiclassical limit, that is continuous electromagnetic fields. To first order the electron-electron interaction is then given by the Coulomb-Breit interaction

$$\hat{g}(1,2) = \hat{g}^{\text{Coulomb}} + \hat{g}^{\text{Breit}} \tag{1.48}$$

The zeroth order term is the Coulomb term

$$\hat{g}^{\text{Coulomb}} = e^2 \frac{I_4 \cdot I_4}{r_{12}}. \tag{1.49}$$

We have inserted 4×4 times identity matrices I_4 to remind the reader that, although at first sight the Coulomb term appears to be identical to the non-relativistic electron-electron interaction, it's physical content is different. Upon reduction to non-relativistic form [13, 4, 68, 45, 91] through a Foldy-Wouthuysen transformation one finds that the relativistic operator contains for instance spin-own orbit interaction in addition to the instantaneous Coulomb interaction.

The first order term is the Breit interaction [9]

$$\hat{g}^{\text{Breit}} = -\frac{e^2}{2c^2 r_{12}} \left[\left(c\boldsymbol{\alpha}_1 \cdot c\boldsymbol{\alpha}_2\right) + \frac{1}{r_{12}^2} \left(c\boldsymbol{\alpha}_1 \cdot \mathbf{r}_{12}\right) \left(c\boldsymbol{\alpha}_2 \cdot \mathbf{r}_{12}\right) \right]. \tag{1.50}$$

We have written the Breit term in a slightly unusual form using explicitly the relativistic velocity operator $c\boldsymbol{\alpha}$. In this manner one easily recognizes (1.48) as the quantum mechanical analogue of the classical expression (1.16). It is important to note that although the Breit term can be derived as the low-frequency limit of the full electron-electron interaction as described by QED, it can equally well [68] be derived from the quantization of (1.16), which is essentially how it was derived by Breit. In this chapter *we will not go beyond* the semiclassical limit of QED, that is we will not consider quantization of the electromagnetic field since this would open up a whole new level of complexity. The Breit term can be rewritten in the form

$$\hat{g}^{\text{Breit}} = \hat{g}^{\text{Gaunt}} + \hat{g}^{\text{gauge}} \tag{1.51}$$

The first term is the Gaunt term

$$\hat{g}^{\text{Gaunt}} = -e^2 \frac{c\boldsymbol{\alpha}_1 \cdot c\boldsymbol{\alpha}_2}{c^2 r_{12}} \tag{1.52}$$

whereas the second term

$$\hat{g}^{\text{gauge}} = -e^2 \frac{(c\boldsymbol{\alpha}_1 \cdot \boldsymbol{\nabla}_1)(c\boldsymbol{\alpha}_2 \cdot \boldsymbol{\nabla}_2) r_{12}}{2c^2}, \tag{1.53}$$

where $\boldsymbol{\nabla}_1$and $\boldsymbol{\nabla}_2$ act *only* on r_{12} and not on the wave function, disappears in Lorentz (Feynman) gauge. By comparison with (1.11) one sees that the Coulomb and Gaunt terms represent charge-charge and current-current interactions, respectively. One can furthermore show by reduction to the non-relativistic form that the Gaunt term carries all spin-other orbit interaction [79].

The experience accumulated so far indicates that the Breit term has rather minor effects on the spectroscopic constants of molecular systems [112, 101, 74] so that the Coulomb term is usually sufficient. However, it is needed to get spin-orbit splittings correctly, in particular for light systems. From a practical point of view the Gaunt term is then preferred since it involves the same two-electron integrals as the Coulomb term (at least in a scalar basis) and provides the full spin-other orbit interaction.

1.3 Second quantization

Creation and annihilation operators were first introduced by Dirac to describe absorption and emission of photons [16], and the formalism was later extended to fermions by Jordan and Wigner [49, 51]. Although originally conceived to describe processes in which the particle number is not conserved, the second quantization formalism [34, 50] is widely used in molecular electronic-structure theory [43] that considers systems with a fixed particle number. This is so because it allows the derivation and expression of the methodology in an elegant manner, in particular when combined with exponential parametrization. In this section we will consider the second quantized form of the many-electron Hamiltonian at the 4-component relativistic level.

The second-quantized Hamiltonian is obtained from the first-quantized form (1.1) upon the introduction of field operators

$$\Psi(1) = \varphi_p(\mathbf{r}_1) a_p = \widetilde{\varphi}_p(\mathbf{r}_1) \widetilde{a}_p. \tag{1.54}$$

In the above expression the field operator has been expanded in two different orbital sets $\{\varphi_p\}$ and $\{\widetilde{\varphi}_p\}$ related by a unitary transformation which then defines two different sets of annihilation operators, $\{a_p\}$ and $\{\widetilde{a}_p\}$, related by the inverse transformation. In QED parlance the orbital set obtained as the solution of the free-particle Dirac equation corresponds to the "free" picture, whereas the one obtained with the molecular field (1.17) leads to the Furry [36] or bound-state interaction picture [93, 94]. The general form of the second-quantized Hamiltonian is

$$\hat{H} = \int \Psi^\dagger(1)\hat{h}(1)\Psi(1)d\tau_1 + \frac{1}{2}\int\int \Psi^\dagger(1)\Psi^\dagger(2)\hat{g}(1,2)\Psi(2)\Psi(1)d\tau_1 d\tau_2 \tag{1.55}$$

where $\hat{g}(1,2)$ is the chosen form of the electron-electron interaction. We will for the moment limit attention to the Coulomb (1.49) and Gaunt term (1.52). At the 4-component level the direct product $\Psi(2)\Psi(1)$ leads to an expansion in terms of vector functions with sixteen components. From this one sees that the two-electron operator $\hat{g}(1,2)$ should be considered a 16×16 matrix operator [68], so that the *dot* products appearing in the Coulomb and Gaunt terms should be replaced by *direct* products. Expanding the field operators in a specific orbital basis the second-quantized Dirac-Coulomb-(Gaunt) Hamiltonian is written as

$$\hat{H} = h_{pq}a_p^\dagger a_q + \frac{1}{4}\mathcal{L}_{pq,rs}a_p^\dagger a_r^\dagger a_s a_q \tag{1.56}$$

in which appear one-electron integrals over the Dirac operator in the molecular field

$$h_{pq} = \int \varphi^\dagger(\mathbf{r}_1)\hat{h}_{D;V}(1)\varphi(\mathbf{r}_1)d\tau_1 \tag{1.57}$$

and where we have introduced antisymmetrized two-electron integrals

$$\mathcal{L}_{pq,rs} = (pq\,|\,rs) - (ps\,|\,rq) = \mathcal{L}_{rs,pq} = -\mathcal{L}_{ps,rq}. \tag{1.58}$$

We write the two-electron integrals as

$$(pq\,|\,rs) = \int\int \frac{\Omega_{pq}(\mathbf{r}_1)\Omega_{rs}(\mathbf{r}_2)}{r_{12}}d\tau_1 d\tau_2 \tag{1.59}$$

where we introduce generalized overlap distributions (including charge)

$$\Omega_{pq}(\mathbf{r}) = \varphi_p^\dagger(\mathbf{r})S_\mu\varphi_q(\mathbf{r});\quad S_\mu = ie\beta\gamma_\mu = -e\left(-i\boldsymbol{\alpha}, I_4\right). \tag{1.60}$$

The time-like part corresponds to standard overlap distributions and contributes to the Coulomb term, whereas the space part corresponds to current distributions and contributes to the Gaunt term.

2. Variational procedures

There has been considerable discussion as to the variational stability of the Dirac-Coulomb-(Gaunt) Hamiltonian. The discussion originated from an argument put forward by Brown and Ravenhall [10]. They considered the interaction of two electrons described by this Hamiltonian. Consider for instance the helium atom. In a perturbational approach one can start with the non-interacting system and thus a Slater-determinant consisting of the 1s orbitals obtained as solutions of the hydrogenic atom with $Z = 2$. This Slater determinant is, however, degenerate with an in principle infinite number of Slater determinants containing one orbital from the positive continuum and one from the negative continuum. When the electron-electron interaction is turned on, these determinants will mix in and lead to what has been called a continuum dissolution, meaning that no bound state is obtained. Brown and Ravenhall suggested the use of projection operators to avoid the mixing in of these continuum determinants, a proposal that has been further explored by Sucher and others (see [54] for a review). In this section we will carefully consider variational approaches based on the Dirac-Coulomb-(Gaunt) Hamiltonian using the second quantization formalism introduced in the previous section. This will allow us to employ an exponential parametrization of the wave function in terms of orbital rotations and furthermore make the connection to QED where a true minimization principle is assumed to exist.

2.1 The standard approach

In the second quantization formalism [43] Slater determinants are replaced by occupation-number vectors in Fock space as the fundamental entity. The antisymmetry of the wave function under particle exchange follows from the anticommutation properties of the creation- and annihilation operators

$$[a_p, a_q]_+ = \left[a_p^\dagger, a_q^\dagger\right]_+ = 0; \quad \left[a_p^\dagger, a_q\right]_+ = \delta_{pq} \tag{1.61}$$

Starting with a given set of orthonormal orbitals $\{\varphi_p\}$ the Fock space equivalent of a Slater determinant is generated by acting with the corresponding creation operators on the vacuum state $|0\rangle$

$$|\Phi\rangle = a_1^\dagger a_2^\dagger \dots a_N^\dagger |0\rangle \tag{1.62}$$

where the vacuum state itself is defined by the relation

$$a_i |0\rangle = 0; \quad \forall a_i. \tag{1.63}$$

The occupation-number vector $|\Phi\rangle$ is an eigenfunction of the *number operator*

$$\hat{N} = a_p^\dagger a_p \tag{1.64}$$

with eigenvalue N corresponding to the number of occupied orbitals. The Fock space formalism goes beyond expansions in Slater determinants by allowing the coupling of vectors with different occupation numbers.

The variational *ansatz* at the Hartree-Fock level of theory is

$$\left|\widetilde{\Phi}\right\rangle = \exp\left[-\widehat{\kappa}\right]|\Phi\rangle \tag{1.65}$$

where $\exp\left[-\widehat{\kappa}\right]$ is an *orbital rotation operator* which is the negative exponential of the single replacement operator

$$\widehat{\kappa} = \kappa_{pq} a_p^\dagger a_q; \quad \kappa_{pq} = -\kappa_{qp}^*. \tag{1.66}$$

It is readily shown that $\widehat{\kappa}$ commutes with the number operator (1.64) which implies that the orbital rotation operator conserves particle number. The antihermiticity of the matrix containing the single replacement amplitudes $\{\kappa_{pq}\}$ guarantees the unitarity of the orbital rotation operator. Using (1.62) we can therefore rewrite the HF wavefunction as

$$\left|\widetilde{\Phi}\right\rangle = \widetilde{a}_1^\dagger \widetilde{a}_2^\dagger \dots \widetilde{a}_N^\dagger |0\rangle \tag{1.67}$$

in which appear transformed creation operators

$$\widetilde{a}_p^\dagger = \exp\left[-\widehat{\kappa}\right] a_p^\dagger \exp\left[\widehat{\kappa}\right] = a_q^\dagger U_{qp}; \quad U = \exp\left[-\kappa\right]. \tag{1.68}$$

To derive (1.67) we have used the result

$$\exp\left[-\widehat{\kappa}\right]|0\rangle = |0\rangle \tag{1.69}$$

which follows from the fact that annihilation operators appear on the right in the first and higher order terms of the exponential expansion of the orbital rotation operator.

The orbital set corresponding to the transformed creation operators (1.68) is given by

$$\widetilde{\varphi}_p = \varphi_q U_{qp}^* \tag{1.70}$$

and demonstrates that the orbital rotation operator provides a means of parameterizing the wave function in such a manner that orthonormality of the orbitals is assured without the need to introduce Lagrange multipliers. The orbital rotation operator furthermore allows the use of unconstrained optimization techniques by choosing a linearly independent set of elements of the κ matrix as variational parameters. Due to the antihermiticity of the κ matrix this is straightforwardly accomplished by choosing e.g. the upper triangle and the (imaginary) diagonal elements. The $\widehat{\kappa}$ operator can then be written as

$$\widehat{\kappa} = \sum_{p<q} \left[\kappa_{pq} a_p^\dagger a_q - \kappa_{pq}^* a_q^\dagger a_p\right]. \tag{1.71}$$

Table 1.1. Classes of variational parameters in the standard 4-component approach compared to QED.

	Generic	Standard	QED
A	κ_{ia}	$\kappa_{ia}^{++}, \kappa_{ia}^{+-}$	$\kappa_{ia}^{++}, \kappa_{ia}^{-+}$
B	κ_{ij}	κ_{ij}^{++}	$\kappa_{ij}^{++}, \kappa_{ij}^{+-}, \kappa_{ij}^{--}$
C	κ_{ab}	$\kappa_{ab}^{++}, \kappa_{ab}^{+-}, \kappa_{ab}^{--}$	κ_{ab}^{++}

We have discarded the diagonal elements since they only introduce a complex phase in the wave function and do not change the energy. Other redundant parameters should also be eliminated from the set of variational parameters. Within the chosen parametrization this is an easy task, as will be seen shortly. For this purpose it is convenient to introduce classes of orbitals and thereby the elements of the κ matrix. We shall use indices i, j, k and l to identify occupied orbitals. We furthermore employ indices a, b, c and d for virtual orbitals and p, q, r and s as general indices. When needed we shall use superscripts $+$ and $-$ to distinguish positive- and negative-energy orbitals. This of course assumes that the actual potential allows this distinction. We can then identify three classes of variational parameters as given in table 1.1 (the column marked QED should be ignored for the moment). It should be noted that the direct use of the elements of the U matrix (1.68) as variational parameters is less advantageous since they are related in a non-linear manner by the unitary condition $U^\dagger U = I$. It is furthermore difficult to identify redundant parameters in this alternative parametrization.

We now gather the chosen, generally complex variational parameters in the vector

$$\mathbf{K} = \left[\begin{array}{c} \kappa \\ \kappa^* \end{array} \right] \tag{1.72}$$

and consider a Taylor expansion of the energy in terms of these. Let $\mathbf{K} = 0$ correspond to the current expansion point, that is the reference determinant (1.62) and the corresponding orbital set. We then have

$$E\left(\mathbf{K}\right) = E^{[0]} + \mathbf{K}^\dagger \mathbf{E}^{[1]} + \frac{1}{2}\mathbf{K}^\dagger E^{[2]}\mathbf{K} + O(\kappa^3). \tag{1.73}$$

The various terms of the Taylor expansion can be easily found by comparing with a Baker-Campbell-Hausdorff (BCH) expansion of the energy expectation value

$$\begin{aligned} E = \left\langle \tilde{\Phi} \left| \hat{H} \right| \tilde{\Phi} \right\rangle &= \left\langle \Phi \left| \exp\left[\hat{\kappa}\right] \hat{H} \exp\left[-\hat{\kappa}\right] \right| \Phi \right\rangle \qquad (1.74) \\ &= \left\langle \Phi \left| \hat{H} \right| \Phi \right\rangle + \left\langle \Phi \left| \left[\hat{\kappa}, \hat{H}\right] \right| \Phi \right\rangle + \frac{1}{2} \left\langle \Phi \left| \left[\hat{\kappa}, \left[\hat{\kappa}, \hat{H}\right]\right] \right| \Phi \right\rangle + O(\kappa^3) \end{aligned}$$

The zeroth order term is the expectation value of the energy with respect to the current Slater determinant

$$E^{[0]} = h_{ii} + \frac{1}{2}\Gamma_{ii}\left[\varphi_j\right]; \quad \Gamma_{pq}\left[\varphi_i\right] = \mathcal{L}_{pqii} \tag{1.75}$$

where we have introduced the mean-field potential $\Gamma\left[\varphi_i\right]$ involving summation over the orbital set $\{\varphi_i\}$. It will turn out to be useful to also consider the explicit form of the energy expression in the algebraic approximation. Consider the expansion of the current orbital set in some set of suitable basis functions (to be discussed in section 3.1) $\varphi_p = \chi_\mu c_{\mu p}$ where Greek indices are used for the AO-basis. The energy (1.75) can then be re-expressed as

$$E = D_{\mu\nu}h_{\nu\mu} + \frac{1}{2}D_{\mu\nu}\mathcal{L}_{\nu\mu\lambda\kappa}D_{\kappa\lambda}; \quad D_{\lambda\kappa} = c_{\lambda i}c^*_{\kappa i} \tag{1.76}$$

in which appears the AO-density matrix D.

The first-order term can be expressed as

$$\begin{aligned}\mathbf{K}^\dagger\mathbf{E}^{[1]} &= \sum_{p<q}\left\{-\kappa^*_{pq}\left\langle\Phi\left|\left[a^\dagger_q a_p, \hat{H}\right]\right|\Phi\right\rangle + \kappa_{pq}\left\langle\Phi\left|\left[a^\dagger_p a_q, \hat{H}\right]\right|\Phi\right\rangle\right\} \\ &= \left(F_{ip}\kappa^*_{pi} - \kappa^*_{ip}F_{pi}\right) + \text{c.c.}\end{aligned} \tag{1.77}$$

in which appears the Fock matrix

$$F_{pq} = h_{pq} + \Gamma_{pq}\left[\varphi_i\right] = h_{pq} + \mathcal{L}_{pq\kappa\lambda}D_{\lambda\kappa}. \tag{1.78}$$

From (1.77) one can easily see that derivatives of the energy at the current expansion point ($\mathbf{K} = 0$) with respect to parameter classes B and C (see table 1.1) are identically zero. Since this will hold for *any* expansion point the two parameter classes can be eliminated as redundant. The only non-redundant parameters are rotations between occupied (+) and virtual (±) orbitals, that is parameter class A, and the $\widehat{\kappa}$ can accordingly be written in terms of non-redundant parameters as

$$\widehat{\kappa} = \kappa_{ia}a^\dagger_a a_i - \kappa^*_{ai}a^\dagger_i a_a \tag{1.79}$$

Note that in order to ensure unitarity of the orbital rotation operator the $\widehat{\kappa}$ operator contains de-excitation operators $\left\{a^\dagger_i a_a\right\}$ in addition to excitation operators $\left\{a^\dagger_a a_i\right\}$. This can be contrasted with the exponential operator in a coupled-cluster expansion which is entirely based on excitation operators. In terms of non-redundant parameters the gradient vector can be written as

$$\mathbf{E}^{[1]} = \left[\begin{array}{c} \mathbf{g} \\ \mathbf{g}^* \end{array} \right]; \quad g_{ai} = \left.\frac{\partial E}{\partial \kappa^*_{ai}}\right|_{K=0} = \left\langle \Phi \left| \left[-a^\dagger_i a_a, \hat{H} \right] \right| \Phi \right\rangle = -F_{ai} \tag{1.80}$$

The current expansion point corresponds to a stationary value of the energy when all the elements of the gradient vector are zero. From the above expression we can see that this corresponds to the occupied-virtual blocks of the Fock matrix being zero. This can be accomplished by diagonalization of the Fock matrix (until convergence of the SCF procedure) and then leads to the canonical Hartree-Fock orbitals with associated orbital energies $\{\epsilon_p\}$. Alternatively one can proceed by second-order methods such as Newton-Raphson. These are generally more expensive since they involve calculation of the Hessian matrix $E^{[2]}$, but are also more robust. The second-order methods lead to a set of orthonormal orbitals not necessarily equal to, but related through a unitary transformation, to the canonical HF orbitals.

To characterize the stationary points we must consider the eigenvalues of the Hessian matrix. It has been defined in such a manner that it is Hermitian and diagonal dominant. It has the structure

$$E^{[2]} = \left[\begin{array}{cc} A & B \\ B^* & A^* \end{array} \right]; \quad \begin{array}{rcl} A_{ai,bj} & = & \delta_{ij} F_{ab} - \delta_{ab} F^*_{ij} - \mathcal{L}_{abij} \\ B_{ai,bj} & = & -\mathcal{L}_{aj,bi} \end{array} \tag{1.81}$$

Consider first a non-interacting system, i.e. $\hat{g}(i,j) = 0$. Then all two-electron terms disappear. In canonical orbitals the Hessian furthermore becomes diagonal with diagonal elements

$$A_{ai,ai} = \epsilon_a - \epsilon_i \left\{ \begin{array}{ll} > 0 & \text{for} \quad \left\{ \kappa^{++}_{ia} \right\} \\ < 0 & \text{for} \quad \left\{ \kappa^{+-}_{ia} \right\} \end{array} \right. \quad \text{(no summation !)} \tag{1.82}$$

One sees that the stationary point is a *minimum* with respect to rotations $\left\{ \kappa^{++}_{ia} \right\}$ between occupied (+) and positive-energy virtual orbitals, but a *maximum* with respect to rotations $\left\{ \kappa^{+-}_{ia} \right\}$ between occupied and (virtual) negative-energy orbitals.

For an interacting system the diagonal elements of the Hessian correspond to energy differences between singly excited determinants $\Phi_{i\rightarrow a}$ and the reference determinant

$$A_{ai,ai} = E\left(\Phi_{i\rightarrow a}\right) - E\left(\Phi\right) = \epsilon_a - \epsilon_i - \mathcal{L}_{aa,ii} \quad \text{(no summation !)} \tag{1.83}$$

Since the excited determinant is calculated in the orbitals optimized for the reference determinant the diagonal elements of the Hessian constitute a rather poor approximation to excitation energies that goes under the name of the single-transition approximation [21]. A better approximation of the energies needed to reach the manifold of singly excited states with respect to the ground state is obtained at the RPA (random-phase approximation) level (coupled Hartree-Fock) where the single excitation energies from a closed-shell ground state are the eigenvalues of the Hessian matrix. In view of this, it is clearly reasonable to claim that the minimax principle (1.82) applies to interacting systems as well.

The minimax principle for non-interacting systems is not identical to the minimax principle for the Dirac equation proposed by Talman [97] which states that the expectation value of the Dirac operator is a minimum with respect to variations in the large components and a maximum with respect to variations in the small components

$$E = \min_{\psi^L}\left[\max_{\psi^S}\frac{\left\langle \psi \left| \hat{h}_D \right| \psi \right\rangle}{\langle \psi \mid \psi \rangle}\right] \tag{1.84}$$

This minimax principle has been justly criticized by Kutzelnigg [56]. The problem is that the coupling between the large and the small components of the Dirac spinor means that some variations of the large component may lead to energies below the exact energy. We shall discuss this further in section 3.1, but one should note that the minimax principle (1.82) is susceptible to the same problem unless the chosen basis set expansion allows the proper coupling of large and small components. There is a significant difference, though. The Talman minimax principle operationally implies the *separate* variation of the large and small components. However, due to their coupling variations must be done in a *concerted* manner and this is achieved by the minimax defined in (1.82).

2.2 Towards QED

The standard approach to 4-component relativistic molecular calculations is not quantum electrodynamics and does not even constitute its semiclassical limit, that is the level of theory in which the electromagnetic field is not quantized. In this section we shall explore a variational description of the proper semiclassical limit of QED. This section is very much inspired by two rarely cited papers by Chaix and Iracane [12] on the transition from quantum electrodynamics to mean-field theory, but whereas these authors employ the elements of the U matrix (1.68) as variational parameters, leading to what they call the Bogoliubov-Dirac-

Fock formalism, we shall employ the more advantageous parametrization in terms of elements of the κ matrix.

In the previous section we have seen that the standard approach to 4-component relativistic molecular calculations leads to a minimax principle at the Hartree-Fock level. This means that the bound electronic states at this level of theory are excited states and may therefore decay through an infinite succession of transition through various states of the negative-energy continuum, thereby causing a radiative catastrophe [40]. To avoid this difficulty Dirac postulated that the negative-energy orbitals are all occupied so that transitions into the negative-energy continuum, "the Dirac sea", is forbidden by the Pauli exclusion principle. On the other hand, the excitation of an electron from the Dirac sea to a positive-energy orbital, requiring an energy on the order of $2mc^2$, leaves a hole with opposite charge that in time was identified as the positron.

The first step towards proper QED is to introduce a *particle-hole formalism.* The field operators (1.54) are rewritten as

$$\Psi = \varphi_p^+ b_p + \varphi_p^- d_p^\dagger \tag{1.85}$$

in which appears *electron* annihilation operators b_p associated with the positive-energy orbitals φ_p^+ and *positron* creation operators $d_p^\dagger$ describing the creation of positrons whose orbitals are obtained by charge conjugating the associated negative-energy orbitals φ_p^-. This distinction leads to a more involved expression for the second quantized form of the Hamiltonian

$$\begin{aligned}
\hat{H} &= h_{pq}^{++} b_p^\dagger b_q + h_{pq}^{+-} b_p^\dagger d_q^\dagger + h_{pq}^{-+} d_p b_q + h_{pq}^{--} d_p d_q^\dagger \qquad (1.86) \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{+++++} b_p^\dagger b_r^\dagger b_s b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{+-++} b_p^\dagger b_r^\dagger b_s d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{+++-} b_p^\dagger b_r^\dagger d_s^\dagger b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{+-+-} b_p^\dagger b_r^\dagger d_s^\dagger d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{++-+} b_p^\dagger d_r b_s b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{+--+} b_p^\dagger d_r b_s d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{++--} b_p^\dagger d_r d_s^\dagger b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{+---} b_p^\dagger d_r d_s^\dagger d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{-+++} d_p b_r^\dagger b_s b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{--++} d_p b_r^\dagger b_s d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{-++-} d_p b_r^\dagger d_s^\dagger b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{--+-} d_p b_r^\dagger d_s^\dagger d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{-+-+} d_p d_r b_s b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{---+} d_p d_r b_s d_q^\dagger \\
&+ \frac{1}{4}\mathcal{L}_{pqrs}^{-+--} d_p d_r d_s^\dagger b_q + \frac{1}{4}\mathcal{L}_{pqrs}^{----} d_p d_r d_s^\dagger d_q^\dagger
\end{aligned}$$

By inspection one immediately sees that the QED Hamiltonian couples occupation-number vectors with different particle number. However, we will demonstrate later that charge is conserved.

As in the previous section we consider the description of bound electronic states in terms of a single Slater determinant or, equivalently, in terms of a single occupation-number vector in Fock space

$$\left|\Phi\right\rangle = b_1^\dagger b_2^\dagger \ldots b_n^\dagger \left|0\right\rangle . \tag{1.87}$$

The vacuum state $\left|0\right\rangle$ is in this formalism defined by

$$\left(b_p \left|0\right\rangle = 0, \quad \forall b_p\right) \quad \text{and} \quad \left(d_p \left|0\right\rangle = 0, \quad \forall d_p\right). \tag{1.88}$$

In analogy with (1.65) we now consider the following variational *ansatz*

$$\left|\widetilde{\Phi}\right\rangle = \exp\left[-\widehat{\kappa}\right]\left|\Phi\right\rangle \tag{1.89}$$

The $\widehat{\kappa}$ operator appearing in the orbital rotation operator now has the form

$$\widehat{\kappa} = \underbrace{\kappa_{pq}^{++} b_p^\dagger b_q}_{\widehat{\kappa}^{++}} + \underbrace{\kappa_{pq}^{+-} b_p^\dagger d_q^\dagger}_{\widehat{\kappa}^{+-}} + \underbrace{\kappa_{pq}^{-+} d_p b_q}_{\widehat{\kappa}^{-+}} + \underbrace{\kappa_{pq}^{--} d_p d_q}_{\widehat{\kappa}^{--}} \tag{1.90}$$

We may next introduce number operators $\hat{N}^e$ and $\hat{N}^p$ for electrons and positrons, respectively

$$\hat{N}^e = b_p^\dagger b_p; \quad \hat{N}^p = d_p^\dagger d_p. \tag{1.91}$$

One finds that the $\hat{\kappa}$ operator commutes with neither number operator

$$\left[\hat{\kappa}, \hat{N}^e\right] = \left[\hat{\kappa}, \hat{N}^p\right] = \hat{\kappa}^{-+} - \hat{\kappa}^{+-}, \tag{1.92}$$

but with the linear combination

$$\hat{Q} = e\left(\hat{N}^p - \hat{N}^e\right) \tag{1.93}$$

which can be identified as the *charge* operator. From this we can conclude that the orbital rotation operator of QED conserves charge but *not* the particle number. The charge operator furthermore commutes with the QED Hamiltonian (1.86), thus demonstrating that the latter conserves charge as well.

Using the unitarity of the orbital rotation operator we may now rewrite the HF ansatz as

$$\left|\widetilde{\Phi}\right\rangle = \widetilde{b}_1^\dagger \widetilde{b}_2^\dagger \ldots \widetilde{b}_n^\dagger \left|\widetilde{0}\right\rangle \tag{1.94}$$

where appear the transformed creation operators

$$\widetilde{b}_p^\dagger = \exp\left[-\widehat{\kappa}\right] b_p^\dagger \exp\left[\widehat{\kappa}\right] = b_q^\dagger U_{qp}; \quad U = \exp\left[-\kappa\right]. \tag{1.95}$$

Note that due to charge conservation we can identify the transformed creation operators with dressed electrons. Equation (1.94) contains not only dressed electrons, but also a *dressed vacuum*

$$\left|\tilde{0}\right\rangle = \exp\left[-\hat{\kappa}\right]|0\rangle = \left\{1 - \kappa_{pq}^{--} d_p d_q - \kappa_{pp}^{--} + O(\kappa^2)\right\}|0\rangle \neq |0\rangle\,. \quad (1.96)$$

This relation may be compared with (1.69) and shows the effect of *vacuum polarization* which is not present in the standard approach described in the previous section.

If we evaluate the expectation value of the Hamiltonian (1.86) with respect to the reference determinant (1.87) we obtain the expression (1.75), but with *all* the negative-energy orbitals included amongst the occupied orbitals, thus leading to an infinite negative energy. In order to avoid working with infinite energies renormalization procedures are introduced in QED. In the present case the infinite negative energy is avoided by writing the Hamiltonian on *reordered* form, that is all creation operators are shifted to the left and all annihilation operators are shifted to the right as if they anticommuted. Using the notation of Chaix and Iracane [12] we write the reordered QED Hamiltonian as

$$\begin{aligned} \hat{H} &= \int N\left[\Psi^\dagger(1)\hat{h}(1)\Psi(1)d\tau_1\right] \qquad (1.97)\\ &+ \frac{1}{2}\int\int N\left[\Psi^\dagger(1)\Psi^\dagger(2)\hat{g}(1,2)\Psi(2)\Psi(1)\right] d\tau_1 d\tau_2 \end{aligned}$$

In the orbital set which diagonalizes $\hat{h}$, the one-electron part, can be written as

$$\begin{aligned} \hat{h} &= h_{pq}^{++} b_p^\dagger b_q + h_{pq}^{+-} b_p^\dagger d_q^\dagger + h_{pq}^{-+} d_p b_q - h_{pq}^{--} d_q^\dagger d_p \\ &= \epsilon_p^+ b_p^\dagger b_p + \left(-\epsilon_p^-\right) d_p^\dagger d_p \qquad (1.98) \end{aligned}$$

One may observe that on this reordered form electrons and positrons both appear with *positive* energies, but with opposite charges.

To relate the original Hamiltonian (1.86) to the reordered form (1.97) one may employ Wick's theorem [114, 35]. For the one-electron part this gives

$$N\left[\Psi^\dagger(1)\Psi(2)\right] = \Psi^\dagger(1)\Psi(2) - \left\langle 0^{(\mathrm{ref})}\left|\Psi^\dagger(1)\Psi(2)\right|0^{(\mathrm{ref}\ )}\right\rangle \qquad (1.99)$$

which shows that normal ordering of a one-electron operator corresponds to the subtraction of its vacuum expectation value. We have inserted a superscript (ref) on the vacuum to remind the reader that the definition of the vacuum and thus the reordering depends on the choice of orbital set in which the field operators are expanded. As reference vacuum the *bare* vacuum is employed, corresponding to the orbital sets generated from the solution of the free-particle Dirac equation (1.7). Normal ordering of the two-electron part gives

$$\begin{aligned} N\left[\Psi^\dagger(1)\Psi^\dagger(2)\Psi(3)\Psi(4)\right] &= \Psi^\dagger(1)\Psi^\dagger(2)\Psi(3)\Psi(4) \qquad (1.100)\\ &- N\left[\Psi^\dagger(1)\Psi(4)\right]\Psi^\dagger(2)\Psi(3)\\ &- N\left[\Psi^\dagger(2)\Psi(3)\right]\Psi^\dagger(1)\Psi(4)\\ &+ N\left[\Psi^\dagger(1)\Psi(3)\right]\Psi^\dagger(2)\Psi(4)\\ &+ N\left[\Psi^\dagger(2)\Psi(4)\right]\Psi^\dagger(1)\Psi(3)\\ &- \left\langle 0^{(\mathrm{ref})}\left|\Psi^\dagger(1)\Psi^\dagger(2)\Psi(3)\Psi(4)\right|0^{(\mathrm{ref}\ )}\right\rangle \end{aligned}$$

which does *not* correspond to simple subtraction of the vacuum expectation value, even though one readily sees that the bare vacuum expectation value of the reordered Hamiltonian (1.97) is zero.

We have already seen that the particle-hole formalism leads to a rather complicated expression for the Hamiltonian (1.86). With the introduction of reordering more complexity is added and manipulations involving the reordered Hamiltonian become extremely tedious. In order to simplify the ensuing manipulations it is therefore advantageous to go back to the form of the field operators (1.54) introduced in the previous section and instead explicitly define the bare vacuum as filled with all the negative-energy solutions of the free-particle Dirac equation

$$\left|0^{(\mathrm{ref}\ )}\right\rangle = a^\dagger_{[-1]}a^\dagger_{[-2]}\ldots a^\dagger_{[-\infty]}\left|\mathrm{empty}\right\rangle. \qquad (1.101)$$

The empty state $|\mathrm{empty}\rangle$ corresponds to the vacuum (1.63) of the standard approach to 4-component relativistic molecular theory. In the above expression as in the following we use *square brackets* (e.g. $h^{--}_{[pq]}$) around indices referring to the negative-energy solutions of the *free-particle* Dirac equation. We can now write the reordered QED Hamiltonian as

$$\hat{H} = \left\{h_{pq} - \Gamma_{pq}\left[\varphi^-_{[i]}\right]\right\}a^\dagger_p a_q + \frac{1}{4}\mathcal{L}_{pqrs}a^\dagger_p a^\dagger_r a_s a_q - h^{--}_{[ii]} + \frac{1}{2}\mathcal{L}^{----}_{[iijj]} \quad (1.102)$$

What we loose by this approach is the physical picture of electron-positron pair creation that was provided by the particle-hole formalism.

Using this reordered Hamiltonian we can now easily find the terms appearing in the Taylor expansion (1.73). The energy at the current expansion point is given by

$$E^{[0]}_{QED} = h^{++}_{ii} + h^{--}_{ii} - h^{--}_{[ii]} + \frac{1}{2}\Gamma^{--}_{[ii]}\left[\varphi^-_{[j]}\right] \qquad (1.103)$$

$$\begin{aligned} &+ \quad \frac{1}{2}\Gamma_{ii}^{++}\left[\varphi_j^+\right] + \frac{1}{2}\Gamma_{ii}^{++}\left[\varphi_j^-\right] + \frac{1}{2}\Gamma_{ii}^{--}\left[\varphi_j^+\right] + \frac{1}{2}\Gamma_{ii}^{--}\left[\varphi_j^-\right] \\ &- \quad \frac{1}{2}\Gamma_{ii}^{++}\left[\varphi_{[j]}^-\right] - \frac{1}{2}\Gamma_{ii}^{--}\left[\varphi_{[j]}^-\right] - \frac{1}{2}\Gamma_{[ii]}^{--}\left[\varphi_j^+\right] - \frac{1}{2}\Gamma_{[ii]}^{--}\left[\varphi_j^-\right] \end{aligned}$$

This rather formidable expression becomes quite a lot more intelligible when expressed in AO-basis

$$E_{QED}^{[0]} = D_{\mu\nu}^{\mathrm{QED}} h_{\nu\mu} + \frac{1}{2} D_{\mu\nu}^{\mathrm{QED}} \mathcal{L}_{\nu\mu\lambda\kappa} D_{\kappa\lambda}^{\mathrm{QED}} \tag{1.104}$$

in which appears the AO-density matrix of QED

$$D_{\kappa\lambda}^{\mathrm{QED}} = D_{\kappa\lambda} + D_{\kappa\lambda}^{\mathrm{pol}}; \quad D_{\kappa\lambda}^{\mathrm{pol}} = \sum_i^{(-)} \left(c_{\kappa i} c_{\kappa i}^* - c_{\kappa[i]} c_{\kappa[i]}^* \right). \tag{1.105}$$

Comparing with (1.76) one sees that the standard AO-density matrix D has been replaced by the QED counterpart D^{QED} obtained by the addition of the *vacuum polarization density* D^{pol} which reflects how the vacuum density is modified with respect to the reference vacuum upon the introduction of the actual potential.

The gradient vector in semiclassical QED is given by a formula identical to (1.80), except that the Fock matrix of the standard approach is to be replaced by its QED counterpart F^{QED} which is obtained by the substitution $D \rightarrow D^{\mathrm{QED}}$ in (1.78). Once again we find that the non-redundant orbital rotations are given by the parameter class A of table 1.1. However, the reader should carefully note that the distribution of the elements of the κ-matrix on the three parameter classes *changes* when going from the standard approach to QED, reflecting that the negative-energy orbitals are now filled and not empty.

The Hessian in semiclassical QED has the same form as in (1.81), but with the substitution $F \rightarrow F^{\mathrm{QED}}$. For a non-interacting system the Hessian becomes diagonal like in the standard approach (1.82)

$$A_{ai,ai} = \epsilon_a - \epsilon_i > 0; \quad \forall \kappa_{ia} \tag{1.106}$$

Following the same line of argument as in section 2.1 we may conclude that the Hessian of the interacting systems has all eigenvalues positive as well. We have thereby shown that the electronic ground state of semiclassical QED is characterized by a *minimization principle* at the Hartree-Fock level of theory.

2.3 Discussion

In the previous two subsections we have developed variational theory at the closed-shell Hartree-Fock level according to the standard 4-component approach and QED in the semiclassical limit. In this section we will summarize and discuss our findings. This will allow us to return to and study in more detail the argument of Brown and Ravenhall. It furthermore allows us to consider the extension to the correlated level of theory in the next subsection.

In QED the *negative-energy orbitals are filled,* in accordance with Dirac's proposal. This allows, at sufficiently large energies, to create electron-positron pairs out of the vacuum. Such processes do not conserve particle number, but do conserve charge. The energies of interaction in chemistry are generally too low for real pair creation processes, but the Dirac sea manifests itself through the phenomenon of vacuum polarization. As we have seen, at the closed-shell Hartree-Fock level the QED electronic ground state corresponds to a *true minimum* of the energy and this allows for instance the relativistic extension of the Hohenberg-Kohn theorem of DFT [75]. In contrast, the electronic ground state in the standard approach is characterized by a *minimax principle.* The vacuum is then *empty,* and the negative-energy orbitals are accordingly treated as an *orthogonal complement* to the electronic orbitals. However, and this is a crucial point, retaining the additional degrees of freedom provided by this orthogonal complement (the positronic degrees of freedom) allows the complete relaxation of the electronic ground state.

With the notation and machinery introduced in the two previous subsections we may now revisit the argument of Brown and Ravenhall. We consider a system of two non-interacting electrons and in the standard approach (std) write the ground state as

$$|\Phi_0\rangle = a_i^\dagger a_{\bar{i}}^\dagger |0\rangle_{\text{std}} \tag{1.107}$$

corresponding to the Slater determinant of the degenerate Kramers partners φ_i and $\overline{\varphi}_i$. The ground state energy is $E_0^{(0)} = \epsilon_i + \epsilon_{\bar{i}} = 2\epsilon_i$. We then turn on the two-electron interaction. By standard Rayleigh-Schrödinger perturbation theory the first order amplitudes of the perturbed wave function are

$$a_n = \frac{\langle \Phi_n | \hat{g}(1,2) | \Phi_0 \rangle}{E_0^{(0)} - E_n^{(0)}}. \tag{1.108}$$

One can now straightforwardly construct doubly-excited determinants

$$|\Phi_n\rangle = \left| \Phi_{i\to a^+}^{\bar{i}\to b^-} \right\rangle = a_a^\dagger a_b^\dagger |0\rangle_{\text{std}} \tag{1.109}$$

with one orbital φ_a^+ from the positive continuum and one orbital φ_b^- from the negative continuum such that the energy $E_n^{(0)} = \epsilon_a^+ + \epsilon_b^-$ becomes identical with $E_0^{(0)}$ and perturbation theory breaks down. The solution proposed by Brown and Ravenhall was to embed the Hamiltonian in projection operators Λ_+ onto positive energy orbitals

$$\hat{H} \quad \rightarrow \quad \Lambda_+ \hat{H} \Lambda_+ . \tag{1.110}$$

This corresponds to retaining only the purely electronic terms of the QED Hamiltonian (1.86)

$$\hat{H}^{\mathrm{no-pair}} = h_{pq}^{++} b_p^\dagger b_q + \frac{1}{4} \mathcal{L}_{pqrs}^{++++} b_p^\dagger b_r^\dagger b_s b_q . \tag{1.111}$$

Pair creation and annihilation is thereby excluded and this approximation has therefore been referred to as the *no-pair* approximation. However, *the no-pair Hamiltonian is not unique* since the distinction between electronic and positronic creation and annihilation operators depends on the orbital set in which the field operators are expanded. One possible choice is the solutions of the free-particle Dirac equation (1.7), giving the "free" picture. Another choice is the solutions of the Dirac equation in the molecular field (1.17) leading to the Furry picture. A third possibility is to continuously update the projection operators through SCF iterations so that they at convergence correspond to the solutions of the combined molecular and mean-field potentials of the Hartree-Fock equations. We shall see that this choice, the "fuzzy" picture, as proposed by Mittleman [67], corresponds to the standard 4-component approach.

Let us, however, first consider the two-electron system discussed by Brown and Ravenhall as described by QED. We write the reference determinant as

$$|\Phi_0\rangle_{\mathrm{QED}} = b_i^\dagger b_{\bar{i}}^\dagger |0\rangle_{\mathrm{QED}} \tag{1.112}$$

where the vacuum is defined by (1.88) and thereby the complete orbital set of the non-interacting system. Using (1.102) we find the QED ground state energy to be

$$E_{\mathrm{QED};0}^{(0)} = 2\epsilon_i^+ + \sum_j^{(-)} \left(\epsilon_j^- - h_{[jj]}^{--} \right) . \tag{1.113}$$

Let us next consider the QED analogues of the troublesome doubly excited determinants (1.109). From the reinterpretation of the field operators (1.85) and the definition of the vacuum (1.88) we immediately obtain

$$|\Phi_n\rangle_{\mathrm{QED}} = b_a^\dagger d_b |0\rangle_{\mathrm{QED}} = 0 \quad (!), \tag{1.114}$$

showing that these determinants simply do not occur since all the negative-energy orbitals are already occupied. One may attempt the alternative form

$$|\Phi_n\rangle_{\mathrm{QED}}' = b_a^\dagger d_b^\dagger |0\rangle_{\mathrm{QED}} \tag{1.115}$$

corresponding to the creation of a true electron-positron pair out of the vacuum. However, this determinant corresponds to a different charge (zero and not -2) and does therefore not interact with the two-electron reference determinant since the QED Hamiltonian conserves charge. Such determinants are thus excluded on physical grounds. We can therefore conclude that at the QED level the Brown-Ravenhall disease is easily cured; the reference determinant (1.112) mixes only with purely electronic determinants *for fixed particle number N.* The two-electron part of the QED Hamiltonian (1.86) allows coupling of the reference occupation-number vector to vectors to which are added one or two electron-positron pairs. These can again couple to vectors with more pairs at higher order in perturbation theory. It is perhaps easier to analyze these interacting occupation-number vectors by going from the particle-hole formalism to the original form of the field operators (1.54) and instead fill up all the negative-energy orbitals of the interacting system. The occupation-number vector of the particle-hole formalism containing one and two electron-positron pair(s) then correspond to determinants

$$|\Phi_{k^-\rightarrow a^+}\rangle \quad \text{and} \quad \left|\Phi^{l^-\rightarrow b^+}_{k^-\rightarrow a^+}\right\rangle, \tag{1.116}$$

respectively. The first class of determinants contain the excitation of an electron from the negative-energy orbital k to the virtual positive-energy orbital a. Using the form (1.102) of the reordered QED Hamiltonian with all two-electron terms deleted we find that the unperturbed energy of the determinant $|\Phi_n\rangle = |\Phi_{k^-\rightarrow a^+}\rangle$ is

$$E^{(0)}_{\text{QED};n} = 2\epsilon_i^+ + \epsilon_a^+ - \epsilon_k^- + \sum_j^{(-)} \left(\epsilon_j^- - h_{[jj]}^{--}\right) \tag{1.117}$$

Using the full Hamiltonian we can determine the transition moment, that is the numerator of (1.108), and thus obtain the following expression for the first-order amplitude of the perturbed wave function

$$a_n = -\frac{\mathcal{L}^{-+--}_{ka[ii]}}{\epsilon_k^- - \epsilon_a^+} \quad \text{(no summation !)} \tag{1.118}$$

The denominator is clearly of order $O(c^2)$. The numerator $\mathcal{L}^{-+--}_{ka[ii]} = (ka|ii) - (ki|ia)$ contains two-electron integrals in which the integration over one electron contains the overlap of one positive-energy and one negative-energy orbital. To determine the order of this contribution let us recall from the discussion in section 1.1.1 that for positive-energy solutions the large component is in an averaged sense a factor c larger than the small component, and that for negative-energy solution this role is reversed. We can then conclude that the numerator is on the order of $O(c^{-1})$ and thus the amplitude (1.118) is of order $O(c^{-3})$. By similar arguments we find that the amplitude of doubly-excited determinants $|\Phi_n\rangle = \left|\Phi^{l^-\to b^+}_{k^-\to a^+}\right\rangle$ is

$$a_n = \frac{\mathcal{L}^{-+-+}_{kalb}}{\epsilon^-_k + \epsilon^-_l - \epsilon^+_a - \epsilon^+_b} \tag{1.119}$$

and is of order $O(c^{-4})$. It is important to note that the determinants of (1.116) contain excitations from *occupied* negative-energy orbitals to virtual positive-energy orbitals. They are therefore *absent* in the standard approach since in this approach the vacuum is *empty* and the negative-energy orbitals only serve as an orthogonal complement. The presence of the determinants (1.116) constitute a pure QED effect and our order analysis shows the minuteness of their contribution.

We have seen that the standard approach differs from QED in the semiclassical limit only by the absence of vacuum polarization. This is a rather minute effect, giving rise to the Uehling effect, which is a minor contribution to the Lamb shift [77], and it would therefore be surprising if the elimination of vacuum polarization should lead to a complete breakdown of the theory. In subsection 2.1 we have seen that the electronic ground state in the standard approach can be found by the minimax principle (1.83). However, in practice the solutions are found by vector selection, that is in each iteration of the SCF cycle vectors for the next iteration are not selected according to an *aufbau* principle, rather one selects the lower *electronic orbitals* that are generally easily identified through the energy gap down to the negative-energy orbitals. This procedure corresponds *precisely* to the use of a no-pair Hamiltonian (1.111) with continuously updated projection operators as proposed by Mittleman. Another way of seeing this procedure is to note that the formulas of the standard approach can be recovered from QED by choosing the reference vacuum as the vacuum defined by the self-consistent mean-field potential. The vacuum polarization density D^{pol} (1.105) then goes to zero. Having identified the electronic Hamiltonian of the standard 4-component approach with the no-pair Hamiltonian in the "fuzzy" picture, we can see that the Brown-Ravenhall disease is cured, since the doubly excited determinants (1.109) leading to continuum dissolution are projected out.

The no-pair Hamiltonian (1.111) depends on the orbital set in which the field operators (1.85) are expanded and thus on the potential generating this orbital. This *non-uniqueness* of the no-pair Hamiltonian is an important point that seems to escape a number of authors. In particular when deriving various approximate 1- or 2-component relativistic Hamiltonians a number of authors embed the Dirac-Coulomb-(Gaunt/Breit) Hamiltonian in projection operators (1.110) *without* specifying the reference orbitals. In an otherwise excellent paper by M. Barysz and A.J.Sadlej [5] one reads: "[...] one should mention that the methods of relativistic quantum chemistry, which are based directly on 4-spinors [...], also neglect the positronic (negative energy) solutions. This is generally known as the *no-pair approximation* [...] and makes the *exact* electronic 2-spinor solutions fully equivalent to four-component electronic solutions. Hence, as long as our knowledge of the pure electronic spectrum of the Dirac equation is sufficient, the 4-spinor formalism becomes obsolete." What the authors miss is the fact that the 4-component methods *update* the projection operators of the no-pair Hamiltonian *until self-consistency* and therefore achieve a complete relaxation of orbitals to the actual potential. The approximate 1- and 2- component methods *freeze* the projection operators before performing an approximate decoupling of the electronic and positronic degrees of freedom. This may lead to excellent approximations that allow relativistic calculations at reduced computational cost, but it is incorrect to state they provide complete equivalence with the 4-component methods.

Just how good these approximations with fixed projection operators are can be easily investigated at the 4-component level in the algebraic approximation. It suffices to generate the orbital set corresponding to the reference potential defining the projection operator and then delete the negative energy vectors from the ensuing calculation in the actual potential. In table 1.2 this procedure is illustrated by 4-component relativistic Hartree-Fock calculations on the radon atom. It can be seen that the use of projection operators defined by the free-particle Dirac equation, the "free" picture, gives rather large deviations, in particular in the core region, compared to the standard approach, the "fuzzy" picture, based on fully relaxed projection operators. On the other hand, the use of projection operators defined by the molecular field (1.17), the Furry picture, compare fairly well with the standard approach. For reference we have also included the results of a 1-component (scalar) second-order Douglas-Kroll calculation in the same basis. Apart from the lack of spin-orbit interaction one can see that the result is rather close to the Furry picture, which can be considered as infinite-order Douglas-Kroll. An at first sight surprising result is obtained by projecting the standard or "fuzzy" occupied orbitals onto the negative-energy free particle solutions in the same basis. On then finds that the negative-energy free particle solutions contribute only 0.0053 to the total density of 86 electrons! However, al-

Table 1.2. The effect of embedding the Dirac-Coulomb Hamiltonian by projection operators onto positive energy orbitals as illustrated by a Hartree-Fock calculation on the radon atom. The calculations were carried out using a point nucleus and an uncontracted 32s32p17d11f family Gaussian large component basis. The small component basis was generated by restricted kinetic balance. For comparison we have included the results of a scalar second-order Douglas-Kroll calculation (DK2) in the same basis.

	Fuzzy	Furry	Free	DK2
E(total)	-23610.98632	-23611.01630	-24153.25409	-23533.64569
$1s_{1/2}$	-3644.74282	-3644.75635	-3859.56081	-3626.81594
$2s_{1/2}$	-669.37883	-669.38209	-694.44791	-667.27997
$2p_{1/2}$	-642.35771	-642.36308	-650.93434	-570.13212
$2p_{3/2}$	-541.08440	-541.08776	-540.98657	idem
$3s_{1/2}$	-166.96715	-166.96783	-172.50957	-166.54445
$3p_{1/2}$	-154.90291	-154.90395	-156.89683	-138.48586
$3p_{3/2}$	-131.72668	-131.72734	-131.61723	idem
$3d_{3/2}$	-112.56321	-112.56374	-112.25150	-109.69442
$3d_{5/2}$	-107.75599	-107.75642	-107.45065	idem
$4s_{1/2}$	-41.34840	-41.34854	-42.76820	-41.25251
$4p_{1/2}$	-36.02132	-36.02153	-36.50780	-31.85484
$4p_{3/2}$	-30.11915	-30.11927	-30.06259	idem
$4d_{3/2}$	-21.54718	-21.54726	-21.44792	-20.89499
$4d_{5/2}$	-20.43800	-20.43805	-20.34020	idem
$4f_{5/2}$	-9.19411	-9.19409	-9.11432	-9.06496
$4f_{7/2}$	-8.92842	-8.92840	-8.85031	idem
$5s_{1/2}$	-8.41670	-8.41671	-8.72474	-8.39894
$5p_{1/2}$	-6.40907	-6.40910	-6.49582	-5.54050
$5p_{3/2}$	-5.17528	-5.17529	-5.14742	idem
$5d_{3/2}$	-2.18932	-2.18932	-2.15957	-2.09081
$5d_{5/2}$	-2.01625	-2.01625	-1.98766	idem
$6s_{1/2}$	-1.07263	-1.07263	-1.12104	-1.06787
$6p_{1/2}$	-0.54033	-0.54033	-0.54932	-0.42777
$6p_{3/2}$	-0.38390	-0.38390	-0.37708	idem

beit minute, one should keep in mind that the negative-energy orbitals contribute very large energies and this explains the large deviations of the "free" picture result from the standard approach.

2.4 The correlated level

Let us now consider the extension to the correlated level. The most general variational parametrization is provided by the MCSCF *ansatz*

$$\left|\Psi^{\mathrm{MCSCF}}\right\rangle = \exp\left[-\widehat{\kappa}\right]\sum_i c_i \left|\Phi_i\right\rangle \tag{1.120}$$

in which the orbital rotation parameters $\{\kappa_{pq}\}$ are supplemented by the coefficients $\{c_i\}$ of the CI expansion in Slater-determinants $\{|\Phi_i\rangle\}$. At the Hartree-Fock level the orbital rotation parameters provide relaxation of orbitals. At the correlated level orbital relaxation, in addition to correlation, is also provided by the CI expansion coefficients. In fact, in the non-relativistic domain the orbital rotation parameters become redundant in the limit of a complete CI-expansion within the given orbital basis, showing that in this domain the exact solution within a given 1-particle basis is provided by full CI.

At the 4-component relativistic level the choice of CI-expansion becomes more difficult since one employs a no-pair Hamiltonian with projection operators that in principle should be allowed to relax completely to the actual potential of the system. At the MCSCF level the relaxation is provided by the orbital rotation parameters that can be employed in conjunction with an expansion in purely electronic determinants, that is containing only positive-energy orbitals. In CI and CC methods the orbital basis is frozen, and the conventional approach is to employ the no-pair Hamiltonian (1.111) defined by Hartree-Fock orbitals. This means that determinantal expansions are restricted to purely electronic determinants generated from this orbital set. Complete relaxation of the projection operators is therefore not possible, although the projection operators defined by the Hartree-Fock orbitals can be expected to constitute a very good approximation. In the limit of full CI the orbital parameters κ_{ia}^{++} describing rotations between occupied and virtual positive energy orbitals become redundant, but the parameters κ_{ia}^{+-} , describing rotations between occupied positive energy orbitals and virtual negative energy orbitals are not accounted for by the CI-expansion. This tells us that at the 4-component relativistic level the exact solution in a given 1-particle basis is not provided by full CI, but by MCSCF.

Bunge *et al. [11]* has advocated 4-component relativistic CI using in addition to purely electronic determinants also mixed determinants, that is Slater determinants containing both positive- and negative-energy orbitals. A subclass of these determinants are precisely the doubly excited determinants that appear in the argument of Brown and Ravenhall analyzed in the previous section. It is our firm conviction that the methods advocated by Bunge *et al.* are plain wrong and our argument against the use of these methods runs as follows: Consider CI at the QED level of theory, or rather in the semiclassical limit that was analyzed in section 2.2. The CI *ansatz* is

$$\left|\Psi^{\mathrm{CI}}\right\rangle = \sum_i c_i \left|\Phi_i\right\rangle \tag{1.121}$$

which is simply the MCSCF *ansatz* (1.120) with the orbital rotation operator deleted. We may assume that we start with the orbital set generated at the Hartree-Fock level. We want to describe a system of N electrons and so our Hartree-Fock reference is simply an occupation-number vector of N electrons as in (1.87). The QED Hamiltonian (1.97) does not conserve particle number and so the CI-expansion (1.121) will contain occupation-number vectors with particle number $N + 2n$, where n is the number of electron-positron pairs generated from the reference determinant. On the other hand, the QED Hamiltonian does conserve charge and so the occupation-number vectors with N particles must all be *purely* electronic since they are the only ones that couple to the reference determinant through the QED Hamiltonian. The occupation-number vectors with more than N particles correspond to determinants (1.116) involving one or more excitations of electrons from occupied negative-energy orbitals to virtual positive-energy orbitals. The inclusion of these determinants will provide complete relaxation of the orbital set. However, as emphasized already in the previous section, these determinants are *absent* in the standard approach since all the negative-energy states are empty. The mixed determinants included by Bunge *et al.* in the CI-expansion simply do not exist at the QED level. They are forbidden by the Pauli exclusion principle since they involve the double occupation of negative-energy orbitals. Their inclusion can precisely lead to the continuum dissolution predicted by Brown and Ravenhall. Bunge *et al.* claims variational control using the Hylleraas-Undheim theorem which implies that the ordered sequence of eigenvalues of a CI-matrix with one determinant added to the expansion are interlaced with those of the original CI-matrix. However, the Hylleraas-Undheim theorem only connects sequences of *eigenvalues* and not *eigenstates.* It is well known in direct-CI methods that upon enlarging the trial vector space *root flipping* can occur, that is two states may change order, and this is precisely the possibility that precludes variational control in the approach advocated by Bunge *et al. [11].*

3. Implementation and Computational Scaling

Now that the necessary general theory has been introduced in the preceding sections we can direct our attention to the application to molecules. This concerns both the implementation of algorithms and their computational scaling in comparison to the "spinfree" algorithms that are used in non-relativistic quantum chemistry. We start by considering basis set expansion techniques.

3.1 The algebraic approximation

While Hartree-Fock equations for atoms can be solved via numerical integration one needs a basis set expansion to apply the method to general molecular systems. In order to obtain a viable scheme one would like to choose functions that are readily integrated, preferably using the same techniques as employed in non-relativistic quantum chemistry, so that one can benefit from the large body of experience and implementations that are available in this field. Non-relativistic orbitals are, however, usually chosen as real functions of the space coordinates only, whereas the Dirac spinors are complex functions of space and spin coordinates. How can one best exploit the available basis set technology in this domain ?

Basis set expansion is usually done in the LCAO approximation where the molecular orbitals are expressed as linear combinations of atomic orbitals. These atomic orbitals are in turn expressed as fixed linear combinations of simpler functions called primitives. Consider the solution of the time-independent Dirac equation for a molecular field (1.41) in the algebraic approximation. We expand the large and small component in two different sets of primitives

$$\psi_p^X = \chi_\mu^X c_{\mu p}^X; \quad X = L, S. \tag{1.122}$$

We then obtain the eigenvalue equation

$$\begin{bmatrix} V^{LL} & c\Pi^{LS} \\ c\Pi^{SL} & V^{SS} - 2mc^2 S^{SS} \end{bmatrix} \begin{bmatrix} \mathbf{c}_p^L \\ \mathbf{c}_p^S \end{bmatrix} = \begin{bmatrix} S^{LL} & 0 \\ 0 & S^{SS} \end{bmatrix} \begin{bmatrix} \mathbf{c}_p^L \\ \mathbf{c}_p^S \end{bmatrix} \epsilon_p \tag{1.123}$$

in which appear the matrix elements

$$S_{\mu\nu}^{XY} = \left\langle \chi_\mu^X \mid \chi_\nu^Y \right\rangle; \quad V_{\mu\nu}^{XY} = \left\langle \chi_\mu^X \left| \hat{V} \right| \chi_\nu^Y \right\rangle; \quad \Pi_{\mu\nu}^{XY} = \left\langle \chi_\mu^X \left| (\boldsymbol{\sigma} \cdot \mathbf{p}) \right| \chi_\nu^Y \right\rangle \tag{1.124}$$

The first attempts along these lines failed rather miserably, even for one-electron systems. The origin of the problem was the neglect of the coupling of the large and small components in the Dirac equation

$$2mc\psi_p^S(\mathbf{r}) = \widehat{R}_p (\boldsymbol{\sigma} \cdot \mathbf{p}) \psi_p^L(\mathbf{r}); \quad \widehat{R}_p = \left[1 + \frac{\epsilon_p - \hat{V}}{2mc^2}\right]^{-1} \tag{1.125}$$

We obtain the finite basis equivalent of this relation by first writing out (1.123) in terms of two matrix equations

$$\begin{aligned} V^{LL}\mathbf{c}_p^L + c\Pi^{LS}\mathbf{c}_p^S &= S^{LL}\mathbf{c}_p^L\epsilon_p && (1.126) \\ c\Pi^{SL} + \left(V^{SS} - 2mc^2 S^{SS}\right)\mathbf{c}_p^S &= S^{SS}\mathbf{c}_p^S\epsilon_p && (1.127) \end{aligned}$$

and solve for the vector $\mathbf{c}_p^S$ of small component expansion coefficients. This gives the relation

$$\mathbf{c}_p^S = \frac{1}{2mc}\left[S^{SS} + \frac{\epsilon_p S^{SS} - V^{SS}}{2mc^2}\right]^{-1} \Pi^{SL}\mathbf{c}_p^L. \tag{1.128}$$

From (1.125) or, alternatively, (1.128) we see that the small component wave function can be regarded as the result of the consecutive action of the operators $(\boldsymbol{\sigma}\cdot\mathbf{p})$ and $\widehat{R}_p$ on the large component wave function. The energy-dependent operator $\widehat{R}_p$ is totally symmetric, but the operator $(\boldsymbol{\sigma}\cdot\mathbf{p})$ couples functions of opposite parity which indicates that separate basis set expansions are needed for the large and small components, as was anticipated by (1.122). We thus see that the coupling of the large and small components generally leads to the use of larger basis sets and thereby increased computational cost at the 4-component relativistic level compared to non-relativistic methods. For instance, at the Hartree-Fock level three classes of integrals over the Coulomb operator (1.49) appear — $(LL\,|\,LL)$, $(SS\,|\,LL)$ and $(SS\,|\,SS)$ — where the first and smaller class (all indices correspond to large component basis functions) constitute the two-electron integrals of a non-relativistic calculation. On the other hand, we shall see that a closer study of the coupling relation provides suggestions as how to reduce computational cost.

Before entering a detailed discussion of the coupling (1.125), let us first consider the choice of primitives in the basis set expansions in relativistic calculations. We know that the exact solutions of the relativistic hydrogenic atom [6] differ from the non-relativistic ones in having a singularity at the nucleus and having a coupling between the spatial and spin coordinates. The singularity is somewhat artificial because it appears in the exact solution for a model in which the nucleus is represented by a point charge. A more realistic finite nucleus model [108] gives the wave functions an approximately Gaussian shape in this region. The second difference, coupling of the spatial and spin degrees of freedom, leads to more complications. An expansion in 2-spinor functions was suggested by the Oxford group[73] who defined basis functions as

$$\begin{aligned}\chi_\mu^L(\mathbf{r}_A) &= N_\mu^L f_\mu^L(r_A)\,\xi_{\kappa_\mu,m_\mu}(\theta_A,\varphi_A)\\ \chi_\mu^S(\mathbf{r}_A) &= N_\mu^S f_\mu^S(r_A)\,\xi_{-\kappa_\mu,m_\mu}(\theta_A,\varphi_A)\end{aligned} \tag{1.129}$$

The radial functions $f_\mu^L(r_A)$ and $f_\mu^S(r_A)$ depend only on the distance to expansion center A, all the angular and spin dependence is carried by the 2-spinor functions $\xi_{\kappa_\mu,m_\mu}(\theta_A,\varphi_A)$ for which the analytic form is known [39]. The problem with this approach is that, due to the changed angular dependence relative to non-relativistic theory, completely new integral evaluation routines need to be developed. This made the actual implementation for molecular systems appear relatively late[73, 117]. Implementations of a more pragmatic approach with scalar expansion functions were available about a decade earlier[1, 23, 62, 24, 81] because they employed non-relativistic integral evaluation packages [65]. In such schemes one chooses expansion functions in which only one component of the 4-spinor is non-zero :

$$\psi_p^X = \left[\begin{array}{cc} \boldsymbol{\chi}^X & 0 \\ 0 & \boldsymbol{\chi}^X \end{array} \right] \left[\begin{array}{c} \mathbf{c}_p^{X\alpha} \\ \mathbf{c}_p^{X\beta} \end{array} \right]; \quad X = L, S \tag{1.130}$$

Here $\boldsymbol{\chi}^X$ is a row vector of primitives and $\mathbf{c}_p^{X\alpha}$ and $\mathbf{c}_p^{X\beta}$ are column vectors of the corresponding expansion coefficients. Note that the same set of primitives is used for the α and β components. The scalar functions $\chi_\mu(\mathbf{r}_A)$ are again chosen as functions of the position of the electron relative to expansion centers A. One may still separate the radial and angular parts by choosing the spherical form

$$\chi_\mu(\mathbf{r}_A) = N'_\mu r_A^{\ell_\mu} f_\mu(r_A) Y_{\ell_\mu, m_\mu}(\theta_A, \varphi_A) \tag{1.131}$$

but this gives now only a marginal advantage over the Cartesian form

$$\chi_\mu(\mathbf{r}_A) = N''_\mu x_A^{n_\mu^x} y_A^{n_\mu^y} z_A^{n_\mu^z} f_\mu(r_A) \tag{1.132}$$

since neither has the correct angular dependence. The correct angular dependence is in this approach of course still achieved at the Hartree-Fock stage once the actual spinors are found.

In all three models — 2-spinor, spherical or Cartesian — one still needs to choose the specific form of the radial expansion functions $f_\mu(r_A)$. Again one can take the exact solutions for the hydrogenic atom as a guideline. Slater functions $f_\mu(r_A) = e^{-\zeta_\mu r_A}$ have the correct long range behavior, but do not have the correct shape close to the nuclei. They neither represent the singularity found in the point nucleus model nor the approximate Gaussian shape appropriate for finite nuclear models. There may thus be an advantage for expansion in Gaussian type functions $f_\mu(r_A) = e^{-\zeta_\mu r_A^2}$ for properties that depend on the precise shape near extended nuclei, while properties that depend on the electron density in the outer regions of the molecule are better described using Slater type functions. In practice, however, decisive is the more efficient evaluation of multi-center integrals that makes Gaussian based expansions the method of choice for both kind of properties. In both the scalar and the two-spinor expansion schemes one can then employ integration schemes developed for Gaussian type functions [43].

As said before, the advantage of the scalar function approach is that they require very little adaptation of existing non-relativistic integral evaluation routines. A disadvantage is, however, that the expansion in N^L large and N^S small component primitives is unnecessarily long. To see why this is so we need to consider the relation (1.125) between the large and the small component part of the wavefunction in more detail. For electronic orbitals *and* non-singular potentials ϕ, e.g. finite nuclei, the coupling reduces in the non-relativistic limit to

$$\lim_{c\to\infty} 2mc\,\psi_p^S(\mathbf{r}) = (\boldsymbol{\sigma}\cdot\mathbf{p})\,\psi_p^L(\mathbf{r}) \tag{1.133}$$

and equivalently, in the algebraic approximation, to

$$\lim_{c\to\infty} 2mc\,\mathbf{c}_p^S = \left[S^{SS}\right]^{-1}\Pi^{SL}\mathbf{c}_p^L \tag{1.134}$$

since the operator $\hat{R}_p$ then goes to unity. In practice one obtains this result for point nuclei as well in the algebraic approximation because Gaussian basis functions are not able to describe the singularities at nuclei. When using basis set expansions, one usually ignores the effect of $\hat{R}_p$ since this operator, even in the relativistic regime, is close to unity due to the large value of $2mc^2$. This is called the kinetic balance procedure [88, 92] since it guarantees proper representation of the operator identity $(\boldsymbol{\sigma}\cdot\mathbf{p})(\boldsymbol{\sigma}\cdot\mathbf{p}) = p^2$ in matrix form. We can see this better by inserting the non-relativistic coupling (1.134) into the matrix equation (1.126). We then obtain

$$V^{LL}\mathbf{c}_p^L + \frac{1}{2m}\Pi^{LS}\left[S^{SS}\right]^{-1}\Pi^{SL}\mathbf{c}_p^L = S^{LL}\mathbf{c}_p^L\epsilon_p \tag{1.135}$$

which gives the matrix representation of the non-relativistic Schrödinger equation provided that the relation

$$\left\langle\chi_\mu^L\left|p^2\right|\chi_\nu^L\right\rangle = \left\langle\chi_\mu^L\left|(\boldsymbol{\sigma}\cdot\mathbf{p})\,\Omega\,(\boldsymbol{\sigma}\cdot\mathbf{p})\right|\chi_\nu^L\right\rangle \tag{1.136}$$

holds. The term

$$\Omega = \sum_{\kappa\lambda}\left|\chi^S\right\rangle_\kappa\left[\left(S^{SS}\right)^{-1}\right]_{\kappa\lambda}\left\langle\chi_\lambda^S\right| \tag{1.137}$$

has the form of the resolution of identity in a non-orthogonal basis. As carefully analyzed by Dyall *et al.* [22] it is not necessary to have a complete small component basis in order for relation (1.136) to hold; it suffices that the small component basis *spans* the result of the operator $(\boldsymbol{\sigma}\cdot\mathbf{p})$ acting on the large component basis. This observation explains the weakness of the Talman minimax principle (1.84) alluded to in section 2.1. Assume that we have the exact solution. If the large component function is now varied by adding new primitives *without* conjointly adding small component primitives that span the effect of $(\boldsymbol{\sigma} \cdot \mathbf{p})$ on these new functions, then relation (1.136) will not hold. The kinetic energy will be underestimated, and one may observe that the energy falls below the exact energy, contrary to the Talman minimax principle. The kinetic balance recipe in practice prevents this so-called variational collapse. It does not mean that kinetically balanced basis sets always provide an upper limit of the true energy since the relation (1.134) does not represent the exact coupling (1.128). The effect of the operator $\hat{R}_p$ can be seen in figure 1.1 where we compare the small component radial function of the $1s_{1/2}$ orbital of the radon atom from a Hartree-Fock calculation using a finite nucleus with the function generated by kinetic balance from the corresponding large component radial function. It can be seen that close to the nucleus there is a marked discrepancy between the two functions, illustrating that the kinetic balance prescription gives a rather poor description of the coupling of the large and small components in this region. However, since this breakdown occurs in the close vicinity of nuclei, well within the radial expectation value of the $1s_{1/2}$ orbital, it is reasonable to assume that it occurs in a region of local atomic symmetry, even for molecular systems. This implies that large component s functions then only couple to small component p functions and not to other angular momentum types. With a sufficiently flexible basis the kinetic balance prescription will therefore allow the establishment of the correct coupling. In most cases energy optimizing a sequence of uncontracted kinetically balanced basis sets shows monotonous convergence from above upon extending the basis. For very large basis sets one sometimes sees that the energy in a kinetic balance basis set expansion is slightly lower than the reference value that is obtained via numeric integration [32]. There is, however, ample numerical evidence that such small deviations do not present a real problem, and the kinetic balance procedure has therefore become the standard approach in developing basis sets for 4-component relativistic calculations.

Since it suffices to span the range of functions $\chi_{\mu}^{SA}(\mathbf{r}) = (\boldsymbol{\sigma} \cdot \mathbf{p})\, \chi_{\mu}^{LA}(\mathbf{r})$ in the small component basis, kinetic balance can be realized in different ways. In 2-component basis sets one can directly include one small component expansion function for each large component function . This 1:1 relation between the large and small component basis functions has been denoted *restricted* kinetic balance. In scalar basis sets one usually considers all three components of the $\mathbf{p}$ separately, which is then denoted *unrestricted* kinetic balance. Still, the separate one-component functions can be recombined in the transformation to the orthogonal basis when the Hartree-Fock matrix equation is solved. If this is done, the final result becomes identical to that of the 2-component procedure, with as only difference the calculation of primitive integrals and the construction of the Fock matrix. In these steps the scalar expansion method needs significantly more primitive functions than needed in the comparable two-spinor expansion. As example we take the representation of the 2p spinors. The large component is expressed as a linear combination of $2p_x$, $2p_y$, and $2p_z$ Cartesian Gaussian functions using the same exponents for the whole set of p-functions. The small component is then expanded in the set of functions that are generated by operation with the three components of the gradient operator on the three scalar 2p functions. This gives seven unique functions ($1s$, $3s$ and $3d$) to balance three large component functions. This overrepresentation can be reduced using so-called dual family basis sets [32] in which the exponents of the d-functions are a subset of those of s-functions, and the exponents of f-functions a subset of those of p-functions, etc. This is efficient because small component scalar functions are then used to balance two large component functions at the same time. Still, one always ends up with a longer expansion than used in an qualitatively equivalent two-spinor expansion. The superfluous linear combinations are usually projected away because they may give problems with linear dependencies. The problem is enhanced in heavier elements because the $p_{1/2}$ and $p_{3/2}$ orbitals then have markedly different radial extent [80]. A scalar expansion scheme does not permit distinction between these subshells which means that the tight functions needed in the expansion set for the $p_{1/2}$ will also be included in the set used for the $p_{3/2}$. This illustrates that a two-spinor expansion is to be preferred on theoretical grounds . It is, however, also a matter of strategy whether this reduction in application time really warrants the additional effort in constructing and maintaining a dedicated integral evaluation implementation for relativistic calculations. An implementation of the scalar expansion scheme that shares most of its inner kernels with a non-relativistic implementation will benefit easily from new advances in non-relativistic integral evaluation techniques, while these need to be rederived and separately implemented in a two-spinor scheme. Another issue is the interface of non-relativistic and relativistic schemes, an approach advocated by Dyall [26, 27, 28, 29], where it also may be easier to work with primitive scalar expansion functions throughout.

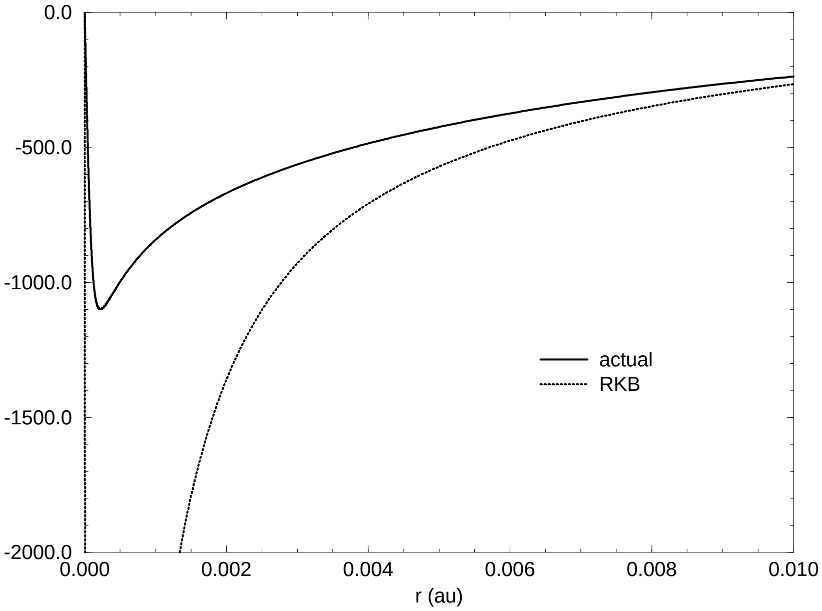
Let us now direct attention to the $\hat{R}$ operator (1.125) that modifies the function $(\boldsymbol{\sigma} \cdot \mathbf{p})\,\psi^L(\mathbf{r})$ in regions where the potential $\phi(\mathbf{r})$ is large. This is the case in the vicinity of nuclei and has important implications for the kinetic balance procedure for contracted basis sets. The kinetic balance prescription has proven to be a valid method to generate primitive small component basis sets. However, it is clear from figure 1.1 that it fails when applied to the large component part of an eigenspinor. The dramatic consequences can easily be verified for hydrogenic systems [100]. It means that the simple kinetic balance procedure can not be applied to heavily contracted basis functions (that approach the exact large component solution) because the generated small component functions do not provide sufficient flexibility to establish the correct coupling. The proper procedure is to take both the large and the small component coefficients directly from uncontracted atomic reference calculations. This has been referred to as atomic balance [100]. As a side remark we note that the sometimes advocated [63] use of non-relativistic functions to expand the large component, combined with application of the kinetic balance prescription for the small component, also prevents the divergences but at the expense of having wrong expansion functions for both the large and the small component. Again, contraction with the atomic spinor coefficients is easier in two-spinor expansion schemes than in scalar expansion schemes because the former makes it possible to define specific contractions for the spin-orbit split subshells ($j = l - 1/2$ and $j = l + 1/2$). In the latter scheme one needs to give one set of contraction coefficients for a given nl shell which makes it necessary to compromise.

The $\hat{R}$-operator is also of interest when studying the long-range behavior of the small component wave function. In regions of negligible potential it reduces to a constant factor of

$$\hat{R}_p = \left[1 + \frac{\epsilon_p}{2mc^2}\right]^{-1} \tag{1.138}$$

Since the amplitude of the large component wave function in this region is dominated by that from the HOMO with a small value of ε_p and only a small gradient, the small component wave function will have nearly zero amplitude in this region. The small component density is therefore rather localized and atomic in nature. This observation has made it possible to calculate spectroscopic constants of molecular systems where the complete set of $(SS \mid SS)$ integrals is eliminated and the potential curve is corrected by a simple Coulombic correction [107]. This constitute a perturbational correction, but more elaborate schemes of integral modeling have been developed in order to reduce computational cost. Recently a scheme was presented in which all overlap between small component basis functions located on different expansion centers was neglected in the evaluation of potential energy matrix elements [48]. The promising results obtained in these pilot calculations indicate that in the long run it will probably suffice to restrict evaluation of potential energy inte-

Figure 1.1. The actual small component radial function $[r^{-1}Q(r)\ ]$ of the $1s_{1/2}$ orbital of radon obtained from a numerical GRASP calculation with Gaussian nucleus, as compared with the radial small component function $[r^{-1}P(r)\ ]$ generated from restricted kinetic balance (RKB). For comparison the radial expectation value of the $1s_{1/2}$ orbital is 0.0015 a.u.



grals in the small component basis to those that share a common center. In this approximation the scaling of integral evaluation then becomes $aN_L^4 + bN_S N_L^2 + cN_S$ where the first term is comparable to that of a non-relativistic all-electron calculation. This means that the preceding discussion on the efficiency of integral evaluation in two-spinor versus scalar expansion schemes looses some of its significance because it will apply mainly to small systems. For larger systems the integral evaluation cost will be dominated by the "non-relativistic" first term and it becomes more important to employ efficient integral-direct or multipole expansion techniques to improve the computational efficiency than to optimize the calculation of integrals over the small component basis.

For small molecules, molecular symmetry may of course also help to increase computational efficiency. In the two-spinors expansion one can resort directly to double group symmetry and apply projection operators or other techniques to create the appropriate symmetric linear combinations of atomic two-spinors. When scalar functions are chosen as expansion basis it is natural to first adapt these using non-relativistic point group symmetry and then combine the resulting functions to obtain double group symmetry adapted functions. In that case one may combine the symmetry adaption of scalar basis functions to single point groups with time reversal symmetry in a quaternion symmetry scheme [82, 83]. The sidestep via non-relativistic symmetry adapted functions also has advantages when approximate, in particular so-called spinfree, algorithms are considered. If one has defined a basis in which the large component scalar functions transform according to the irreps of the appropriate single point group and the small component functions are related by the kinetic balance relation, it becomes possible to identify spin-orbit couplings as arising due to the off-diagonal matrix elements of the Fock operator. Neglecting these contributions becomes then identical to solving the spinfree modified Dirac equation of section 1.1.4 [24, 104] . While this is not very important in the Hartree-Fock stage it offers major saving in the electron correlation procedure. It means that correlation calculations can be carried out using non-relativistic algorithms and implementations. This will be discussed in more detail in the next section.

3.2 Electron correlation methods

The electron correlation methods available for 4-component methods are derived from non-relativistic counterparts. As discussed in section 2.4 a no-pair Hamiltonian (1.110) with fully relaxed projection operators is accessible only at the MCSCF level, where orbital rotations are included.

At the CI or CC level the "best" choice of no-pair Hamiltonian is the one defined by the Hartree-Fock (or MCSCF) orbitals. The generation of the matrix elements appearing in the no-pair Hamiltonian (1.111) is rather costly since it involves summation of integrals over both the large and small component type basis sets. In the scalar expansion scheme this is conveniently expressed in quaternion algebra [110] and gives the following expression in terms of electron repulsion integrals over scalar functions

$$g_{pqrs}^{\Lambda_{12}\Lambda_{34}} = \sum_{X}^{L,S}\sum_{Y}^{L,S}\sum_{\mu,\nu}^{N_X}\sum_{\kappa,\lambda}^{N_Y} B_{\mu\nu,pq}^{XX,\Lambda_{12}} g_{\mu\nu\kappa\lambda}^{XXYY} B_{\kappa\lambda,rs}^{YY,\Lambda_{34}} \quad (\Lambda_{12},\Lambda_{34}=0,1,2,3) \tag{1.139}$$

where the B-matrices are quaternion density matrices

$$B_{\mu\nu,pq}^{XX,\Lambda_{12}} e_{\Lambda_{12}} = \sum_{\Lambda_1=0}^{3}\sum_{\Lambda_1=0}^{3} c_{\mu p}^{X,\Lambda_1} c_{\nu q}^{X,\Lambda_2} e_{\Lambda_1}^{*} e_{\Lambda_2}. \tag{1.140}$$

Each quaternion integral $g_{pqrs}^{\Lambda_{12}\Lambda_{34}}$ can be expressed as a sum of 16 individual real numbers multiplied by a quaternion phase $e_{\Lambda_{12}}$ for electron 1 and a quaternion phase $e_{\Lambda_{34}}$ for electron 2. This is fully equivalent to the more conventional notation in terms of barred and unbarred partners of Kramers pairs. One can see that the effect of spin-orbit coupling translates into making the density matrices complex instead of real, while the use of a 4-component instead of a 2-component formalism leads to the additional summation over the small component basis functions. Together this makes the 4-index formation step, though still scaling as a fifth power with the number of basis functions, much more costly than the same step in non-relativistic calculations. Let M be the number of active Kramers pairs (or spatial orbitals in the non-relativistic case) in a correlated calculation. Pernpointner et al. [70] take the realistic assumption that the primitive small component basis is about twice the size of the primitive large component basis, that is $N_S \approx 2N_L$ and arrive using $M = N_L$ at a non-relativistic/relativistic operation count ratio of 1:130 for the first halftransformation and a ratio of 1:88 for the second halftransformation. The increase in the first steps is largely due to the presence of the small component basis set, while that in the last steps is caused by the coupling of the spin and spatial degrees of freedom. The computational scaling in the first steps can be reduced by using two-spinor expansion functions (so that $N_S = N_L$) and/or by using one-center approximations and it is probable that this will not present a major problem in the near future. The limiting ratio will then be that of a two-spinor algorithm that gives a scaling of 1:10 in the first halftransformation and 1:24 in the second. The later steps remain more demanding due to the fact that spin-orbit couplings, that are neglected in a purely non-relativistic theory, are taken into account. This gives a shift from real to complex or quaternion algebra and a loss of permutational symmetry. Here one may only improve the scaling in cases that these couplings are so small that approximate algorithms can be used. This is similar to the situation with 2-component methods where spin-orbit coupling are often neglected in the Hartree-Fock procedure and introduced in a CI step. Precisely the same treatment with similar computational gains (and loss of accuracy in some cases) is possible for the 4-component scheme. After the index transformation we end up with a second quantized Hamiltonian that can be used in various correlation treatments. The computations are then identical to those necessary in 2-component calculations.

A number of algorithms and implementations have been developed that tackle the electron correlation problem in the 2- or 4-component no-pair approximation. We will here only consider the algorithms that assume true 4- or 2-spinors and not the methods that neglect the effect on spin-orbit couplings in the Hartree-Fock stage. For more complete descriptions of the algorithms we refer to a recent overview by one of us [111]. In this review we will focus on the computational scaling of these methods.

The computationally most efficient treatment is given by many-body perturbation theory, in particular MP2. Since one only needs to sum transformed integrals divided by the orbital energy differences, this method allows for integral-direct implementations, thus opening up for application to larger systems. The method has as drawback that it is only applicable in cases were a single determinant reference already gives a reasonable description of the system. The computational scaling is identical to that of the index transformation step because the summation of the transformed integrals themselves takes a negligible amount of time.

The CI-type methods are more flexible but computationally less efficient and, more importantly, lack the correct scaling of energy with system size. This restricts their application to relatively small model systems in which they can give results close enough to the full CI limit. As orbital generator usually an average-of-configuration Hartree-Fock procedure is used, but work MCSCF algorithms is underway [98]. The computational scaling depends much on the actual implementation and on the type of CI that is used.

The last class of *ab initio* correlated method are the coupled cluster type approaches. They share with the perturbation theory type methods the features of size-extensivity and reasonable computational efficiency, but also the requirement that the reference wave function should be simple. Application of these type of relativistic methods to atoms has been pioneered by the groups of Lindgren [59] and Kaldor [52, 53] who have shown that almost arbitrary accuracy can be reached once a sufficient number of one-particle functions (up to i-type functions) and excitation level is used. While such accurate treatments are feasible due to the high symmetry and relatively few electrons to be correlated in atomic systems, this is still out of reach for molecular applications. It is possible to formulate the theory entirely in terms of Kramers' pairs instead of individual spinors but the corresponding algorithms have not yet been fully developed for the general case. The unrestricted CCSD(T) algorithm that has been implemented can be routinely applied to diatomics but larger systems and basis sets beyond triple or quadruple zeta level are usually still too demanding. These systems are smaller than feasible with efficient non-relativistic implementations of CC methods, which is mainly due to the limiting steps in the necessary index transformation that make this step often more expensive or cumbersome (due to the necessary diskspace and I/O) than the coupled cluster step itself. This situation will improve due to the increasing performance of computer hardware and the development of more efficient (parallel) algorithms, but unless approximations are made that break the notorious seventh power scaling with system size the coupled cluster algorithms will remain only applicable to relatively small systems.

The latter problem is in fact shared among all the correlation methods and is no different from the situation in non-relativistic quantum chemistry. However, while in non-relativistic quantum chemistry rewriting of the algorithms in the atomic orbital basis combined with approximation of long-range interactions via multipole expansions [113, 85] permits the development of algorithms that scale much better with system size [87], this is cumbersome in the relativistic case. In order to make AO-direct algorithms feasible one also needs to take into account the large difference in size between the active spinor set and the complete basis set. It is quite common to correlate only 10 % of the electrons in a system, using only the lowest lying virtual spinors. This makes the AO list of functions much longer than MO-list and it takes larger systems before the gain due to approximate treatment of long-range effects starts to pay off. This leaves the implementers of methods with a difficult choice : In the long run one will see that computations for system sizes for which the effort of using AO-based algorithms pays off are easily feasible and that this then enables much larger computations. In the current situation it is, however, still more efficient to use MO-based algorithms.

The observations regarding the computational scaling of the various steps in a conventional *ab initio* calculation (using scalar basis functions) are summarized in table 1.3. The first steps show a linear dependence on the number of primitive integrals that need to be calculated and contracted with the density matrix. The factors a, b and c depend on the type of basis functions contained in the integrals. Factor c will in general be larger then a because the small component basis contains higher angular momentum type functions than the large component basis. This is the main reason that relativistic Hartree-Fock calculations for small molecules are so much more expensive than comparable non-relativistic calculations. The difference in algebra (quaternion instead of real) furthermore reduces the permutational symmetry of the density matrix making the Fock matrix building and diagonalization procedure more expensive. This algebra difference also shows up in two-spinor methods where the presence of two-electron spin-orbit integrals increases the computation time by a constant factor relative to non-relativistic calculations. Index transformation of two-electron integrals exhibits the well-known fifth order scaling with the number of basis functions. In the transformation of the first two indices one sees mainly the effects of the small component basis set. In the second halftransformation the effect of spin-orbit couplings start to dominate making the scaling of full Dirac-Coulomb and two-spinor approaches comparable. They become fully equivalent in the last step (taken here as a CCSD calculation since this algorithm has been analyzed in detail previously [105]) where the variational inclusion of spin-orbit coupling leads to a 32-fold increase in operation count. The numbers presented here are theoretical estimates and actual measurements may give a somewhat different picture depending on efficiency of implementation and convergence of iterative procedures. Still, we think that it is useful to have such estimates, — both as a guideline for the efficiency of implementation and for the development of a long term strategy. We can for instance deduce that molecules may as well be treated with relativistic coupled cluster methods based on the spinfree Dirac-Coulomb equation than by other scalar relativistic or non-relativistic counterparts since the rate-determining step is the CCSD step which outscales the preliminary Hartree-Fock or index transformation steps. Including spin-orbit coupling increases the computational time, regardless of whether this is done in a 2- or a 4-spinor algorithm. This trend that is already visible in large basis set calculation on diatomics will become even stronger once more powerful computers that allow larger molecules to be treated become available. On the other hand, if we move to the large systems still inaccessible by current coupled cluster algorithms and use only Hartree-Fock or DFT methods we see that the inclusion of spin-orbit couplings is less crucial. The effort should go into the efficient evaluation or approximation of integrals over the small component basis set. For such larger systems one can then use

Table 1.3. Theoretical operation counts of steps in a correlated calculation for different approximation of the Dirac Coulomb no-pair Hamiltonian. 2-Spinor is any method that works with frozen no-pair projection operators but keeps the spin-orbit couplings, Spinfree is any method that varies the projection operators but neglects spin-orbit couplings. N is the number of basis functions (with subscripts referring to Large or Small component where appropriate). M is the number of active orbitals or Kramers' pairs (with the subscripts for the CCSD algorithm referring to the subsets of Occupied or Virtual orbitals). The steps considered are: A)Integral evaluation/Hartree-Fock/DFT, B)Index transformation, step 1, C)Index transformation, step 2, D) CCSD

	Non-Relativistic	Spinfree
A	$\frac{1}{8}aN^4$	$\frac{1}{8}\left(aN_L^4 + 4bN_S^2N_L^2 + 4cN_S^4\right)$
B	$\frac{1}{2}MN^4 + \frac{1}{2}M^2N^3$	$\frac{1}{2}M\left(N_L^4 + 4N_S^2N_L^2 + 4N_S^4\right) + \frac{1}{2}M^2\left(N_L^3 + 4N_SN_L^2 + 4N_S^3\right)$
C	$\frac{1}{2}M^3N^2 + \frac{1}{2}M^4N$	$\frac{1}{2}M^3\left(N_L^2 + 4N_S^2\right) + \frac{1}{2}M^3\left(N_L + 4N_S\right)$
D	$\frac{1}{4}M_o^4M_v^2 + 4M_o^4M_v^2 + \frac{1}{2}M_o^2M_v^4$	$\frac{1}{4}M_o^4M_v^2 + 4M_o^4M_v^2 + \frac{1}{2}M_o^2M_v^4$
	2-spinor	Dirac-Coulomb
A	$\frac{1}{2}aN^4$	$\frac{1}{2}\left(aN_L^4 + bN_S^2N_L^2 + cN_S^4\right)$
B	$2MN^4 + 8M^2N^3$	$2M\left(N_L^4 + N_S^2N_L^2 + N_S^4\right) + 8M^2\left(N_L^3 + N_SN_L^2 + N_S^3\right)$
C	$4M^3N^2 + 16M^4N$	$4M^3\left(N_L^2 + N_S^2\right) + 16M^4\left(N_L + N_S\right)$
D	$8M_o^4M_v^2 + 128M_o^4M_v^2 + 16M_o^2M_v^4$	$8M_o^4M_v^2 + 128M_o^4M_v^2 + 16M_o^2M_v^4$

the effect of the locality of the small component wave function making integrals prescreening and/or one-center expansion methods take effect.

4. Conclusion

In this chapter we have discussed 4-component relativistic methods and in particular the challenges that arise when extending the application of these methods from atomic to molecular systems, notably arising from the introduction of the algebraic approximation. We have analyzed in detail the variational stability of the Dirac-Coulomb-(Gaunt/Breit) Hamiltonian by comparing the standard approach to 4-component relativistic molecular calculations with QED in the semiclassical limit. We find that we recover the formulas of the standard approach by deleting vacuum polarization from semiclassical QED. The effect is, however, that the minimization principle of QED is replaced by a minimax principle in the standard approach, due to the fact that in QED all negative-energy orbitals are filled whereas they are empty and treated as an orthogonal complement in the standard approach. The standard approach employs a (and not "the" !) no-pair Hamiltonian which corresponds to surrounding the relativistic many-electron Hamiltonian by projection operators. Contrary to approximate 1-or 2-component approximations the 4-component methods allows a continuous update of the projection operators and thereby of the no-pair Hamiltonian and thus allows a complete relaxation of the electronic wave function to the actual potential of the system.

We insist on the distinction between Hamiltonians and methods. It is then easier to see that the difference in computational cost of relativistic and non-relativistic calculations is a difference in prefactor rather than order, and so it is not like comparing DFT with CCSD. Through a careful analysis of computational cost we furthermore show that one must distinguish the extra computational cost arising from the introduction of larger basis sets, notably the separate expansion of the large and small components, from the cost arising from the transition from non-relativistic to relativistic symmetry, that is the introduction of spin-orbit coupling. This latter contribution is identical at the 2- and 4-component level of theory. We consider how the computational cost can be reduced by exploiting symmetry, in particular time reversal symmetry, and the atomic nature of the small component density. We also outline the dilemma facing the programmer on whether he should choose a scalar basis expansion which allows him to benefit from the continuous development of (integral) codes in the non-relativistic domain or whether he should choose the more natural expansion in terms of 2-spinors which requires a more dedicated programming effort. The area of 4-component relativistic molecular methods continues to be an area of challenge and promise.

References

[1] P. J. C. Aerts, Ph. D. thesis, University of Groningen, (1986).

[2] C. D. Anderson, Phys.Rev. **41**, 405(1932).

[3] T. Aoyama, H. Yamakawa and O. Matsuoka, J. Chem. Phys. **73**, 1329 (1980).

[4] W. A. Barker and F. N. Glover, Phys.Rev. **99**, 317 (1955).

[5] M. Barysz and A. Sadlej, J. Mol. Struct.(Theochem) **573**, 181(2001).

[6] H. A. Bethe and E. E. Salpeter, *Quantum mechanics of one- and two-electron atoms*, Springer, Berlin (1957).

[7] F. Bloch., in F. Bopp (ed.), *W. Heisenberg und die Physik unserer Zeit*, Vieweg & Sohn, Braunschweig (1961).

[8] M. Born and J.R. Oppenheimer, Ann. Phys. **84**, 457 (1927).

[9] G. Breit, Phys.Rev. **34**, 553 (1929).

[10] G. E. Brown and D. G. Ravenhall, Proc. Roy. Soc. London A **208**, 552 (1951).

[11] C. F. Bunge, R. Jauregui,and E. Ley-Koo, Int. J. Quant. Chem. **70**, 805 (1998).

[12] P. Chaix and D. Iracane, J. Phys. B: At. Mol. Opt. Phys. **22**, 3791 (1989); ibid. **22**, 3815 (1989).

[13] Z. V. Chraplyvy, Phys.Rev. **91**, 388 (1953); ibid. **92** 1310 (1953)

[14] C. Cohen-Tannoudji, J. Dupont-Roc and G. Grynberg, *Photons et atomes*, Savoirs Actuels, New York (1987).

[15] C. G. Darwin, Phil.Mag. **39**, 537 (1920).

[16] P. A. M. Dirac, Proc. Roy. Soc. London A **113**, 243 (1927).

[17] P. A. M. Dirac, Proc. Roy. Soc. London A**133**, 60(1932).

[18] P. A. M. Dirac, Proc. Roy. Soc. London A**117**, 714(1928).

[19] P. A. M. Dirac, Proc. Roy. Soc. London A**118** 351 (1928).

[20] P. A. M. Dirac, Proc. Roy. Soc. London A**126**, 360 (1930).

[21] T. H. Dunning and V. McKoy, J. Chem. Phys. **47**, 1735 (1967).

[22] K. G. Dyall, I. P. Grant and S. Wilson, J. Phys. B **17**, 493 (1984).

[23] K. G. Dyall, P. R. Taylor, K. Faegri jr. and H. Partridge, J. Chem. Phys. **95**, 2583 (1991).

[24] K. G. Dyall, J. Chem. Phys. **100**, 2118 (1994).

[25] K. G. Dyall, Chem. Phys. Lett. **224**, 186 (1994).

[26] K. G. Dyall, J. Chem. Phys. **106**, 9618 (1997).

[27] K. G. Dyall, J. Chem. Phys. **109**, 4201 (1998).

[28] K. G. Dyall and T. Enevoldsen, J. Chem. Phys. **111**, 10000 (1999).

[29] K. G. Dyall, J. Chem. Phys. **115**, 9136 (2001).

[30] E. Engel, Int. J. Quant. Chem. **56**, 217 (1995).

[31] K. Faegri jr. and T. Saue, J. Chem. Phys. **115**, 2456 (2001).

[32] K. Faegri jr., Theor. Chem. Acc. **105**, 252 (2001).

[33] C. F. Fischer, Mol. Phys. **98**, 1043 (2000).

[34] V. Fock, Zeit. Phys., **75**, 622 (1932).

[35] P. Fulde, *Electron Correlation in Molecules and Solids*, Springer, Berlin (1995)

[36] W. H. Furry, Phys.Rev. **81**, 115 (1951).

[37] M. Gell-Mann, Nuovo Cimento Suppl. **4**, 848 (1956).

[38] I. P. Grant, Proc. R. Soc. London A **262**, 555 (1961).

[39] I. P. Grant and H. M. Quiney, Adv. At. Mol. Phys. **23**, 37 (1988).

[40] W. Greiner, *Relativistic Quantum Mechanics*, Springer-Verlag, Berlin (1990).

[41] D. J. Griffith, *Introduction to Electrodynamics*, Prentice-Hall, Upper Saddle River, New Jersey (1999).

[42] O. Heaviside, Philos. Mag., Ser.5 **27**, 324 (1889).

[43] T. Helgaker, P. Jørgensen and J. Olsen, *Molecular Electronic Structure Theory*, John Wiley & Sons, Ltd, Chichester (2000).

[44] A.C. Hennum, W. Klopper and T. Helgaker, J. Chem. Phys. **115**, 7356 (2001).

[45] T. Itoh, Rev. Mod. Phys. **37**, 159 (1965).

[46] J. D. Jackson. and L. B. Okun, Rev. Mod. Phys. **73**, 663 (2001).

[47] H. J. Aa. Jensen, K. G. Dyall, T. Saue and K. Faegri jr., J. Chem. Phys. **104**, 4083 (1996).

[48] G. Th. de Jong and L. Visscher, Theor. Chem. Acc., published online at http://dx.doi.org/10.1007/s002140020335.

[49] P. Jordan, Zeit. Phys. **44**, 473(1927).

[50] P. Jordan, Zeit. Phys. **75**, 648 (1932).

[51] P. Jordan and E. Wigner, Zeit. Phys. **47**, 631 (1928).

[52] U. Kaldor, Phys. Rev. A **38**, 6013 (1988); J. Chem. Phys. **88**, 5248 (1998).

[53] U. Kaldor, Theor. Chim. Acta **80**, 427 (1991).

[54] W. Kutzelnigg, Phys. Scr. **36**, 416 (1987).

[55] W. Kutzelnigg, Z. Phys. D. **11**, 15 (1989).

[56] W. Kutzelnigg, Chem. Phys. **225**, 203 (1997).

[57] J. K. Laerdahl, T. Saue and K. Faegri jr., Theor. Chem. Acc. **97**, 177 (1997)

[58] J.-M. Lévy-Leblond, Commun. Math. Phys. **6**, 286 (1967).

[59] I. Lindgren and J. Morrison, *Atomic Many-Body Theory*, Springer-Verlag, Berlin (1986)

[60] W. J. Liu, G. Y. Hong, D. D. Dai, L. M. Li, and M. Dolg, Theor. Chem. Acc. **96**, 75 (1997).

[61] O. Matsuoka, N. Suzuki, T. Aoyama and G. Malli, J. Chem. Phys. **73**, 1320 (1980).

[62] O. Matsuoka, J. Chem. Phys., **97** 2271 (1992).

[63] O. Matsuoka, Chem. Phys. Lett. **195**, 184 (1992).

[64] A. D. McClean and Y. S. Lee, in R. Carbo (ed.), *Current Aspects of Quantum Chemistry* Elsevier (1981).

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

[66] R. McWeeny, *Methods of Molecular Quantum Mechanics*, Academic Press, London (1992).

[67] M. H. Mittleman, Phys. Rev. A **24**, 1167 (1981).

[68] R. E. Moss, *Advanced Molecular Quantum Mechanics*, Chapmann and Hall, London (1973).

[69] W. Pauli, Z. Phys. **43**, 601 (1927), see footnote 2 on page 607.

[70] M. Pernpointner, L. Visscher, W. A. de Jong and R. Broer, J. Comp. Chem. **21**, 1176 (2000).

[71] L. Pisani and E. Clementi, J. Comp. Chem. **15**, 466 (1994).

[72] C. P. Poole and H. A. Farach, Found. Phys. **12**, 719 (1982).

[73] H. M. Quiney, H. Skaane, I. P. Grant, J. Phys. B **30**, L829 (1997).

[74] H. M. Quiney, H. Skaane and I. P. Grant, Chem. Phys. Lett. **290** 473 (1998).

[75] A. K. Rajagopal and J. Callaway, Phys. Rev. B **7**, 1912 (1973).

[76] B. O. Roos (ed.) *Lecture Notes in Quantum Chemistry - European Summer School in Quantum Chemistry.*, Lecture Notes in Chemistry 58. Springer, Berlin (1992); *Lecture Notes in Quantum Chemistry II - European Summer School in Quantum Chemistry.*, Lecture Notes in Chemistry 64. Springer, Berlin (1994).

[77] J. J. Sakurai, *Advanced Quantum Mechanics*, Addison-Wesley, Reading, Massachusetts (1967).

[78] T. Saue, in P. Schwerdtfeger (ed.), *Relativistic Electronic Structure Theory. Part 1. Fundamental Aspects*, Elsevier, to be published.

[79] T. Saue, Ph. D. thesis, University of Oslo (1996).

[80] T. Saue, K. Faegri jr. and O. Gropen, Chem. Phys. Lett. **263**, 360 (1996).

[81] T. Saue, K. Faegri jr., T. Helgaker, and O. Gropen, Mol. Phys. **91**, 937 (1997).

[82] T. Saue and H. J. Aa. Jensen, J. Chem. Phys. **111**, 6211 (1999).

[83] T. Saue and H. J. Aa. Jensen, in M. Defrancheschi and C. Le Bris (eds.), *Mathematical Methods for Ab Initio Quantum Chemistry*, Lecture Notes in Chemistry, Springer, Berlin (2000).

[84] T. Saue and T. Helgaker, J. Comp. Chem. **23**, 814 (2002).

[85] E. Schwegler and M. Challacombe, J. Chem. Phys. **105**, 2726 (1996).

[86] E. Schrödinger, Sitzungsber. Phys. Math. Kl., 418 (1930).

[87] M. Schutz and H. J. Werner, J. Chem. Phys. **114**, 661 (2001).

[88] W. H. E. Schwarz and H. Wallmeier, Mol. Phys. **46**, 1045 (1982).

[89] W. H. E. Schwarz and E. Wechsel-Trakowski, Chem. Phys. Lett. **85**, 94 (1984).

[90] K. Schwarzschild, Gött. Nach. Math.-Phys. Kl., 126 (1903).

[91] M. Sjovoll, H. Fagerli, O. Gropen, J. Almlof, T. Saue, J. Olsen, and T. Helgaker, J. Chem. Phys. **107**, 5496 (1997).

[92] R. E. Stanton, and S. Havriliak, J. Chem. Phys. **81**, 1910 (1984).

[93] J. Sucher, Phys. Rev. **22**, 348 (1980).

[94] J. Sucher, Int. J. Quant. Chem.: Quant. Chem. Symp. **25**, 3 (1984).

[95] B. Swirles, Proc. Roy. Soc. (London) A **152**, 625 (1935).

[96] A. Szabo and N. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*, McGraw-Hill, New York (1989).

[97] J. D. Talman, Phys. Rev. Lett. **57**, 1091 (1986).

[98] J. Thyssen, Ph. D. thesis, University of Southern Denmark (2001).

[99] S. Varga, E. Engel, W.-D. Sepp, and B. Fricke, Phys. Rev. A **59**, 4288 (1999).

[100] L. Visscher, P. J. C. Aerts, O. Visser and W. C. Nieuwpoort, Int. J. Quant. Chem.: Quant. Chem. Symp. **25**, 131 (1991).

[101] L. Visscher, T. Saue, W. C. Nieuwpoort, K. Fægri jr. and O. Gropen, J. Chem. Phys. **99**, 6704 (1993).

[102] L. Visscher, O. Visser, P. J. C. Aerts, H. Merenga and W. C. Nieuwpoort, Comp. Phys. Comm. **81**, 120 (1994).

[103] L. Visscher and E. van Lenthe, J. Chem. Phys. **306**, 357 (1999).

[104] L. Visscher and T. Saue, J. Chem. Phys. **113**, 3996 (2000).

[105] L. Visscher, K. G. Dyall, and T. J. Lee, Int. Journal of Quant. Chem. : Quant. Chem. Symp. **29**, 411 (1995).

[106] L. Visscher, T. J. Lee and K. G. Dyall, J. Chem. Phys. **105**, 8769 (1996).

[107] L. Visscher, Theor. Chem. Acc. **98**, 68 (1997).

[108] L. Visscher and K. G. Dyall, At. Data Nucl. Data Tables **67**, 207 (1997).

[109] L. Visscher, E. Eliav and U. Kaldor, J. Chem. Phys. **115**, 759 (2001).

[110] L. Visscher, J. Comp. Chem. **23**, 759 (2002).

[111] L. Visscher, in P. Schwerdtfeger(ed.), *Relativistic Electronic Structure Theory. Part 1. Fundamental Aspects*, Elsevier, to be published.

[112] O. Visser, L. Visscher, P. J. C Aerts and W. C. Nieuwpoort, Theor. Chim. Acta **81**, 405 (1992).

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

[114] G. C. Wick, Phys. Rev. **80**, 268 (1950).

[115] E. Wigner, Nachrichten der Akad. der Wissensch. zu Göttingen,II, 546 (1932).

[116] S. Wilson, J. Mol. Struct. (Theochem) **547**, 279 (2001).

[117] T. Yanai, T. Nakajima, Y. Ishikawa and K. Hirao, J. Chem. Phys. **114**, 6526 (2001).

[118] T. Yanai, H. Iikura, T. Nakajima, Y. Ishikawa and K. Hirao, J. Chem. Phys. **115**, 8267 (2001).